

# La lematización en español: una aplicación para la recuperación de información

Raquel Gómez Díaz

EDICIONES TREA, S. L.

## ÍNDICE

Prólogo.....	21
Introducción.....	25
<b>PARTE I:</b>	
<b>LA RECUPERACIÓN DE LA INFORMACIÓN .....</b>	<b>31</b>
<b>1. Concepto de recuperación de información .....</b>	<b>33</b>
1.1. Distinción entre recuperación de información, recuperación de datos y recuperación de documentos .....	36
<b>2. Historia de la R. I.....</b>	<b>37</b>
2.1. Foros de encuentro más importantes de la R. I.....	39
2.1.1. TREC: Text Retrieval Conference .....	39
2.1.2. CLEF: Cross Language Evaluation Forum .....	41
2.1.3. Otros foros de encuentro .....	42
2.1.3.1. CIKM: Conference on Information and Knowledge Management .....	42
2.1.3.2. CIVR: International Conference on Image and Video Retrieval .....	43
2.1.3.3. ISMIR: The International Conferences on Music Information Retrieval and Related Activities..	43
2.1.3.4. ECIR: European Conference on Information Retrieval Research .....	43
2.1.3.5. MLDM: International Conference on Machine Learning and Data Mining .....	44

2.2.	Grupos de trabajo de R. I.....	44
2.2.1.	Grupos de investigación españoles.....	44
2.2.2.	Grupos de investigación europeos.....	45
2.2.3.	Grupos de investigación estadounidenses.....	46
<b>3.</b>	<b>Los modelos de R. I.....</b>	<b>49</b>
3.1.	Evolución de los sistemas de recuperación.....	49
3.2.	Clasificaciones de los modelos de recuperación.....	50
3.2.1.	La clasificación de Belkin.....	50
3.2.2.	La clasificación de Baeza-Yates y Ribeiro Neto.....	52
3.2.3.	La clasificación de Dominich.....	53
3.2.4.	La clasificación de Heting Chu.....	53
3.3.	Los modelos de recuperación.....	54
3.3.1.	Modelo booleano.....	54
3.3.2.	Lógica difusa.....	55
3.3.3.	Modelo booleano extendido.....	56
3.3.4.	Modelo de espacio vectorial.....	56
3.3.5.	Modelo de espacio vectorial generalizado.....	62
3.3.6.	Indización semántica latente.....	62
3.3.7.	Modelo basado en la idea del «Cluster».....	63
3.3.8.	Spreading dissemination.....	64
3.3.9.	Redes neuronales.....	65
3.3.10.	Modelo probabilístico.....	66
3.3.11.	Modelo lógico.....	68
3.3.12.	Modelo gráfico.....	69
3.3.13.	Browsing.....	69
3.3.14.	Los algoritmos genéticos.....	70
3.3.15.	Modelos relacionados con el procesamiento del lenguaje natural.....	72
3.3.15.1.	Definición de P.L.N.....	72
3.3.15.2.	Niveles del P.L.N.....	73
3.3.15.3.	La R. I. y el P.L.N.....	74
3.3.15.4.	Líneas de investigación aplicadas a la R. I. ...	75
3.3.15.5.	Algunas aplicaciones de P.L.N. a la R. I. ....	77
3.3.16.	Los sistemas expertos.....	79
<b>4.</b>	<b>Las palabras vacías en la R. I. ....</b>	<b>81</b>
<b>5.</b>	<b>La evaluación de la R. I.....</b>	<b>83</b>

5.1.	La relevancia .....	85
5.1.1.	El concepto de relevancia .....	85
5.1.2.	El cálculo de la relevancia .....	87
5.2.	Las colecciones experimentales .....	88
5.2.1.	El text de Cranfield .....	89
5.2.2.	Las colecciones TREC y CLEF.....	90
5.3.	Principales medidas de evaluación.....	91
5.3.1.	Medidas orientadas a los procesos .....	91
5.3.1.1.	Evaluación de la base de datos .....	91
5.3.1.2.	Evaluación de la consulta .....	91
5.3.2.	Medidas orientadas al resultado .....	92
5.3.2.1.	La precisión .....	93
5.3.2.2.	La exhaustividad .....	95
5.3.2.3.	Relación entre precisión y exhaustividad .....	97
5.3.2.4.	Medidas complementarias a la precisión y la exhaustividad.....	99
5.3.2.5.	Medidas relacionadas con el usuario .....	102
6.	<b>La R. I. en español: experimentos más significativos</b> .....	105
6.1.	La R. I. en español en las TREC .....	106
6.2.	La R. I. en español en el CLEF .....	109

## PARTE II:

<b>LA LEMATIZACIÓN</b> .....	115
<b>7. La lematización</b> .....	117
7.1. Introducción .....	117
7.2. Definición y problema de uso del término .....	117
7.3. La necesidad de lematizar .....	122
<b>8. Tipos de algoritmos de lematización</b> .....	125
8.1. Según el tipo de sufijos que aplican: flexivos y derivativos .....	125
8.2. Por el modo de establecer la lematización .....	126
8.3. Por el modo de establecer la confluación.....	126
8.4. En función del conocimiento lingüístico .....	129
<b>9. Problemas de la lematización</b> .....	131
<b>10. La evaluación de los sistemas de lematización</b> .....	133

10.1.	La corrección de la lematización .....	133
10.2.	La correcta ejecución de la compresión .....	134
10.3.	La efectividad en la recuperación .....	134
10.4.	El tiempo .....	135
11.	<b>Principales algoritmos de lematización para el inglés.....</b>	<b>137</b>
11.1.	Algoritmo de Lovins .....	138
11.2.	Algoritmo de Salton .....	138
11.3.	Algoritmo de Dawson .....	138
11.4.	Algoritmo de Porter .....	139
11.5.	Algoritmo de Krovetz.....	141
11.6.	Comparación de algoritmos para el inglés.....	142
12.	<b>La lematización en idiomas distintos del inglés.....</b>	<b>143</b>
13.	<b>La lematización en español.....</b>	<b>147</b>

### PARTE III:

UN LEMATIZADOR PARA EL ESPAÑOL .....	149
14. El lematizador para el español.....	151
14.1. Introducción .....	151
14.2. Objetivos .....	151
14.3. Antecedentes del trabajo .....	152
15. La formación de palabras en español .....	153
15.1. Dificultades del estudio de la derivación en español .....	154
15.2. Clasificación de los sufijos .....	156
15.3. Procesos de sufijación .....	157
15.4. Reglas de sufijación.....	157
16. Los autómatas de estados finitos .....	161
16.1. Definición de autómata y máquina de estados finitos.....	161
16.2. Diagramas y tablas de transiciones.....	162
16.3. Tipos de autómatas y máquinas de estados finitos .....	163
17. Consideraciones previas a la creación del lematizador.....	165
17.1. Los acentos .....	165
17.2. Los prefijos .....	166
17.3. La estructura de las palabras.....	166
17.4. La elección de los sufijos .....	168

17.5. Los criterios de selección de los lemas .....	169
<b>18. La creación del lematizador .....</b>	<b>171</b>
18.1. La creación de las reglas.....	171
18.2. La lematización manual .....	175
18.3. El funcionamiento del lematizador .....	175
<b>19. Aplicación del lematizador a la R. I.....</b>	<b>183</b>
19.1. La creación de la colección experimental .....	183
19.2. Las palabras vacías en los experimentos.....	184
19.3. El sistema de recuperación .....	185
19.3.1. Proceso de lematización .....	185
19.3.2. Proceso de indización.....	186
19.3.3. Proceso de recuperación.....	186
19.4. Los experimentos .....	187
19.4.1. Sin lematizar .....	187
19.4.2. Lematización derivativa .....	188
19.4.3. Lematización flexiva .....	188
<b>20. La evaluación de la lematización .....</b>	<b>191</b>
20.1. La corrección de la lematización .....	191
20.2. La compresión .....	191
20.3. La evaluación de la recuperación .....	192
20.3.1. Precisión .....	193
20.3.2. Exhaustividad .....	195
20.3.3. Precisión-exhaustividad .....	196
20.4. Conclusiones .....	198
<b>21. Comparación del lematizador del español con otros lematizadores .....</b>	<b>201</b>
<b>22. Otras aplicaciones de la lematización .....</b>	<b>205</b>
<b>Glosario de términos .....</b>	<b>207</b>

## ANEXOS

<b>Anexo I. Terminaciones flexivas y derivativas.....</b>	<b>215</b>
<b>Anexo II. Terminaciones flexivas .....</b>	<b>217</b>
<b>Anexo III. Lista de palabras vacías según categorías gramaticales (vacías leve). .....</b>	<b>218</b>

<b>Anexo IV.</b> <i>Lista de palabras vacías según categorías gramaticales y alta frecuencia de aparición (vacías fuerte)</i> .....	221
<b>Bibliografía</b> .....	227

## PRÓLOGO

La recuperación de información es un tema candente. No hay más que ver cómo aumenta el número de libros nuevos sobre el tema; cómo aparecen nuevas revistas dedicadas expresamente a la recuperación de información, o cómo incluso otras más genéricas cada vez incluyen más artículos sobre esta; cómo, en fin, los congresos, simposios, seminarios y similares se hacen ya imposibles de seguir, salvo que uno tuviera el don de la ubicuidad.

La causa es clara: además de la consabida explosión documental (ya se sabe, la cantidad de documentos crece, no sé si exponencialmente, pero en cualquier caso, muchísimo), fundamentalmente obedece a dos factores: la disponibilidad creciente de documentos en formato electrónico, y la formidable capacidad de difusión que proporciona Internet. Esa disponibilidad de documentos íntegramente en formato electrónico posibilita su procesamiento automático, con programas informáticos, y ello hace que podamos plantearnos la total automatización del proceso de recuperación.

Esta automatización supone, entre otras cosas, la superación de la indización manual; esto es, la construcción manual de representaciones más o menos formales del contenido de cada uno de los documentos. Supone también la superación de los lenguajes controlados y afines, utilizados para construir dichas representaciones; antes bien, la automatización del proceso tiene mucho que ver con el uso del lenguaje natural en sus dos vertientes: la del propio documento, por supuesto, y la del usuario cuando trata de expresar sus necesidades de información.

De manera que la recuperación de información contemporánea tiene poco que ver con tesauros, encabezamientos de materias o códigos de diversa especie; y mucho menos con la asignación manual de estos para describir el contenido de documentos que están en su integridad en formato electrónico y a los que se puede



acceder directamente a través de Internet. Si tuviéramos que esperar a que se realizase la indización manual de tales documentos para poder recuperarlos, no los encontraríamos nunca, por la sencilla razón de que la velocidad de aparición de documentos es mucho mayor que nuestra capacidad ideal de indizarlos manualmente. Esto, sin entrar en cosas como la inconsistencia de la indización manual, bien conocida, por otra parte, desde hace años.

Así que lo que necesitamos son programas informáticos que sean capaces de llevar a cabo estas tareas de forma automática. La mayor parte de estos programas utilizan los términos o palabras que componen los documentos para representar sus contenidos; y, también, las palabras que utiliza el usuario para expresar su necesidad informativa, esto es, cuando compone una consulta en lenguaje natural. Sucede que, con frecuencia, una misma palabra toma distintas formas muy parecidas, pero no exactamente iguales; los ordenadores no son inteligentes como las personas, y tienen cierta dificultad en entender que *biblioteca* y *bibliotecas*, por ejemplo, son casi lo mismo; y que, probablemente, si uno busca información sobre *catalogación* y en un documento se habla de *catálogos* o de *catalogadores*, probablemente trate de aquello que estábamos buscando.

Además, muchos de esos programas que venimos comentando basan su eficacia en que son capaces de discriminar la importancia de unos términos frente a otros. En un documento concreto (o en una consulta de un usuario), los términos no representan por igual el contenido temático de dicho documento (o consulta); unas palabras son más representativas que otras. Los programas asignan más o menos importancia o representatividad a cada término mediante diversas técnicas, pero la mayor parte de estas aplican cálculos que utilizan las frecuencias de aparición de las palabras en diversos contextos. Naturalmente, esas frecuencias no resultan igual si consideramos *biblioteca* y *bibliotecas*, o *catalogación*, *catálogo*, *catalogador* palabras distintas.

Desde hace ya bastantes años se pensó que se podrían mejorar los resultados de estos programas de recuperación si se conseguía agrupar bajo una única forma esas palabras que, con ligeras diferencias, hacían referencia a un mismo concepto. Los investigadores llamaron a esta operación *stemming*, en alusión a esa raíz común a las palabras que se agrupan. No hay una buena traducción al español, aunque muchos hemos optado por la expresión *lematización*; esta opción, por cierto, no tiene una aceptación unánime, y algunos de nuestros colegas se enfadan cuando la encuentran en artículos, ponencias y demás (hay quien piensa que el problema está en que algunos de nuestros colegas son demasiado propensos al enfado).

Sin embargo, se llame como se llame, no es una operación fácil para un programa informático. Se han ensayado numerosas técnicas, la mayor parte dependientes o válidas para un idioma concreto (el inglés en la mayoría de los casos, naturalmente). Para el caso del español ha habido pocos intentos, y de eso trata este libro: de la lematización en español.

Si comparamos el español (y, en general, las lenguas procedentes del latín) con otras, veremos que nos enfrentamos a una riqueza y complejidad morfológicas importantes. El ejemplo más evidente es el de los verbos irregulares; estos (y sus conjugaciones) son la pesadilla de cualquier estudiante de español como lengua extranjera, y en cualquier diccionario encontraremos varios millares de ellos, cada uno de los cuales puede dar lugar a una buena cantidad de flexiones (modos, tiempos, personas y número).

Del mismo modo, nos encontramos con la gran variedad de palabras derivadas que se forman añadiendo determinados sufijos a una raíz dada. El problema es que, además del número de sufijos que tenemos en la lengua española, estos se pegan a las raíces de las más variadas formas, modificando incluso dichas raíces hasta dejarlas, a veces, difícilmente reconocibles. De manera que no es posible abordar la lematización en español sin codificar el complejo y abundante conocimiento lingüístico que explica tanto las variaciones de las palabras a través de las flexiones morfológicas como a través de la derivación. Y esto es, precisamente, una parte importante de lo que Raquel Gómez nos presenta en este trabajo.

Pero es que, además, la presunción de que la lematización mejora los resultados de la recuperación no es más que eso: una mera presunción. Presunción, por cierto, que, aunque fundada, ha sido puesta en discusión en numerosas ocasiones. Nuevamente estamos ante una cuestión dependiente del idioma, de manera que lo que produce determinados resultados recuperando documentos en inglés no tiene por qué funcionar de la misma manera recuperando documentos en español. Así que no se trata solamente de lematizar, sino de aplicar la lematización a la recuperación de información. Esta es otra parte importante del trabajo de Raquel Gómez: la investigación experimental, a fin de evaluar el impacto o incidencia de la lematización en los resultados de la recuperación en español.

Y termino refiriéndome a las otras partes de este trabajo que, sin ser tan novedosas, resultan también inestimables. Estoy hablando de las partes previas o introductorias en las que, en pocas páginas, se nos presenta de una forma realmente