PROPUESTA PARA UN ANALISIS AUTOMATICO DEL CONTENIDO EN DOCUMENTACION DE EST

Félix de Moya (Universidad de Granada)
José Manuel Muñoz (Universidad de Córdoba)
Pedro Hípola (Universidad de Granada)

Moya Anegón, F.; Muñoz Muñoz, J. M.; Hípola, P. (1988). Propuesta para un análisis automático del contenido en documentación de EST. In M. Laurén, Christer; Nordman (Ed.), *Special Language: From Humans Thinking To Thinking Machines* (pp. 448-452). Clevedon: Multilingual Matters Ltd. Clevedon.

http://books.google.es/books?id=oKRVbYGv3fEC&lpg=PA3&ots=r82HshWzX4&lr&hl=es&pg=PA448#v=onepage&q&f=false

Las caracter�sticas del sistema que presentamos en estas p�ginas han determinado la participaci�n en �l de un grupo de trabajo con car�cter marcadamente interdisciplinar. Se ha tratado de hacer confluir la tarea de estudio sobre materiales de EST para fines docentes con las t�cnicas usuales en indizaci�n autom�tica. Los componentes del grupo coinciden en unos intereses de trabajo comunes, encaminados a obtener el aislamiento de t�rminos especialmente relevantes dentro de textos de Ciencia y Tecnolog�a escritos en ingl�s, que sirvan en primer lugar para el establecimiento y puesta al d�a de vocabularios utilizables en la ense�anza del ingl�s con fines espec�ficos, los cuales a su vez puedan ser utilizados dentro de una aplicaci�n que se encargue de realizar la indizaci�n documental de forma autom�tica sobre palabras clave extra�das de los textos escritos que se presenten al sistema.

La metodolog�a empleada para cubrir los objetivos definidos se basa en procedimientos usuales dentro del �mbito de la estad�stica ling��stica, especialmente los m�s extendidos en el tratamiento de unidades l�xicas, completados con la utilizaci�n del an�lisis morfogr�fico necesario. En la base del funcionamiento del sistema se encuentra el recurso a la comparaci�n de las frecuencias del l�xico empleado en textos ingleses de Ciencia y Tecnolog�a con las frecuencias del vocabulario de la lengua standard, para efectuar as� el aislamiento de los t�rminos que se desea localizar. Establecidos de manera semiautom�tica repertorios l�xicos pertenecientes a dominios tem�ticos espec�ficos, los vocabularios obtenidos pueden ser utilizados como objeto de estudio ling��stico, como auxiliar en la ense�anza de "English for Specific Purposes" y adem�s como t�rmino de comparaci�n<1> para el siguiente objetivo, el de la indizaci�n autom�tica.

Por este tipo de procedimiento se facilita en primer lugar que la selecci�n del vocabulario que se va a usar en la ense�anza de ESP se realiza no s�lo a trav�s de un criterio de estimaci�n subjetiva por parte del profesor, sino con la ayuda de un

estudio objetivo, que le orienta en la determinaci�n de qu� tipo de material se debe emplear y adem�s le ofrece datos para establecer prioridades en el l�xico que se ense�a. Por otra parte, desde el punto de vista documental el an�lisis que se produce podr�a ser clasificado, dentro del esquema que ha se�alado Van Slype<2>, como "indizaci�n autom�tica selectiva en lenguaje natural".

Se ha recurrido a algoritmos y criterios de ordenaci�n del material basados en la ley de Zipf y en la f�rmula del test de Pearson, adem�s de diversos c�lculos de probabilidades. Las herramientas inform�ticas empleadas est�n constituidas por un sistema de gesti�n de base de datos relacional. Para preparar las rutinas de tratamiento del material ling��stico se ha utilizado el lenguaje de comandos incorporado en el mismo gestor.

Presentamos en esta comunicaci�n el proceso de preparaci�n de uno de los vocabularios de tema espec�fico, el correspondiente a textos sobre temas inform�ticos, realizado a partir de un corpus de m�s de doscientos cincuenta mil tokens, del que, desechadas f�rmulas, nombres de variables, etc., quedaron aproximadamente unas ciento cincuenta mil formas l�xicas. El inter�s de contar con estudios serios de vocabulario es especialmente patente en el �rea tem�tica de la computaci�n, por el alto nivel de especificidad l�xica que caracteriza a los textos de este �mbito, su considerable ritmo en la creaci�n de nuevos vocablos y por la gran cantidad de pr�stamos que en este dominio cient�fico suelen tomar del ingl�s otras lenguas, entre ellas el espa�ol.

Tras llevar a cabo la recopilaci�n de los textos que componen el corpus, se realiz� la carga de los mismos y, previa depuraci�n formal de diversos caracteres, se efectu� el recuento de frecuencias de cada una de las formas l�xicas<3>. Independientemente de esto, se ha procedido a la carga del diccionario de frecuencias elaborado por Kucera y Francis<4>. Los datos as� extra�dos fueron incorporados al sistema completo de base de datos. Por otra parte, de forma manual, se ha introducido en cada registro un campo que contiene un puntero que remite, si procede, a la forma can�nica correspondiente, o, en caso contrario, una marca que indica que se trata de un t�rmino lexical can�nico. El sistema de punteros no est� s�lo concebido para agrupar flexiones de un mismo lema, sino tambi�n con vistas a compilar grupos de palabras que, generalmente por estar construidas en torno a la misma lex�a, constituyen un subconjunto l�xicosem�ntico uniforme cuya reducci�n simplifica el proceso de indizaci�n documental<5>. Por otra parte, ese mismo campo se ha utilizado para anotar diversas informaciones de tipo gramatical y lexicogr�fico tales como marcas para identificar acr�nimos, palabras "vac�as", etc.

En lo que se refiere al establecimiento del antidiccionario de palabras vac�as, se han seguido diversos criterios. El planteamiento de la cuesti�n parti� de considerar qu� tipo de sistema se deseaba implementar. En primer lugar fueron marcadas como vac�as las formas habitualmente consideradas como de valor primordialmente gramatical o al menos de escasa relevancia significativa: preposiciones, conjunciones, determinantes, verbos auxiliares y modales, c�pulas, etc. As� mismo se ha intentado se�alar el mayor n�mero posible de palabras "atem�ticas", de sentido impreciso o excesivamente general, utilizando tambi�n como punto de referencia

criterios estad�sticos. Resulta delicado establecer el grupo de t�rminos "omnibus" que deben formar parte del antidiccionario, pues es dif�cil determinar el grado de generalidad de las palabras. En nuestro caso, cuando se ha planteado una duda sobre si considerar o no vac�o alg�n t�rmino, en la mayor�a de las ocasiones se ha preferido optar por incluirlo en el antidiccionario, pues el tipo de indizaci�n que se pretende alcanzar por ahora es de tipo "indicativo", m�s que "informativo". Esto es, se trata de obtener un conjunto de palabras clave lo m�s reducido posible que consigne los aspectos m�s esenciales del contenido del documento que se procese, no de hacer una valoraci�n exhaustiva. As�, se han considerado "atem�ticos" t�rminos como _get_, _go_, _take_, _give_, _make_, etc., e incluso otros de car�cter polis�mico, como _even_, una de cuyas acepciones (_even number_) puede resultar relevante en el an�lisis de determinados documentos.

A la hora de evaluar el sistema se han observado determinadas limitaciones, especialmente en lo que se refiere a la gesti�n de homograf�as -y por tanto a la red de reenv�os entre flexiones y lexicales can�nicas-, que son origen de diversos ruidos en el proceso. Otro tipo de problemas surgen del diccionario de frecuencias que se ha seleccionado como t�rmino de comparaci�n, pues por las especiales caracter�sticas del corpus procesado no resulta un reflejo exacto del ingl�s standard. Sin embargo, nos parece que los resultados alcanzados a trav�s de la metodolog�a que hemos empleado se pueden considerar satisfactorios, pues se obtienen unas indizaciones con un grado de ruido no muy elevado y de suficiente exhaustividad, todo ello por procedimientos autom�ticos, lo que supone una soluci�n a la falta de coherencia -y a la carga econ�mica- que suele caracterizar el an�lisis de los indizadores humanos. Dadas las caracter�sticas del sistema, su actualizaci�n resulta posible sin excesiva intervenci�n humana. Y esto aparece tanto m�s necesario cuanto m�s se considera cu�l es la rapidez con la que se modifica el subsistema ling��stico de la Ciencia y la Tecnolog�a. La creciente disponibilidad de material escrito en soporte legible por ordenador aumenta las facilidades para hacerlo. La opini�n de los autores es que la investigaci�n actual, en lo que se refiere a los proyectos de la inteligencia artificial, no deber�a marginar, a la hora de preparar bases de conocimientos para sistemas int�rpretes de lenguaje natural, los avances cl�sicos de la estad�stica ling��stica.

NOTAS

(1) La utilizaci&#65533;n de este tipo de t&#65533;cnicas ha sido muy experimentada durante las &#65533;ltimas cuatro d&#65533;cadas con vistas a desarrollar una "estil&#65533;stica computacional". Puede verse el reciente art&#65533;culo de YANG HUIZHONG, "A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts", en _Literary and Linguistic Computing_, vol. 1, n. 2, Oxford University Press, 1986.

(2) Cfr. VAN SLYPE, G., _Les langages d'indexation_, Paris, 1987, p&#65533;g. 168 y ss. La indizaci&#65533;n selectiva, bien en lenguaje natural, bien en lenguaje controlado, se caracteriza porque "toma en cuenta s&#65533;lo ciertos t&#65533;rminos, considerados por el algoritmo del sistema como los m&#65533;s representativos del contenido del documento" (p&#65533;g. 169).

(3) Puede encontrarse una información más detallada en: Moya Anegón, F.; Muñoz Muñoz, J. M.; Hípola, P."Análisis automático de documentación técnica informática", V Congreso de la AESLA, Pamplona, 1987.

(4) KUCERA, H. y FRANCIS, W. N., _Computational Analysis of Present-day American English_, Rhode Island, 1967.

(5) Deweze utiliza el t&#65533;rmino "forma sem&#65533;ntica" para referirse a este tipo de agrupamientos. Cfr. A. DEWEZE, _Informatique documentaire_, Paris, 1985, p&#65533;g. 129 y ss.