Información colocacional y Recuperación de la Información

Margarita Alonso Ramos
Universidade da Coruña
Campus de Elviña s/n, 15071 CORUÑA
lxalonso@udc.es

Leo Wanner
Universty of Stuttgart/ Univeristat Pompeu Fabra
wanner@informatik.uni-stuttgart.de

Título: *Optimización de la indexación semántica por medio de información colocacional*
Entidad: Xunta de Galicia (PGIDIT02PXIB30501PR)
Grupos participantes: Grupo DICE (Universidade da Coruña) y Colaborador externo Leo Wanner (University of Sttugart y Universitat Pompeu Fabra)
Investigador responsable: Margarita Alonso Ramos (lxalonso@udc.es)

1 Objetivo del proyecto

Este proyecto tiene como objetivo investigar la significación de la información colocacional en Recuperación de la información (RI). Una *colocación* es una relación entre dos unidades léxicas (UL) L_1 y L_2 tal que para expresar un significado específico en relación con L_1 la elección de L_2 no es libre (Mel'čuk 1995). Así, para expresar el sentido 'hacer lo que está previsto que se debe hacer' en relación con la UL CARGO, podemos decir *desempeñar un cargo*. El mismo sentido predicado de PROMESA se expresaría por CUMPLIR y en combinación con SECRETO, por GUARDAR. El peso semántico de L_1 y L_2 no está equilibrado: mientras L_1 (la *base* de la colocación) guarda el sentido que tiene cuando no está en colocación, el significado de L_2 (el *colocativo*) está reducido o modificado cuando aparece en colocación

Las siguientes propiedades de las colocaciones las hacen interesantes para RI: 1) A la hora de indexar un texto, los elementos de una colocación no pueden ser considerados como términos aislados: es evidente que la aportación de ABANDONAR en *abandonó al niño*, donde la UL es usada libremente, no es la misma que la aportada en la colocación *abandonó el cargo*; 2) Una base dada puede formar varias colocaciones con el mismo significado. Así, el nombre CARGO no sólo selecciona ABANDONAR, sino también RENUNCIAR, DIMITIR (de) o CESAR (en). 3) Las colocaciones pueden der clasificadas según una tipología semánticamente motivada.

Nuestra hipótesis es que estas propiedades deben tener ciertos efectos en los resultados de la RI. A pesar de que en los últimos años se registra una fuerte tendencia a utilizar información lingüística para RI, no se han explorado hasta ahora las colocaciones. La razón se debe a que no existía ningún procedimiento fiable de reconocimiento automático de colocaciones en los documentos. Sin embargo, algunos trabajos previos basados en la tipología de las *funciones léxicas* (FL) (Wanner y Alonso 2001 y Wanner en preparación) han mostrado la posibilidad de identificar automáticamente colocaciones verbo-nombre. Actualmente hemos extendido el mismo enfoque para reconocer colocacaciones nombre-adjetivo y verbo-adverbio. En nuestra investigación sobre RI, usamos esta estrategia para la identificación de colocaciones tanto en los documentos como en las consultas de los usuarios y pretendemos usar la información colocacional para mejorar los resultados de RI.

2 Colocaciones en los documentos

La aparición de colocaciones en los documentos no es un fenómeno excepcional. En la colección de documentos CLEF 2002 que estamos anotando colocacionalmente (Peters 2002), hemos comprobado su gran abundancia. A modo de ilustración, en cien documentos sobre la crisis gubernamental en España durante los años ochenta, el nombre *crisis* aparece cien veces de las cuales sesenta en colocación. En otros términos, el 60% de veces, el significado del término que funciona como colocativo es diferente de su significado aisladamente. La tabla siguiente lista los colocativos que coocurren con el nombre CRISIS en nuestra colección y el tipo de colocación en términos de FL.

FL	colocativo
Magn	grave, profunda
$CausFunc_0$	provocar, desatar, crear, abrir
$Liqu\ Func_0$	atajar, combatir, superar, resolver, solucionar, poner fin, zanjar
$Oper_1$	sufrir, vivir, pasar, atravesar
$CausOper_1$	sumir
$LiquOper_1$	sacar
$Real_2$	hacer frente
$CausContFunc_0$	prolongar
CausPredPlus	agravar
$Incep\ Func_0$	estallar

Tabla 1: Colocaciones con CRISIS

Otro dato importante concierne a la cuasi-sinonimia de las colocaciones. En nuestra colección de documentos, encontramos *renunciar al escaño/ al cargo/ a la presidencia*; *abandonar la presidencia/ el puesto/ el escaño/ el cargo*; *dimitir del cargo/ del escaño/ como presidente*. El significado de todas estas colocaciones es aproximadamente el mismo y puede ser etiquetado por la FL $FinReal_1$, cuya glosa sería 'X ya no hace lo que está previsto que debe hacer'. Estos datos muestran en qué medida una buena descripción del contenido de un documento depende de una indexación que tenga en cuenta las colocaciones.

3 Enfoque y estado del proyecto

Usamos la información colocacional durante la indexación y el tratamiento de la consulta. Para registrar la información en el índice, procedemos del modo siguiente: **(1)** Todos los valores de una FL **f** con la misma base B son reducidos al término **f**(B). Así, todas las apariciones de *renunciar*, *dimitir*, *abandonar*, *dimitir* y *cesar* con *cargo* son representadas en el índice como $FinReal_1$(*cargo*). **(2)** Para no perder los valores individuales de la FL, los listamos también en el índice pero guardando la información de que son colocativos: *renunciar* <$FinReal_1$(*cargo*)>. **(3)** Las bases son indexadas como términos ordinarios. Entre diferentes bases de la misma FL que aparecen en contextos similares se establecen *relaciones semánticas latentes* con un procedimiento similar al de Deerwester et al. (1990).

Las colocaciones pueden ser explotadas también para las consultas del usuario. Distinguimos dos tipos de operaciones: (i) detectar las colocaciones en la consulta y reducirlas a la forma **f**(B); (ii) expandir ciertos términos aislados de la consulta por las colocaciones adecuadas. Con respecto al primer aspecto, la detección de la colocación *abandonar la presidencia* en una consulta y su etiquetación por $FinReal_1$(*presidencia*) discriminaría la búsqueda de documentos y filtraría documentos que hablen de abandonos de niños, pongamos por caso. Con respecto a la expansión, se trata de expandir un término aislado como *dimisión* por una colocación de verbo soporte del tipo $Oper_1$ como *presentar la dimisión*. Actualmente, estamos desarrollando experimentos para evaluar el impacto que la consideración de la información colocacional tiene en la calidad de la recuperación de documentos. Esperamos que los experimentos muestren que gracias a la información colocacional, somos capaces de (i) recuperar documentos que no contengan todos los términos de la consulta, pero que sean relevantes para el tema; (ii) evitar recuperar documentos que contengan algunos de los términos de la consulta, pero que no sean relevantes.

Referencias

Deerwester, S., S.T.Dumais, G.W.Furnas, T.K. Landauer y R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 321-407.

Mel'čuk, I. 1995. Phrasemes in Language and Phraseology in Linguistics. En M. Everaert et al. (eds.), *Idioms*, Lawerence Erlbaum, Hillsdale, pp. 167-232.

Peters, C. (ed.).2002. *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*. Roma.

Wanner, L. y M. Alonso. 2001. Vers une approche sémantique pour l'identification des collocations en corpus. En *Actes: Journée d'études de l'ATALA: la Collocation*, Paris.

Wanner, L. (en preparación), Towards automatic fine-grained semantic classification of verb-noun collocations.