Desarrollo de herramienta para lexicógrafo con detección automática de relaciones semánticas implícitas

Wilbert A. Olán Cristobal, Alexander Gelbukh, Grigori Sidorov

Laboratorio de lenguaje natural,
Centro de Investigación en Computación,
Instituto Politécnico Nacional,
Av. Juan de Dios Batiz s/n, esq. Mendizabal, Zacatenco, 07738,
México, D. F.
wolan@correo.cic.ipn.mx, {gelbukh,sidorov}@cic.ipn.mx

Resumen. Uno de los problemas más importantes en lexicografía es la existencia de círculos viciosos en definiciones en los diccionarios. Se presenta una herramienta que ayuda a encontrar los ciclos mencionados, escoger algunas palabras como primitivas o miembros del vocabulario definidor y analizar el impacto de esta selección a los círculos viciosos, es decir, encontrar las relaciones semánticas implícitas y depurarlas.

1 Introducción

A partir de los sistemas de diccionarios explicativos existentes actualmente, existe una parte de ellos la cual no tiene alguna solución práctica. Referimos a un problema en los diccionarios explicativos que es la presencia de los círculos viciosos en definiciones. Por ejemplo,

Gallina: hembra de gallo
Gallo: macho de gallina

Este es un problema en los sistemas de definiciones ya que lo anterior equivale a decir que *gallina es hembra de macho de gallina*, y esto no ayuda a entender lo que es *una gallina* sin saberlo de antemano.

Este es el círculo vicioso de longitud 1, hay círculos más largos. Sin embargo, el verdadero problema es que no se puede evitar los círculos de este tipo porque todo conjunto de palabras se define a través de mismo conjunto. Existen dos posibles soluciones

- Tolerar los círculos y solamente tratar de obtener los círculos de mayor longitud, lo que da la ventaja para un lector humano, porque se aumenta la probabilidad de conocer alguna de las palabras en el círculo, y
- Declarar algunas palabras como las palabras "primitivas" y no dar definiciones algunas para ellas. Es el enfoque más aceptable para las computadoras.

A. Gelbukh, G. Sidorov, W. Olán, J. Vera (Eds.) Recientes avances en la ciencia de la computación en México, pp.99-104.

que se expresa por el verbo *to want* (*querer*), mientras *feel like* (*tener ganas de*), por otra parte, lo debilita.

Así se puede establecer que *to want* (*querer*) no es la palabra semánticamente simple. Además, las acumulaciones semánticas las cuales complican el sentido de propio *to want* son específicas en el idioma inglés.

Se propone la siguiente definición de las palabras primitivas: La palabra se considera como una primitiva si el lenguaje dado no tiene ningún conjunto de las palabras vía las cuales puede ser explicada. Es decir, en este caso tanto *to want*, como *to wish* son las palabras primitivas, y existe una cantidad significativa de las palabras primitivas evaluada en más de 5000 elementos, porque solo las palabras que pueden ser definidas de modo claro se consideran no primitivas.

2.3 Vocabulario definidor

Antes que nada, se representa el diccionario explicativo como un grafo dirigido — cada palabra que tiene definición es un vértice, y las palabras que se encuentra en la definición están conectadas con ella. En su turno, la palabra puede formar definiciones de otras palabras.

Esta idea de representar el diccionario como un grafo es con la finalidad de verlo como una red semántica, y no es nueva, es una idea desarrollada por autores como Evens (1988) y Fellbaum (1990). Kozima y Furugori (1993) también analizan una red semántica, en este caso para saber que palabras "se activan" empezando de alguna palabra determinada.

Obviamente, el diccionario no contiene en sus definiciones las palabras que no son definidas en el mismo diccionario.

En este grafo se puede elegir algunas palabras como las primitivas semánticas. Las primitivas semánticas son las palabras las cuales pertenecen a conjunto de vértices y se marcan de tal modo que existe una ruta en el grafo de cualquier longitud para definir las demás palabras.

A diferencia con (Gelbukh and Sidorov, 2002) vamos a distinguir las primitivas semánticas y el vocabulario definidor.

El vocabulario definidor ese define de la misma forma con la única diferencia que la ruta debe tener longitud uno. Es decir, vocabulario definidor es mucho más interesante que las puras primitivas semánticas. De hecho, vocabulario definidor es un conjunto de las primitivas con cierta propiedad adicional.

Existen algunos diccionarios para inglés donde el vocabulario definidor se forma manualmente, por ejemplo, los de Oxford o de Collins. El número total de los elementos de vocabulario definidor es alrededor de 2000-3000 palabras. El número de primitivas muy parecido se reporta en (Gelbukh and Sidorov, 2002), (Rivera, 2003) para la detección automática de las primitivas semánticas –alrededor de 2000 palabras primitivas.

3 Características de la herramienta

En base del algoritmo descrito en (Gelbukh and Sidorov, 2002), (Rivera, 2003), se desarrolló la herramienta de ayuda a un lexicógrafo a la investigación de las estructuras del diccionario con el fin de detectar y corregir los círculos viciosos cortos a través de búsqueda automática de las primitivas semánticas y del vocabulario definidor.

La herramienta proporciona la siguiente información:

- Muestra un visor del Diccionario (en nuestro caso usamos el diccionario del español del grupo Anaya) de manera electrónica, donde muestra entre otras cosas la palabra del diccionario, su definición, la definición normalizada con las partes de oración de todas las palabras y, como una opción, también con sentidos de las palabras marcados. Nótese que el diccionario fue preprocesado usando la herramienta de análisis morfológico automático (Gelbukh and Sidorov, 2003).
- En este mismo visor nos muestra de las palabras existentes en la base de datos, el número de homónimos y de significados que tienen y cual es la parte de oración que generalmente juega.
- En este mismo apartado cuenta con los servicios de búsqueda, ya sea de manera común insertando la palabra a buscar en un cuadro de texto o bien búsquedas avanzadas en formato SQL, donde se puede dar los criterios de la búsqueda como, por ejemplo, parte de oración, etc.
- La funcionalidad de este modulo consiste en poder añadir más palabras con su significado o corregir la definición existente. En base del analizador morfológico de manera automática se transforman las palabras en definición en la forma normalizada. Existe la posibilidad de corregir manualmente los resultados de análisis automático.
- Para cada palabra se muestra la información si es la palabra primitiva o la palabra de vocabulario definidor. Se puede agregar o quitar la palabra seleccionada en la lista de definidores y observar la estadística de los círculos viciosos para todo el diccionario y para la palabra seleccionada. El algoritmo que se aplica es el algoritmo presentado en (Gelbukh and Sidorov, 2002), (Rivera, 2003).

Ejemplos de los ciclos presentados:

1) acción<117> -> impacto<6563> -> impresión<6601> -> cuerpo<3445> -> objeto<8643> -> ejercicio<4458> -> acción<117>
2) acción<119> -> obra<8666> -> acción<119>
3) aceite<126> -> bacalao<1418> -> aceite<126> etc.

4 Conclusiones

Se presentó el sistema que da la flexibilidad a un lexicógrafo de poder el mismo manualmente establecer a algunas palabras candidatas como primitivas semánticas puras o miembros del vocabulario definidor, esto con un análisis y criterios válidos

para poder considerarla como tal. El sistema permite agregar las palabras con sus definiciones al diccionario explicativo y corregir las definiciones existentes con análisis automático de los círculos viciosos. Sin embargo, las decisiones finales se toman por el lexicógrafo en base de la información proporcionada por el sistema.

Referencias

1. Apresjan, J. (2000) *Semantic Lexicography*, Oxford University Press. New York. 286 pp..
2. García Quesada, M. (2001). *Estudios de lingüística Española*. Universidad de Granada. Volumen 14 (2001) http://elies.rediris.es/elies14/index.html.
3. Gelbukh, Alexander and Grigori Sidorov. (2003) Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: *Computational Linguistics and Intelligent Text Processing. Proc. CICLing-2003, 4th International Conference on Intelligent Text Processing and Computational Linguistics*, February 15–22, 2003, Mexico City. Lecture Notes in Computer Science N 2588, Springer-Verlag, pp. 215–220.
4. Gelbukh, Alexander and Grigori Sidorov. (2002) Automatic Selection of Defining Vocabulary in an Explanatory Dictionary. *Proc. CICLing-2002, Conference on Intelligent Text Processing and Computational Linguistics*, February 16–23, 2001, Mexico City. Lecture Notes in Computer Science N 2276, Springer-Verlag, pp. 300–303.
5. Evens, M. N. (ed.), (1988). *Relational models of lexicon: Representing knowledge in semantic network.* Cambridge: Cambridge University Press.
6. Kozima, H. and Furugori, T. (1993). Similarity between words computed by spreading activation on an English dictionary. Proc. *6th conf. of the European chapter of ACL*, pp. 232–239.
7. Rivera, G. (2003) *Selección automática de primitivas semánticas para un diccionario explicativo del idioma español*, Tesis de Maestría (director A. Gelbukh, co-director G. Sidorov), CIC, IPN, México D.F.
8. Wierzbicka, A. (1980), *Lingua Mentalis: The semantics of natural language*. New York: Academic Press.
9. Wierzbicka, A. (1990), Prototypes save: on the uses and abuses of the notion of "prototypes" in linguistics and related fields, in S. L. Tsohatzidis (ed.), *Meanings and Prototypes: Studies in Linguistic Categorization*. London: Routledge & Kegan Paul.
10. Wierzbicka, A. (1996), *Semantics: Primes and Universals*. Oxford: Oxford University Press.