Polibits

ISSN: 1870-9044

polibits@nlp.cic.ipn.mx

Instituto Politécnico Nacional

México

Castro-Sánchez, Noé Alejandro; Sidorov, Grigori

Extracción automática de los patrones de rección de verbos de los diccionarios explicativos

Polibits, vol. 45, 2012, pp. 67-74

Instituto Politécnico Nacional

Distrito Federal, México

Extracción automática de los patrones de rección de verbos de los diccionarios explicativos

Noé Alejandro Castro-Sánchez y Grigori Sidorov

***Resumen*—En este trabajo se propone el uso de métodos simbólicos para la extracción de las valencias semánticas de verbos describiéndolas bajo el concepto de patrones de rección de la teoría Significado ⇔ Texto. El método se basa en el procesamiento automático de las definiciones de verbos contenidas en diccionarios explicativos y en el análisis de relaciones semánticas, principalmente de inclusión y de sinonimia, establecidas entre ellos. Partimos de la hipótesis de que las definiciones lexicográficas existentes en diccionarios explicativos deben proporcionar la suficiente información para identificar los actantes de verbos. Los resultados obtenidos demuestran que, a pesar de que en muchas de las definiciones no es posible encontrar información relativa a la estructura argumental de los verbos, es posible deducirla identificando y analizando las definiciones con las que existan relaciones sinonímicas y de inclusión.**

***Palabras clave*—Actantes, sinónimos, marcos de subcategorización, valencias, diccionarios explicativos.**

Automatic Extraction of Semantic Valences of Verbs from Explanatory Dictionaries

Noé Alejandro Castro-Sánchez and Grigori Sidorov

***Abstract*—In this work we propose the application of symbolic methods for extraction of semantic valences of the verbs describing them under the Government Pattern concept of the Meaning ⇔ Text Theory. The method is based on the automatic processing of the definitions of verbs used in Explanatory Dictionaries and the analysis of semantic relationships, as inclusion and synonymy, given among them. We believe that lexicographic definitions of Explanatory Dictionaries supply enough information for identifying verb actants. The obtained results show that even when it is not possible to find information related to the argument structure of verbs in the definitions, it is possible to deduce it identifying and analyzing other definitions which semantic relationships are established.**

***Index terms*—Actants, synonyms, subcategorization frames, valences, explanatory dictionaries.**

Manuscrito recibido el 14 de febrero de 2012, manuscrito aceptado el 7 de mayo de 2012.

Los autores trabajan en el Centro de Investigación en Computación, Instituto Politécnico Nacional, México DF (email: noe.acastro@gmail.com, sidorov@cic.ipn.mx).

I. Introducción

La gramática tradicional considera a la oración como una estructura bimembre, formado por sujeto y predicado. Sin embargo, el lingüista francés Lucien Tesnière propuso en 1959 [25] representar a la oración como una estructura jerárquica, y no binaria, donde el verbo ocupa la posición central, determinando los papeles que desempeñan el resto de elementos en la oración.

Todas las palabras que conforman la oración, establecen relaciones en donde algunas de ellas establecen o determinan propiedades de otras. Estas relaciones son denominadas "relaciones de rección". Los elementos regidos por un (o dependientes del) verbo se consideran complementos en la construcción del significado del verbo.

El hecho de regir o requerir una o varias palabras, se le denomina "régimen". De esta manera tenemos el régimen verbal, y el régimen preposicional: el primero hace referencia a la exigencia del verbo de ir o no acompañado por un elemento subordinado (régimen transitivo, y régimen intransitivo respectivamente), y el segundo señala la exigencia de una forma específica de la preposición a utilizar: "inducir a", "convertirse en", "depender de", etc.

La manera de nombrar a estos elementos que se espera acompañen a un verbo para lograr construir una oración gramatical e inteligible, varía de acuerdo al formalismo teórico que los procesa. En el enfoque teórico de constituyentes, se conocen más ampliamente con el nombre de 'marcos de subcategorización' (en inglés, *subcategorization frames* o SCF). Dentro del formalismo de dependencias, se conocen como "actantes". Bajo este formalismo pero en la "Teoría Significado ⇔ Texto" son conocidos bajo el nombre de "patrones de rección" [14].

En este trabajo de investigación identificamos de manera automática los actantes de los verbos a través del procesamiento automático de las definiciones contenidas en diccionarios explicativos y del análisis de las relaciones semánticas que ocurren entre éstos, apoyándonos en el enfoque teórico de la teoría 'Significado ⇔ Texto'.

En las secciones II y III haremos una revisión sobre los trabajos que se han desarrollado para la identificación automática de la valencia verbal y explicamos la metodología general que se ha utilizado. En la sección IV explicamos en qué se basa el método y en las secciones V, VI y VII abordamos a grandes rasgos los algoritmos que implementamos. Finalmente en la sección VIII hacemos una

descripción de los resultados obtenidos. Al final presentamos las conclusiones.

II. Trabajos relacionados

La recopilación de información de los complementos de los verbos fue una idea originalmente sugerida por el lingüista Noam Chomsky, y que se ha ido implementado por las teorías sintácticas subsecuentes.

El diseño pionero de la extracción automática de esta información, corresponde a Michael Brent [4], quien propone el desarrollo de un programa que toma texto de un corpus no etiquetado como única entrada para identificar SCF, extrayendo primeramente los verbos contenidos en él, y a continuación, frases que representen a los argumentos de los verbos.

En este trabajo Brent identificó cinco SCF, utilizando una técnica basada en el "Filtro de Casos de Rouvret y Vergnaud". A través de este filtro se identificaron los verbos potenciales, buscando, por ejemplo, palabras que contengan o carezcan del sufijo *–ing* (equivalente en español del gerundio *–ando* y *-endo*) o que sigan a un determinante o una preposición diferente a *to*. Por ejemplo, *was walking* (*estaba caminando*) se puede considerar como verbo, pero *a talk* (*una plática*) no.

En un segundo trabajo [5], identificó seis marcos sintácticos. En éste Brent incorporó un modelo estadístico en el cual se mide la frecuencia de aparición de claves con los verbos para cada uno de los marcos, así como el número de veces que cada verbo ocurre.

Posteriormente Ushioda [26] propone hacer uso de sentencias parseadas no completamente, derivadas de un corpus etiquetado. El sistema que elaboró es capaz de reconocer y calcular las frecuencias relativas de 6 marcos de subcategorización, los mismos trabajados por Brent. El proceso consiste en extraer del Corpus etiquetado las sentencias que contienen un verbo y dividir el sintagma nominal en pequeños fragmentos (*chunks*) utilizando un parseador de estados finitos, así como el resto de palabras usando un conjunto de 16 símbolos y categorías frasales. A estas sentencias les es aplicado un conjunto de reglas de extracción de marcos de subcategorización. Estas reglas están escritas como expresiones regulares y se obtienen a través de la extracción de ocurrencias de una pequeña muestra de verbos en un texto de entrenamiento.

Manning [16] propone un sistema más ambicioso capaz de reconocer 19 marcos sintácticos diferentes. Los marcos sintácticos se obtienen a través de un programa que procesa la salida de un etiquetador estocástico de partes de la oración (*part-of-speech tagger*) ejecutado sobre el corpus a analizar. El programa consta de dos partes: un parseador de estados finitos que analiza el texto etiquetado buscando un verbo, y que al encontrarlo, divide toda la información que lo sigue en pequeños componentes o chunks, hasta encontrar algún elemento reconocido como terminador de argumentos subcategorizados.

La segunda parte del programa, consiste en la reducción del ruido, para lo cual se utilizó el mismo filtro estadístico usado por Brent: el ruido (o pistas falsas), puede ser eliminado observando qué marcos aparecen con un verbo en una frecuencia razonablemente superior a la que pudiera considerarse casualidad (adjuntos) o errores en la detección.

Monedero *et al*. [19], inspirados en el trabajo de Brent y Manning, desarrollaron una herramienta para obtener marcos sintácticos de verbos en español.

El trabajo realizado, denominado SOAMAS, consistió en generar tres gramáticas: la primera de ellas encargada de identificar verbos principales y auxiliares, así como posibles conjunciones y preposiciones. La segunda realizada con el fin de reconocer sintagmas nominales, adjetivos y preposicionales. La tercera consistió en ser la encargada de identificar los complementos verbales.

El principal problema enfrentado para entonces, consistió en la carencia de corpus etiquetados para el español suficientemente extensos (dispusieron sólo de 10,000 palabras etiquetadas), lo que imposibilitó llegar a resultados confiables.

III. Metodología usada en trabajos previos

Los trabajos antes mencionados siguen una metodología de procesamiento como la expuesta en [7] y [22], en la que es posible distinguir los siguientes puntos:

1. *Selección y preparación del corpus*: indica la elección del corpus en el que se va a realizar la identificación de SCF, y, en caso de no estar anotado, el tipo de etiquetado que se le realizará (gramatical, sintáctico, etc).
2. *Detección de marcos*: establece el método computacional a seguir para identificar los SCF.
3. *Filtrado estadístico*: determina el método para eliminar el posible ruido obtenido en el paso previo.

En la *selección y preparación del corpus* se trabaja en considerar tanto el tipo como el tamaño de los corpus a procesar, pues estos factores pueden provocar variaciones en cuanto a los resultados que se obtienen. En general, los investigadores prefieren contar con la mayor cantidad de información (texto) posible, ya que de esta manera aseguran una muestra más representativa del idioma en el que se esté trabajando.

En [20] y [21] se expone cómo diferentes géneros de corpus provocan variaciones en las frecuencias de SCF. En [21] se estudiaron cinco corpus diferentes, dos de los cuales fueron obtenidos de fuentes psicológicas (caracterizados principalmente por contener sentencias aisladas), y los tres restantes fueron el "Brown corpus", "Wall Street Journal corpus" y el "Switchboard corpus". Las diferencias reportadas se encontraron tanto en los tipos de SCF como las frecuencias de los tipos de SCF.

La presentación del corpus tocante a la anotación de información lingüística, determinará la manera en que se procederá para ejecutar la tarea de extracción de SCF. Brent utiliza un corpus no anotado al cual aplica claves morfosintácticas para detectar verbos y sus posibles marcos. Ushioda propone utilizar sentencias parseadas sólo parcialmente, derivadas de un corpus ya etiquetado, y a las

cuales les es aplicado reglas escritas como expresiones regulares. Manning aplica un etiquetador estocástico sobre el corpus a analizar y así extraer todas aquellos componentes de la oración que tengan elementos reconocidos como terminadores de marcos. Gahl extrae subcorporas a través de la ejecución de expresiones regulares sobre el BNC para detectar en ellos a los posibles marcos.

La *detección de marcos* en general se ha realizado a través del *emparejamiento de patrones*, que consiste en definir a priori información gramatical que pudiera considerarse relevante para identificar alguna combinación de elementos léxicos como candidados a SCF. Posteriormente se busca en el corpus información que pudiera emparejarse con los patrones predefinidos.

La adquisición de los posibles marcos realizada por el proceso previo, no está exenta de errores, como es de esperarse. La información obtenida contiene ruido que puede derivarse de errores en la fase de etiquetado gramatical, por ejemplo, o incluso, errores en la fase de detección de SCF provocada por una ineficiencia en la discriminación de adjuntos.

Para remover toda la información no deseada, se realiza un procesamiento estadístico. En suma, se busca determinar si un candidato a SCF de un verbo en particular debe realmente considerarse como tal o no. Los métodos estadísticos para realizar el filtrado de información se hacen usualmente con la "prueba de hipótesis" (*hypothesis test*). Esta prueba consiste en establecer una hipótesis nula H_0, como verdadera, a menos que los datos sugieran lo contrario, lo cual provoca que se rechace la hipótesis y entonces se acepta como verdadera una hipótesis alternativa H_1. En el contexto de la adquisición de SCF, H_0 se considera como una falta de asociación entre un determinado verbo y un SCF, y H_1 como la afirmación a dicha asociación. Se establece la prueba como de *una cola*, dado de que la hipótesis alternativa establece una dirección, en este caso la correlación positiva entre el verbo y el marco. En seguida se calcula el valor estadístico de prueba con los datos de la muestra, lo que sirve para decidir si H_0 es verdadera o falsa. Esto se realiza comparando la probabilidad esperada de que exista correlación si H_0 es verdadera, con la probabilidad observada de coocurrencia. Si esta última es mayor que la primera, la hipótesis H_0 es rechazada.

IV. Método propuesto

La identificación de actantes de verbos se basa comúnmente en la aplicación de métodos estadísticos aplicados a corpus, analizando patrones de ocurrencia de eventos de acuerdo a la frecuencia de uso en el lenguaje.

En este trabajo se propone el uso del diccionario explicativo para su procesamiento, empleando una serie de heurísticas basadas en observaciones a priori de la naturaleza y comportamiento de los datos contenidos en las definiciones lexicográficas para la identificación de la valencia verbal. De la variedad de diccionarios que existen, nos enfocamos en aquellos dirigidos a los hablantes nativos de un idioma (monolingües), que no presentan restricciones de dominio (generales) y que presentan la definición semántica de los vocablos que contienen (explicativos). En particular se eligió el Diccionario de la Real Academia de la Lengua Española (DRAE), considerado como el de mayor resonancia en países hispanohablantes.

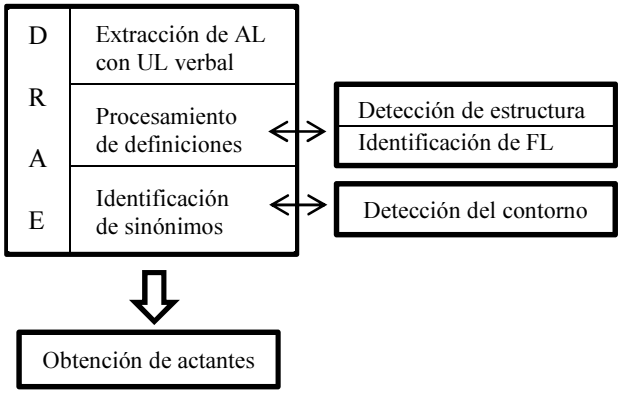

Fig. 1. Arquitectura del proyecto.

En general, los diccionarios presentan secciones textuales dispuestas ordenadamente denominadas "Artículos Lexicográficos" (AL), conformados por una entrada o "Unidad Léxica" (UL) y la información que la define o describe. Esta información se presenta ya sea como definición propia o perifrástica, cuando expresa el significado de las entradas en cuanto a su contenido léxico-semántico, o como definición impropia, cuando se utiliza para describir o explicar el funcionamiento y empleo de palabras funcionales, debido a su falta de un verdadero significado léxico, véase fig. 1.

La estructura de las definiciones propias suele seguir la norma establecida por la llamada definición aristotélica, la cual consiste de un enunciado encabezado por un término genérico o hiperónimo inmediato (genus), seguido de una diferencia específica (differentia), o conjunto de rasgos y características que permiten distinguir el término definido de otros que se agrupan bajo el mismo hiperónimo.

Esta posición predecible de los elementos que conforman las definiciones propias, genus + differentia, permite utilizar heurísticas que puedan aplicarse para identificarlos de manera automática. Además de estos elementos, se sabe ([6]) que en algunas definiciones lexicográficas es posible identificar elementos que proporcionan información sobre la estructura argumental de las UL's, relacionada con restricciones contextuales o algunos usos sintácticos. Esta información se denomina "contorno de la definición".

El contorno no siempre es señalado explícitamente en los diccionarios lexicográficos, posiblemente porque resultaría redundante para los nativos del idioma, aunque suele ser importante para saber hacer un uso correcto de la UL definida. El diccionario "Salamanca de la Lengua Española", por ejemplo, hace un señalamiento del contorno en sus definiciones:

Derivar. *Ser < una cosa > consecuencia de [otra cosa]*

Los sujetos son rodeados con los signos mayor y menor que, y los complementos entre corchetes.

Por otro lado, en el DRAE el contorno se indica encabezando la definición con la fórmula "Dicho de":

***Convalecer**. Dicho de una persona: salir del estado de postración o peligro...*

Esto indica que el verbo *Convalecer* selecciona como sujeto alguna *persona*. Lamentablemente, en este diccionario no todas las definiciones de verbos vienen acompañadas por la especificación del sujeto, y definitivamente no es posible encontrar indicaciones sobre el resto de complementos en las definiciones. Por ejemplo:

***Bajar.** Poner **algo** en un lugar inferior a aquel en que estaba*

En este caso, el objeto directo del verbo (*algo*) aparece en la definición, pero no es acompañado por alguna marca que logre identificarlo como complemento del verbo.

En algunos otros casos, algunos o todos los complementos del verbo definido no forman parte de la redacción de la definición:

***Controlar.** Ejercer el control*

Este tipo de casos representan un verdadero reto para la identificación automática del contorno en las definiciones.

V. Identificación de los componentes de las definiciones

En primer lugar se realizó un preprocesamiento de datos que consistió en extraer únicamente los artículos lexicográficos que eran de relevancia para este trabajo, es decir, unidades léxicas verbales y sus respectivas definiciones. Se utilizó la herramienta de análisis de texto de código abierto para varios idiomas, Freeling [2], como etiquetador de partes de la oración (POST, por sus siglas en inglés) para conocer la categoría gramatical de cada palabra de los datos seleccionados.

Con esta información fue posible hacer un primer intento para identificar el genus de la diferencia específica. Se hizo un primer análisis de tipo manual para identificar algunos patrones que pudieran ayudar a automatizar el proceso para atender todos los casos del dicionário. Esto arrojó las diferentes maneras en que es posible encontrar el genus, lo cual puede resumirse en lo siguiente:

1) *Verbos individuales:*
   a. Con un solo verbo. Ejemplo:
      ***Cotizar**. Pagar una cuota.*
   b. Dos o más verbos enlazados por conjunciones y/o disyunciones. Ejemplo:
      ***Armonizar**. Escoger y escribir los acordes correspondientes a una melodía.*
      ***Aballar**. Amortiguar, desvanecer o esfumar las líneas y colores de una pintura.*
2) *Como cláusula subordinada en infinitivo cumpliendo la función de complemento directo. Ejemplo:*
   ***Gallear**. Pretender sobresalir entre otros con presunción o jactancia.*
3) *Como Función Léxica. Ejemplo:*
   ***Anunciar**. Dar publicidad a algo con fines de propaganda comercial.*

Cada caso particular requiere un tratamiento diferente que permita su correcta identificación. En el caso 1 y 2, todo verbo existente como cabecera de la definición se considera genus de la UL definida. En 3) se requiere un procesamiento más complejo: los verbos que vienen acompañados por un sustantivo son Funciones Léxicas (FL) potenciales. Las FL se definen como una función que asocia una palabra denominada "base", la cual aporta su significado literal a la expresión, a otra llamada "colocador", que adquiere un significado diferente de su significado típico, de tal manera que el significado del conjunto incluye el significado de una de las palabras (base), pero no del otro (colocador) [12]. De esta manera, el genus en una definición que es encabezada por una FL no puede ser el colocador.

Siendo posible identificar el genus en la definición, el resto de elementos que la constituyen automáticamente son tomados como parte de la diferencia específica.

VI. Procesamiento del contorno

Observaciones de las definiciones mostraron que el contorno se conforma por sustantivos comunes (NC) y pronombres indefinidos (PI), lo que condujo a la elaboración de una heurística que identifica los segmentos de las definiciones constituidos por palabras con estas categorías gramaticales.

El algoritmo desarrollado se basa en una serie de reglas que reflejan la estructura básica de las definiciones, más concretamente, de la diferencia específica, que permiten capturar incluso el contexto sintáctico que delimita cada elemento del contorno, ayudando a conocer por ejemplo las preposiciones con las que puede acompañarse.

Las reglas quedan definidas de la siguiente manera:

1) *La nomenclatura utilizada se define en la tabla I.*

Tabla I. Definición de los símbolos utilizados en la gramática

Símbolo utilizado	Significado
Diff	Differentia
Cont	Contorno
Nuc	Núcleo del contorno (PI ó NC)
EleIzq	Elementos a la izquierda
EleDer	Elementos a la derecha
EIzq	Elemento izquierdo
EDer	Elemento derecho
DA, DI, DO, DP, CS, RG, Z, AQ, RN, CC, FC	Etiquetas asignadas a palabras para indicar su información morfológica, propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus

2) *El lado izquierda de la primera producción, es el símbolo inicial.*

3) *Reglas:*
Diff → Cont
Cont→ Nuc | EleIzq Nuc | EleIzq Nuc EleDer | Nuc EleDer | Cont Liga Cont
Nuc → PI | NC
EleIzq → EIzq | EIzq EleIzq
EleDer→ EDer | EDer EleDer
EIzq→ DO | DA | DI | DP | DD | SP | CS | RG | Z | AQ
EDer → AQ | RN
Liga → CC | FC

Estas reglas no se utilizan en la producción de oraciones (pues podrían generar oraciones agramaticales como un nombre común acompañado por una sucesión ininterrumpida de preposiciones), sino en la segmentación de las definiciones, donde cada segmento está conformado por un único candidato a elemento del contorno.

Un ejemplo de la aplicación de estas reglas en las definiciones, es el siguiente:

Poner. *Colocar en un sitio o lugar a alguien o algo*
Segmentación: *en un sitio o lugar* | *a alguien o algo*

VII. Adquisición de sinónimos

Para redactar las definiciones de verbos, probablemente los lexicógrafos no toman un criterio unificado sobre el uso o no del contorno asociado a los verbos, ni sobre el número de elementos del contorno que se puedan utilizarse en las definiciones. Es decir, existirán definiciones que aporten mayor información en este rubro, que otras. Debido a esto, lo que hemos propuesto es utilizar las definiciones de otros verbos para complementar la información faltante en casos donde sea necesario. Esta selección de verbos no se realiza de manera aleatoria, sino que se basa en las relaciones semánticas dadas entre verbos, como la sinonimia y las relaciones de inclusión.

La identificación de los verbos relacionados entre sí por sinonimia, se realiza de la siguiente manera: el diccionario de la RAE emplea el tipo de definición sinonímica recurrentemente, la cual consiste en utilizar como definición una o varias palabras con la misma categoría gramatical que la UL definida. Por ejemplo, el verbo "Coger" se define como:

Coger. *Asir, agarrar o tomar*

Lo que significa que la definición de "Coger" puede encontrarse en la definición de los verbos "asir", "agarrar" o "tomar". Este tipo de definiciones puede provocar círculos viciosos, lo cual es considerado como un defecto por los lexicógrafos, sin embargo, este comportamiento beneficia nuestra tarea. Un ejemplo de círculo vicioso es el conformado entre los verbos "coger, asir, agarrar y tomar", mostrado en la siguiente gráfica. El inicio de cada flecha indica la UL definida, y el nodo al que apunta la UL que se utiliza como sinónimo en su definición.

Vemos un ejemplo de los círculos viciosos. Las definiciones que componen cada verbo, son las siguientes:

- ***Coger***. *Asir, agarrar o tomar*
- ***Agarrar***. *Coger, tomar.*
- ***Tomar***. *Coger o asir con la mano algo.*
- ***Asir***. *Tomar o coger con la mano, y, en general, tomar, coger, prender.*

Al ser considerados estos verbos como sinónimos, significa que pueden sustituirse indistintamente en al menos algún sentido de los varios que tienen atribuidos. Siendo así, se deberían cumplir los siguientes dos supuestos:

1. El número de actantes de cada verbo es el mismo para cada uno de sus sinónimos (en al menos un sentido),
2. Las restricciones semánticas que un verbo impone a sus actantes, son las mismas que el resto de sus sinónimos (en al menos un sentido).

De cumplirse los puntos previos, permitiría subsanar en la medida de lo posible la falta de información referente al contorno que suele existir en las definiciones de verbos en el diccionario de la RAE, combinando el contorno de las definiciones que aparecen en un conjunto de sinónimos. Para ello, en primer lugar, se debe distinguir qué sentido en específico logra la relación sinonímica de los verbos. Por ejemplo, el verbo "abatir" en el sentido 6 incluye como sinónimos en su definición los verbos "desarmar" y "descomponer". Ambos verbos disponen de varios sentidos, de entre los cuales es necesario distinguir cuáles son los que los relacionan como sinónimos. La solución que en este trabajo se implementó consiste en buscar en las definiciones algún hiperónimo común a los verbos, lo que indicaría que existe relación semántica en ese sentido en específico.

En las tablas II y III, se muestran los hiperónimos de los primeros 5 sentidos de los verbos "desarmar" y "descomponer", respectivamente. Se observa que el sentido

TABLA II.HIPERÓNIMOS DEL VERBO *DESARMAR*

Num. sentido	Hiperónimo
1	*Quitar, hacer entregar*
2	*Desnudar o desceñir*
3	*Reducir*
4	*Dejar*
5	*Desunir, separar*

TABLA III. HIPERÓNIMOS DEL VERBO *DESCOMPONER*

Num. sentido	Hiperónimo
1	*Desordenar y desbaratar*
2	*Separar*
3	*Indisponer*
4	*Averiar, estropear, deteriorar*
5	*Corromperse*
…	…

5 de “desarmar” y el sentido 2 de “descomponer” comparten el mismo hiperónimo.

Teniendo identificados los sentidos relacionados semánticamente, pueden combinarse los contornos de las definiciones para complementar la información faltante en algunas de ellas. En esta tarea pueden identificarse los siguientes casos:

1) No existe información alguna del contorno en alguna definición, pero sí en las otras. Retomando las definiciones de los verbos “coger”, “agarrar”, “asir” y “tomar”, observamos que la definición del verbo “coger” sólo incluye sinónimos, sin hacer mención alguna del contorno. Sin embargo, la definición del verbo “tomar” incluye dicha información. El resultado de la obtención de segmentos de la definición es:

Tomar. *Coger o asir con la mano algo*
Segmentación*: con la mano* | *algo*

Por lo tanto, el contorno del verbo “tomar” se considera también perteneciente al verbo “coger”.

2) Algunas definiciones incluyen segmentos que no pertenecen al contorno. Este es el caso más común, y es complicado lograr una correcta discriminación de segmentos. Por ejemplo:

Llevar. *Conducir algo desde un lugar a otro alejado de aquel en que se habla o se sitúa mentalmente la persona que emplea este verbo.*
Segmentación: *algo* | *desde un lugar* | *a otro* | *mentalmente la persona* | *este verbo*

En esta definición, los segmentos “mentalmente la persona” y “este verbo”, no son elementos que formen parte del contorno y que por lo tanto reflejen a los actantes del verbo.

VIII. Resultados experimenatales

Después de procesar todas las definiciones de verbos encontramos poco más de 6,000 definiciones sinonímicas. Estas 6,000 definiciones se procesaron para identificar si existía algún genus común a las definiciones de los verbos agrupados y así precisar el número del sentido en que se relacionaban. Esto llevó a la identificación de un aproximado de 6,500 grupos de sinónimos en donde se identificaron explícitamente los sentidos. Por ejemplo, el verbo “amparar” en su sentido 4 se define como: “Defenderse, guarecerse”. Estos verbos usados en la definición, ambos en su sentido 2, se definen como:

Defender (2): *Mantener, conservar, sostener algo contra el dictamen ajeno.*
Guarecer (2): *Guardar, conservar y asegurar algo*

Ambas definiciones comparten el verbo *conservar*, por lo que en ese sentido en particular conforman un grupo de sinónimos con sentido identificado. Sin embargo, observamos también que *defender* en su sentido 1 y *guarecer* en su sentido 4 se definen como:

Defender (1): *Amparar, librar, proteger*
Guarecer (4): *Socorrer, amparar, ayudar.*

Conformarían un nuevo grupo en dichos sentidos bajo el verbo *amparar*. Del ahora total aproximado de 6,500 grupos conformados por verbos en un sentido en particular, en 3,000 agrupaciones no se lograron identificar los sentidos que relacionaban a los verbos siguiendo el criterio del genus común. De los 6,500 grupos, cerca de 500 grupos no ofrecen ningún candidato a contorno. Por ejemplo:

Abrasar (3): *Calentar demasiado.*
Quemar (2): *Calentar mucho.*

Por otro lado, considerando que no todos los sustantivos comunes y pronombres indefinidos que aparecen en una definición pueden ser catalogados como elementos del contorno (ver apartado 5.1), decidimos procesar aquellas definiciones cuyos candidatos a elementos del contorno estuvieran conformados únicamente por los pronombres indefinidos “algo, alguien”, y los sustantivos comunes “cosa, persona, animal, lugar” y “parte”, ya que al realizar una medición de las categorías gramaticales de palabras funcionales más frecuentemente utilizadas en las definiciones, las palabras antes mencionadas tuvieron mayor presencia (Tabla IV).

Tabla IV. Elementos de controno más frecuentes

Palabra	Frecuencia
Algo	3,000
Alguien	2,000
Otro	900
Cosa	800
Parte	500
Persona	400
Lugar	350
Cuerpo, acción, fuerza, agua, tierra, …	< 300

Por otro lado, estas palabras representarían en cualquier ontología el nivel más alto o abstracto de los grupos que la componen.

El procesamiento de estos datos nos arrojó un total de 420 grupos de sinónimos que contienen dichas palabras en sus funciones.

La cantidad de verbos que se lograron detectar en este último grupo, fue de 280 verbos, y de estos, se lograron identificar 390 sentidos en total.

Varios grupos incluyen el mismo sentido de algún verbo. Al existir intersección entre ellos, podemos proceder a la unión de grupos, y así muy posiblemente, complementar de manera más precisa la información de los diferentes verbos y sobre todo de su contorno.

TABLA V. ESTADÍSTICAS DE SINÓNIMOS

Elemento evaluado	Cantidad
Definiciones sinonímicas	6,000
Grupos de sinónimos con sentidos de verbos identificados	6,500
Grupos de sinónimos donde no se identificaron los sentidos de verbos	3,000
Grupos de sinónimos donde no se identificaron candidatos a contorno	500
Grupos de sinónimos con candidatos a contorno más abstractos	420

Por ejemplo, consideremos el siguiente grupo de sinónimos tomados de la definición del verbo "maliciar" en su primer sentido:

Maliciar (1): *Recelar, sospechar, presumir algo con malicia*

Los verbos "recelar" y "sospechar" coinciden en usar el mismo genus en sus sentidos 1 y 2 respectivamente:

Recelar (1): *Temer,* ***desconfiar*** *y sospechar*
Sospechar (2): Desconfiar*, dudar, recelar de alguien*

Combinamos las definiciones de ambos verbos en los sentidos antes indicados y el contorno resultante es "de alguien".

Por otro lado, "recelar" y "dudar" son también sinónimos según el segundo sentido de "sospechar". Ambos verbos son definidos en los sentidos abajo indicados, nuevamente bajo el genus "desconfiar", de la siguiente manera:

Recelar (1): *Temer,* ***desconfiar*** *y sospechar*
Dudar (2)*:* ***Desconfiar****, sospechar de alguien o algo*

La identificación del contorno en ambas definiciones sería "de alguien o algo". Como ambos grupos de sinónimos incorporan el verbo "recelar" en un mismo sentido (1), entonces los unimos para conformar un solo grupo. De esta manera, tenemos que los verbos "recelar" (en 1), "sospechar" (en 2) y "dudar" (en 2) comparten el contorno "de alguien o algo".

En suma, por cada verbo en un sentido en particular, unimos todos los grupos de sinónimos que lo incluían y combinamos los contornos identificados.

La evaluación manual de los resultados dio 83% de precisión del método. Se evaluaron manualmente los contornos de 115 verbos.

IX. CONCLUSIONES

En este trabajo propusimos un método para la extracción de los actantes de verbos para el idioma español, basándonos en el análisis de las definiciones en diccionarios explicativos. Dado que la redacción de los artículos lexicográficos se apega a estructuras bien establecidas, es posible crear heurísticas para el análisis y extracción de información de ellos. Cada uno de los elementos que conforman estas estructuras, aportó datos relevantes para el cumplimiento de los objetivos propuestos.

En particular, el contorno de las definiciones de los verbos, al indicar condiciones sintagmáticas del verbo y recoger las restricciones de tipo semántico que sus argumentos requieren, se consideran una imagen de la valencia verbal. Así, la extracción del contorno se traduce en la obtención de información sobre los actantes del verbo.

La falta de una especificación rigurosa del contorno en la mayoría de las definiciones de los verbos, imposibilita conocer de manera certera sus valencias. Sin embargo, encontramos un recurso para complementar esta escasa información apoyándonos en las definiciones de otros verbos. Esto se hizo atendiendo las relaciones léxicas de inclusión (hiperonimia/hiponimia) establecidas entre los genus y los artículos lexicográficos y las relaciones de sinonimia que pueden encontrarse en las llamadas definiciones sinonímicas de las que hace uso el diccionario. A través de estas relaciones pudimos identificar qué sentidos de los verbos establecían relaciones de sinonimia con otros.

Considerando que los sinónimos pueden sustituirse mutuamente en cualquier contexto (bajo sentidos en específico) fue posible afirmar que bajo estas condiciones existe una coincidencia en la valencia verbal. Esta obtención de ciclos nos ayudó a completar la lista de actantes de cada verbo complementando la información que cada definición manejaba.

AGRADECIMIENTOS

El trabajo fue realizado con el apoyo parcial del gobierno de México (proyectos CONACYT 50206-H y 83270, SNI) e Instituto Politécnico Nacional, México (proyectos SIP 20111146, 20113295, 20120418, COFAA, PIFI), Gobierno del DF (ICYT-DF proyecto PICCO10-120) y la Comisión Europea (proyecto 269180).

REFERENCIAS

[1] Ch. Aone and D. MacKee, "Acquiring Predicate-Argument Mapping Information from Multilingual Texts," in *Corpus processing for lexical acquisition*, pp. 191–202, 1996.

[2] J. Atserias, B. Casas, E. Comelles, M. Gonzáles, and L. Padró, "FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library," in *Fifth international conference on Language Resources and Evaluation*, Genoa, Italy nlp/freeling, http://www.lsi.upc.edu/nlp/freeling, 2006.

[3] I. Bolshakov, A. Gelbukh, *Computational Linguistics: Models, Resources, Applications*, 2004.

[4] M. Brent, "Automatic acquisition of subcategorization frames from untagged text," in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA., pp. 209-214, 1991.

[5] M. Brent, "From grammar to lexicon: unsupervised learning of lexical syntax," *Computational Linguistics* 19(3): 243-262, 1993.

[6] M. Cordero, "Diccionario de la lengua española secundaria (DILES): Planta para su elaboración con algunos apuntes básicos de metalexicografía," *Káñina, Rev. Artes y Letras*, Univ. Costa Rica. XXXI (1): 167-195, ISSN: 0378-0473, 2007.

[7] R. Dale, H. Moisl, and H. Somers. *Handbook of Natural Language Processing*, ISBN: 0-8247-9000-6, 2000.

[8] *Diccionario de la Lengua Española*, Edición vigésimo segunda. www.rae.es, 2001.

[9] J. Fernández, *Rektion. Rección/Régimen.* http://culturitalia.uibk.ac.at, Hispanoteca, 2002.
[10] S. Fujita and F. Bond, "An Automatic Method of Creating Valency Entries using Plain Bilingual Dictionaries," in *The tenth conference on theoretical and methodological issues in machine translation*, Baltimore, Maryland, pp. 55-64, 2004.
[11] S. Gahl, "Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus," in *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada., pp. 428-432, 1998.
[12] A. Gelbukh, O. Kolesnikova. *Semantic Analysis of Verbal Collocations with Lexical Functions*. Studies in Computational Intelligence, N 414, Springer, 2012.
[13] D. Ienco, S. Villata., and C. Bosco, "Automatic Extraction of Subcategorization Frames for Italian," in *International Conference on Language Resources and Evaluation LREC*, 2008.
[14] S. Kahane, "Meaning-text theory," in Ágel, Vilmos et al. (eds.): *Dependency and Valency. An International Handbook of Contemporary Research*. Berlin, 2003.
[15] D. Kawahara and S. Kurohashi, "Case frame compilation from the web using high-performance computing," in *Proceedings of LREC2006*, 2006.
[16] C. Manning, "Automatic acquisition of a large subcategorization dictionary from corpora," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 235- 242, 1993.
[17] S. Marinov and C. Hamming, *Automatic Extraction of Subcategorization Frames from the Bulgarian Tree Bank*, 2004.
[18] A. Mendikoetxea, "En busca de los primitivos léxicos y su realización sintáctica: del léxico a la sintaxis y viceversa," *2º Xarxa Temàtica de Gramàtica Teòrica*, Barcelona, UAB, 2004.
[19] J. Monedero, J. González, J. Goñi, C. Iglesias, and A. Nieto, "Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS," *Procesamiento del lenguaje natural*, boletín 17, 1995.
[20] D. Roland, D. Jurafsky, "How Verb Subcategorization Frequencies Are Affected By Corpus Choice," in *Proc. of COLING/ACL-98*, pp. 1122-1128, 1998.
[21] D. Roland and D. Jurafsky, "Verb Sense and Verb Subcategorization Probabilities," in Stevenson, Suzanne, and Paola Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam: John Benjamins, pp. 325-346, 2002.
[22] S. Sabine, "The Induction of Verb Frames and Verb Classes from Corpora," in *Corpus Linguistics. An International Handbook*. Anke Lüdeling and Merja Kytö (eds). Mouton de Gruyter, Berlin, pp. 952–972. eBook ISBN: 978-3-11-021388-1. Print ISBN: 978-3-11-020733-0, 2009.
[23] A. Sarkar and D. Zeman, "Automatic Extraction of Subcategorization Frames for Czech," in *Proc. of the 18th International Conference on Computational Linguistics*, 2000.
[24] A. Séreny, E. Simon, and A. Babarczy, "Automatic Acquisition of Hungarian Subcategorization Frames," in *9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics CINTI*, 2008.
[25] L. Tesnière, *Éléments de syntaxe structurale* (*Elementos de sintaxis estructural*) 1959.
[26] A. Ushioda, D. Evans, T. Gibson, and A. Waibel, "The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora," in Boguraev, B. and Pustejovsky, J. eds. *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text*. Columbus, Ohio, pp. 95-106, 1993.
[27] E. Uzun, Y. Kılıçaslan, H.V. Agun, and E. Uçar, "Web-based Acquisition of Subcategorization Frames for Turkish," in *Computational Intelligence: Methods and Applications*, IEEE Computational Intelligence Society, 2008.