

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla

Extracción de Opiniones sobre Características: Un Enfoque Práctico Adaptable al Dominio

Fermín L. Cruz Mata, 31719167M
fcruz@us.es

Dirigido por Prof. Dr. José A. Troyano



Memoria de Tesis Doctoral

Índice general

I	Preámbulo	17
1.	Introducción	19
1.1.	Motivación y contexto	19
1.2.	Hipótesis	21
1.3.	Resumen de la propuesta	22
1.3.1.	Resumen de aportaciones	22
1.4.	Estructura del documento	23
II	Antecedentes	25
2.	Procesamiento de textos subjetivos	27
2.1.	Introducción	27
2.1.1.	La web como fuente de opiniones	29
2.2.	Análisis del sentimiento y minería de opiniones	29
2.2.1.	Orígenes	30
2.2.2.	Aplicaciones	31
2.2.3.	Tareas	34
2.2.4.	Recursos	41
2.3.	Experiencias previas en el área	43
2.3.1.	Clasificación binaria de documentos basada en la opinión	44
2.3.2.	Construcción de lexicones de opinión orientados al do- minio	48
2.3.3.	Resumen automático de textos de opinión	56
3.	Extracción de opiniones sobre características	61
3.1.	Introducción	61
3.2.	Principales trabajos relacionados	62
3.2.1.	Primeros trabajos	62

3.2.2. La extracción de opiniones desde la perspectiva de la extracción de información	66
3.2.3. Otros trabajos	68
3.3. Resumen comparativo de trabajos	70
3.3.1. Extracción de características	70
3.3.2. Búsqueda de palabras de opinión	71
3.3.3. Clasificación de opiniones	72
3.3.4. Supervisado vs. no supervisado e influencia del dominio	73
3.3.5. Analizadores y recursos lingüísticos	74
3.3.6. Evaluación de la tarea	75
III Una propuesta para la extracción de opiniones	79
4. Extracción de opiniones sobre características adaptable al dominio	81
4.1. Introducción	81
4.2. Definición de la tarea	84
4.2.1. Definiciones previas	84
4.2.2. Alcance de la tarea	90
4.2.3. Definición formal de la tarea	91
4.2.4. Retos y dificultades de la tarea	94
4.3. Resumen de la propuesta	96
4.3.1. Sistema de extracción genérico	97
4.3.2. Recursos específicos del dominio	98
5. Recursos para la extracción de opiniones	101
5.1. Introducción	101
5.2. Corpus	102
5.2.1. Calidad de los documentos	102
5.2.2. Tamaño del corpus	103
5.2.3. Representatividad de los documentos	104
5.2.4. Definición del recurso	105
5.2.5. Sintaxis del recurso	106
5.3. Taxonomía de características	108
5.3.1. Definición del recurso	108
5.3.2. Sintaxis del recurso	108
5.4. Corpus anotado	110
5.4.1. Definición del recurso	110

5.4.2. Sintaxis de las anotaciones	110
5.5. Lexicón de opiniones	111
5.5.1. Definición del recurso	111
5.5.2. Sintaxis del recurso	113
5.6. Indicadores de características implícitas	114
5.6.1. Definición del recurso	114
5.6.2. Sintaxis del recurso	115
5.7. Patrones de dependencias	116
5.7.1. Definición del recurso	118
5.7.2. Sintaxis del recurso	121
5.8. Listas de expresiones especiales	122
5.8.1. Expresiones de negación	122
5.8.2. Expresiones de no-negación	123
5.8.3. Expresiones de polaridad dominante	124
6. Metodología para la generación de los recursos	125
6.1. Introducción	125
6.2. Recolección y procesado del corpus	126
6.2.1. Extracción de los documentos	127
6.2.2. Selección de los documentos	128
6.2.3. Procesado de los documentos	129
6.3. Extracción semiautomática de <i>feature words</i>	129
6.3.1. Descripción del algoritmo	130
6.3.2. Simulación de la extracción de <i>feature words</i>	132
6.3.3. Influencia de las restricciones morfosintácticas	133
6.3.4. Influencia del número de candidatos aceptados o rechazados	135
6.3.5. Influencia del número de nuevas semillas de opinión por iteración	137
6.3.6. Resultados de la extracción de <i>feature words</i> para el dominio <i>headphones</i>	138
6.4. Construcción de la taxonomía de características previa	141
6.4.1. Taxonomía de características previa para <i>headphones</i>	141
6.5. Anotación y validación de evidencias de opinión	144
6.5.1. Validación de sintaxis	145
6.5.2. Validación de las características	145
6.5.3. Validación de la polaridad	145
6.5.4. Validación de la cobertura	147

6.5.5. Corpus anotado para <i>headphones</i>	147
6.6. Refinamiento de la taxonomía de características	148
6.6.1. Refinamiento de la taxonomía de características para <i>headphones</i>	150
6.7. Inducción del lexicón de opiniones	153
6.7.1. Descripción del algoritmo	153
6.7.2. Inducción del lexicón de opiniones para <i>headphones</i>	156
6.8. Ampliación del lexicón de opiniones	158
6.8.1. Descripción del método de ampliación	159
6.8.2. Ampliación del lexicón de opiniones para <i>headphones</i>	160
6.9. Inducción de los indicadores de opiniones implícitas	166
6.9.1. Descripción del algoritmo	166
6.9.2. Inducción de los indicadores de opinión implícita para <i>headphones</i>	168
6.10. Inducción de los patrones de dependencias	172
6.10.1. Descripción del algoritmo	172
6.10.2. Influencia de las restricciones morfosintácticas en los patrones de dependencias	176
6.10.3. Patrones de dependencias para <i>headphones</i>	176

IV Un sistema de extracción de opiniones 181

7. TOES:

Un Sistema Extractor de Opiniones sobre Características Adaptable al Dominio	183
7.1. Introducción	183
7.2. Descripción de la arquitectura del sistema	184
7.2.1. Entidades participantes	185
7.3. Definición de componentes abstractos y concretos	186
7.3.1. Anotador de palabras de característica	187
7.3.2. Anotador de características implícitas	189
7.3.3. Enlazador de palabras de opinión	190
7.3.4. Separador de opiniones	194
7.3.5. Solucionador de opiniones superpuestas	196

7.3.6. Clasificador de opiniones	198
7.3.7. Componentes <i>simulados</i>	203
7.3.8. <i>Pipelines</i> de procesado	204
8. TOES: Evaluación y ajuste de parámetros	207
8.1. Introducción	207
8.2. Metodología de las evaluaciones	208
8.3. Evaluación individual de los componentes	209
8.3.1. Evaluación del anotador de palabras de característica basado en la taxonomía	209
8.3.2. Evaluación del anotador de características implícitas basado en PMI-IR	210
8.3.3. Evaluación del anotador de características implícitas basado en los indicadores	211
8.3.4. Evaluación del enlazador de palabras de opinión y de expresiones especiales basados en ventana	212
8.3.5. Evaluación del enlazador de palabras de opinión y expresiones especiales basado en dependencias	213
8.3.6. Evaluación del clasificador de opiniones basado en el lexicón	216
8.3.7. Evaluación de los clasificadores de opiniones basados en WordNet, SentiWordNet y PMI-IR	217
8.3.8. Evaluación del clasificador de opiniones basado en conjunciones	218
8.3.9. Evaluación de la combinación de clasificadores de opiniones	218
8.4. Evaluación del sistema completo	218
8.4.1. Evaluación de <i>pipelines</i> ligeros de recursos	219
8.4.2. Evaluación de <i>pipelines</i> basados en recursos	221
8.5. Resumen y análisis de los resultados	222
8.5.1. Importancia de los recursos específicos del dominio	223
8.5.2. Estimación de coste de la anotación y relación con la mejora del sistema	226
8.5.3. Dificultad de las opiniones implícitas	227
8.5.4. Dificultad de determinadas características	228
8.5.5. Precisión <i>vs.</i> cobertura	228
8.5.6. Resultados finales y dificultad de la tarea	228
8.5.7. Utilidad práctica del sistema	230

V Casos de aplicación y consideraciones finales	233
9. Casos de aplicación: hoteles y coches	235
9.1. Introducción	235
9.2. Generación de recursos	236
9.2.1. Corpus anotado y taxonomía de características	236
9.2.2. Inducción de recursos	239
9.3. Evaluación del sistema de extracción	244
9.3.1. Evaluación de <i>pipelines</i> ligeros de recursos	244
9.3.2. Evaluación de <i>pipelines</i> basados en recursos	245
9.4. Resumen de resultados	246
10. Conclusiones	247
VI Apéndices	267
A. Ejemplos de entradas y salidas de los componentes concretos de TOES	269
A.1. Anotador de palabras de característica basado en la taxonomía	269
A.2. Anotador de características implícitas basado en los indicadores	270
A.3. Anotador de características implícitas basado en PMI-IR . . .	272
A.4. Enlazador de palabras de opinión basado en ventana	272
A.5. Enlazador de palabras de opinión basado en dependencias . .	274
A.6. Enlazador de expresiones especiales basado en ventana	276
A.7. Enlazador de expresiones especiales basado en dependencias .	277
A.8. Separador de opiniones basado en conjunciones	278
A.9. Clasificador de opiniones basado en el lexicón de opiniones .	279
A.10. Clasificador de opiniones basado en PMI-IR	282
A.11. Clasificador de opiniones basado en WordNet	283
A.12. Clasificador de opiniones basado en SentiWordNet	283
A.13. Clasificador de opiniones basado en conjunciones	284
A.14. Solucionador de opiniones explícitas superpuestas	285
A.15. Solucionador de opiniones implícita-explícita superpuestas .	286
B. PolarityRank: justificación algebraica y convergencia	289
B.1. Definición	289
B.2. Justificación algebraica	290
B.3. Convergencia	293

Índice de figuras

2.1. Arquitectura de nuestro sistema de resumen	58
3.1. Ejemplo de resumen a partir de opiniones sobre características	63
3.2. Ejemplo de resumen a partir de opiniones sobre características	65
4.1. Algunos ejemplos de evidencias de opinión	93
4.2. Esquema conceptual de nuestra propuesta	97
5.1. Porción de la taxonomía de características para <i>headphones</i>	109
5.2. Árbol de dependencias de oración de ejemplo	118
6.1. Proceso de generación de los recursos	126
6.2. Extracción interactiva de <i>feature words</i>	130
6.3. Simulación del proceso interactivo de extracción de <i>feature words</i> : influencia de las restricciones morfosintácticas.	134
6.4. Simulación del proceso interactivo de extracción de <i>feature words</i> : influencia del número de candidatos aceptados antes de iterar.	135
6.5. Simulación del proceso interactivo de extracción de <i>feature words</i> : influencia del número de candidatos rechazados antes de iterar.	136
6.6. Simulación del proceso interactivo de extracción de <i>feature words</i> : comparativa entre iterar al aceptar, al rechazar o al vaciar lista de candidatos.	138
6.7. Simulación del proceso interactivo de extracción de <i>feature words</i> : influencia del número de nuevas semillas de opinión en cada iteración.	139
6.8. Taxonomía de características previa para <i>headphones</i>	143
6.9. Proceso de validación de anotaciones	148
6.10. Nuevas características añadidas a la taxonomía características tras anotaciones para <i>headphones</i> (en negrita)	150
6.11. Taxonomía final de características para <i>headphones</i>	152

6.12.	Porcentaje de aparición de las características en las anotaciones para <i>headphones</i>	152
6.13.	Ejemplo de grafo de orientaciones semánticas	159
6.14.	Resultados experimentos de ampliación del léxico de opiniones para <i>headphones</i> : parámetro <i>minAbsWeightArcs</i>	164
6.15.	Resultados experimentos de ampliación del léxico de opiniones para <i>headphones</i> : parámetro <i>minSupport</i>	165
6.16.	Histograma de probabilidades no condicionadas de los indicadores de opinión implícita para <i>headphones</i>	170
7.1.	Arquitectura conceptual de TOES	185
7.2.	Árbol de dependencias de la oración de ejemplo	192
7.3.	Ejemplo de <i>pipeline</i>	204
8.1.	Evaluación individual del anotador de características implícitas basado en los indicadores para <i>headphones</i> : influencia del umbral de probabilidad	212
8.2.	Evaluación individual del anotador de características implícitas basado en los indicadores para <i>headphones</i> : influencia del umbral de soporte	213
8.3.	Evaluación individual del enlazador de palabras de opinión basado en ventana para <i>headphones</i> : influencia del tamaño de las ventanas y las restricciones morfosintácticas	214
8.4.	Evaluación individual del enlazador de palabras de opinión basado en patrones de dependencias de tipo 2, 3, 4 y 5 para <i>headphones</i> : influencia del umbral de precisión mínima de los patrones utilizados	216
8.5.	<i>Pipelines</i> ligeros de recursos	220
8.6.	<i>Pipelines</i> basados en recursos	221
8.7.	Resumen comparativo de resultados obtenidos por cada uno de los <i>pipelines</i>	224
8.8.	Evolución del rendimiento del sistema (<i>pipeline</i> mixto) a medida que se aumenta el número de documentos anotados utilizados en la inducción de los recursos.	225
8.9.	Desglose de valores de F_1 obtenidos por el <i>pipeline</i> mixto para cada característica de la taxonomía	229
8.10.	Ejemplo de agregación de opiniones para un modelo determinado de auriculares (Sony MDR-V700DJ)	232
9.1.	Taxonomía de características del dominio <i>hotels</i>	237
9.2.	Taxonomía de características del dominio <i>cars</i>	238

Índice de cuadros

2.1.	Distribución según puntuaciones del corpus de críticas de cine.	44
2.2.	Patrones morfosintácticos para la extracción de bigramas.	45
2.3.	Resultados de la clasificación de críticas de cine en español.	47
2.4.	Evaluación de los experimentos de construcción del lexicón usando todos los documentos del corpus.	55
2.5.	Valores de $\tau_{\frac{1}{2}}$ obtenidos para los experimentos de construcción de lexicones sobre grafos equilibrados.	56
2.6.	Evaluación del sistema presentado al <i>TAC 2008 Opinion Summarization Task</i>	60
3.1.	Resumen de características de trabajos relacionados con la extracción de opiniones: extracción de características	76
3.2.	Resumen de características de trabajos relacionados con la extracción de opiniones: búsqueda de palabras de opinión y clasificación de opiniones	77
6.1.	Algunos datos del corpus obtenido para el dominio <i>headphones</i>	128
6.2.	Resumen de la ejecución de la herramienta de extracción de palabras de característica para el dominio <i>headphones</i>	140
6.3.	Lista de <i>feature words</i> obtenidas para el dominio <i>headphones</i> .	140
6.4.	Características obtenidas por agrupación de las <i>feature words</i> para <i>headphones</i>	142
6.5.	Estadísticas del corpus anotado del dominio <i>headphones</i>	149
6.6.	<i>Feature words</i> para cada una de las características de la taxonomía tras anotaciones para <i>headphones</i>	151
6.7.	Estimaciones para el término <i>flat</i> en el lexicón inducido para <i>headphones</i>	156
6.8.	Estimaciones para el término <i>awful</i> en el lexicón inducido para <i>headphones</i>	158
6.9.	Resultados de la ampliación del lexicón de opiniones para <i>headphones</i>	163

6.10. Algunas entradas del recurso de indicadores de opinión implícita para el dominio <i>headphones</i>	168
6.11. Algunas estimaciones contenidas en el recurso de indicadores de característica implícita inducido para <i>headphones</i>	169
6.12. Términos del recurso de indicadores de opinión implícita para <i>headphones</i> con probabilidad igual a 1	171
6.13. Resultados de los experimentos para medir la influencia de las restricciones morfosintácticas en los patrones de dependencias para <i>headphones</i>	177
6.14. Algunos patrones de dependencias de tipo 1 para <i>headphones</i> .	178
6.15. Algunos patrones de dependencias de tipo 2, 3, 4 y 5 para <i>headphones</i>	179
8.1. Evaluación individual del anotador de palabras de característica basado en la taxonomía para <i>headphones</i>	210
8.2. Evaluación individual del anotador de características implícitas basado en PMI-IR para <i>headphones</i>	211
8.3. Evaluación individual del anotador de características implícitas basado en los indicadores para <i>headphones</i>	211
8.4. Evaluación individual del enlazador de palabras de opinión basado en ventana para <i>headphones</i>	214
8.5. Evaluación individual del enlazador de palabras de opinión y el enlazador de expresiones especiales basados en ventana para <i>headphones</i>	215
8.6. Evaluación individual del enlazador de palabras de opinión basado en dependencias.	215
8.7. Evaluación individual del clasificador de opiniones basado en el lexicón para <i>headphones</i>	217
8.8. Evaluación comparativa de los clasificadores de opiniones basados en WordNet, SentiWordNet y PMI-IR para <i>headphones</i> . 217	
8.9. Evaluación comparativa de los clasificadores de opiniones para <i>headphones</i> , al concatenar a cada clasificador un clasificador basado en conjunciones.	218
8.10. Evaluación comparativa de la combinación de un clasificador basado en lexicón con cada uno de los demás clasificadores para <i>headphones</i>	219
8.11. Resultados obtenidos por los <i>pipelines</i> ligeros de recursos (optimizados para F_1)	220
8.12. Resultados obtenidos por los <i>pipelines</i> ligeros de recursos (optimizados para $F_{\frac{1}{2}}$)	220

8.13. Resultados obtenidos por los <i>pipelines</i> basados en recursos (optimizados para F_1)	222
8.14. Resultados obtenidos por los <i>pipelines</i> basados en recursos (optimizados para $F_{\frac{1}{2}}$)	222
8.15. Resumen comparativo de los mejores resultados conseguidos para cada una de las subtareas en que hemos dividido la tarea de reconocimiento y clasificación de opiniones, utilizando o sin utilizar los recursos del dominio.	223
8.16. Estimación de coste temporal y resultados del sistema para conjuntos de documentos anotados de distintos tamaños	227
8.17. Desglose de resultados obtenidos por el <i>pipeline mixto</i> para opiniones explícitas e implícitas (tarea <i>Opinion Recognition</i>) . .	227
9.1. Estadísticas del corpus para los dominios <i>hotels</i> y <i>cars</i>	236
9.2. Estadísticas del lexicón de opiniones para el dominio <i>cars</i>	240
9.3. Estadísticas del lexicón de opiniones para el dominio <i>hotels</i>	240
9.4. Términos del recurso de indicadores de opinión implícita con probabilidad igual a 1 para el dominio <i>hotels</i>	241
9.5. Términos del recurso de indicadores de opinión implícita con probabilidad igual a 1 para el dominio <i>cars</i>	242
9.6. Estadísticas de los patrones de dependencias inducidos para el dominio <i>hotels</i>	243
9.7. Estadísticas de los patrones de dependencias inducidos para el dominio <i>cars</i>	243
9.8. Resultados obtenidos por los <i>pipelines</i> ligeros de recursos para los dominios <i>hotels</i> y <i>cars</i> (optimizados para F_1)	244
9.9. Resultados obtenidos por los <i>pipelines</i> ligeros de recursos para los dominios <i>hotels</i> y <i>cars</i> (optimizados para $F_{\frac{1}{2}}$)	244
9.10. Resultados obtenidos por los <i>pipelines</i> basados en recursos para los dominios <i>hotels</i> y <i>cars</i> (optimizados para F_1)	245
9.11. Resultados obtenidos por los <i>pipelines</i> basados en recursos para los dominios <i>hotels</i> y <i>cars</i> (optimizados para $F_{\frac{1}{2}}$)	245
9.12. Resumen de resultados del sistema TOES en los distintos dominios(<i>pipelines</i> optimizados para F_1)	246
9.13. Resumen de resultados del sistema TOES en los distintos dominios(<i>pipelines</i> optimizados para $F_{\frac{1}{2}}$)	246

Agradecimientos

Quiero expresar mi agradecimiento a las personas que han participado directa o indirectamente en el desarrollo de esta tesis. En primer lugar, a José Antonio Troyano Jiménez, director del trabajo, por aportar su experiencia y dedicación absoluta, y especialmente por hacer de contrapeso anímico en los momentos más bajos. A Fernando Enríquez de Salamanca Ros y Francisco Javier Ortega Rodríguez por su colaboración y apoyo en gran parte del trabajo. A Carlos García Vallejo por su inestimable ayuda en la justificación algebraica del algoritmo PolarityRank. A Jorge González López y su equipo en el Centro de Tecnologías del Habla de IBM en Sevilla, por ofrecerme la oportunidad de aprender de ellos y meterme de paso en el cuerpo el “gusanillo” de la investigación. A Horacio Rodríguez Hontoria por acogerme por unas semanas en su grupo de trabajo y hacerme partícipe del mismo, y también por sus comentarios y revisiones sobre la tesis. A Francisco Fernández Ruiz por el diseño de la portada de la edición impresa. A Pablo Montoya Julián y Juan Manuel Nieto Moreno por su trabajo en la anotación del corpus y la implementación de la demo técnica, respectivamente. Al resto de compañeros y amigos del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla, por su apoyo en distintos aspectos y por los buenos momentos vividos.

A mi familia.

Resumen

En el contexto de la Web 2.0, las opiniones volcadas por los usuarios a través de las redes sociales, foros y otros servicios, acerca de productos, política u otras temáticas, conforman una interesantísima información con un gran potencial práctico de cara a las empresas, las administraciones y los ciudadanos. Siendo el texto libre el principal vehículo de dicha información en Internet, las Tecnologías del Lenguaje y el Procesamiento del Lenguaje Natural ocupan un papel protagonista de cara al tratamiento y análisis automático de la misma. De manera más concreta, recientemente diversos investigadores se han venido ocupando del tratamiento computacional de las opiniones, los sentimientos y otros fenómenos subjetivos del lenguaje. Dentro de esta disciplina, en el presente trabajo de tesis abordamos el problema de la extracción de opiniones sobre características, tarea cercana a la extracción de información y consistente en extraer representaciones estructuradas de las opiniones individuales contenidas en los textos, incluyendo la identificación de las características concretas del objeto sobre las que se vuelcan las opiniones, y la polaridad positiva o negativa de dichas opiniones. Los aspectos fundamentales de nuestro acercamiento son dos: la consideración del dominio de aplicación en cada una de las fases de la resolución del problema, y el uso de taxonomías de características, que permiten que el conjunto de opiniones extraídas sean fácilmente agregables y visualizables.

En la presente memoria, definiremos un conjunto de recursos de apoyo a la tarea de extracción de opiniones sobre características, adaptables al dominio; propondremos una metodología para la generación de los mismos, proporcionando las herramientas y algoritmos necesarios para minimizar la participación manual en el proceso; y describiremos un sistema de extracción de opiniones sobre características modular y adaptable al dominio, al que hemos denominado TOES (*Taxonomy-based Opinion Extraction System*). Los resultados de la experimentación que llevaremos a cabo nos permitirán afirmar la importancia del dominio en el problema que nos ocupa, y la utilidad de las opiniones extraídas de cara a una aplicación de agregación y visualización de opiniones.

Parte I

Preámbulo

Capítulo 1

Introducción

Resumen: El presente capítulo pretende, a modo de introducción, motivar y contextualizar el trabajo expuesto en esta memoria, plantear algunas hipótesis iniciales y esbozar las líneas generales de nuestra aportación. La última sección sirve de guía de lectura del resto del documento, ya que incluye un breve resumen del contenido de cada uno de los capítulos y apéndices que lo componen.

1.1. Motivación y contexto

Tradicionalmente, las opiniones de los actores involucrados en una determinada actividad siempre han sido clave en los procesos de toma de decisiones. Por ejemplo, los políticos están interesados en las opiniones de sus potenciales votantes; las empresas, por su parte, están interesadas en las opiniones de sus potenciales clientes, y los mismos clientes suelen evaluar las opiniones de algunos de sus conocidos a la hora de decidir entre varios productos. En el contexto actual, Internet se ha convertido en una fuente masiva y continua de opiniones. Estas opiniones suelen estar contenidas en textos escritos en lenguaje natural, ya sea en los *reviews* de productos escritos por clientes o por analistas profesionales, en los foros públicos existentes de diversas temáticas, en los blogs o más recientemente en determinados servicios de *microblogging* y redes sociales. El acceso y la explotación de estas nuevas fuentes de opinión poseen un indudable atractivo para las administraciones, las empresas y los clientes.

En los últimos años, diversos investigadores del campo del Procesamiento del Lenguaje Natural se han centrado en estudiar el tratamiento computacional de las opiniones, los sentimientos y otros fenómenos subjetivos contenidos

en este tipo de textos. Esta nueva disciplina, conocida como Minería de Opiniones o Análisis del Sentimiento, es de gran utilidad en el contexto que acabamos de plantear. Muchas son las tareas que se han ido definiendo en estos últimos años: clasificación de documentos basada en la opinión contenida en los mismos, detección de la subjetividad las emociones expresadas en los textos, clasificación de documentos basada en la perspectiva política del autor, etcétera.

En nuestro trabajo, nos centramos en la extracción de opiniones sobre características, tarea cercana a la extracción de información consistente en obtener representaciones de las opiniones individuales contenidas en documentos de opinión. Se parte de documentos que contienen opiniones sobre una entidad determinada (por ejemplo, un producto), debiendo decidirse para cada opinión la característica concreta sobre la que se centra (por ejemplo, el precio) y la polaridad (positiva o negativa). Ante la observación de que el dominio concreto de aplicación influye de manera determinante en las características del lenguaje empleado (palabras empleadas para expresar las opiniones, polaridad de dichas palabras, construcciones sintácticas utilizadas, etc.), pretendemos abordar la construcción de sistemas de extracción de opiniones adaptados al dominio. Esta orientación al dominio implica un esfuerzo adicional a la hora de adaptar el sistema a un nuevo dominio; es por ello que también estamos interesados en definir una metodología, así como proporcionar las herramientas y algoritmos necesarios, que faciliten dicha adaptación. Nuestra visión del problema será eminentemente práctica, por lo que trataremos de delimitar el tipo de opiniones que abordaremos, para permitir la construcción de sistemas de extracción útiles. Además, para facilitar la agregación, visualización u otros tratamientos posteriores de las opiniones extraídas, representaremos las características opinables de los objetos mediante taxonomías. Estas taxonomías, además de facilitar el tratamiento estadístico de las opiniones extraídas, constituyen un formalismo para la captura de requisitos, ya que a través de las mismas se establece cuáles son las características del objeto en las que estamos interesados.

1.2. Hipótesis

Nuestra hipótesis principal es la siguiente:

Hipótesis 1: *La disponibilidad de recursos específicos del dominio permitirá llevar a cabo la extracción de opiniones sobre características con mayor efectividad.*

Esta hipótesis se sustenta en las siguientes observaciones:

- Cada tipo de objeto (y por tanto, cada dominio) posee un conjunto distinto de características opinables.
- El vocabulario y expresiones utilizadas para expresar opiniones es sustancialmente distinto dependiendo del dominio. Por ejemplo, en algunos dominios se utiliza un lenguaje más formal, y en otros un lenguaje más coloquial.
- La orientación semántica (es decir, las implicaciones positivas o negativas de determinadas palabras o expresiones) es dependiente del dominio. Por ejemplo, el adjetivo *predictable* suele tener connotaciones negativas cuando se utiliza en el contexto de una crítica de cine, y connotaciones positivas cuando se refiere a la respuesta a la conducción de un vehículo. Más aún, demostraremos que la orientación semántica también depende de la característica concreta de la entidad a la que se aplica.

Por tanto, la definición de recursos de apoyo que recojan estas y otras particularidades del dominio de aplicación debería facilitar la extracción de opiniones. Para corroborar la hipótesis anterior, diseñaremos un sistema de extracción de opiniones modular, que permita disponer de distintas implementaciones de los componentes que resuelven cada una de los subproblemas en que dividiremos la tarea de extracción. De esta forma, podremos experimentar con sistemas formados a partir de componentes que hagan uso, o no, de los recursos del dominio, y comparar los resultados obtenidos.

La generación de los recursos se llevará a cabo de manera semiautomática, a partir de un conjunto de documentos del dominio cuyas opiniones han sido manualmente anotadas. Pretendemos confirmar una segunda hipótesis:

Hipótesis 2: *La definición de un esquema de anotación de opiniones sobre características y su aplicación a un conjunto de documentos de opinión de un dominio dado permitirá, mediante la definición de las herramientas y los algoritmos correspondientes, inducir los recursos necesarios*

para la consecución de la extracción de opiniones sobre dicho dominio sin apenas participación manual en el proceso.

De cara a que las opiniones extraídas sean fácilmente acumulables y visualizables, y representen intuitivamente al objeto analizado, se emplearán representaciones taxonómicas de las características del objeto:

Hipótesis 3: *La representación de las características del dominio mediante una taxonomía permitirá la mejor inducción de los recursos a nivel de característica, y facilitará la agregación y visualización de las opiniones extraídas.*

1.3. Resumen de la propuesta

Nuestra propuesta consiste en el desarrollo de un sistema genérico de extracción de opiniones sobre características, que puede ser adaptado a un dominio concreto mediante la generación de una serie de recursos que capturan el conocimiento específico de dicho dominio. Aportaremos una metodología para la generación de dichos recursos, incluyendo un flujo de trabajo y una serie de herramientas y algoritmos que permitan automatizar en lo posible la tarea.

Un recurso fundamental es la taxonomía de características, que contiene el conjunto de características opinables del dominio sobre las que el sistema procederá a extraer opiniones. La utilización de dichas taxonomías permitirá desarrollar sistemas de extracción dirigidos a los aspectos del dominio que se consideren más importantes, y facilitará la explotación de las opiniones extraídas. La taxonomía de características puede ser construida utilizando una serie de herramientas de apoyo que trabajan a partir de una colección de documentos del dominio, o bien puede ser definida por un experto en el dominio. En este caso, la taxonomía se convierte en un formalismo para la captura de requisitos, permitiendo decidir cuáles son las características del objeto en las que estamos interesados.

1.3.1. Resumen de aportaciones

A continuación adelantamos las principales aportaciones expuestas en la presente memoria:

1. Definiremos un conjunto de **recursos** de apoyo a la tarea de extracción de opiniones sobre características.

2. Propondremos una **metodología** para la generación de los recursos de apoyo, incluyendo un flujo de trabajo y un conjunto de herramientas y algoritmos que minimizan la participación manual en el proceso.
3. Diseñaremos un **sistema** de extracción sobre características modular y adaptable al dominio, al que denominaremos TOES (*Taxonomy-based Opinion Extraction System*).

A partir de los resultados obtenidos en un desarrollo experimental llevado a cabo utilizando el sistema anterior, demostraremos las hipótesis anteriormente enunciadas, entre otras conclusiones. Además de las aportaciones principales, en el transcurso de la investigación hemos desarrollado un algoritmo de *ranking* sobre grafos con aristas negativas, llamado PolarityRank, que puede ser empleado en diversos problemas de propagación de información, así como un método de expansión automática de lexicones de opinión. También hemos puesto a disposición de la comunidad investigadora un corpus de documentos de opinión anotados a nivel de características, que contiene más de 2.500 documentos de tres dominios distintos¹.

1.4. Estructura del documento

La estructura del documento es como sigue. En el capítulo 2 introducimos al lector en el subárea del Procesamiento del Lenguaje Natural que trata con textos subjetivos, recorriendo las principales tareas y aplicaciones relacionadas, y resumiendo algunas aportaciones propias previas a la realización de la propuesta principal de este trabajo de tesis. En el capítulo 3 realizamos un repaso bibliográfico de los trabajos relacionados directamente con la extracción de opiniones sobre características, incluyendo un estudio comparativo de los aspectos principales de las distintas propuestas. En el capítulo 4 se presenta nuestra propuesta, tratando de acotar el alcance de la tarea abordada, concretando la definición formal de la misma y de las entidades participantes, y dando una visión general de la propuesta. En el capítulo 5 se definen los recursos específicos del dominio, cuya objetivo es la captura de conocimiento útil para la extracción de opiniones sobre características, mostrando además la sintaxis utilizada para representarlos. En el capítulo 6 proponemos una metodología para la generación de los recursos para un dominio específico, incluyendo un flujo de trabajo y presentando una serie de herramientas y algoritmos encaminados a automatizar el proceso en la mayor medida posible. El capítulo 7 describe nuestro sistema de extracción de opiniones sobre

¹<http://www.lsi.us.es/~fermin/index.php/Datasets>

características, al que hemos denominado TOES. En el capítulo 8 llevamos a cabo una serie de experimentos con la intención de evaluar el sistema y, a partir de los resultados obtenidos, planteamos algunas conclusiones sobre la tarea abordada y sobre el sistema, incluyendo un ejemplo de aplicación de agregación y visualización de las opiniones extraídas. En el capítulo 9 se recogen datos y resultados de la aplicación de la metodología de generación de los recursos y la ejecución del sistema de extracción de opiniones sobre dos dominios distintos al utilizado en el resto del trabajo. Finalmente, en el capítulo 10 resumimos las aportaciones del trabajo, las conclusiones de los experimentos realizados y los resultados de la investigación.

Al final del documento, hemos incluido dos apéndices. En el apéndice A se aportan ejemplos concretos de la ejecución de cada uno de los componentes del sistema, con la intención de servir de apoyo a la comprensión de los mismos. En el apéndice B se expone la justificación algebraica y de la convergencia del algoritmo PolarityRank, que como veremos es utilizado en nuestro trabajo para expandir automáticamente los lexicones de opinión.

Parte II

Antecedentes

Capítulo 2

Procesamiento de textos subjetivos

Resumen: En este capítulo introducimos al lector en el subárea del Procesamiento del Lenguaje Natural que trata con textos subjetivos. En concreto, nos centraremos en la disciplina conocida como Análisis del Sentimiento (*Sentiment Analysis*) o Minería de Opiniones (*Opinion Mining*). Presentaremos las principales tareas y aplicaciones del área, y resumiremos algunas aportaciones propias llevadas a cabo de manera previa al desarrollo de la propuesta principal de este trabajo de tesis.

2.1. Introducción

En las últimas décadas, los investigadores que trabajan en el campo del Procesamiento del Lenguaje Natural (PLN) han intentado establecer mecanismos eficaces que permitan el tratamiento computacional de textos escritos en lenguaje natural y el acceso a la información contenida en ellos. Siendo el lenguaje natural una representación del pensamiento humano, ésta ha sido y sigue siendo un área que involucra grandes dificultades, y en la que participan multitud de problemas transversales a las distintas tareas: resolución de la ambigüedad inherente al lenguaje en los distintos niveles de abstracción (léxico, sintáctico, semántico y pragmático), reconocimiento de las entidades y las relaciones entre ellas, análisis del discurso, representación del conocimiento, razonamiento automático, etc. La resolución de muchos de estos problemas está relacionada con áreas como la lingüística, la psicología, la

inteligencia artificial, la minería de datos (conocida como minería de textos en este contexto) o el aprendizaje automático, entre otros.

Uno de los objetivos fundamentales del PLN es permitir el acceso a la información contenida en los textos escritos en lenguaje natural. En ocasiones, este tipo de información es conocida como información no estructurada, en contraposición al tipo de información contenida en una base de datos o más genéricamente en cualquier sistema informático. En cierto modo, los mecanismos del PLN permiten convertir la información no estructurada en información estructurada, que puede ser consumida computacionalmente para habilitar multitud de aplicaciones. La irrupción de Internet y particularmente la World Wide Web (WWW) supusieron un punto de inflexión, en tanto que la mayor parte de la información contenida en la misma es información no estructurada, principalmente en forma de textos en distintos idiomas. El acceso a la información por medio de motores de recuperación de información o sistemas de búsqueda de respuestas, la clasificación automática de los contenidos de las páginas web, o la extracción automática de información y conocimiento contenidos en las mismas, son algunas de las tareas relacionadas con Internet y en las que el PLN tiene mucho que aportar.

En los últimos años, el uso de Internet se ha ido generalizando, y los usuarios han ido ganando cada vez mayor protagonismo. El que antes era un usuario pasivo, que en la mayoría de los casos se limitaba a acceder y consultar la información contenida en la WWW, se ha convertido en un usuario activo: además de consultar información, llevar a cabo diversas gestiones administrativas o comunicarse con sus conocidos, un gran porcentaje de los internautas actuales generan contenidos. Esta es la base de lo que se conoce como Web 2.0. Inicialmente, los internautas comenzaron participando en foros de debate, en los que podían mantener conversaciones temáticas más o menos públicas con otros internautas, o escribiendo en blogs, bitácoras personales en las que el autor escribe periódicamente acerca de algún tema o de manera completamente libre. En ambos casos, las aportaciones de unos usuarios se ven enriquecidas por las aportaciones de otros, en forma de respuestas en los foros o comentarios en los blogs. Poco a poco, se han ido incrementando los servicios disponibles relacionados con los contenidos generados por usuarios: agregadores de *reviews* de productos escritos por usuarios (*Epinions, CNet*), servicios de *microblogging* (*Twitter*), redes sociales de carácter general (*Facebook, MySpace*) o especializadas (profesionales como *LinkedIn*, de fotografía como *Flickr*, de vídeo como *YouTube*, y muchas otras), En resumen, todas estas nuevas formas de participación en Internet suponen una ingente fuente de información (en gran parte no estructurada), continuamente cambiante y virtualmente imposible de procesar manualmente, lo que supone nuevas oportunidades de aplicación para el PLN.

2.1.1. La web como fuente de opiniones

Una gran parte de los contenidos generados por los usuarios en forma de textos escritos en lenguaje natural tiene un carácter subjetivo: los *reviews* de productos o servicios, las entradas de algunos blogs o los comentarios hechos por otros usuarios, los hilos de discusión en los foros públicos (acerca de política, cultura, economía o cualquier otro tema), los mensajes escritos en servicios de microblogging como *Twitter* o en redes sociales como *Facebook*. Todos estos son textos en los que los internautas pueden expresar sus opiniones, puntos de vista y estados de ánimo. Al igual que es frecuente apoyarse en las opiniones de nuestros conocidos a la hora de tomar determinadas decisiones (por ejemplo, qué modelo comprar de un determinado producto, o qué película ir a ver al cine), es cada vez más frecuente basar dichas decisiones en las opiniones que los internautas vuelcan en la red. Por otro lado, son evidentes las implicaciones prácticas de este tipo de contenidos para las compañías o administraciones públicas; por ejemplo, hoy día muchas compañías disponen de personal que se dedica a monitorizar las opiniones aparecidas en Internet acerca de los productos y servicios de la misma, de manera que puedan interferir activamente evitando la propagación de estas opiniones y sus posibles consecuencias negativas de cara a las ventas o a la imagen de la marca. Pero independientemente del objetivo buscado, el número de contenidos individuales que deberían ser considerados es de tal magnitud que se hacen imprescindibles ciertos mecanismos de procesamiento automático de los mismos. Si bien estamos de nuevo hablando del campo de actuación del PLN, el carácter subjetivo del tipo de contenidos que se pretende procesar presenta determinadas peculiaridades y da lugar a nuevos retos, poco estudiados tradicionalmente por el PLN. Se abre así una nueva subdisciplina conocida como Análisis del Sentimiento o Minería de Opiniones que viene a ocuparse del procesamiento computacional de este tipo de contenidos subjetivos.

2.2. Análisis del sentimiento y minería de opiniones

El análisis del sentimiento (*sentiment analysis*) o minería de opiniones (*opinion mining*) abarca aquellas tareas relacionadas con el tratamiento computacional de las opiniones, los sentimientos y otros fenómenos subjetivos del lenguaje natural. Algunas de estas tareas son la clasificación de documentos de opinión según el carácter positivo o negativo de las opiniones, la extracción de representaciones estructuradas de opiniones, o el resumen de textos de opinión, entre otras. Estas tareas tienen puntos en común con otras ta-

reas clásicas del PLN, como pueden ser la clasificación de documentos, la extracción de información y el resumen automático. Sin embargo, el carácter subjetivo de los textos analizados añade ciertas peculiaridades a las tareas anteriores y las hace especialmente difíciles. Por ejemplo, la clasificación de documentos basada en la opinión es sensiblemente más difícil que la clasificación de documentos según la temática. En efecto, la aplicación de las técnicas habitualmente empleadas para esta última no conduce a buenos resultados cuando se aplica a la clasificación basada en la opinión (Pang et al, 2002).

2.2.1. Orígenes

Las dos denominaciones anteriores son las más habituales en la bibliografía, siendo el término *sentiment analysis* más empleado por grupos cercanos al PLN y *opinion mining*, por su parte, en contextos relacionados con la recuperación de información y la minería de textos. El término *sentiment* fue usado por primera vez por Das & Chen (2001) y Tong (2001), trabajos en los que se llevaba a cabo un análisis automático de textos de opinión con la intención de predecir la “confianza del mercado” (*market sentiment*). Algunos trabajos posteriores hicieron uso del término *sentiment* (Turney, 2002a; Pang et al, 2002) para referirse a la carga evaluativa positiva o negativa que poseen algunas palabras o expresiones. Finalmente el término *sentiment analysis* apareció en un par de trabajos en 2003 (Nasukawa & Yi, 2003; Yi et al, 2003), momento a partir del cual se popularizó enormemente y es utilizado en multitud de trabajos. En cuanto al término *opinion mining*, apareció por primera vez en (Dave et al, 2003). Aunque inicialmente se refería a la subtarea de extracción y agregación de opiniones que describiremos más adelante, actualmente el término se considera apropiado para describir los distintos tipos de tratamientos y análisis que se llevan a cabo con textos subjetivos, es decir, como término intercambiable por *sentiment analysis* (Ding et al, 2008a). Aún así, este último está más extendido como denominación general del área.

Aunque existen muchos trabajos a lo largo de los últimos 35 años relacionados de una u otra forma con el tratamiento computacional de los fenómenos subjetivos del lenguaje natural, es entre los años 2001 y 2003 cuando se produce un aumento repentino de trabajos relacionados con el análisis del sentimiento (Cardie et al, 2003; Das & Chen, 2001; Dave et al, 2003; Dini & Mazzini, 2002; Liu et al, 2003; Morinaga et al, 2002; Nasukawa & Yi, 2003; Pang et al, 2002; Tateishi et al, 2001; Turney, 2002a; Wiebe et al, 2003; Yu & Hatzivassiloglou, 2003). Esta explosión de trabajos está relacionada con los siguientes factores, según Pang & Lee (2008):

- Incremento en el uso de métodos basados en aprendizaje automático en el PLN.
- Disponibilidad de conjuntos de datos sobre los que entrenar modelos de aprendizaje automático, debido a la aparición de determinadas webs de agregación de *reviews*, principalmente.
- Descubrimiento por parte de la comunidad de los retos y aplicaciones potenciales que ofrece el área.

Desde nuestro punto de vista, la popularización de las redes sociales y sus implicaciones en las relaciones entre las compañías y sus clientes (o entre los políticos y sus votantes) también guarda relación con el creciente interés en el área.

2.2.2. Aplicaciones

Lo que hace realmente interesante la investigación en temas relacionados con el análisis del sentimiento, además de los retos intelectuales que plantea, son las innumerables aplicaciones. Algunas de estas aplicaciones son evidentes, y han sido señaladas en múltiples ocasiones por los distintos autores desde la explosión de trabajos de 2001; otras muchas continúan apareciendo, a medida que Internet y la sociedad de la información evolucionan. Podemos dividir las aplicaciones según el actor que se vería beneficiado: usuarios de productos y servicios, compañías, actores financieros, administraciones públicas y partidos políticos.

Aplicaciones para clientes

Desde el punto de vista de los usuarios de productos y servicios, serían útiles todos aquellos sistemas que les permitieran acceder a las opiniones de otros usuarios de manera más potente que la actual. Si actualmente uno puede consultar en sitios web especializados *reviews* o críticas de otros usuarios acerca de los productos o servicios que se deseé, la utilización de técnicas de minería de opiniones permitiría visualizar de manera resumida la opinión de cientos o miles de usuarios acerca del producto en cuestión; se reduciría con ello enormemente el tiempo de consulta necesario, y se aumentaría la cobertura de las distintas opiniones consideradas. Además de permitir un mejor acceso a las opiniones de los demás, las propias opiniones que expresemos en distintos servicios de Internet podrían ser utilizadas para configurar un perfil de nuestros gustos e intereses. En esta situación, el análisis del sentimiento

vendría a enriquecer los sistemas de recomendación (Adomavicius & Tuzhilin, 2005), permitiendo a los mismos hacer uso de la información subjetiva contenida en las aportaciones textuales de sus usuarios. Algunos autores han trabajado en aplicaciones relacionadas con la detección de lenguaje ofensivo (Spertus, 1997), algo de lo que los usuarios de servicios de correo electrónico o foros públicos se verían beneficiados.

Aplicaciones para compañías

Desde el punto de vista de las compañías, el análisis del sentimiento aporta herramientas que permiten simplificar y hacer más eficiente la monitorización de las opiniones de sus clientes, trabajo que ya llevan a cabo muchas de ellas manualmente. Se trataría por tanto de disponer de *robots* que continuamente estarían consultando los contenidos generados por usuarios, capaces de detectar menciones explícitas o implícitas a la compañía y sus productos o servicios, y de reconocer y clasificar adecuadamente las opiniones vertidas. Esto permitiría generar gráficas de valoración y otras representaciones útiles, descubrir nuevas demandas de los clientes, o definir sistemas de alarma automáticos que avisen a los responsables del sistema ante la aparición de opiniones “peligrosas” para la compañía. Una de las fuentes de información principales de este tipo de sistemas son las redes sociales; es por esto que también es interesante conocer a qué usuarios pertenecen las opiniones reconocidas, y qué rol juegan dichos usuarios dentro de su red de influencia, ya que las opiniones volcadas por ciertos usuarios tendrán un mayor impacto en la red que las opiniones de otros. Aquí entran en juego otro tipo de tecnologías relacionadas con el análisis de grafos y redes sociales, que si bien no se consideran parte del análisis del sentimiento, sí están en estrecha relación en algunas aplicaciones como la anterior. La existencia en los últimos años de numerosos talleres¹ y números especiales de revistas² que aglutan trabajos de análisis del sentimiento y análisis de redes sociales confirma esta relación. Otra aplicación interesante para las compañías que ha sido señalada en la bibliografía es la personalización del contenido o la ubicación de la publicidad *online* (Jin et al, 2007).

¹ Workshop on Search and Mining User-generated Contents (CIKM 2010, CIKM 2011), Workshop on Dynamic Networks and Knowledge Discovery (ECML PKDD 2011), Workshop on Link Analysis in Heterogeneous Information Networks (IJCAI 2011)

² Special Issue on Search and Mining User Generated Contents (*Transactions on Intelligent Systems and Technology*), Special Issue on Analysis of Short Texts on the Web (*Language Resources and Evaluation Journal*), Special Issue on Social Networks & Social Web Mining (*World Wide Web Journal*)

Aplicaciones en la política y e-administración

Sistemas de monitorización de opiniones como los descritos anteriormente también serían útiles a los partidos políticos, en este caso para conocer la valoración de los ciudadanos acerca de sus miembros o de las medidas tomadas desde las administraciones. Servirían aquí las herramientas proporcionadas por el análisis del sentimiento como complemento a los estudios demoscópicos habituales, basados en encuestas personales. Además de para fines demoscópicos, se ha sugerido en la bibliografía la posibilidad de monitorizar la actividad en determinados foros “calientes” y otras fuentes de comunicación para detectar posibles hostilidades extranjeras o actividades terroristas, como herramienta para los servicios de inteligencia de los gobiernos (Abbasí, 2007).

Aplicaciones en las finanzas

El análisis del sentimiento también es aplicable a contextos financieros, hasta el punto de que los trabajos que introdujeron el término *sentiment* para referirse al análisis automático de textos de opinión (Das & Chen, 2001; Tong, 2001) se centraban precisamente en dicho contexto (análisis de la confianza del mercado o *market sentiment*). Es conocida la extraordinaria permeabilidad de los mercados bursátiles hacia los estados cambiantes de opinión. Es por esto que una de las preocupaciones de los inversores y demás actores financieros es conocer en todo momento dichos estados de opinión (podemos incluso hablar de estados anímicos), ya que pueden servir de heurística para prever los movimientos ascendentes o descendentes de los mercados. La importancia de la rapidez de reacción ante cambios en los estados de opinión y la amplitud de las fuentes de información que es necesario consultar para medir dicho estado hacen de éste un campo de aplicación evidente del análisis del sentimiento.

Otros campos de aplicación

Otros campos de aplicación serían el académico (el análisis de citas en artículos (Piao et al, 2007) y de la reputación (Taboada et al, 2006) se verían beneficiados por una clasificación de las citas basada en la opinión), y el sociológico. En las ciencias sociales se utilizan técnicas de análisis textuales, en las que se trabaja sobre transcripciones de entrevistas con sujetos, artículos de prensa, libros de texto y otras fuentes de información textual. Generalmente, los investigadores se apoyan en algunas herramientas más o menos sofisticadas de minería de textos para procesar la información contenida en dichos documentos. La adición de técnicas de minería de opiniones a estas

herramientas facilitaría a los investigadores la detección y contabilización de las opiniones y los sentimientos aparecidos en dichos textos.

2.2.3. Tareas

Son muchas las tareas concretas en las que se ha centrado la investigación en el área del análisis del sentimiento. A continuación hacemos un repaso por algunas de las fundamentales, tratando de dar una descripción breve de la definición y las características de cada una. No pretendemos hacer un repaso pormenorizado de las distintas soluciones propuestas por los autores, sino dar una visión general de la amplitud de objetivos y problemas incluidos en el campo del análisis del sentimiento.

Detección de la subjetividad

Algunos de los primeros trabajos aparecidos en el contexto del tratamiento de textos subjetivos están relacionados con el reconocimiento de dicho tipo de textos, y su diferenciación con respecto a los textos objetivos (Hatzivassiloglou & Wiebe, 2000). Existe gran cantidad de trabajos relacionados con esta problemática, incluyendo a los participantes en el *2006 Blog track* de la Text Retrieval Conference (TREC) (Ounis et al, 2006), que estuvo centrado en la misma. La forma más inmediata de atacar este problema es la clasificación binaria de oraciones u otras unidades textuales menores en dos categorías (objetivo/subjetivo) (Wiebe et al, 2004). Otros trabajos llevan a cabo la clasificación a nivel de documento, lo cual está relacionado con la tarea de clasificación basada en el género (Yu & Hatzivassiloglou, 2003) (por ejemplo, distinguir de entre un conjunto de textos periodísticos los correspondientes a la sección de noticias de los de la sección de editorial u opinión). Este tipo de clasificación, independientemente de la granularidad de los textos a clasificar, puede ser considerada un paso previo a otras tareas del análisis del sentimiento como la clasificación basada en la opinión o la extracción de opiniones. Dichas tareas trabajan sobre textos que se presuponen subjetivos, por lo que un sistema que las llevara a cabo se vería beneficiado por la existencia de un módulo previo de detección de subjetividad que filtrara los textos a procesar (Mihalcea et al, 2007).

La clasificación binaria no es la única manera de afrontar la detección de la subjetividad. Algunos trabajos afrontan la tarea desde un punto de vista regresivo: se trata en este caso de determinar la existencia o no de subjetividad según una escala progresiva (Wilson et al, 2004) (por ejemplo, 0: totalmente objetivo, 5: totalmente subjetivo). Este enfoque responde a la ambigüedad detectada por los investigadores en las tareas asociadas a la

detección de la subjetividad, que se refleja en la imposibilidad por parte de los anotadores humanos de llegar a un acuerdo sobre el carácter objetivo o subjetivo de algunos textos de entrada (Yu & Hatzivassiloglou, 2003).

Detección de emociones

Se trata a priori de uno de los retos más difíciles del área del análisis del sentimiento: la detección de expresiones de afectos en los textos subjetivos. Si mencionábamos antes la ambigüedad intrínseca de la tarea consistente en distinguir entre un texto objetivo y uno subjetivo, ser capaz de identificar en el texto la aparición de emociones y distinguir además de qué emoción se trata se antoja una misión extremadamente difícil. Aún así, las múltiples aplicaciones posibles y conexiones con diversas áreas de la informática (por ejemplo, con el diseño de interfaces hombre-máquina sensibles al estado anímico de los usuarios (Liscombe et al, 2005)) ha animado a muchos investigadores a tratar de resolver el reto.

En un trabajo de 1982 en el que se abordaba la relación entre las expresiones faciales y las emociones fueron definidas seis emociones universales: ira, disgusto, miedo, alegría, tristeza y sorpresa (Ekman, 1982). Esta categorización de las emociones humanas ha sido adquirida casi como un estándar de facto por los trabajos en el área del análisis del sentimiento que se ocupan de detectar la aparición de emociones en los textos (Alm et al, 2005; Liu et al, 2003; Subasic & Huettner, 2001). Aunque existen algunos investigadores que han empleado una categorización de las emociones de un mayor nivel de detalle; por ejemplo, en (Boldrini et al, 2010) se describe un amplio esquema de anotación de fenómenos subjetivos en el que, entre otros fenómenos, se distinguen hasta 15 emociones distintas.

Cálculo de la orientación semántica

La orientación semántica de una palabra o expresión se define como una medida de las implicaciones afectivas, positivas o negativas, que tiene dicha palabra o expresión cuando es usada en un contexto evaluativo (Lehrer, 1974; Battistella, 1990; Hatzivassiloglou & McKeown, 1997). Así, por ejemplo, diremos que la palabra “excelente” tiene una orientación semántica positiva, y que la palabra “péssimo” tiene una orientación semántica negativa. Algunos autores emplean una definición binaria de la orientación semántica, contemplando únicamente dos posibles orientaciones semánticas contrapuestas: positiva y negativa (Hatzivassiloglou & McKeown, 1997; Esuli & Sebastiani, 2005; Takamura et al, 2005; Hu & Liu, 2004a; Kim & Hovy, 2004; Esuli & Sebastiani, 2006a, 2005; Andreevskaia & Bergler, 2006). Es común emplear

el nombre *polaridad semántica*, o simplemente *polaridad*, para referirse a esta definición de orientación semántica. Otros autores consideran un mayor conjunto de valores posibles, ya sea de manera discreta (por ejemplo, estableciendo una escala de 1 a 5) o de manera continua (por ejemplo, utilizando un número real entre -1 y 1) (Turney, 2002a; Turney & Littman, 2003a; Kamps et al, 2004a; Esuli & Sebastiani, 2007, 2006b; Baccianella et al, 2010). En este caso, la orientación semántica no sólo representa la polaridad del término en cuestión, sino también la intensidad de las implicaciones afectivas. Por ejemplo, podríamos asignar a la palabra “excelente” un valor de orientación semántica igual a 1.0, indicando que su polaridad es positiva con una gran intensidad, y valores menores pero igualmente positivos a otras palabras, como “bueno” o “correcto”, que aún expresando polaridad positiva no lo hagan con tanta vehemencia.

Aunque existen algunos intentos de construcción de listas de términos positivos y negativos de manera manual o semi-automática (Stone, 1966; Huettner & Subasic, 2000; Das & Chen, 2001; Tong, 2001; Cerini et al, 2007), la mayoría de los trabajos citados tratan de generar automáticamente y de manera no supervisada lexicones de opinión en los que se aglutan palabras y expresiones de carácter subjetivo junto a estimaciones de sus orientaciones semánticas; se trata de recursos de aplicación general e independientes del dominio, en los que se presupone que el valor de orientación semántica de los términos es único e independiente del contexto. Sin embargo, esto no es así en todos los casos: la orientación semántica de determinadas palabras y expresiones puede cambiar de intensidad e incluso polaridad en función del contexto de aplicación. Por ejemplo, no es deseable que una película sea *predictible*, pero sí lo es que lo sea la conducción de un coche. Existen algunos trabajos que tienen en cuenta la influencia del contexto en el cálculo de la orientación semántica (Wilson et al, 2005; Popescu & Etzioni, 2005; Kanayama & Nasukawa, 2006; Ding et al, 2008b; Qiu et al, 2011); nosotros hemos trabajado en la inducción de lexicones de opinión en español orientados al dominio (Cruz et al, 2009b) (ver sección 2.3.2), demostrando la ventaja de utilizar valores de orientación semántica contextualizados. Dicho resultado condicionó en gran medida nuestra aproximación al problema de la extracción de opiniones, desde la perspectiva de la consideración del dominio de aplicación como uno de los elementos fundamentales de la propuesta.

Clasificación basada en la polaridad

Una de las tareas que más interés ha despertado en la comunidad de investigadores es la clasificación de documentos basada en la polaridad (Turney, 2002a; Pang et al, 2002), quizás por su cercanía a una tarea clásica dentro

del PLN: la clasificación de documentos. Dado un documento de opinión (por ejemplo, una crítica de cine o literaria), la clasificación de documentos basada en la polaridad consiste en clasificar el documento como positivo o negativo, en función de las opiniones contenidas en el documento. Se trata por tanto de una clasificación binaria, y en la cual se llevan a cabo algunas suposiciones y simplificaciones. En primer lugar, se supone que el documento en cuestión contiene opiniones acerca de una única entidad. Además, se considera que la mayoría de las opiniones contenidas son de la misma polaridad, posibilitando elegir la categoría positiva o negativa del documento sin existencia de ambigüedad. Por supuesto, estas suposiciones no se cumplen en todos los casos: es común que, aún tratándose de un documento de opinión, algunas pasajes del documento no sean textos subjetivos (por ejemplo, es común que las críticas de cine contengan descripciones del argumento de la película), o que se incluyan opiniones acerca de otros objetos distintos al principal (por ejemplo, en una crítica acerca de una determinada película pueden aparecer referencias a otras películas); y, sobre todo, es común encontrar documentos de opinión que no podemos considerar positivos o negativos, puesto que alaban determinados aspectos del objeto de análisis y critican otros. La falta de aplicabilidad práctica de la tarea ha ocasionado una disminución en el interés de los investigadores y el número de trabajos en la misma en los últimos años, junto a un aumento en el interés de la clasificación de unidades textuales menores (pasajes u oraciones) (Yu & Hatzivassiloglou, 2003; Kim & Hovy, 2004; Wilson et al, 2005; Zhao et al, 2008) o incluso en otras tareas relacionadas pero con una mayor cercanía la extracción de información, como la extracción de opiniones basada en características en la que se centra el presente trabajo de tesis.

Aún así, la clasificación de documentos basada en la polaridad ha servido como tarea de acceso al campo del análisis del sentimiento para muchos grupos de investigación. Su estudio permite entrar en contacto con aspectos transversales al resto de tareas del área, como el concepto de orientación semántica o las consideraciones acerca de la influencia del contexto y de los elementos discursivos. Una observación importante es la realizada en (Pang et al, 2002), trabajo en el que se aplican técnicas clásicas de clasificación de documentos. Una de las conclusiones del trabajo es que la clasificación de documentos basada en la polaridad es más difícil que otros tipos de clasificación de documentos (por ejemplo, la basada en el tema o el género). En estos otros tipos de clasificación, las técnicas basadas en bolsas de palabras funcionan relativamente bien, debido a las diferencias existentes en el vocabulario utilizado en las distintas clases consideradas; sin embargo, acercamientos similares no funcionan igual de bien cuando son aplicados a la clasificación basada en la polaridad, por la mayor importancia de los elementos discursivos.

vos a la hora de distinguir entre opiniones positivas y negativas, y el mayor solapamiento entre las terminologías empleadas en las dos clases.

La cercanía de la tarea a la clasificación de documentos y la ausencia de trabajos realizados para textos en español nos motivó a afrontarla en nuestro primer trabajo en el área del análisis del sentimiento (Cruz et al, 2008) (ver sección 2.3.1).

Inferencia de la puntuación de documentos de evaluación

Esta tarea consiste en determinar la puntuación que el autor de un *review* otorga al objeto de análisis según alguna escala discreta (por ejemplo, es bastante común en webs especializadas en *reviews* utilizar una escala de 1 a 5). Este problema puede verse como una generalización de la clasificación basada en la polaridad, en la que en lugar de tener dos clases de salida (positiva y negativa) se dispone de más de dos (en el ejemplo anterior, cinco clases). Pero dado que existe una relación de orden entre las clases, algunos trabajos abordan esta tarea como un problema de regresión (Pang & Lee, 2005; Goldberg & Zhu, 2006). Al igual que ocurría en el cálculo de la orientación semántica, algunos autores señalan la ambigüedad inherente de la tarea: según algunos experimentos, un anotador humano sólo es capaz de acertar exactamente la puntuación en una escala de 1 a 5 en poco más de la mitad de las veces (Pang & Lee, 2005). Parece que, a la hora de evaluar, cada individuo otorga una semántica sensiblemente distinta a los distintos valores de la escala, de manera que, por ejemplo, las valoraciones de algunas personas están sesgadas hacia puntuaciones más altas y las de otras hacia puntuaciones más bajas (Amatriain et al, 2009).

Existen pocos trabajos que aborden esta tarea, quizás en parte por su dificultad. Sin embargo, parece que la utilidad de la misma al ser aplicada a los sistemas de recomendación está haciendo resurgir el interés (Szomszor et al, 2007; Golbeck, 2006; Ganu et al, 2009; Leung et al, 2011).

Clasificación basada en la perspectiva

Aunque la mayoría de los trabajos de clasificación de textos subjetivos están enfocados a la clasificación basada en la polaridad, existen otras aproximaciones. Por ejemplo, algunos trabajos tratan de clasificar transcripciones de debates políticos según estén a favor o en contra del asunto que se esté debatiendo (Bansal et al, 2008; Thomas et al, 2006). Otros trabajos intentan clasificar textos de opinión en función de la tendencia política (progresista frente a conservadora) (Mullen & Malouf, 2006, 2008; Malouf & Mullen, 2008), o de si el autor pertenece a una u otra parte de interés en algún con-

flicto (por ejemplo, si se trata de un texto pro-israelí o pro-palestino (Lin et al, 2006)). En (Tumasjan et al, 2010) se intenta predecir el resultado de unas elecciones a partir del análisis de los comentarios políticos encontrados en Twitter.

En general, estos trabajos tienen en común que tratan de clasificar textos de carácter subjetivo en dos clases contrapuestas, existiendo una cierta correlación entre cada una de las clases y un conjunto determinado de opiniones y afectos hacia determinados temas. Un aspecto interesante que plantean algunos de estos trabajos, en los que se analizan textos que forman parte de un debate (es decir, hay intervenciones y réplicas), es la posibilidad de llevar a cabo la clasificación de todos los textos de manera global, teniendo en cuenta las referencias de unos participantes a otros y el tipo de lenguaje utilizado en dichas referencias (apoyo o crítica, detección de ironía o sarcasmo, etc.), lo que da pie a la representación del problema mediante grafos y por tanto a la aplicación de técnicas de análisis de los mismos (Mullen & Malouf, 2008; Malouf & Mullen, 2008).

Extracción de opiniones

Partiendo de la tarea de clasificación basada en la polaridad, y aumentando progresivamente la granularidad (documentos, pasajes, oraciones), algunos trabajos llegaron a plantear la tarea a nivel de las opiniones individuales contenidas en los textos. Es decir, más que clasificar un documento u otra unidad textual menor en función de la polaridad media de las opiniones contenidas, se trata ahora de identificar pequeños trozos de textos que contengan opiniones individuales, y clasificar dichas opiniones. Se supera de esta forma una de las simplificaciones de la tarea de clasificación, según la cual todas las opiniones contenidas en un documento debían ser del mismo signo; al trabajar a nivel de opiniones, la clasificación es más real, y la salida obtenida más útil en la práctica. La tarea está ahora más cercana a la extracción de información que a la clasificación de documentos, ya que hay que identificar la aparición de un tipo determinado de entidad (las opiniones), representada por distintos campos de información relacionados entre sí que hay que extraer del texto (por ejemplo, el objeto sobre el que versa la opinión y las palabras utilizadas para calificarlo). Una vez que se han obtenido dichas entidades, se lleva a cabo un proceso de clasificación de las mismas, la mayor parte de las veces basada en la polaridad, para lo que se utilizan técnicas similares a las utilizadas en los trabajos de clasificación anteriormente comentados.

La mayoría de los trabajos de extracción de opiniones se centran en extraer opiniones de productos a partir de *reviews*, principalmente debido a la disponibilidad de páginas webs especializadas a partir de las cuáles generar

conjuntos de datos para entrenamiento y evaluación (Dave et al, 2003; Hu & Liu, 2004b,a; Popescu & Etzioni, 2005; Ding et al, 2008b). En estos trabajos, las opiniones son extraídas a nivel de *característica*, es decir, se pretende identificar el aspecto concreto del producto sobre el que versa la opinión (por ejemplo, la *calidad de imagen* de una cámara de fotos); esta tarea se conoce como extracción de opiniones sobre características (*feature-based opinion extraction*), y fue definida por primera vez en (Hu & Liu, 2004b).

Dado que es esta la tarea en la que se centra el presente trabajo de tesis, faremos un estudio más pormenorizado de las distintas soluciones propuestas en la bibliografía en el capítulo 3.

Resumen de textos de opinión

Al igual que el resumen automático clásico, esta tarea consiste en obtener versiones resumidas a partir de uno o muchos documentos, de manera que se reduzca la redundancia perjudicando lo mínimo posible la legibilidad y tratando de incluir la información más importante en el resumen. La particularidad con respecto al resumen automático de textos está en el carácter subjetivo de los textos a resumir. Estos textos pueden ser, por ejemplo, un conjunto de documentos de opinión acerca de un producto determinado recuperados de la web. El resumen de textos de opinión es una tarea más aplicada que las anteriores, en tanto que la salida de la misma es directamente de utilidad para muchas de las aplicaciones que comentamos en la sección 2.2.2.

Las características de los textos sobre los que se trabaja influyen en las posibles maneras de llevar a cabo el resumen. Por ejemplo, en un resumen multidocumento se podría tratar de dar una cobertura similar a los documentos con opiniones positivas y a los documentos con opiniones negativas. O también se podría construir el resumen mostrando de manera diferenciada las opiniones positivas y negativas que más aparecen. En ambos casos, la tarea de resumen se apoyaría en alguna de las tareas previamente comentadas (clasificación de documentos basada en la opinión, extracción de opiniones).

En comparación con el resumen clásico, el resumen de textos subjetivos da pie en mayor medida a plantear distintos tipos de resúmenes no textuales, basados en representaciones gráficas. Por ejemplo, algunos trabajos proponen representaciones en forma de grafo conectando a las personas o entidades que enuncian opiniones con los objetos concretos sobre los que opinan (Cardie et al, 2003), representaciones gráficas de las opiniones extraídas de *reviews* de productos a nivel de características para un producto determinado (Gamon et al, 2005; Carenini et al, 2006) o para varios productos de manera comparativa (Liu et al, 2005; Yi & Niblack, 2005).

2.2.4. Recursos

A continuación enumeramos algunos recursos útiles para la resolución y evaluación de las distintas tareas del área.

Colecciones de *reviews*

El género de los *reviews* de productos o servicios es el que dispone de una mayor cantidad de recursos en forma de conjuntos de datos, anotados a distintos niveles. Bing Liu y su equipo ofrecen distintos conjuntos de datos relacionados con el género³. Los conjuntos de datos *customer review datasets* y *additional review datasets* contienen *reviews* acerca de 5 y 9 productos respectivamente, anotados a nivel de oración; para cada oración que contiene una opinión, se indica la característica del objeto sobre la que trata la opinión y la polaridad de la misma. Los conjuntos de datos *comparative sentence datasets* (se trata de dos conjuntos de datos distintos, aunque con el mismo nombre) consisten en oraciones que contienen juicios comparativos entre distintos productos. También ofrecen conjuntos de datos enfocados a la detección de *spam* en *reviews* (*Amazon Product Review Data*) y a la detección del género de autores de blogs (*Blog author gender classification dataset*).

*Cornell movie-review datasets*⁴ son un conjunto de datos en el dominio de las críticas de cine. Se incluyen cuatro conjuntos de datos. En primer lugar, un conjunto de 1.000 críticas positivas y 1.000 críticas negativas. En segundo lugar, un conjunto de 5331 oraciones negativas y otras tantas oraciones positivas, extraídas de las críticas de cine. Estos dos conjuntos son útiles para quienes estén interesados en desarrollar trabajos relacionados con la clasificación basada en la opinión, ya sea a nivel de documento o de oración. En tercer lugar, un conjunto de críticas de cine que incluyen un valor numérico correspondiente a la puntuación asignada a la película por parte del autor, aplicable por tanto en trabajos de *rating inference*. Y por último, un conjunto de 5000 oraciones objetivas y otras tantas subjetivas, útil para la realización de trabajos sobre detección de subjetividad.

*Spanish Movie Reviews*⁵ es un corpus de críticas de cine en español, acompañadas de una puntuación, que construimos en el contexto de un trabajo sobre clasificación basada en la opinión (ver sección 2.3.1). Se trata de uno de los pocos recursos que existen actualmente en español. *Multi-Domain, Taxonomy-based Opinion Dataset* es un conjunto de *reviews* en inglés de

³<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

⁴<http://www.cs.cornell.edu/People/pabo/movie-review-data/>

⁵<http://www.lsi.us.es/~fermin/index.php/Datasets>

tres dominios distintos (auriculares, coches y hoteles), que incluyen opiniones anotadas a nivel de característica. Para cada dominio, se incluye además una taxonomía de características. Es fruto del presente trabajo de tesis, y se encuentra disponible en la misma dirección que el anterior.

Otros conjuntos de datos de *reviews* de productos son el *multi-aspect restaurant reviews*⁶, que para cada documento (*reviews* de restaurantes) incluye puntuaciones de 5 aspectos (comida, ambiente, servicio, precio y general); y el *multi-domain sentiment dataset*⁷, que incluye *reviews* de múltiples dominios junto a una puntuación de 1 a 5 para cada documento.

Otras colecciones de datos

Existen conjuntos de documentos con textos subjetivos de otros géneros, como pueden ser el género blog, el periodístico o el político. *Blog06*⁸ es un enorme conjunto de textos (con un tamaño de 25GB) extraídos de blogs de diversas temáticas. Se incluyen algunos blogs con *spam*, lo que hace interesante este conjunto de datos para aquellos grupos interesados en la detección de *spam* en contextos subjetivos.

*Congressional floor-debate transcripts*⁹ son un conjunto de transcripciones de debates parlamentarios sobre leyes en el congreso de Estados Unidos. Se incluye información acerca de si los distintos participantes están a favor o en contra de la ley que se está debatiendo, así como del orden cronológico de las participaciones y las distintas menciones de unos participantes a otros. Se trata por tanto de un conjunto de datos con gran utilidad para desarrollar trabajos de clasificación basada en la perspectiva, sobretodo aquellos que pretendan hacer uso de la información relativa a la estructura del debate.

El *MPQA Opinion Corpus*¹⁰ contiene noticias periodísticas anotadas con un gran nivel de detalle; no sólo se anotan opiniones, sino también emociones, creencias, y otros fenómenos subjetivos. Actualmente se encuentra en su versión 2.0.

El corpus de la competición *NTCIR-6 Multilingual Opinion-Analysis Task*¹¹ contiene artículos de prensa en tres idiomas (japonés, chino e inglés), con anotaciones relativas a las opiniones contenidas, incluyendo la persona que emite la opinión (*opinion holder*) y la polaridad. Sucesivas ediciones del congreso (*NTCIR-7* y *NTCIR-8*) también incluyeron competiciones relacionadas con

⁶<http://people.csail.mit.edu/bsnyder/naacl07>

⁷<http://www.cs.jhu.edu/~mdredze/datasets/sentiment>

⁸http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

⁹<http://www.cs.cornell.edu/home/llee/data/convote.html>

¹⁰<http://www.cs.pitt.edu/mpqa/databaserelease/>

¹¹<http://research.nii.ac.jp/ntcir/ntcir-ws6/data-en.html>

el tratamiento de opiniones, y disponen de otros recursos en forma de conjuntos de datos¹².

Recursos léxicos

Existen diversos recursos léxicos que contienen información de carácter subjetivo acerca de palabras o expresiones. *General Inquirer*¹³ es una herramienta de análisis de contenido cuyo origen se remonta a la década de los 60. Entre los recursos que incorpora, se encuentran listas de palabras y expresiones etiquetadas con orientaciones semánticas positivas y negativas. Se trata del primer recurso léxico en contener este tipo de información. *Subjectivity Lexicon*¹⁴ es una lista de palabras y expresiones que suelen aparecer en construcciones subjetivas; es por tanto un recurso útil para desempeñar tareas de detección de la subjetividad.

Existen diversos recursos léxicos basados en WordNet (Fellbaum, 1998). SentiWordNet¹⁵ es un recurso de generación semiautomática en el que a cada *synset* de WordNet se le asignan tres puntuaciones: una de *positividad*, otra de *negatividad* y una tercera de *objetividad*. MicroWN-Op¹⁶, por su parte, etiqueta un subconjunto de *synsets*, distinguiendo también entre términos positivos, negativos y objetivos; en este caso, se trata de un recurso elaborado manualmente, y por tanto más fiable que el anterior, aunque con una cobertura considerablemente menor. WordNet-Affect¹⁷, además de asignar etiquetas de positividad y negatividad a los *synsets*, incorpora otros tipos de información subjetiva, representando las connotaciones emocionales de los términos y sus implicaciones afectivas.

2.3. Experiencias previas en el área

En la presente sección vamos a repasar los trabajos previos que hemos realizado en el área del análisis del sentimiento, explicando las técnicas y algoritmos propios o de la bibliografía que fueron utilizados en los mismos. Estos trabajos han tenido una repercusión en mayor o menor medida en nuestro sistema de extracción de opiniones, ya sea por la incorporación de algunas de las conclusiones obtenidas en la motivación de la propuesta, o por el uso directo de algunas de las técnicas y algoritmos desarrollados.

¹²<http://research.nii.ac.jp/ntcir/data/data-en.html>

¹³<http://www.wjh.harvard.edu/~inquirer/>

¹⁴<http://www.cs.pitt.edu/mpqa/>

¹⁵<http://sentiwordnet.isti.cnr.it/>

¹⁶<http://www-3.unipv.it/wnop/>

¹⁷<http://wndomains.fbk.eu/wnaffect.html>

2.3.1. Clasificación binaria de documentos basada en la opinión

La clasificación binaria de documentos basada en la opinión consiste en decidir la categoría positiva o negativa de una serie de documentos de opinión, en función de la opinión media expresada en el documento. Basándonos en los trabajos de Peter D. Turney (Turney, 2002a; Turney & Littman, 2003a), llevamos a cabo algunos experimentos de clasificación binaria de críticas de cine en español (Cruz et al, 2008). Se trata del primer trabajo de clasificación basada en la opinión de documentos en español. El corpus utilizado en los experimentos está disponible para uso público¹⁸, y ha sido empleado en algunos trabajos de otros investigadores (del Hoyo et al, 2009). En la tabla 2.1 se muestra la distribución de los documentos del corpus según las puntuaciones asignadas por los autores a cada crítica.

Puntuación	Nº de documentos
1	351
2	923
3	1.253
4	890
5	461
Total	3.878

Cuadro 2.1: Distribución según puntuaciones del corpus de críticas de cine.

En (Turney, 2002a) se describe un clasificador no supervisado basado en la opinión. Dicho clasificador decide el carácter positivo o negativo de los documentos en base a la suma de las orientaciones semánticas de ciertos bigramas de los documentos. La orientación semántica de los bigramas se calcula mediante un algoritmo llamado *PointWise Mutual Information - Information Retrieval (PMI-IR)*. A pesar de tratarse de un planteamiento relativamente simple y no supervisado, el sistema clasifica correctamente un 84 % de los documentos de un pequeño corpus de *reviews* de coches utilizado por el autor. Sin embargo, cuando se trata de clasificar críticas de cine, la precisión obtenida es sólo del 65 %, de lo que se deduce que la clasificación de críticas de cine es una tarea especialmente difícil. Esta diferencia de dificultad de la tarea de clasificación basada en la opinión según el dominio ha sido constatada por otros artículos posteriores, señalando algunos autores que en el caso concreto de las críticas de cine o de libros algunas partes del texto

¹⁸<http://www.lsi.us.es/~fermin/index.php/Datasets>

se dedican a narrar una sinopsis de la obra; estas secciones pueden dar lugar a errores en los resultados de la clasificación si no son excluidas del análisis (Pang & Lee, 2004).

De manera resumida, el método de clasificación utilizado en (Turney, 2002a) y adaptado en nuestro trabajo (Cruz et al, 2008) consta de los siguientes pasos:

- Dado un documento, se extraen bigramas utilizando una serie de patrones morfosintácticos simples. En nuestro trabajo, adaptamos el conjunto de patrones para ser empleados en textos en español (ver tabla 2.2), teniendo en cuenta las diferencias con el inglés en el orden de aparición de adjetivos y nombres en los sintagmas nominales, y de verbos y adverbios en los sintagmas verbales. Se postula que este conjunto contiene al menos algunos bigramas que expresan opinión (aunque también muchos otros bigramas que no indican opinión).
- Para cada uno de los bigramas extraídos, se estima la orientación semántica como un valor real, entre -1.0 y 1.0, mediante el algoritmo *PMI-IR*, que explicaremos en la siguiente sección.
- A partir de la suma de las orientaciones semánticas obtenidas, se clasifica la crítica como positiva si el valor obtenido es mayor o igual que cero, y negativa en caso contrario. En nuestra propuesta, además de este procedimiento no supervisado, llevamos a cabo algunos experimentos en los que se utilizaron algunos documentos como corpus de entrenamiento para encontrar un valor real, no necesariamente cero, que sirviera de umbral entre ambas clases.

	Primera palabra	Segunda palabra	Tercera palabra (no se extrae)
1.	adjetivo	nombre	cualquiera
2.	nombre	adjetivo	no es nombre
3.	adverbio	adjetivo	no es nombre
4.	adverbio	verbo	cualquiera
5.	verbo	adverbio	cualquiera

Cuadro 2.2: Patrones morfosintácticos para la extracción de bigramas.

Algoritmo *PointWise Mutual Information - Information Retrieval*

Para calcular la orientación semántica se utiliza el algoritmo *PMI-IR*, que consiste en estimar la *información mutua puntual* (*pointwise mutual information*) (Church & Hanks, 1989) entre el término en cuestión y un par de palabras semilla (*excellent* y *poor*) que sirven de representantes inequívocos de orientación semántica positiva y negativa, haciendo uso de un buscador de páginas web para llevar a cabo dicha estimación. La *información mutua puntual* se define entre dos palabras w_1 y w_2 y mide estadísticamente la información que obtenemos sobre la posible aparición de un término a partir de la aparición de otro término:

$$PMI(w_1, w_2) = \log_2 \left(\frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \right) \quad (2.1)$$

A partir de esta medida estadística la orientación semántica de un término t se calcula de la siguiente manera:

$$SO(t) = PMI(t, "excellent") - PMI(t, "poor") \quad (2.2)$$

Para estimar la medida *PMI* se utilizan búsquedas en la web, de manera que las probabilidades de coaparición de dos términos que constan en el numerador de la fórmula 2.1 se aproxima mediante el número de páginas web en las que ambos términos aparecen uno cercano al otro. Tras algunas transformaciones algebraicas, la fórmula final para el cálculo de la orientación semántica de un término ($SO(t)$) propuesta por Turney es la siguiente:

$$\log_2 \left(\frac{hits(t \text{ NEAR } excellent)hits(poor)}{hits(t \text{ NEAR } poor)hits(excellent)} \right) \quad (2.3)$$

, donde $hits(t)$ indica el número de páginas devueltas por el buscador *AltaVista*¹⁹ al buscar t . El operador *NEAR* de *AltaVista*²⁰ devuelve las páginas en las que aparecen ambos términos a una distancia máxima de 10 palabras. Al número de páginas obtenido se le suma 0,01 para evitar una posible división por cero.

De manera intuitiva, la idea detrás de este cálculo de la orientación semántica es que expresiones que indiquen una opinión positiva aparecerán con mayor frecuencia cerca de una palabra con claras connotaciones positivas

¹⁹URL: <http://www.altavista.com>

²⁰El operador *NEAR* no está disponible en otros buscadores como *Google*, razón por la que el autor se decanta por utilizar *AltaVista*. Sin embargo, se pueden utilizar otras *queries* similares en Google, haciendo uso del comodín “*”.

Semilla	No supervisada			Supervisada			
	pos.	neg.	total	umbral	pos.	neg.	total
Simple	35,5 %	91,5 %	63,5 %	13,0	72,5 %	82,5 %	77,5 %
Múltiple	70 %	69 %	69,5 %	-2,25	75 %	72,5 %	73,75 %

Cuadro 2.3: Resultados de la clasificación de críticas de cine en español.

como *excellent* y con mucha menor frecuencia cerca de una palabra con connotaciones negativas como *poor*. En sucesivos trabajos (Turney & Littman, 2003a), el autor utiliza varias semillas de cada tipo en lugar de una sola.

La adaptación del algoritmo *PMI-IR* al español se reduce a escoger dos semillas apropiadas para el español. Las semillas escogidas fueron *excelente* y *malo*. También realizamos experimentos utilizando un conjunto de semillas en lugar de una palabra aislada. El conjunto de semillas positivas y negativas utilizado fue el siguiente:

- **Positivas:** excelente, buenísimo, buenísima, superior, extraordinario, extraordinaria, magnífico, magnífica, exquisito, exquisita
- **Negativas:** malo, mala, pésimo, pésima, deplorable, detestable, atroz, fatal

Resultados y conclusiones

En la tabla 2.3 se muestran los resultados obtenidos en los experimentos usando por un lado el método no supervisado y por otro el supervisado (utilizando el 80 % de los documentos del corpus para optimizar el valor umbral y el 20 % restante para la evaluación). Los resultados para la clasificación no supervisada (63,5 %) están a la par de los comunicados en (Turney, 2002a) para el corpus de críticas de cine en inglés (65,83 %), mejorando sensiblemente al utilizarse varias semillas para estimar las orientaciones semánticas (69,5 %). La utilización de algunos documentos para buscar un valor umbral distinto de cero entre las clases positiva y negativa supone una mejora considerable; sorprendentemente, la versión supervisada de la clasificación obtiene mejores resultados usando semillas simples para la estimación de las orientaciones semánticas (77,5 %) que usando semillas múltiples (73,75 %).

Más allá de los resultados obtenidos para la tarea de clasificación, este trabajo nos permitió adentrarnos en el área y tomar contacto con conceptos transversales como la orientación semántica. En concreto, el algoritmo *PMI-IR* ha sido incluido en un par de componentes de nuestro sistema de extracción de opiniones (ver secciones 7.3.2 y 7.3.6).

2.3.2. Construcción de lexicones de opinión orientados al dominio

Una de las conclusiones a las que llegamos en el trabajo de clasificación de documentos fue la necesidad de disponer de mecanismos alternativos de cálculo de las orientaciones semánticas. Dos eran las debilidades principales del algoritmo *PMI-IR*: en primer lugar, la utilización de *AltaVista*, que además de suponer una dependencia con la web, ralentiza el proceso de cómputo de las orientaciones semánticas de nuevos términos, lo que hace que la técnica sea poco apropiada para un sistema en producción; en segundo lugar, las estimaciones obtenidas no tienen en cuenta el dominio de aplicación. Como hemos expuesto anteriormente, algunos términos tienen connotaciones positivas en unos dominios y negativas en otros, o bien connotaciones del mismo signo pero distinta intensidad. Esta dependencia de la orientación semántica con el contexto de aplicación ha sido señalada por diversos trabajos en los últimos años (Wilson et al, 2005; Popescu & Etzioni, 2005; Kanayama & Nasukawa, 2006; Ding et al, 2008b; Qiu et al, 2011). En nuestro trabajo (Cruz et al, 2009b) planteamos un método de construcción de lexicones orientados al dominio. Una versión modificada del método aquí planteado nos fue de utilidad posteriormente para la ampliación automática de los lexicones de opinión que utilizamos en nuestra propuesta de extracción de opiniones (ver sección 6.8). Además, el algoritmo de *ranking* sobre grafos utilizado en dicho método ha resultado ser aplicable a problemas de distinta naturaleza (por ejemplo, a la selección de instancias en minería de datos (Vallejo et al, 2010), el cálculo de la confianza y la reputación en redes sociales (Ortega et al, 2011) y la detección de *spam* (Ortega et al, 2010)).

A continuación realizamos un repaso por las aportaciones de otros autores relacionadas con la construcción automática de lexicones de opinión, y resumimos nuestro trabajo, incluyendo una descripción del algoritmo de *ranking* sobre grafos *PolarityRank*.

Trabajos relacionados

Repasando en la bibliografía los distintos algoritmos propuestos para el cálculo de orientaciones semánticas encontramos distintas propuestas basadas en diccionarios léxicos (principalmente WordNet (Fellbaum, 1998)). Por ejemplo, en (Kamps et al, 2004a) se define una función de distancia d entre palabras usando las relaciones de sinonimia de WordNet, de manera que la orientación semántica de una palabra se calcula a partir de la distancia de la misma a una semilla positiva (*good*) y una semilla negativa (*bad*):

$$EVA(w) = \frac{d(w, bad) - d(w, good)}{d(good, bad)} \quad (2.4)$$

Se obtiene así un valor real entre -1.0 y 1.0 que codifica la polaridad y la intensidad de la orientación semántica. La evaluación del método sobre 349 palabras arrojó una tasa de acierto del 68.19 %, un resultado algo pobre, teniendo en cuenta el *baseline* del 50 % atribuible a una clasificación binaria aleatoria.

Otros autores se basan en la misma idea, pero empleando un conjunto mayor de semillas y añadiendo más relaciones semánticas de WordNet (por ejemplo, las relaciones de antonimia) (Hu & Liu, 2004a; Kim & Hovy, 2004). En otros trabajos llevados a cabo por Esuli y Sebastiani (Esuli & Sebastiani, 2006a,b, 2005), la idea de partida es que si la orientación semántica de una palabra tiene un signo determinado, entonces las palabras que aparecen en su glosa (pequeñas definiciones textuales de las palabras incluidas en WordNet) tenderán a tener orientaciones semánticas del mismo signo. Se construyen inicialmente dos grandes conjuntos de palabras positivas y negativas, ampliando dos pequeños conjuntos de semillas positivas y negativas mediante las relaciones de sinonimia y antonimia de WordNet. Después, para cada una de las palabras del conjunto, se obtiene una representación textual en forma de bolsa de palabras, a partir de todas las glosas de la palabra; estas representaciones son transformadas en vectores utilizando las técnicas habituales de clasificación de documentos. Por último, un clasificador binario es entrenado con estos vectores, de manera que ante una nueva palabra se estima su orientación semántica (en este caso se trata de un valor de polaridad) mediante dicho clasificador.

Otro trabajo de los mismos autores (Esuli & Sebastiani, 2007) se apoya en la misma idea de base (la relación entre las orientaciones de las palabras que aparecen en las glosas) para construir una representación de las palabras de WordNet en forma de grafo. Empleando un algoritmo de *ranking* de grafos similar a PageRank (Page et al, 1998) y un conjunto de semillas positivas y negativas, se calculan dos puntuaciones para cada palabra (positividad y negatividad). A diferencia de los trabajos previos de los autores, los valores así obtenidos son números reales. Como resultado del trabajo, se publicó un lexicón de opiniones llamado *SentiWordNet* (Esuli & Sebastiani, 2006b), al que han seguido versiones posteriores generadas a partir de modificaciones al algoritmo original (Baccianella et al, 2010); actualmente, la última versión del recurso es considerado el estado del arte en lexicones de opinión independientes del dominio.

La principal debilidad de los acercamientos basados en diccionarios léxicos es que no tienen en cuenta la dependencia de la orientación semántica con

el dominio de aplicación. Otras propuestas se basan en grandes conjuntos de documentos, y calculan la orientación semántica de manera no supervisada a partir de ellos (Hatzivassiloglou & McKeown, 1997; Turney, 2002a; Turney & Littman, 2003a; Yu & Hatzivassiloglou, 2003). Todos ellos parten de algunas semillas positivas y negativas, y calculan la orientación semántica de las palabras aplicando determinadas reglas de propagación de las orientaciones semánticas. Los métodos basados en corpus permiten hacer estimaciones de las orientaciones semánticas dependientes del dominio, siempre que se utilicen conjuntos de documentos del dominio en cuestión.

El método que propusimos en (Cruz et al, 2009b) está basado en corpus, y utiliza la misma idea de partida de (Hatzivassiloglou & McKeown, 1997), según la cual los adjetivos que participan en determinadas construcciones conjuntivas tienden a compartir el signo de la orientación semántica (por ejemplo, si aparecen coordinados mediante la conjunción *and*), y los que participan en otro tipo de construcciones conjuntivas suelen tener orientaciones opuestas (por ejemplo, si se utiliza la conjunción *but*). Sin embargo, en lugar de resolver un problema de *clustering* a partir de las restricciones anteriores (como se hace en (Hatzivassiloglou & McKeown, 1997)), en nuestro trabajo aplicamos un algoritmo de *ranking* sobre grafos, inspirados por el trabajo de Esuli y Sebastiani (Esuli & Sebastiani, 2007); sin embargo, nuestro algoritmo de *ranking*, a diferencia del algoritmo usado en (Esuli & Sebastiani, 2007), es capaz de trabajar en grafos con aristas positivas y negativas, de una sola pasada.

Descripción del método

Nuestro método consiste en generar, para cada dominio considerado, un grafo de relaciones conjuntivas entre adjetivos. En dicho grafo, los nodos son los adjetivos que aparecen en los documentos disponibles del dominio, y aquellos nodos correspondientes a palabras que aparecen en construcciones conjuntivas son conectados mediante aristas²¹. Los pesos de dichas aristas son proporcionales al número de relaciones conjuntivas observadas, y su signo indica si las relaciones son de tipo *directo* o *inverso*. Cada nodo tiene asociado un valor de *positividad* y otro de *negatividad*; en principio, ambos valores son nulos para todos los nodos, excepto para algunos nodos que actúan de semillas positivas y negativas. Aplicando un algoritmo propio de *ranking* sobre grafos, se consigue propagar la información aportada por las semillas al resto de los nodos, obteniéndose a partir de los valores obtenidos estimaciones de la orientación semántica de las palabras representadas en el grafo. Explicamos

²¹Dado que el algoritmo que aplicaremos posteriormente trabaja sobre grafos dirigidos, en realidad añadimos dos aristas, una en cada sentido, entre cada par de nodos en cuestión.

a continuación cómo se realiza la construcción del grafo y el funcionamiento del algoritmo de *ranking* utilizado, que hemos llamado *PolarityRank*.

Construcción del grafo

Dos adjetivos participan en una construcción conjuntiva si forman parte de un mismo sintagma, relacionados a través de una conjunción. Diremos que la relación entre los adjetivos participantes en la construcción es *directa* o *inversa* si, a la vista de la conjunción utilizada y de las partículas negativas participantes, es probable que las orientaciones semántica de ambos adjetivos sean del mismo signo o de signo opuesto, respectivamente. Algunos ejemplos de relaciones directas e inversas serían los siguientes:

■ Relación directa

The camera has a **bright and accurate** len.
It is a **marvellous, really entertaining** movie.
... **clear and easy to use** interface.
... **easy to get information, user-friendly** interface.

■ Relación inversa

The camera has a **bright but inaccurate** len.
It is a **entertaining but typical** film.
The driving is **soft and not aggressive**.

Las palabras que aparecen en negrita corresponden a los términos que serán incluidos en el lexicón. Como puede verse, además de adjetivos se permiten construcciones formadas por adjetivos de la clase *easy/difficult* seguidos de un infinitivo y las palabras que completan el significado del verbo (el comportamiento sintáctico de estas construcciones y los adjetivos que forman parte de la clase *easy/difficult* se encuentran estudiados en (Nanni, 1980)).

Basándonos en las relaciones conjuntivas extraídas a partir del conjunto de documentos del dominio, asignamos un peso a cada arista que conecta dos nodos. Este peso es igual al número de apariciones de los adjetivos en relaciones conjuntivas directas menos el número de apariciones en relaciones conjuntivas inversas (también experimentamos con una versión normalizada de los pesos, en el intervalo [-1.0,1.0]). Por tanto, dos nodos conectados mediante una arista positiva deberían obtener un valor similar de orientación semántica, tanto más cuanto mayor sea el valor del peso. Igualmente, dos nodos conectados mediante una arista negativa deberían obtener valores opuestos de orientación semántica, con mayor probabilidad cuanto mayor sea el módulo del peso de la arista.

PolarityRank

El algoritmo de *ranking* que propusimos es una versión modificada del algoritmo PageRank (Page et al, 1998). PageRank es utilizado por Google para medir la importancia de cualquier página web en Internet en función de los enlaces que dicha página recibe. Se entiende que cuando una página web enlaza a otra página web está emitiendo un voto positivo sobre la página enlazada (la está *recomendando*). Por tanto, aquellas páginas que sean recomendadas por más (y mejores) páginas web deberían ser consideradas más importantes o relevantes. Generalizando su aplicación, PageRank permite obtener una puntuación para cada nodo de un grafo dirigido. Esta puntuación mide la relevancia de dicho nodo en la red en función de las aristas entrantes y de las puntuaciones obtenidas mediante el propio algoritmo por los nodos vecinos participantes de esas aristas.

PolarityRank generaliza el concepto de votación o recomendación, permitiendo aristas con pesos positivos y negativos. Una arista positiva sigue significando un voto positivo, por el cual un nodo *recomienda* otro nodo, con más vehemencia cuanto mayor sea el peso de la arista. Una arista negativa, por contra, representa un voto negativo, por el cual un nodo *censura* a otro nodo, con más fuerza cuanto mayor sea el valor absoluto del peso de la arista. Partiendo de un grafo como éste, PolarityRank calcula dos puntuaciones para cada nodo, una positiva y otra negativa, siguiendo una aritmética particular. Según esta aritmética, la puntuación positiva de un nodo n se ve incrementada proporcionalmente a la puntuación positiva de los nodos conectados a n mediante aristas de pesos positivos. Pero además, la puntuación positiva de n también se ve incrementada de manera proporcional a la puntuación negativa de aquellos nodos conectados a n mediante aristas de pesos negativos. Los mismos principios se aplicarían a los valores de puntuación negativa de los nodos.

La definición del algoritmo es la siguiente. Sea un grafo dirigido $G = (V, E)$ donde V es un conjunto de nodos y E un conjunto de aristas dirigidas entre dos nodos. Cada arista de E contiene un valor real asociado o peso, distinto de cero, siendo p_{ji} el peso asociado a la arista que va del nodo v_j al v_i . Se define la operación $Out(v_i)$, que devuelve el conjunto de índices de los nodos para los que existe una arista desde v_i . Se definen las operaciones $In^+(v_i)$ y $In^-(v_i)$, que devuelven los conjuntos de índices de los nodos para los que existe una arista hacia v_i cuyo peso sea positivo o negativo, respectivamente. Definimos el PolarityRank positivo (PR^+) y negativo (PR^-) de un nodo v_i (fórmula 2.5), donde los valores de e^+ son mayores que cero para ciertos nodos que actúan como semillas positivas y cero para el resto de nodos, y los valores de e^- son mayores que cero para ciertos nodos que actúan

como semillas negativas y cero para el resto de nodos.

$$\begin{aligned}
 PR^+(v_i) &= (1 - d)e_i^+ + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^+(v_j) + \right. \\
 &\quad \left. + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^-(v_j) \right) \\
 PR^-(v_i) &= (1 - d)e_i^- + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^-(v_j) + \right. \\
 &\quad \left. + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^+(v_j) \right)
 \end{aligned} \tag{2.5}$$

La constante d es un factor de amortiguación, necesario para la convergencia de la función. En todos los casos usamos el valor 0,85 para dicho factor, tal como se propone en la definición original de PageRank. Por otro lado, la suma de los valores de e^+ por un lado y de e^- por otro debe ser igual al número de nodos del grafo. El cálculo de la ecuación anterior se realiza mediante un proceso iterativo, asignando inicialmente valores constantes a PR^+ y PR^- y calculando consecutivamente nuevas versiones de los valores, hasta que la diferencia entre los valores de una iteración y la siguiente es menor que un cierto valor de precisión.

En el apéndice B se incluye la justificación algebraica y la demostración de la convergencia del algoritmo.

Cálculo de las orientaciones semánticas

Una vez construido el grafo para cada dominio y asignados los valores adecuados de e^+ y e^- a las semillas (utilizamos las propuestas por Turney y Littman en (Turney & Littman, 2003b)²²), se procede al cálculo de los valores de PR^+ y PR^- de cada nodo. A partir de estos valores, se estima la orientación semántica del término representado por cada nodo v_i en cada uno de los dominios mediante la siguiente fórmula:

$$SO(v_i) = \frac{PR^+(v_i) - PR^-(v_i)}{PR^+(v_i) + PR^-(v_i)} \tag{2.6}$$

²²Semillas positivas: good, nice, excellent, positive, fortunate, correct, superior. Semillas negativas: bad, nasty, poor, negative, unfortunate, wrong, inferior.

Los valores así calculados pertenecen al intervalo $[-1, 0; 1, 0]$, indicando el módulo la intensidad y el signo la polaridad de la orientación semántica del término.

Resultados y conclusiones

Los experimentos fueron realizados aplicando el método sobre un conjunto de más de 234.000 *reviews* de 82 tipos de productos distintos, extraídos de la web *Epinions.com*. Aplicamos el método sobre los documentos de cada uno de los dominios, obteniendo 82 estimaciones de orientación semántica para cada uno de los términos observados. La evaluación de los valores obtenidos para las orientaciones semánticas en cada uno de los dominios planteados supone un problema en sí mismo. En otros trabajos, la evaluación se realiza comparando las orientaciones semánticas obtenidas con algunos recursos generados de manera manual o semiautomática; algunos de estos recursos son SentiWordNet (Esuli & Sebastiani, 2006b) , General Inquirer (Stone, 1966) o Micro-WNOp (Cerini et al, 2007). Dichos recursos contienen valores para la orientación semántica independientes del dominio, pero ¿cómo distinguir entre las divergencias entre los valores de orientación semántica fruto de imperfecciones de nuestro método y aquellas inherentes a peculiaridades del dominio en el que nos encontramos?

Para evitar este problema, construimos un grafo global a partir del corpus completo para generar un lexicón independiente del dominio, comparando los valores obtenidos con los del recurso Micro-WNOp. De esta manera pretendíamos medir la bondad del método propuesto. El recurso Micro-WNOp consiste en una muestra de aproximadamente 1100 synsets de WordNet, a cada uno de los cuales les fueron asignados manualmente tres valores reales entre 0 y 1 indicando la positividad, negatividad y neutralidad del mismo. Ordenando los adjetivos contenidos en el recursos según el resultado de la resta entre los valores de positividad y negatividad, obtuvimos un *ranking* de 433 adjetivos, que comparamos con el *ranking* obtenido de la aplicación de nuestro método automático de construcción del lexicón al corpus completo. En cada uno de los experimentos, usamos como *gold standard* sólo aquellos adjetivos de Micro-WNOp encontrados en nuestro lexicón.

Para comparar ambos *rankings* utilizamos la distancia τ de *Kendall* (τ_p) (Fagin et al, 2004), que mide la similitud entre un *ranking* modelo o *gold standard* y otro *ranking* candidato. Cuanto más cercano a cero sea el valor de esta medida, más parecido es el *ranking* obtenido de nuestro lexicón al *ranking* aportado por Micro-WNOp. En la tabla 2.4 se muestran los resultados de dos experimentos: en el primero de ellos, se descartaron las aristas negativas y se empleó el algoritmo de PageRank en lugar del PolarityRank;

en el segundo caso, se utilizó el algoritmo PolarityRank sobre el grafo completo, incluyendo aristas negativas. Ambos resultados mejoran los obtenidos por Esuli y Sebastiani (Esuli & Sebastiani, 2007)²³.

Método	$\tau_{\frac{1}{2}}$
<i>Esuli</i> ⁺	0,325
<i>Esuli</i> ⁻	0,284
PageRank	0,235
PolarityRank	0,225

Cuadro 2.4: Evaluación de los experimentos de construcción del lexicón usando todos los documentos del corpus.

La ganancia obtenida por la utilización de la información aportada por las aristas negativas y la aplicación de nuestra versión del algoritmo de *ranking* es relativamente pequeña. Esto se debe a la baja proporción de aristas negativas que contiene el grafo. En experimentos posteriores no incluidos en (Cruz et al, 2009b), eliminamos aleatoriamente aristas positivas del grafo anterior, consiguiendo una versión “equilibrada” del mismo. A partir de ese grafo, calculamos las orientaciones semánticas de los nodos participantes (usando PageRank y las aristas positivas, por un lado, y usando PolarityRank y todas las aristas, por el otro) y evaluamos los resultados siguiendo el método anterior. Repetimos el proceso completo diez veces (eliminación aleatoria de aristas positivas, cálculo de orientaciones semánticas mediante PageRank/PolarityRank) y obtuvimos las medias de $\tau_{\frac{1}{2}}$ que se muestran en la tabla 2.5. Como puede observarse, en este caso la ganancia al emplear PolarityRank, que contempla la información aportada por las aristas negativas, es mayor.

Como veremos en el capítulo 5, hemos empleado un método similar al de este trabajo para llevar a cabo una expansión automática de los diccionarios de opinión empleados en nuestra propuesta de extracción de opiniones. Además, creemos que el algoritmo *PolarityRank* es de utilidad en otros problemas de propagación de conocimiento, siempre que se disponga de una pequeña cantidad de información sobre algunas de las entidades participantes en el problema, y sobre las similitudes y diferencias existentes entre el total de las entidades. Por ejemplo, hemos aplicado el algoritmo a la selección de

²³Dado que en (Esuli & Sebastiani, 2007) se calculan valores independientes de *positividad* y *negatividad*, en la tabla 2.4 se muestran los valores de τ_p para ambos *rankings* (*Esuli*⁺ y *Esuli*⁻)

Relaciones utilizadas	PageRank	PolarityRank
Directas: 15467		
Inversas: 15467	0.402	0.362

Cuadro 2.5: Valores de $\tau_{\frac{1}{2}}$ obtenidos para los experimentos de construcción de lexicones sobre grafos equilibrados.

instancias en minería de datos (Vallejo et al, 2010) y al cálculo de la confianza y reputación en redes sociales (Ortega et al, 2011).

2.3.3. Resumen automático de textos de opinión

Tal como hemos comentado en la sección 2.2.3, el resumen automático de textos de opinión se puede enfocar de diversas maneras. Para empezar, podemos distinguir entre resúmenes en forma de texto y resúmenes gráficos. En el primer caso, se trata de una tarea que lleva siendo estudiada desde hace años en el campo del PLN, consistente en obtener pequeños resúmenes textuales a partir de uno o muchos documentos, de manera que se dé cabida a la información más importante de los documentos, disminuyendo en lo posible la redundancia y el tamaño del texto original. En el caso del resumen de textos de opinión, la tarea se puede llevar a cabo de maneras específicas; por ejemplo, se podría tratar de dar una cobertura similar a los documentos con opiniones positivas y a los documentos con opiniones negativas, construir el resumen mostrando de manera diferenciada las opiniones positivas y negativas que más aparecen, o incluso utilizar sólo aquellas secciones del texto que contengan opiniones acerca de un determinado objeto.

Esta última opción (la construcción de resúmenes textuales de documentos de opinión que se centren exclusivamente en las opiniones acerca de un tema determinado), fue objeto de una competición organizada por el *National Institute of Standards and Technology (NIST)* en el marco de la *Text Analysis Conference (TAC)* de 2008. La tarea, de nombre *Opinion Summarization*, consistió en generar resúmenes con la mayor calidad posible de las opiniones acerca de determinados objetivos contenidas en un conjunto de documentos extraídos de blogs. Se proporcionó como entrada una serie de preguntas de opinión (lo suficientemente complejas como para no poder ser contestadas mediante pocas palabras), de manera que el resumen generado debía servir de respuesta a dicha pregunta. Algunos ejemplos de las preguntas que se pretendían responder son las siguientes:

- *What did American voters admire about Rudy Giuliani?*
- *What qualities did not endear Rudy Giuliani to some American voters?*
- *Why did readers support Time's inclusion of Bono for Person of the Year?*
- *Why did people enjoy the movie "Good Night and Good Luck"?*

Además de las preguntas y los documentos de los que realizar el resumen, también se disponía de un conjunto de *snippets* (pequeños trozos de texto) generados por aquellos grupos de investigación que participaron en otra tarea paralela de *question answering*, de manera que estos *snippets* podían ser utilizados para encontrar en los documentos secciones de textos a resumir, sin tener que afrontar toda la problemática asociada al *question answering*.

Ya que en el momento de la celebración de la competición estábamos comenzando nuestra investigación en el área del análisis del sentimiento, decidimos participar en la misma presentando un sistema (Cruz et al, 2009a). Nuestra propuesta obtuvo buenos resultados.

Nuestro sistema en el *TAC 2008 Opinion Summarization Task*

Nuestro sistema se basó en la combinación de los snippets proporcionados para la construcción del resumen. Para aumentar la legibilidad y completar la información, se buscan en los documentos las oraciones más relevantes relacionadas con los *snippets*, y se trabaja con ellas para construir el resumen.

En la figura 2.1 se muestra la arquitectura del sistema. A partir de los documentos html se generan documentos de texto plano, separados por oraciones y convenientemente *tokenizados* (empleamos la herramienta *OpenNLP*²⁴). Todas las oraciones obtenidas son indexadas mediante *Lucene*²⁵.

Para cada uno de los *snippets* de una pregunta dada, buscamos la oración del documento asociado más relevante. Suponemos que los *snippets* han sido extraídos (casi) totalmente de manera literal del texto de los documentos; pretendemos por tanto recuperar esa oración como unidad mínima para componer el resumen. Para recuperar la oración más relevante, seguimos una estrategia de relajación progresiva de la exigencia. Primero, buscamos de manera literal el *snippet* en el documento. Si no se encuentra ninguna oración, buscamos alguna oración que contenga todos los sintagmas que componen el *snippet*. Si aún no tenemos resultado, vamos eliminando progresivamente de la búsqueda los sintagmas compuestos por menos palabras (los de una, los de

²⁴<http://opennlp.sourceforge.net>

²⁵<http://lucene.apache.org/>

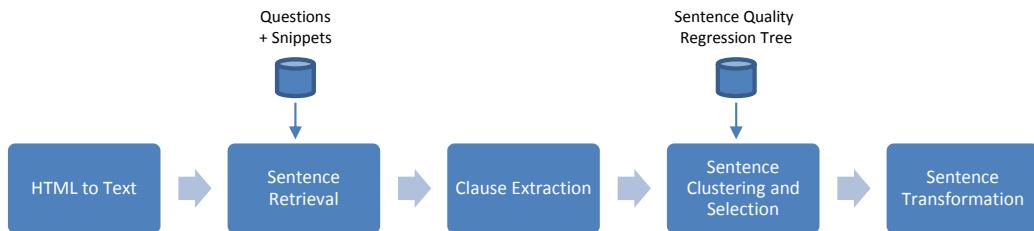


Figura 2.1: Arquitectura de nuestro sistema de resumen

dos,...). Si finalmente ninguna búsqueda consigue resultado, o el resultado obtenido tiene un *score* asignado por *Lucene* menor a 0,75, la recuperación ha fallado. En este caso, utilizamos el *snippet* tal cual como unidad de composición del resumen.

Para cada una de las oraciones obtenidas en la fase de recuperación, buscamos si contiene alguna cláusula menor que la oración completa que contenga todas las palabras no huecas del *snippet* correspondiente. Si encontramos dicha cláusula, la utilizamos en lugar de la oración completa. De esta forma, pretendemos minimizar incoherencias en la estructura del discurso del resumen resultante, manteniendo la información esencial. Sólo en el caso de que excedamos el tamaño máximo exigido por la organización, llevamos a cabo un proceso de selección de oraciones, basándonos en la redundancia. Utilizamos *LingPipe*²⁶ para llevar a cabo un proceso de *clustering* con las oraciones, utilizando la coocurrencia de términos como medida de distancia entre ellas. Para cada uno de los *clusters* generados, escogemos la oración de mayor calidad. Para medir la calidad de las oraciones, desde el punto de vista de si es o no buena candidata a formar parte de nuestro resumen, entrenamos un árbol de regresión en Weka²⁷. Dos integrantes del grupo de trabajo etiquetaron con un valor entero de 1 a 5 la calidad de 84 oraciones extraídas a partir de los *snippets* puestos a disposición como ejemplo por la organización. Se pretendía puntuar negativamente aquellas oraciones que por cuestiones de forma fuera preferible que quedaran fuera del resumen. Las anotaciones con más divergencia entre los anotadores fueron discutidas por estos, y en el resto se utilizó la media entre las dos puntuaciones asignadas. Cada oración fue vectorizada utilizando *features* basados en el árbol sintáctico y en información morfológica.

Una vez obtenidas las oraciones que formarían el resumen, se llevaron a ca-

²⁶<http://alias-i.com/lingpipe/>

²⁷www.cs.waikato.ac.nz/ml/weka/

bo algunas transformaciones sencillas, como añadir puntos de final de oración, eliminar algunas coletillas frecuentes a final de frase (“, *I think.*”, “*huh?*”), o cambiar la persona de algunas oraciones (por ejemplo, “*I think...*” se cambia por “*People think...*”).

Evaluación

La evaluación se llevó a cabo mediante distintas métricas manuales, confeccionadas por 8 asesores. La primera de ellas, conocida como *pyramid evaluation* (Passonneau et al, 2005), utiliza pequeñas porciones de información o *nuggets* que deberían contener los resúmenes para evaluar el contenido de los mismos; el valor que se obtiene es una media armónica entre la precisión y la cobertura del resumen en relación a los *nuggets*, llamada F piramidal. Además del valor obtenido mediante este método, se utilizaron dos medidas más, llamadas *overall fluency/readability* y *overall responsiveness*, obtenidas como la media de las puntuaciones asignadas manualmente por los asesores. En el primer caso, se trata de puntuar la estructura y calidad del resumen desde el punto de vista de la corrección gramatical (*grammaticality*), la no redundancia (*non-redundancy*) y la coherencia estructural (*structure/coherence*). La métrica *overall responsiveness*, por su parte, evalúa lo adecuado del contenido del resumen en relación a las preguntas a las que se debía dar respuesta.

Resultados y conclusiones

Participaron 36 equipos en la competición. Nuestro sistema obtuvo muy buenos resultados en las medidas relacionadas con la cobertura informativa de los resúmenes generados (tabla 2.6). Dos ejecuciones del sistema fueron evaluadas. La única diferencia entre ambas fue que en la segunda ejecución imponemos una calidad mínima a las oraciones que formarán parte del resumen, utilizando para ello el valor generado por el árbol de regresión. Nuestras dos ejecuciones consiguieron el segundo y el tercer mejor resultado en la medida F piramidal, y el primer y segundo mejor resultado en *overall responsiveness*. Creemos que el buen resultado en cuanto a la cobertura de la información relevante de nuestros resúmenes se debió a que pusimos el énfasis en la recuperación de las oraciones más relevantes. Sin embargo, en la métrica *overall fluency/readability* obtuvimos un peor resultado (décimo lugar). En particular, la peor puntuación la obtuvimos en la métrica relacionada con la redundancia, lo que nos hace pensar que deberíamos haber aplicado la fase de *clustering* en todos los casos (sólo la aplicamos en el caso de obtener un resumen demasiado largo). La segunda ejecución obtuvo peores resultados

	Run 1	pos.	Run 2	pos.
Pyramid F-score	0.490	2	0.489	3
Grammaticality	5.591	10	5.545	12
Non-redundancy	5.318	29	5.364	28
Structure/Cohherence	3.273	9	2.682	19
Overall fluency/readability	3.909	10	3.591	19
Overall responsiveness	5.773	1	5.409	2

Cuadro 2.6: Evaluación del sistema presentado al *TAC 2008 Opinion Summarization Task*

en todas las medidas. Creemos que esto pudo deberse a la escasez de datos de entrenamiento con los que se construyó el árbol de regresión para medir la calidad de las oraciones.

La participación en la competición nos condujo a algunas conclusiones interesantes. Si bien la tarea propuesta resultó estar más alejada del área del análisis del sentimiento de lo que en un primer momento pensamos (quizás la tarea paralela de respuesta a preguntas de opinión estaba más próxima), sí nos hizo plantearnos la utilidad práctica de los sistemas de clasificación basada en la opinión en los que veníamos trabajando. De cara a la resolución de la tarea de la competición, habría sido más útil disponer de un sistema que fuese capaz de reconocer las opiniones individuales contenidas en los documentos, no sólo clasificando dichas opiniones de acuerdo a su polaridad sino identificando también los objetos concretos sobre los que tratan dichas opiniones. Centramos en ese momento nuestros esfuerzos en el tema central del presente trabajo de tesis: la extracción de opiniones.

Capítulo 3

Extracción de opiniones sobre características

Resumen: En este capítulo realizamos un repaso bibliográfico de los trabajos relacionados directamente con la extracción de opiniones sobre características. En la parte final del capítulo proponemos un resumen comparativo de las características principales de las distintas aproximaciones al problema, que servirá de base para introducir nuestra propuesta en el capítulo 4.

3.1. Introducción

La extracción de opiniones sobre características es definida de manera distinta según los autores, aunque todos ellos persiguen un mismo objetivo: la obtención de representaciones estructuradas de las opiniones contenidas en textos subjetivos acerca de las distintas características de un objeto determinado, a partir de las cuales puedan construirse representaciones resumidas de los documentos de entrada en base a la agregación de las opiniones extraídas. Se trata de una tarea que engloba múltiples problemas a resolver, desde la detección de menciones (explícitas o implícitas) a las distintas características de los objetos en análisis, hasta la clasificación basada en la polaridad de las oraciones que contienen las opiniones o de las propias representaciones de las opiniones.

3.2. Principales trabajos relacionados

A continuación repasamos los trabajos fundamentales relacionados con la tarea, siguiendo en lo posible un orden cronológico. En la siguiente sección llevaremos a cabo un estudio comparativo de los trabajos más interesantes.

3.2.1. Primeros trabajos

El primer trabajo que sugirió la extracción de opiniones individuales como medio para construir un resumen de documentos de análisis de productos fue llevado a cabo por *Dave et al.* en 2003 (Dave et al., 2003). En realidad, el trabajo plantea la construcción de un clasificador binario supervisado de documentos basado en la opinión. Se experimenta con diversas técnicas de aprendizaje automático (máxima entropía, *support vector machines*, *expectation maximization* y *Naive Bayes*), y se compara con un clasificador basado en representación vectorial de los documentos. Para este último caso, se proponen distintas *features* para representar los documentos (unigramas, bigramas, trigramas, sufijos/prefijos, transformaciones lingüísticas, ...). Es precisamente a partir de estas *features* que el trabajo propone la construcción de resúmenes multidocumento agrupando las *features* encontradas con mayor ganancia de información para las clases positiva y negativa. Aunque aún no se define claramente la extracción de opiniones sobre características (nótese que las *features* en el trabajo se refieren a características lingüísticas utilizadas por el clasificador, más que a características de los productos), se trata del primer trabajo que propone la identificación de opiniones individuales para la construcción de un resumen multidocumento, lo que sin duda dio pie a los sucesivos trabajos en dicha dirección y a la definición de la tarea de extracción de opiniones.

Hu y Liu definieron por primera vez la tarea de extracción de opiniones sobre características de productos en 2004 (Hu & Liu, 2004b,a), desde la perspectiva de la construcción de un resumen multidocumento de las opiniones vertidas sobre un determinado producto. Se trata de un resumen estructurado, en el que aparecerán el número de opiniones positivas y negativas contenidas en los documentos para cada una de las características opinables del producto (ver figura 3.1). Para conseguir generar este tipo de resúmenes, se define la tarea de extracción de opiniones sobre características, que consta de los siguientes pasos:

1. Encontrar en los documentos menciones a características de los productos.

Digital_camera_1:

Feature: **picture quality**

Positive: 253

<individual review sentences>

Negative: 6

<individual review sentences>

Feature: **size**

Positive: 134

<individual review sentences>

Negative: 10

<individual review sentences>

...

Figura 3.1: Ejemplo de resumen a partir de opiniones sobre características (propuesto por *Hu* y *Liu* (Hu & Liu, 2004a))

2. Identificar oraciones en las que se vuelcan opiniones sobre las características anteriores, y decidir si se trata de opiniones positivas o negativas.

En realidad, cada uno de los pasos anteriores puede ser descompuesto en más pasos (por ejemplo, en el paso 2 son problemas distintos decidir las características sobre las que se vuelcan las opiniones o clasificar las opiniones como positivas o negativas); la división planteada responde a la publicación del trabajo en dos artículos distintos, uno explicando el primer punto (Hu & Liu, 2004b) y otro el segundo (Hu & Liu, 2004a).

Resumimos a continuación las técnicas utilizadas para llevar a cabo cada una de las subtareas. En primer lugar, la detección de las características se lleva a cabo de manera no supervisada, utilizando una técnica conocida como *association rule mining* (Agrawal & Srikant, 1994). Las características obtenidas son sintagmas nominales de un máximo de 3 palabras, extraídas de entre los sintagmas nominales más frecuentemente aparecidos en los documentos. Se lleva a cabo un proceso de selección, eliminando los candidatos redundantes o aquellos que no tienen ningún sentido. Para conseguir ampliar la cobertura, se propone un método de *bootstrapping* consistente en buscar adjetivos frecuentes que aparezcan en el contexto de las características anteriormente descubiertas, y buscar a su vez nuevas características a partir de los sintagmas nominales que aparecen más frecuentemente en el contexto de dichos adjetivos. Los resultados obtenidos al evaluar el método sobre un

corpus de *reviews* de 5 productos distintos fueron de un 80 % de cobertura y un 72 % de precisión. La mayor desventaja de la propuesta es la enorme cantidad de características obtenidas (hasta 96 distintas para una cámara de fotos); muchos de los sintagmas nominales extraídos son en realidad distintos sinónimos de una misma característica, sin que el sistema sea capaz de detectarlo.

El segundo paso se lleva a cabo de la siguiente manera. En primer lugar se buscan las *palabras de opinión* (palabras con carga subjetiva), que serán aquellos adjetivos que aparezcan en el contexto de alguna de las características detectadas en la etapa anterior. Aquellas oraciones que contienen menciones a características y palabras de opinión son consideradas *oraciones de opinión*. Para poder decidir la polaridad positiva o negativa de dichas oraciones, previamente se calculan las orientaciones semánticas de las palabras de opinión participantes. Esto se hace mediante una técnica bastante sencilla (y no demasiado fiable, por otro lado), consistente en basarse en las relaciones de sinonimia y antonimia entre adjetivos en WordNet para, a partir de un conjunto de 30 semillas positivas y negativas, inferir la polaridad de las palabras de opinión. La hipótesis en este caso es que dos adjetivos que aparezcan como sinónimos en WordNet compartirán la misma polaridad, de igual modo que dos adjetivos que aparezcan como antónimos tendrán polaridades opuestas. Las oraciones de opinión son clasificadas a partir de la suma de las orientaciones semánticas de las palabras de opinión contenidas (usando +1 para las positivas y -1 para las negativas); se tienen en cuenta las posibles apariciones de negaciones modificando a las palabras de opinión, invirtiéndose el signo de la polaridad de las mismas. Cuando no se consigue clasificar una oración de opinión, se utiliza la polaridad de la última oración de opinión, a partir de la premisa de que las opiniones de un mismo signo tienden a aparecer en un contexto próximo. Usando el mismo corpus anterior, el método descrito consiguió reconocer el 69,3 % de las oraciones de opinión, con una precisión del 64,2 %; el 84,2 % de las oraciones de opinión reconocidas fueron correctamente clasificadas. Estos resultados son relativamente buenos, dada la dificultad de la tarea y la simplicidad del método propuesto, lo que junto a la utilidad de la salida obtenida convirtió a la extracción de opiniones sobre características en una tarea de gran interés para la comunidad investigadora.

Son varias las limitaciones de la propuesta. En primer lugar, la clasificación de polaridad se realiza a nivel de oración, y no de opiniones. Esto implica que no se contemplan los casos en los que en una misma oración aparecen opiniones de signo contrario acerca de distintas características. La evaluación se hace a nivel de oraciones, lo que no se ajusta bien a la definición de la tarea, que plantea la extracción de opiniones a nivel de característica para la construcción del resumen mostrado en la figura 3.1. El trabajo no

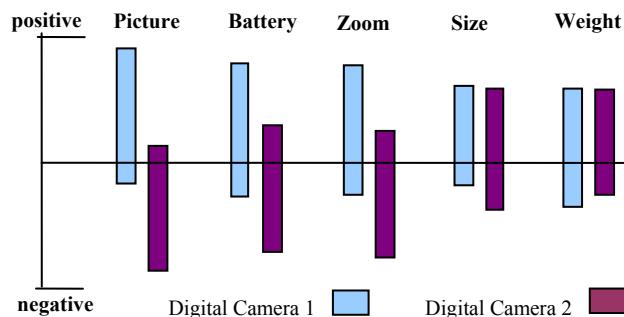


Figura 3.2: Ejemplo de comparativa entre productos propuesta por *Liu et al.* (Liu et al., 2005)

aborda la extracción de opiniones sobre características implícitas (opiniones en las que la característica no aparece mencionada directamente). Por otro lado, las únicas palabras de opinión consideradas son los adjetivos, cuando existen otras categorías morfosintácticas que pueden expresar opiniones (por ejemplo, verbos como “gustar” o nombres como “basura”). Además, no se consideran posibles expresiones de opinión formadas por más de una palabra (por ejemplo, “obra maestra” o “demasiado grande”). El cálculo de la orientación semántica de los adjetivos es demasiado simplista, no teniendo en cuenta las posibles dependencias con el dominio o con la característica. Por último, la asociación entre las características y las palabras de opinión se realiza a partir de la aparición consecutiva o muy cercana entre ambas, quedando por tanto fuera otras posibles construcciones (considérese por ejemplo la oración *“The amount of detail in these headphones, that I bought a month ago, is stunning”*).

La propuesta fue modificada y ampliada en sucesivos trabajos de *Liu* y otros. En (Liu et al, 2005), abordan la extracción de opiniones a partir de la sección de pros y contras de los *reviews*. El trabajo se centra en la extracción de características (la clasificación de las opiniones está implícita, considerándose positivas aquellas que aparecen en la sección “pros” y negativas las que aparecen en la sección “contras”). Se trata por ello de un trabajo menos completo que los anteriores, pero con algunas aportaciones interesantes, como la aparición por primera vez de un tratamiento para las características implícitas (mediante un etiquetado manual de las relaciones entre algunos términos y dichas características), la agrupación de algunas de las características obtenidas mediante el uso de relaciones de sinonimia en WordNet (método no del todo satisfactorio, entre otras cosas porque no se consideran relaciones de sinonimia entre características multipalabra), o la participación optativa de analistas humanos que corrijan los errores del

sistema. El trabajo también propone algunas representaciones interesantes de comparativas de productos basadas en características (ver figura 3.2). En (Ding et al, 2008b), se plantea una función para el cálculo de las orientaciones semánticas de las opiniones a nivel de características (en lugar de a nivel de oraciones como anteriormente), basada en la suma de las orientaciones semánticas de las palabras de opinión de la oración donde aparece una característica divididas por las distancias entre las palabras de características y de opinión. Además se proponen algunas reglas de contexto que pretenden matizar las orientaciones semánticas de las palabras de opinión; sigue sin considerarse de manera completa la dependencia entre la orientación semántica y el dominio de aplicación o las características, pero al menos se trata de un paso claro en esa dirección. En otros trabajos más recientes, los autores han tratado de resolver el problema de la agrupación de características sinónimas mediante la aplicación de técnicas de clustering a posteriori (Zhai et al, 2010).

3.2.2. La extracción de opiniones desde la perspectiva de la extracción de información

La tarea de extracción de opiniones sobre características puede entenderse como una tarea de extracción de información, ya que se trata de encontrar en los textos ciertas entidades (menciones a características, palabras de opinión) y relaciones entre las mismas. Es desde este punto de vista desde el que afrontan el problema Popescu y Etzioni (Popescu & Etzioni, 2005; Popescu et al, 2005), adaptando un sistema de extracción de información llamado *Know-ItAll* (Etzioni et al, 2004). En lugar de trabajar a nivel de las oraciones o sintagmas donde aparecen menciones a características, e inducir la orientación semántica de las opiniones a partir de las palabras de opinión que aparecen en las mismas, en este trabajo se enlazan características individuales a las palabras de opinión individuales que las modifican, obteniéndose con ello representaciones de opiniones individuales. La captura de estas relaciones se realiza de acuerdo a 10 patrones sintácticos confeccionados manualmente y basados en relaciones de dependencias (para una introducción al análisis de dependencias y sus particularidades con respecto al análisis sintáctico completo tradicional, ver (Mel'čuk, 1988)).

La extracción de las características explícitas se lleva a cabo mediante el cálculo de la información mutua puntual (*Pointwise Mutual Information*, el mismo algoritmo utilizado en (Turney, 2002a) para el cálculo de la orientación semántica de términos de opinión) entre los sintagmas nominales más

frecuentes del texto de los documentos y determinadas expresiones que los autores llaman *meronymy discriminators*. Por ejemplo, para encontrar características de un escáner, se calculan los valores de PMI entre los candidatos y expresiones como “*of scanner*” o “*scanner has*”, y se seleccionan aquellos candidatos que obtengan un valor mayor a un umbral. El artículo no describe todas las expresiones de meronimia empleadas, ni el método para fijar el valor umbral de PMI. Una aportación interesante con respecto a las características es la diferenciación de varios tipos de características: se distingue entre *partes* y *propiedades* (por ejemplo, el *tamaño* frente a la *tapa* de un escáner), características de las partes (por ejemplo, la *duración* de la *batería* de una cámara), conceptos relacionados (por ejemplo, la *imagen* en un escáner) y características de conceptos relacionados (por ejemplo, la *calidad* de la *imagen* de un escáner). El sistema distingue unos tipos u otros de características mediante el empleo de distintas expresiones de meronimia en cada caso, y el uso de las relaciones IS-A de WordNet (para el caso de las partes). Más allá de la división concreta de tipos de característica propuesta, es interesante el planteamiento de una cierta jerarquía entre las características de un producto, lo que deja entrever la posible utilidad del uso de *taxonomías de características*, que será una de las bases de nuestra propuesta. En (Popescu et al, 2005) se propone un método para la extracción de características implícitas (aunque está muy poco documentado en el artículo), basado en realizar un *clustering* de las palabras de opinión y asignándoles algunas de las características de tipo propiedad previamente extraídas. Los autores dicen obtener con el tratamiento de las características implícitas una mejora de alrededor de dos puntos porcentuales en la precisión de la extracción de características y de seis puntos en la cobertura.

Para el cómputo de la orientación semántica de las palabras de opinión, se parte de estimaciones iniciales realizadas mediante PMI (de forma similar a (Turney, 2002a)). Estos valores iniciales son contextualizados, mediante el establecimiento de ciertas restricciones locales que afectan a las palabras de opinión; por ejemplo, las palabras de opinión que ocurren en contextos cercanos o las que participan en construcciones conjuntivas directas tienden a compartir la misma polaridad. Todas las restricciones participantes son tenidas en cuenta a través de una técnica de clasificación no supervisada llamada *relaxation labeling* (Hummel & Zucker, 1983). Al igual que ocurre con las técnicas de contextualización de las orientaciones semánticas propuestas en (Ding et al, 2008b), se trata de un intento algo limitado de calcular valores de orientación semántica orientados al dominio y a las características de aplicación.

3.2.3. Otros trabajos

Aunque los trabajos anteriormente expuestos son con diferencia los que mayor número de citas reúnen, existen otros trabajos a lo largo de los últimos años que merece la pena reseñar. En (Carenini et al, 2005), se propone un método no supervisado de extracción de características explícitas similar al propuesto por *Hu y Liu* (Hu & Liu, 2004b), pero con la novedad de que las características extraídas son vinculadas a una taxonomía de características previamente disponible (se emplean taxonomías de productos desarrolladas por la organización *ConsumerReports.org*). Esta vinculación se realiza de acuerdo a la similitud entre las palabras utilizadas para nombrar la característica en el texto y en la taxonomía. Aunque los resultados del método son modestos, y no se contemplan características implícitas ni se aportan métodos para la construcción de las taxonomías para nuevos dominios, el trabajo es el primero en señalar la utilidad de las taxonomías de características en las tareas de extracción y resumen de opiniones sobre características.

El interés en el área de las grandes corporaciones con negocios en la red como *IBM* o *Google* queda de manifiesto por algunas publicaciones como (Yi & Niblack, 2005) o (Blair-Goldensohn et al, 2008). En el caso de (Yi & Niblack, 2005) se trata de una propuesta práctica, independiente del dominio y basada en técnicas muy simples. La extracción de características explícitas (no se tienen en cuenta las características implícitas) se hace capturando sintagmas nominales a comienzo de oración que comiencen por el determinante “*the*” (los autores argumentan que cuando se comienza a hablar de una característica determinada se suele comenzar de esta manera). La búsqueda de expresiones de opinión en las que participen las características anteriores se lleva a cabo con patrones sintácticos simples elaborados a mano. Igualmente, el lexicón de orientaciones semánticas de las palabras de opinión utilizado es de construcción manual.

El trabajo descrito en (Blair-Goldensohn et al, 2008), por su parte, se centra en la extracción de opiniones en dos dominios concretos (hoteles y restaurantes). Se trata de una propuesta orientada a dominio, en la que se contempla la definición manual de ciertas características abstractas de amplio significado propias del dominio considerado (por ejemplo, *food*, *decor*, *service*, *value* y *other* para el dominio de los restaurantes). Las características explícitas habituales, extraídas según el método de *Hu y Liu* (Hu & Liu, 2004b), son asociadas a estas características abstractas de manera automática, utilizando un clasificador entrenado a partir de un conjunto de oraciones anotadas manualmente. Los autores argumentan que el esfuerzo requerido para estas anotaciones es pequeño comparado con el beneficio obtenido en la calidad de las opiniones extraídas por el sistema y su adaptación a un

dominio concreto de aplicación. En el trabajo se emplea el término *aspects* (aspectos) en lugar de *features* (características). Se trata de un término que se adapta quizás mejor al dominio de aplicación elegido (servicios locales). Este cambio en la nomenclatura se ha mantenido en otros trabajos recientes (Titov & McDonald, 2008b,a; Brody & Elhadad, 2010; Zhao et al, 2010; Jo & Oh, 2011). Estos trabajos emplean distintas modificaciones de un modelo no supervisado de detección de la temática o *topic* en textos denominado *Latent Dirichlet Allocation* (LDA) (Blei et al, 2003). Esta técnica parte de la idea de que un documento está generado a partir de una mezcla de temas, y es capaz de identificar conjuntos de palabras de los documentos relacionadas con distintas temáticas. Las variaciones propuestas en el contexto de la extracción de opiniones van encaminadas a detectar características abstractas (o aspectos, según la nomenclatura utilizada en los trabajos en cuestión), definidas a partir de conjunto de palabras. Por ejemplo, en el dominio de los hoteles se puede identificar un *aspecto* determinado por palabras como *breakfast*, *coffee*, *continental*, *morning* o *free* (podríamos asignarle el nombre *breakfast* a dicho aspecto). Aunque estas palabras tienen cierta relación con las palabras de característica utilizadas en las propuestas más cercanas a la extracción de información (como (Popescu & Etzioni, 2005)), no se trata en todos los casos de menciones a características, propiamente dichas. Por ejemplo, *morning* no es una mención de la característica *breakfast*, aunque su aparición en el texto esté relacionada con la misma. Incluso pueden aparecer palabras de opinión (como *free*, en el ejemplo anterior). Además, los términos extraídos mediante LDA y técnicas derivadas son siempre palabras individuales. Por tanto, los trabajos que se basan en estas técnicas no afrontan la tarea de extracción de opiniones como una tarea de extracción de información; más bien, tratan de detectar oraciones u otras unidades de texto menor en las que se haga referencia a algún aspecto del objeto que está siendo evaluado, para posteriormente clasificar dichas porciones de texto según la polaridad. Este enfoque de la tarea queda alejado del que seguiremos en nuestra propuesta, que está más cercana a la definición de la tarea planteada por Popescu y Etzioni (Popescu & Etzioni, 2005; Popescu et al, 2005).

Son reseñables también los trabajos de Alexandra Balahur (Balahur & Montoyo, 2008a,b), que determinan las características de un producto determinado a partir de un conjunto de características comunes a todos los productos, y haciendo uso de las relaciones de sinonimia, meronimia e hipónimia de WordNet. Se proponen también métodos para encontrar adjetivos relacionados con cada característica, usando las relaciones “*has attributes*” de WordNet y las relaciones *PropertyOf* y *CapableOf* de ConceptNet(Havasi et al, 2007). La búsqueda de opiniones se lleva a cabo entonces para un texto de entrada a partir de la aparición de los términos anteriores. Los adjetivos

vos relacionados con dichas apariciones mediante determinadas relaciones de dependencia sintáctica son utilizados para clasificar las opiniones. Esta clasificación se lleva a cabo mediante un clasificador SVM que trabaja sobre representaciones vectoriales de los términos anteriores, calculadas mediante distancias a una serie de términos paradigmáticos mediante *Normalized Google Distance*(Cilibrasi & Vitanyi, 2005). Éste es quizás el punto más débil de la propuesta de cara a su implementación práctica, puesto que la obtención de esta representación vectorial depende de consultas al buscador Google, lo cual no es viable en un sistema en explotación.

3.3. Resumen comparativo de trabajos

En las tablas 3.1 y 3.2 hemos resumido las características principales de las distintas propuestas de trabajos relacionados con la extracción de opiniones. Aunque cada propuesta tiene un enfoque y alcance distintos, hemos utilizado una división de la tarea en tres subproblemas que permite abarcárlas a todas: la extracción de las características, que consiste en identificar las características opinables del producto, así como las menciones concretas a las mismas en los documentos de entrada; la búsqueda de palabras de opinión en los documentos de entrada, generalmente relacionadas con las menciones a características anteriormente detectadas (aunque hay trabajos que lo abordan de manera independiente); y por último la clasificación de las opiniones. Sólo seis de los 14 trabajos considerados abordan los tres subproblemas.

3.3.1. Extracción de características

En la extracción de características distinguimos entre los trabajos que no necesitan de participación humana (automáticos) y los que sí (semiautomáticos). En general, todos los trabajos que están orientados al dominio en mayor o menor medida requieren de una cierta participación humana, ya sea para anotar algunos documentos de entrenamiento, para elaborar una taxonomía de características o para poner nombre a los conjuntos de palabras de característica encontrados (en el caso de los trabajos que trabajan con aspectos). El algoritmo más utilizado es *Association Rule Mining* (ARM)(Agrawal & Srikant, 1994), empleado en primer lugar en (Hu & Liu, 2004b) y del que otros trabajos proponen algunas modificaciones. Los trabajos de *Popescu et al.* (Popescu & Etzioni, 2005; Popescu et al, 2005) emplean el algoritmo *Pointwise Mutual Information* (PMI) para medir la correlación entre los candidatos a característica y los indicadores de meronimia, y *Yi* y *Niblack* (Yi & Niblack, 2005) emplean la medida estadística *likelihood ratio*

(LR). Los trabajos enfocados a aspectos utilizan diversas adaptaciones del algoritmo *Latent Dirichlet Allocation* (LDA) (Blei et al, 2003), tales como *Multi-Grain LDA* (MG-LDA), *Local-LDA* o *Aspect and Sentiment Unification Model* (ASUM). Este último algoritmo no sólo extrae características (aspectos), sino que también clasifica las menciones encontradas en los documentos según la polaridad. Sólo dos trabajos se ocupan de la extracción de características implícitas, aparte de los trabajos orientados a aspectos; en este último caso, las palabras extraídas como aspectos incluyen en ocasiones palabras de opinión relacionadas con una característica, lo cual está relacionado con el concepto de característica implícita de los trabajos orientados a características.

Una vez obtenida la lista de menciones a características, algunos trabajos llevan a cabo tratamientos automáticos posteriores, encaminados a agrupar los términos que hacen referencia a la misma característica, diferenciar las características según el tipo (por ejemplo, distinguir si se trata de una parte o una propiedad del producto), o enlazar las características a una taxonomía previamente disponible. En los trabajos orientados a aspectos, es común la asignación manual de un nombre a cada uno de los *clusters* de aspectos obtenidos.

3.3.2. Búsqueda de palabras de opinión

El acercamiento más simple para la búsqueda de palabras y términos de opinión es la utilización del contexto de las menciones de características previamente detectadas, ya sea mediante ventanas de palabras o usando directamente todas las palabras de la oración. Otros trabajos utilizan la información proporcionada por un *parser*, ya sea sintáctico o de dependencias, para buscar las palabras de opinión relacionadas con las menciones de características a partir de patrones predefinidos. Una propuesta diferente es la de *Zhao et al.* (Zhao et al, 2010), que utiliza un clasificador de máxima entropía entrenado a partir de un conjunto de documentos anotados para distinguir, de entre el conjunto de aspectos obtenidos previamente mediante el algoritmo LDA, cuáles son menciones a característica y cuáles palabras de opinión

En algunos casos, sólo se tienen en cuenta determinadas palabras, en función de la categoría morfosintáctica de las mismas; por ejemplo, es común considerar sólo los adjetivos, ya que la mayor parte de las palabras de opinión lo son. Otros trabajos no establecen este tipo de restricciones, permitiendo la captura de palabras de opinión de otras categorías morfosintácticas (verbos como *like*, nombres como *master-piece*, . . .).

3.3.3. Clasificación de opiniones

No todos los trabajos llevan a cabo la clasificación de opiniones individuales (pares característica-palabras de opinión). Algunos clasifican las oraciones que contienen menciones a características, lo que implica una simplificación en tanto que algunas oraciones contienen varias opiniones individuales. Los trabajos orientados a aspectos no suelen afrontar la clasificación de las opiniones (salvo *Blair et al.*(Blair-Goldensohn et al, 2008)), aunque algunos llevan a cabo experimentos de clasificación a nivel de aspectos (Brody & Elhadad, 2010) o de los documentos completos (Jo & Oh, 2011). Independientemente de la unidad a clasificar, la mayor parte de los trabajos lleva a cabo una clasificación binaria basada en la polaridad de las opiniones, aunque también existen trabajos que distinguen tres categorías (positiva, negativa y neutra), o más (en la tabla 3.2, hemos indicado este último caso como *rating*). El trabajo de *Popescu y Etzioni*(Popescu & Etzioni, 2005) propone una clasificación binaria de las opiniones, pero posteriormente las ordena de mayor a menor intensidad.

La mayoría de los trabajos basan la clasificación en los valores de orientación semántica estimados para los términos de opinión. Esta estimación se realiza mediante técnicas basadas en semillas, como el algoritmo *Pointwise Mutual Information* (PMI) o algoritmos basados en grafos¹ construidos a partir de las relaciones de WordNet o de las expresiones conjuntivas entre adjetivos encontradas en un corpus (de manera similar a (Hatzivassiloglou & McKeown, 1997)). Algunos trabajos emplean lexicones construidos manualmente. Un caso especial es el trabajo de *Jo et al.*(Jo & Oh, 2011), en el que el modelo utilizado para la extracción de las menciones a aspectos obtiene de manera conjunta la polaridad de dichas apariciones.

A partir de las estimaciones individuales anteriores, la mayoría de los trabajos llevan a cabo la clasificación según el resultado de la suma de los valores obtenidos (en algún caso, ponderando dichos valores por las distancias entre las palabras de opinión y la característica en cuestión), o mediante el empleo de técnicas que contextualizan dichas estimaciones (por ejemplo, mediante el empleo de clasificadores supervisados de máxima entropía o usando el algoritmo *Relaxation Labeling* (Hummel & Zucker, 1983)).

¹Algunos trabajos aplican a los grafos un algoritmo de propagación de etiquetas (Zhu & Ghahramani, 2002).

3.3.4. Supervisado vs. no supervisado e influencia del dominio

Decimos que un algoritmo es supervisado si se apoya en cierto conocimiento que debe ser previamente construido. Por ejemplo, un clasificador supervisado es aquel que se apoya en un modelo construido a partir de un conjunto de documentos clasificados manualmente. Esta dependencia con respecto a recursos que contengan conocimiento sobre el problema es al mismo tiempo una virtud y un inconveniente, ya que permite a los algoritmos adaptarse mejor a los problemas a cambio de necesitar un esfuerzo adicional para la elaboración de los recursos.

En un intento por construir soluciones lo más genéricas posible, la mayoría de las propuestas estudiadas utilizan métodos no supervisados para resolver cada uno de los subproblemas anteriormente descritos. Existen sin embargo excepciones. Por ejemplo, en *Liu et al.*(Liu et al, 2005) se etiquetan las menciones de características en una serie de documentos para a partir de los mismos obtener patrones morfosintácticos para detectar características en nuevos textos de entrada. Además, se permite a través de una aplicación la corrección manual de las características detectadas a partir de los patrones. En *Blair et al.*(Blair-Goldensohn et al, 2008) se utiliza un corpus anotado para entrenar un clasificador capaz de decidir si en una oración de entrada se habla de algún aspecto del producto, y otro para determinar la polaridad de dichas oraciones. En *Zhao et al.*(Zhao et al, 2010) también se entrena un clasificador, en este caso para decidir si una determinada palabra relacionada con un aspecto es una mención al mismo o una palabra de opinión relacionada.

Algunas de las técnicas utilizadas, aún no siendo estrictamente supervisadas, se apoyan en recursos construidos manualmente o requieren de cierta participación humana. Por ejemplo, la mayoría de los trabajos orientados a aspectos requieren que se asignen nombres manualmente a los *clusters* de aspectos obtenidos automáticamente. Algunos trabajos utilizan lexicones de opinión construidos manualmente, en los que constan las orientaciones semánticas de un conjunto de palabras y expresiones de opinión. Por otra parte, los trabajos que utilizan patrones sintácticos o de dependencias para buscar las palabras de opinión relacionadas con una mención de característica determinada disponen de conjuntos de patrones elaborados manualmente.

Una de nuestras hipótesis de partida es que la utilización de algunos recursos que proporcionen conocimiento para la resolución de cada uno de los subproblemas puede suponer una considerable ventaja, especialmente si conocemos previamente el dominio de aplicación en el que se utilizará el sistema de extracción de opiniones y generamos los recursos específicamente para

dicho dominio. Efectivamente, existe una relación clara entre la utilización de métodos supervisados y la orientación al dominio del sistema. Por ejemplo, si utilizamos un clasificador supervisado para estimar las orientaciones semánticas de las palabras de opinión, y entrenamos dicho clasificador a partir de un conjunto anotado de documentos de un dominio determinado, las orientaciones semánticas obtenidas serán específicas del dominio. En nuestra propuesta, el dominio de aplicación será tenido en cuenta en la resolución de cada una de los subproblemas, desde la extracción de características (que se apoyará en taxonomías de características específicas del dominio) hasta la clasificación de las opiniones (utilizando un lexicón específico del dominio), pasando por la detección de características implícitas específicas del dominio o por la búsqueda de palabras de opinión (utilizando patrones de dependencias inducidos a partir de documentos del dominio). Ninguno de los trabajos analizados que llevan a cabo clasificación a nivel de las opiniones utiliza valores de orientación semántica orientados al dominio, aunque algunos plantean métodos de contextualización de los mismos, permitiendo matizar sobre la marcha los valores concretos de un término a partir de los valores de otros términos del contexto.

3.3.5. Analizadores y recursos lingüísticos

Todos los trabajos hacen uso en mayor o menor medida de analizadores lingüísticos como parte del preprocesado de los textos. Además de la separación en oraciones, todos utilizan etiquetadores morfosintácticos. Las distintas técnicas planteadas, tanto para la extracción de las características como para la detección de palabras de opinión y la clasificación de las mismas, requieren que dicha información esté disponible. La mayoría de los trabajos emplean técnicas de generalización léxica como el *stemming* o la lematización. El análisis sintáctico superficial o *chunking* también es usado en muchas propuestas, permitiendo trabajar a nivel de sintagmas. Algunos trabajos hacen uso de los sintagmas como unidad de análisis, llevando a cabo la clasificación de polaridad de aquellos sintagmas que contienen opiniones. Otros utilizan los sintagmas junto a la información morfosintáctica para definir patrones sintácticos que permiten capturar relaciones entre las características y posibles palabras de opinión. Sólo unos pocos trabajos hacen uso de analizadores sintácticos completos o analizadores de dependencias. Este tipo de análisis permite la captura de relaciones más complejas y entre constituyentes más alejados en las oraciones, superando a las soluciones basadas en ventanas contextuales de palabras. Algunos trabajos puntuales hacen uso de herramientas propias más sofisticadas de recuperación y extracción de información; es el caso de Yi & Niblack (2005), que hacen uso de *WebFountain*, y de Popescu

& Etzioni (2005), que utilizan la herramienta *KnowItAll*.

Además de analizadores, es común en todos los trabajos el empleo de diversos recursos lingüísticos, siendo el principal WordNet(Fellbaum, 1998), que es utilizado para estimar las polaridades de las palabras de opinión mediante distintos algoritmos y para agrupar características extraídas similares, entre otras cosas. Otros recursos que han sido empleados son General Inquirer (Stone, 1966), Micro-WNOp (Cerini et al, 2007) y SentiWordNet (Baccianella & Sebastiani, 2010).

3.3.6. Evaluación de la tarea

Creemos que uno de los puntos débiles de la mayoría de los trabajos que hemos analizado es el referente a la evaluación de las distintas propuestas. Cada trabajo define sus propias métricas de evaluación, las cuales generalmente no abarcan la tarea completa (extracción de características, búsqueda de palabras de opinión y clasificación de las opiniones), sino que evalúan distintos aspectos de las propuestas. Por citar algunos ejemplos, el trabajo de Popescu & Etzioni (2005) evalúa: (1) la extracción de características explícitas, (2) la clasificación de palabras de opinión individuales, dadas la oración y la característica como entradas, y (3) la extracción y clasificación de expresiones de opinión (pares formados por menciones a característica y palabras de opinión relacionadas), también con la oración y la característica como entradas. Los trabajos de Hu & Liu (2004a), por su parte, evalúan la extracción de características explícitas, por un lado, la detección de oraciones que contienen opiniones, por otro, y la clasificación de dichas oraciones, finalmente. Todas estas evaluaciones son simplificaciones de la tarea completa.

En nuestro trabajo, hemos tratado de evaluar la propuesta de manera completa y definiendo el alcance de la tarea de la forma más precisa posible. Además, los documentos anotados utilizados en dicha evaluación han sido puestos a disposición de la comunidad, para facilitar la comparación con otras propuestas.

Cuadro 3.1: Resumen de características de trabajos relacionados con la extracción de opiniones: extracción de características

Referencia	Tipo	Algoritmo	Superv.	Extracción de características		
				Orientado a dominio	Caract. Implíc.	Tratamiento posterior
(Hu & Liu, 2004b)	Automát.	ARM	No	No	No	Ninguno
(Hu & Liu, 2004a)	Automát.	ARM	No	No	No	Ninguno
(Liu et al, 2005)	Semiaut.	ARM	Sí	Sí	Sí	Agrupación de sinónimos (WordNet)
(Popescu & Etzioni, 2005)	Automát.	PMI	No	No	No	Diferenciación de tipos de características (parts, properties, ...)
(Popescu et al, 2005)	Automát.	PMI	No	No	Sí	Diferenciación de tipos de características (parts, properties, ...)
(Carenini et al, 2005)	Semiaut.	ARM	No	En parte (taxonomías)	No	Mapeo a taxonomía (<i>similarity matching</i>)
(Yi & Niblack, 2005)	Automát.	LR	No	No	No	Ninguno
(Ding et al, 2008b)	-	-	-	-	-	-
(Blair-Goldensohn et al, 2008)	Semiaut.	ARM + ME	Sí	Sí	No	Construcción y <i>ranking</i> de lista combinada de características dinámicas y estáticas
(Titov & McDonald, 2008b)	Semiaut.	MG-LDA	En parte	En parte	Sí* *aspectos	Asignación manual de nombres a aspectos
(Titov & McDonald, 2008a)	Automát.	MG-LDA	No	En parte	Sí* *aspectos	Asignación automática de las características obtenidas a los aspectos puntuados en los reviews del dominio elegido
(Brody & Elhadad, 2010)	Semiaut.	Local-LDA	En parte	En parte	Sí* *aspectos	Asignación manual de nombres a aspectos
(Zhao et al, 2010)	Semiaut.	LDA	En parte	En parte	Sí* *aspectos	Asignación manual de nombres a aspectos
(Jo & Oh, 2011)	Semiaut.	ASUM	En parte	En parte	Sí* *aspectos	Asignación manual de nombres a aspectos
Nuestra propuesta	Semiaut.	Bootstrap	Sí	Sí	Sí	Construcción de taxonomía

3.3. Resumen comparativo de trabajos

77

Cuadro 3.2: Resumen de características de trabajos relacionados con la extracción de opiniones: búsqueda de palabras de opinión y clasificación de opiniones

Referencia	Búsqueda de palabras de opinión					Clasificación de opiniones					
	Algoritmo	Superv.	Orientado a dominio	Tipo palabras	Entidades clasificadas	Cálculo SO	Superv.	Orientado a dominio	Agregación SO	Tipo	
(Hu & Liu, 2004b)	-	-	-	-	-	-	-	-	-	-	-
(Hu & Liu, 2004a)	Contexto (oración)	No	No	Adjetivos	Oraciones	WordNet	No	No	Suma	Ternaria	
(Liu et al, 2005)	-	-	-	-	-	-	-	-	-	-	
(Popescu & Etzioni, 2005)	Patrones dependencias	En parte	No	Todas	Opiniones	PMI	No	En parte (contexto)	<i>Relaxation labeling</i>	Binaria	
(Popescu et al, 2005)	Patrones dependencias	En parte	No	Todas	Opiniones	PMI	No	En parte (contexto)	<i>Relaxation labeling</i>	Binaria + Ranking	
(Carenini et al, 2005)	-	-	-	-	-	-	-	-	-	-	
(Yi & Niblack, 2005)	Patrones sintácticos simples	En parte	No	Adjetivos y nombres	Opiniones	Lexicón manual	En parte	No	Patrones sintácticos simples	Binaria	
(Ding et al, 2008b)	Contexto (oración)	No	No	Todas	Opiniones	WordNet + lexicón manual	No	En parte (contexto)	Suma ponderada + reglas de contextualización	Binaria	
(Blair-Goldensohn et al, 2008)	Contexto (oración)	No	No	Todas	Oraciones	WordNet + <i>label propagation</i>	Sí* *clasific.	En parte (contexto)	Suma + clasificador máxima entropía	Ternaria	
(Titov & McDonald, 2008b)	-	-	-	-	-	-	-	-	-	-	
(Titov & McDonald, 2008a)	-	-	-	-	-	-	-	-	-	-	
(Brody & Elhadad, 2010)	Patrones sintácticos simples	En parte	Sí (aspectos)	Adjetivos	Aspectos	Conjunciones + <i>label propagation</i>	No	Sí (aspectos)	-	<i>Rating</i>	
(Zhao et al, 2010)	MaxEnt-LDA	Sí	Sí	Todas	-	-	-	-	-	-	
(Jo & Oh, 2011)	-	-	-	-	<i>Reviews</i>	ASUM	No	Sí	<i>Probabilistic sentiment distribution</i>	Binaria	
Nuestra propuesta	Contexto, patrones dependencias	Sí* *patrones	Sí* *patrones	Configurable	Opiniones	WordNet, SWN, PMI, Lexicón	Sí* *lexicon	Sí	Umbral prob. + suma + expresiones especiales	Binaria	

Parte III

Una propuesta para la extracción de opiniones

Capítulo 4

Extracción de opiniones sobre características adaptable al dominio

Resumen: En el presente capítulo presentamos nuestra propuesta para la extracción de opiniones sobre características, cuyas bases fundamentales son la adaptación al dominio y el uso de taxonomías para la representación de las características. En primer lugar, trataremos de acotar el alcance de la tarea, concretando la definición formal de la misma y de las entidades participantes: opiniones, características, dominios y objetos, entre otras. Una vez claro el alcance de la tarea, daremos una visión general de nuestra propuesta, que será convenientemente desglosada, descrita y evaluada en los siguientes capítulos.

4.1. Introducción

Aunque lo definiremos más formalmente en la siguiente sección, en términos generales la extracción de opiniones sobre características consiste en detectar las opiniones vertidas en los textos de entrada acerca de las características de un objeto determinado. Por ejemplo, ante la siguiente oración extraída de un texto de opinión sobre una compañía telefónica:

“The customer service is terrible”

, habría que detectar que la oración contiene una opinión negativa acerca del servicio al cliente de la compañía. Parece evidente que la participación

de la palabra “*terrible*” es importante, puesto que su semántica nos permite afirmar que la oración expresa una opinión, e incluso el carácter negativo de la misma. Sin embargo, la mera aparición de dicha palabra no es suficiente. También el contexto en que aparece es importante. Por ejemplo, la relación sintáctica de la palabra con el resto de componentes de la oración, o la temática del texto del que se ha extraído la oración. Determinadas palabras pueden ser indicativas de la existencia de una opinión positiva en un contexto, y serlo de una opinión negativa en otro, como ocurre con la expresión “*hard to hear*” en el siguiente ejemplo:

“It is really hard to hear”

Si la oración pertenece a un texto de opinión acerca de un electrodoméstico (una lavadora o un lavavajillas), parece lógico pensar que la opinión es positiva. Sin embargo, en el contexto de una crítica de un teléfono móvil o un navegador GPS, las connotaciones del término son claramente negativas. Esta dependencia de la orientación semántica de algunos términos con respecto al dominio de aplicación está ampliamente documentada en la bibliografía. En nuestra propuesta, trataremos de capturar esta dependencia a través de la definición y generación de determinados recursos específicos del dominio. Estos recursos capturarán no sólo la dependencia que acabamos de precisar, sino también otras particularidades del dominio con respecto al resto de participantes en el problema (por ejemplo, las construcciones sintácticas específicas utilizadas en cada dominio).

Más allá de lo afirmado en la literatura en referencia a la influencia del contexto en la tarea de extracción de opiniones, en nuestra propuesta trataremos de capturar también la posible influencia de la característica concreta sobre la que se vuelcan las opiniones. Por ejemplo, en la siguiente oración, extraída de un texto de opinión acerca de un hotel:

“It was a good hotel, very clean and cheap”.

, la palabra “*cheap*” tiene connotaciones positivas, refiriéndose al precio del hotel. Sin embargo, en la siguiente oración del mismo dominio:

“The bathroom was very spacious, but they installed such a cheap bathtub that creaked and felt as if it would break open at any moment”.

, la misma palabra tiene connotaciones negativas, al ser utilizada para calificar un aspecto distinto del hotel (el baño). Nuestra propuesta tendrá en cuenta este tipo de dependencias a nivel de característica, de mayor granularidad que la utilizada en otros trabajos.

Además de la adaptación al dominio y la captura de las dependencias a nivel de característica, nuestra propuesta se basa en la utilización de *taxonomías de características*, que representan los aspectos opinables de un dominio determinado, y las relaciones de generalización/especialización que se establecen entre ellos. En muchas ocasiones, una característica o aspecto opinable puede ser nombrado de diversas maneras en los textos. Por ejemplo, el equipo de sonido de un coche a veces será mencionado como *sound system* y otras como *audio system*. Las opiniones en las que se utilicen uno u otro término deberán ser extraídas y asociadas al mismo elemento de la taxonomía. Esto dificulta la tarea con respecto a otras propuestas de la bibliografía, en las que únicamente se señalan las palabras que representan la característica sobre la que trata una opinión (se suele utilizar el término *opinion target*). Téngase en cuenta que pueden existir ambigüedades respecto a los términos utilizados para nombrar determinadas características. Por ejemplo, la palabra “*range*” es utilizada en algunos textos sobre auriculares inalámbricos para referirse al alcance del receptor, y en otros al rango de frecuencias que reproducen.

Nuestra definición de la tarea, basada en la taxonomía de características, aún dificultando la resolución de la misma, genera una salida más útil, en términos de agregación, resumen o visualización de las opiniones generadas. En esto también tendrá importancia la jerarquía establecida en dicha taxonomía, que nos permitirá, por ejemplo, utilizar las opiniones extraídas acerca de los altavoces o el amplificador a la hora de resumir las opiniones acerca del equipo de sonido de un coche.

De manera resumida, podemos decir que nuestro acercamiento a la extracción de opiniones consiste en el desarrollo de un sistema genérico que puede ser adaptado a un dominio concreto mediante la utilización de una serie de recursos que capturan el conocimiento específico de dicho dominio a nivel de características. Dichos recursos permiten lidiar con las cuestiones que acabamos de exponer informalmente, acerca de la influencia del contexto en la identificación y tratamiento de las distintas entidades participantes en el problema.

Uno de los recursos es la taxonomía de características, que contiene el conjunto de características opinables del dominio sobre las que el sistema procederá a extraer opiniones. La utilización de dichas taxonomías permitirá desarrollar sistemas de extracción dirigidos a los aspectos del dominio que se consideren más importantes, y facilitará la explotación de las opiniones extraídas. La taxonomía de característica puede ser construida utilizando una serie de herramientas de apoyo que trabajan a partir de una colección

de documentos del dominio, o bien puede ser definida por un experto en el dominio. En este caso, la taxonomía se convierte en un formalismo para la captura de requisitos: a través de la misma, se comunica al sistema de extracción cuáles son las características del objeto en las que estamos interesados.

4.2. Definición de la tarea

En la presente sección vamos a definir la tarea concreta que pretendemos abordar. La definición que utilizaremos es cercana a la utilizada en la bibliografía, pero con algunas diferencias. Aún pudiendo parecer sutiles, dichas diferencias cambian considerablemente la tarea a abordar y, desde nuestro punto de vista, la hacen más útil de cara a la implementación práctica de un hipotético sistema de extracción de opiniones. En la definición de la tarea participan distintas entidades que también serán convenientemente precisadas. Por ejemplo, cuando hablamos de *opiniones*, ¿a qué nos estaremos refiriendo exactamente? Concretar este y otros términos nos permitirá precisar el alcance real de la tarea y, consecuentemente, nos facilitará su resolución.

Comenzaremos definiendo las distintas entidades participantes en el problema. Después, y dada la amplitud del concepto de opinión, haremos algunas consideraciones sobre el alcance del problema, estableciendo qué tipos de opiniones pretendemos extraer y cuáles quedan fuera de nuestro estudio. Pasaremos entonces a definir formalmente la tarea, para acabar la sección planteando los retos y dificultades que a priori presenta la resolución de la misma.

4.2.1. Definiciones previas

Opiniones sobre características

Según la primera acepción del término en el Diccionario de la Real Academia de la Lengua Española, una *opinión* es un “*dictamen o juicio que se forma de algo cuestionable*”. El diccionario inglés Merriam-Webster de la Encyclopedia Británica, la define como “*a view, judgment, or appraisal formed in the mind about a particular matter*”, y en otra acepción como “*a formal expression of judgment or advice by an expert*”. Según estas definiciones, las opiniones tienen un carácter inequívocamente subjetivo y una intención evaluativa. Una opinión así entendida es una idea abstracta, situada en la mente de la persona que la posee, e imposible de representar de manera precisa como una entidad estructurada (es decir, de manera que sea procesable por una máquina). Necesitamos claramente simplificar y concretar esta definición

para conseguir que una opinión pueda ser representada como una entidad estructurada, con determinados atributos y rangos de valores definidos para los mismos. Pero también creemos necesario ampliar o generalizar las definiciones anteriores, para incluir no sólo evaluaciones o juicios subjetivos, sino también aquellos enunciados que describan *objetivamente* características del objeto en cuestión, aunque con claras implicaciones positivas o negativas.

Situémonos en el contexto de un documento de revisión de un producto o reseña (quizás estemos más habituados al término inglés, *review*). Más que situar el foco de la definición de *opinión* en la persona que escribe el documento, lo haremos en la persona que lo lee. Los mensajes que va recibiendo ocasionan que se vaya formando una opinión favorable o desfavorable acerca de distintos aspectos o *características* del producto en cuestión. Algunos de estos mensajes se adecúan efectivamente a las definiciones anteriores de opinión: son juicios o evaluaciones subjetivas expresadas por el autor de la reseña. Como tales, ocasionan la formación de opiniones en el lector, en principio similares, aunque esto depende de la actitud crítica del lector y del grado de confianza que atribuya a lo que está leyendo. Pero también hay otros mensajes que quizás no encajan tan directamente en las definiciones anteriores de opinión, pero que evocan igualmente *sentimientos* positivos o negativos en el lector acerca de las características del objeto. Por ejemplo, si el autor de la reseña proclama que *le encanta* el diseño del producto, aunque no se trate estrictamente de una evaluación sino más bien de un sentimiento afectivo, las implicaciones positivas son evidentes. De todas formas, podríamos considerar que este tipo de enunciados son realmente juicios, y que por tanto quedan recogidos en las definiciones anteriores: cuando el autor dice que *le gusta* el diseño del producto, está diciendo en cierto modo que cree que *el diseño es bueno*. Un ejemplo que se saldría más claramente de las definiciones anteriores sería el siguiente enunciado:

“The body of the camera is made of plastic”

En este caso, estamos ante una descripción completamente objetiva de una característica del objeto. Pero dicha descripción tiene probables implicaciones negativas, que incitan al lector a formarse una opinión negativa. Queremos que nuestra definición de opinión también incluya este tipo de enunciados.

Definición 1 Una **opinión** es una porción de texto lo más pequeña posible con implicaciones positivas o negativas sobre alguna de las características de un objeto o sobre el objeto mismo.

La definición anterior es intencionadamente genérica, abarcando los casos considerados anteriormente (evaluaciones subjetivas, declaraciones afectivas y descripciones objetivas), siempre que tengan implicaciones positivas

o negativas acerca de alguna característica del objeto. Al mismo tiempo que genérica, la definición 1 permite concretar una representación estructurada de las opiniones, que en su mínima expresión consistiría en indicar la *polardad* (positiva o negativa) y la *característica*. Definamos a qué nos referimos con este último término:

Definición 2 *Una característica (feature) es cualquier propiedad, componente o aspecto de un objeto. El propio objeto se considera una característica, así como cualquier propiedad, componente o aspecto de cualquier característica es también en sí mismo una característica.*

La definición de característica es de nuevo intencionadamente genérica. Para un ordenador, por ejemplo, cada uno de los componentes es una característica: procesador, disco duro, memoria, tarjeta gráfica, etc. Pero también lo son el diseño, el precio o la apariencia, por poner algunos ejemplos de características no tangibles. Además, cada una de las propiedades, subcomponentes o aspectos de estas características serán también características; por ejemplo, la memoria caché, la capacidad o la velocidad del disco duro. El carácter recursivo de esta definición hace apropiada la utilización de taxonomías para representar las características de un objeto, como veremos más adelante.

Es importante la puntualización acerca del tamaño del texto en la definición 1 (*... una porción de texto lo más pequeña posible ...*). Considérese la siguiente oración de ejemplo:

“This car is efficient and fun to drive”

En esta oración, si escogemos una porción de texto lo más pequeña posible, existen dos opiniones sobre la misma característica. Por un lado, tenemos la porción de texto “*This car is efficient*”, y por otro “*This car is ... fun to drive*”. Se extraen varias conclusiones de este ejemplo. En primer lugar, en una oración pueden aparecer varias opiniones, incluso sobre la misma característica. En segundo lugar, la porción de texto correspondiente a cada opinión no tiene porqué estar formada por palabras consecutivas de la oración.

Objetos y dominios

Aunque en los ejemplos expuestos hasta el momento el objeto que está siendo analizado siempre es un producto, queremos señalar que nuestras definiciones no se limitan a dichos objetos.

Definición 3 *Un **objeto** es cualquier entidad susceptible de ser analizada, ya sea un producto, un servicio, una persona o una corporación.*

Por tanto, las soluciones que mostraremos a lo largo de este trabajo de tesis pueden ser aplicadas a cualquier tipo de objeto de análisis, siempre que podamos confeccionar un conjunto de características opinables del mismo. Por supuesto, este conjunto de características será más sencillo de definir en unos casos (por ejemplo, para un producto) que en otros (por ejemplo, para una persona). Es importante diferenciar el objeto concreto sobre el que versa un documento concreto de opinión (por ejemplo, un modelo de cámara de fotos) del tipo de objeto, en el que se incluyen todos aquellos objetos concretos con un conjunto similar de características (por ejemplo, las cámaras de fotos).

Definición 4 *Un **dominio** es un conjunto de objetos concretos con una finalidad y un conjunto de características comunes que lo definen.*

Ejemplos de dominios serían cámaras de fotos, hoteles o coches, pero también políticos, compañías de seguros o supermercados, por ejemplo. Podemos definir dominios con distintos niveles de granularidad. Por ejemplo, podemos definir el dominio de los teléfonos móviles, o el dominio más concreto de los teléfono móviles inteligentes (*smartphones*). El dominio es fundamental en nuestra propuesta, puesto que los sistemas de extracción de opiniones que pretendemos construir están orientados a extraer opiniones en un dominio concreto, previamente elegido. La elección de dominios más o menos generales repercutirá en la dificultad para concretar el conjunto de características del dominio, y también en los resultados que obtendrá el sistema: a mayor generalidad del dominio, la información recogida en los recursos del dominio que utiliza nuestra propuesta será menos específica, y por tanto, la extracción de opiniones será de menor calidad.

Palabras de opinión y de característica

En muchas de las oraciones utilizadas para expresar opiniones, observamos que un reducido número de palabras concentran la carga semántica necesaria para dictaminar la polaridad de la opinión. Por ejemplo, en la oración:

“All I have to say is that this camera is excellent”.

la palabra “*excellent*” contiene suficiente información para deducir que se trata de una opinión positiva. A estas palabras, que contienen la carga semántica responsable de la polaridad de la orientación semántica de la opinión, las llamaremos *palabras de opinión*.

Definición 5 *Las palabras de opinión (opinion words) son el conjunto mínimo de palabras participantes en una opinión a partir de las cuales es posible decidir la polaridad de la misma.*

Además de las palabras de opinión, serán importantes en nuestra propuesta las *palabras de característica*:

Definición 6 *Las palabras de característica (feature words) son el conjunto de palabras participantes en una opinión que hacen mención explícita a la característica sobre la que se está opinando.*

En las siguientes oraciones de ejemplo, las palabras de opinión aparecen en cursiva y las palabras de característica en negrita:

1. “*Perfect sound quality*”.
2. “The **beds** are *large* and *comfortable*, but the **decoration** is *too traditional*”.
3. “These headphones are really *expensive*, but they *sound great*”.
4. “Do *not buy them!*”.

En el primer ejemplo, “*perfect*” es la palabra de opinión, puesto que es dicha palabra la que nos permite afirmar que se trata de una opinión positiva. Lo mismo ocurre con “*large*”, “*comfortable*” y “*too traditional*” en la segunda oración, que serían las correspondientes palabras de opinión de las tres opiniones contenidas en la oración. Nótese que en el tercer ejemplo hemos señalado la palabra “*expensive*” como palabra de opinión, excluyendo a la palabra previa “*really*”. Esto es porque según la definición 5 debemos escoger el *mínimo* conjunto de palabras que nos permitan decidir la polaridad de la opinión. En este caso, “*really*” puede afectar a la intensidad de la opinión, pero no a la polaridad. La elección como palabras de opinión del *mínimo* conjunto de palabras necesarias para decidir la polaridad, nos permitirá generar recursos de mayor calidad a partir de documentos anotados, al conseguirse una mayor representatividad estadística en las anotaciones.

Sin embargo, en la segunda oración de ejemplo, la palabra “*too*” fue incluida como palabra de opinión junto a “*traditional*”, al ser determinante para decidir la polaridad de la opinión. Igualmente ocurre con el adverbio *not* en el cuarto ejemplo, que invierte la polaridad de la otra palabra de opinión participante.

En cuanto a las palabras de característica, es necesario hacer algunas consideraciones. En primer lugar, las palabras de característica siempre forman parte de un sintagma nominal, en el que actúan como núcleo. En segundo lugar, no todas las menciones a características son consideradas palabras de característica, puesto que no siempre constituyen una mención a una característica sobre la que se está opinando. Por ejemplo, en el tercer ejemplo “*headphones*” no ha sido marcada como palabra de característica, puesto que se considera que la característica sobre la que se está opinando es el precio. Por último, no en todos los casos la característica sobre la que se está opinando aparece mencionada explícitamente en el texto, como es el caso de este último ejemplo, en el que aparecen dos opiniones sobre las características *precio* y *calidad de sonido*. Esto nos lleva a la siguiente definición.

Definición 7 *Una opinión sobre característica implícita es una opinión en la que no se menciona explícitamente la característica sobre la que se opina.*

A lo largo del trabajo, nos referiremos en ocasiones a estas opiniones con el término abreviado *opiniones implícitas*. Nótese que en este tipo de opiniones, las palabras de opinión no sólo informan de la polaridad de la opinión, sino también de la característica. Por ejemplo, la palabra de opinión “*expensive*” está relacionada con la característica *precio*. No deben confundirse las opiniones sobre característica implícita con aquellas opiniones en las que la característica aparece mencionada anafóricamente por un pronombre, como en el ejemplo 4. En este caso, consideramos que dicho pronombre constituye una mención explícita a la característica sobre la que se opina (en el caso del ejemplo, el propio producto).

4.2.2. Alcance de la tarea

Existen opiniones en las que, aún adecuándose a la definición 1, es difícil señalar cuáles son las palabras de opinión y de característica. Obsérvese por ejemplo la siguiente oración, extraída de un *review* sobre una cámara de fotos:

“If you’re not careful , the battery will fall out when you try and open the door to switch cards”.

En esta oración, se nos informa de una característica poco deseable de una cámara de fotos determinada. Se trataría de una opinión negativa sobre la característica *diseño*. Para ser conscientes de ello, es necesario entender de forma profunda el significado de la oración, en estrecha relación con una construcción sintáctica relativamente compleja, además de tener un conocimiento del mundo apropiado: hay que comprender que la caída de la batería puede ocasionar su rotura, y que esto es algo indeseable. Además, habría que percatarse de que el problema está relacionado con el *diseño* de la cámara, y no con la *batería* en sí (no sería contradictorio que existiesen buenas opiniones en cuanto a la batería de la cámara). En este tipo de oraciones, se hace muy difícil elegir el conjunto de palabras de opinión, e incluso las palabras de característica. Dada la motivación práctica de nuestra propuesta, que pretende hacer uso exclusivamente de herramientas de análisis léxico y sintáctico, opiniones como la mostrada quedarán fuera de nuestro foco de estudio. Nos centraremos en aquellas opiniones en las cuales seamos capaces de identificar las palabras de opinión y en su caso de característica que concentran la carga semántica de la opinión. Además, exigiremos que dichos elementos estén sintácticamente relacionados entre sí de manera abordable por los analizadores sintácticos disponibles.

Otro tipo de opiniones que no consideraremos son aquellas en las cuales se establece una comparación entre varios productos. Por ejemplo, considérese la oración siguiente:

“The body of the camera is very similar to that of the ELPH in terms of durability”.

En ella se expresa una opinión acerca de la durabilidad de una cámara. Pero para decidir la polaridad de la opinión es necesario que conozcamos la opinión del autor acerca de la durabilidad de otra cámara. Sí trataremos aquellas oraciones en las que se utilizan adjetivos superlativos, como en:

“This is the best camera that you can buy!”.

, dado que en este caso no se presenta el problema anterior: a la vista de la palabra “*best*”, podemos decidir que se trata de una opinión positiva.

4.2.3. Definición formal de la tarea

Taxonomías de características

Dado un dominio D , según la definición 4, llamaremos $F_D = \{f_1, f_2, \dots, f_n\}$ al conjunto de características opinables del dominio D , según la definición 2. Según esta definición, cada característica f_i puede tener a su vez un conjunto de subcaracterísticas, que podemos denotar por $F_{f_i} = \{f_{i1}, f_{i2}, \dots, f_{in}\}$, cada una de las cuales puede tener sus propias subcaracterísticas, y así sucesivamente. Por tanto, más que de un conjunto, hablamos de una *taxonomía de características*. Esta taxonomía es un árbol, cuya raíz es el propio dominio (recuérdese que según la definición 2, el propio objeto es una característica opinable). Los hijos de un nodo determinado son subcaracterísticas de la característica a la que representa dicho nodo. De todas formas, a efectos de definición de la tarea, llamaremos F_D al conjunto de todas las características del dominio, incluyendo las de todos los niveles de la taxonomía. Veremos ejemplos de taxonomías de características en los capítulos 6 y 9.

La definición de una taxonomía de características para el dominio permitirá obtener opiniones que serán más fácilmente agregables, como veremos al final de la sección 8.5.7. Además, la estructura de la taxonomía será tenida en cuenta en los algoritmos de generación de recursos que presentaremos en el capítulo 6.

Reconocimiento y clasificación de opiniones sobre características

Sea $R_D = \{r_1, r_2, \dots, r_n\}$ un conjunto de documentos de análisis de un objeto concreto (o varios) del dominio D , donde cada documento r_i está formado por un conjunto de oraciones $\{s_1, s_2, \dots, s_n\}$. Representaremos mediante la tupla $o_k = (f_i, s_j, polarity)$ a una opinión sobre la característica $f_i \in F_D$ con polaridad positiva (*polarity* = +) o negativa (*polarity* = -), contenida en la oración s_j .

El objetivo principal de la extracción de opiniones sobre características es descubrir $O_{R_D} = \{o_1, o_2, \dots, o_n\}$, el conjunto de opiniones o_k sobre cualquier característica f_i de F_D , que aparecen en cualquiera de las oraciones de los documentos de R_D . Basándonos en la nomenclatura utilizada en la tarea

de reconocimiento y clasificación de entidades con nombre, clásica del PLN, dividimos la resolución del problema en dos subproblemas: *reconocimiento de opiniones* y *clasificación de opiniones*.

Definición 8 *El reconocimiento de opiniones sobre características consiste en identificar las opiniones existentes en R_D sobre alguna característica de F_D .*

Definición 9 *La clasificación de opiniones sobre características consiste en determinar la polaridad de las opiniones previamente reconocidas.*

Nótese que la tarea se define en función del conjunto de características F_D , que podemos considerar una entrada más del problema. Por tanto, aquellas opiniones sobre características que no estén incluidas en el conjunto anterior, no serán extraídas. Ésta es una diferencia fundamental con respecto a otras definiciones de la tarea en la bibliografía. La elección de un conjunto u otro de características dependerá de la aplicación deseada del sistema. En nuestra propuesta plantearemos métodos semiautomáticos para la construcción de la taxonomía de características, a partir de un conjunto de documentos del dominio (ver sección 6).

Evidencias de opinión

Aunque los componentes fundamentales de las opiniones que pretendemos extraer son la característica sobre la que se opina y la polaridad de la opinión, nuestro sistema de extracción utiliza una abstracción más especializada de las opiniones, en la que se incluyen las palabras de característica y las palabras de opinión. Llamaremos *evidencia de opinión* a esta entidad más específica. Nuestro sistema de extracción tratará de descubrir el conjunto de evidencias de opinión $OE = \{oe_1, oe_2, \dots, oe_n\}$, siendo $oe_k = (o_k = (f_i, s_j, polarity), fW, opW)$, donde fW es un conjunto de palabras de característica que hacen referencia a la característica f_i , y opW es un conjunto de palabras de opinión a partir de las cuáles se infiere la polaridad indicada en *polarity*. En algunos casos, fW puede ser vacío (opiniones implícitas).

Las evidencias de opinión sólo son una representación interna de las opiniones utilizada por el sistema. Si queremos evaluar lo bien o mal que resuelve nuestro sistema la tarea de reconocimiento y clasificación de opiniones, lo haremos utilizando exclusivamente la característica y la polaridad de las evidencias extraídas. Sin embargo, en el proceso de anotación manual de documentos que describiremos en la sección 5.4, se anotarán evidencias de opinión, puesto que la finalidad del corpus anotado no será exclusivamente

la evaluación del sistema, sino también la inducción de los recursos del dominio, proceso en el cuál las palabras de característica y de opinión aportan una valiosa información.

Oración 1:

The sound quality is not impressive, with extremely powerful low frequencies but unclean, not well-defined high-end.

	(Feature,	Feature words,	Opinion words,	Polarity)
<i>oe₁:</i>	(sound quality,	sound quality,	not impressive,	negative)
<i>oe₂:</i>	(bass,	low frequencies,	powerful,	positive)
<i>oe₃:</i>	(treble,	high-end,	unclean,	negative)
<i>oe₄:</i>	(treble,	high-end,	not well-defined,	negative)

Oración 2:

I love them, they are lightweight and look cool!

	(Feature,	Feature words,	Opinion words,	Polarity)
<i>oe₅:</i>	(headphones,	them,	love,	positive)
<i>oe₆:</i>	(size,	-,	lightweight,	positive)
<i>oe₇:</i>	(appearance,	-,	look cool,	positive)

Oración 3:

The cord is durable but too long: I always have to untwist it every time I get them out off my pocket.

	(Feature,	Feature words,	Opinion words,	Polarity)
<i>oe₈:</i>	(cord,	cord,	durable,	positive)
<i>oe₉:</i>	(cord,	cord,	too long,	negative)

Figura 4.1: Algunos ejemplos de evidencias de opinión

En la figura 4.1 se muestran algunas oraciones de ejemplo extraídas de *reviews* de auriculares, y las evidencias de opinión que contienen. Las características corresponden a las de la taxonomía del dominio *headphones* cuya construcción será descrita en el capítulo 5. Las evidencias *oe₆* y *oe₇* son ejemplos de evidencias de opinión implícitas, en las que la característica debe ser inferida directamente de las palabras de opinión. En la última oración puede observarse una opinión que no es representable mediante una evidencia de opinión (“*I always have to untwist it every time I get them out off my pocket*”), puesto que es imposible identificar un pequeño conjunto de palabras a partir de las cuáles se pueda deducir la polaridad de la opinión; por tanto, las evidencias de opinión representan otra manera de establecer el alcance

de nuestra propuesta: el sistema que trataremos de construir pretende extraer todas aquellas opiniones sobre características que sean representables mediante una evidencia de opinión.

4.2.4. Retos y dificultades de la tarea

La tarea de reconocimiento y clasificación de opiniones sobre características, tal como la hemos definido, presenta numerosos retos y dificultades que debemos tener en cuenta de cara a proponer una solución. Vamos a comentar los más interesantes.

Calidad de los textos

Aunque la definición de la tarea no especifica cuál será el origen de los textos que serán procesados, una de las motivaciones que fueron comentadas en el capítulo 1 era explotar la ingente cantidad de contenidos generados por usuarios en la Web 2.0. Por tanto, la aplicación más directa de un sistema de extracción de opiniones será sobre textos escritos por usuarios. Esto implica, por lo general, una baja calidad de los textos: faltas de ortografía, oraciones sintácticamente incorrectas, ausencia de signos de puntuación, etc. La consecuencia inmediata es que las herramientas de análisis lingüístico que utilizaremos en nuestro sistema (detectores de oraciones, analizadores morfosintácticos, lematizadores y analizadores de dependencias sintácticas) obtendrán una mayor tasa de errores, que pueden verse propagados y ocasionar errores en el proceso de extracción de opiniones. Además, la utilización de términos coloquiales puede repercutir en la menor eficacia de los métodos de clasificación de opiniones que se basen en recursos léxicos formales como WordNet. En nuestra propuesta plantearemos soluciones que se basen en los mismos textos escritos por usuarios para generar recursos que incluyan los posibles términos coloquiales empleados en el dominio con más frecuencia.

Opiniones acerca de otros objetos

La tarea que pretendemos llevar a cabo consiste en la extracción de las opiniones contenidas en una serie de documentos de texto acerca de un objeto concreto. A pesar de que suponemos que los documentos de entrada al sistema contienen fundamentalmente opiniones acerca del objeto en cuestión, es muy común que se intercalen opiniones acerca de otros objetos distintos. Por ejemplo, en un *review* sobre un teléfono móvil es común que el usuario dedique parte del texto a hablar de otros teléfonos distintos, para establecer comparaciones. Éste es un problema de difícil solución: si en el texto nos

encontramos una opinión, ¿cómo podemos estar seguros de que se trata de una opinión sobre el objeto concreto en el que estamos interesados?

Ambigüedades en la determinación de la característica

La definición de la tarea de reconocimiento de opiniones implica determinar correctamente las características concretas sobre las que tratan las opiniones. Como veremos en los capítulos siguientes, es frecuente tener un alto número de características (por lo general, más de 30). En algunos casos, pueden existir ambigüedades entre dichas características, ya sea por la polisemia de algunas de las palabras de característica (por ejemplo, “*range*” en un *review* sobre unos auriculares inalámbricos se puede referir al rango de frecuencias de sonido, o al alcance de la conexión inalámbrica), o por la correlación de algunas palabras de opinión con más de una característica (por ejemplo, en el dominio *headphones* hemos observado experimentalmente que la palabra de opinión “*cheap*” está correlacionada con opiniones implícitas sobre las características *price*, *durability*, *appearance* y *design*).

Ambigüedades en las palabras de característica

La búsqueda de palabras de característica en los textos puede ser un buen punto de inicio para encontrar posibles opiniones. Sin embargo, muchas de las menciones a características se hacen fuera del contexto de una opinión. En ocasiones se trata de oraciones en las que se están describiendo objetivamente las características sin implicaciones positivas o negativas. Además, la ambigüedad semántica de algunas de las palabras de característica también puede jugarnos malas pasadas. Por ejemplo, la palabra de característica *look* es utilizada en diversos dominios en opiniones acerca de la apariencia de un producto, pero al mismo tiempo aparece frecuentemente en otras construcciones que no son opiniones.

Dependencia de la polaridad con respecto al dominio y a la característica

La polaridad de determinadas palabras de opinión puede depender en ocasiones tanto del dominio de aplicación como de la característica concreta a la que se aplique. Por ejemplo, la palabra de opinión “*unpredictable*” tiene connotaciones negativas en el contexto de la conducción, y connotaciones positivas en el contexto de una crítica de cine. Esto está ampliamente documentado en diversos trabajos de la bibliografía; sin embargo, la dependencia de la polaridad con la característica concreta no ha sido tenida en

cuenta en ningún trabajo previo. Por ejemplo, dentro de un dominio determinado que incluya las características *price* y *appearance*, la palabra de opinión “*cheap*” tiene connotaciones positivas al ser aplicada a la primera característica, y negativas al ser aplicada a la segunda característica. Nuestra propuesta tendrá en cuenta esta dependencia entre la polaridad y la característica a la hora de plantear la definición y generación de los recursos del dominio.

Atomicidad de las opiniones

Según la definición 1, consideramos las opiniones formadas por la “*porción de texto más pequeña posible*”; así por ejemplo, en aquellas oraciones en las que se utilicen varios adjetivos para calificar a una característica (“*The screen is big, bright and vivid*”), o en las que se califiquen de igual forma varias características (“*The sound quality and volume are great*”), nuestro sistema debería extraer opiniones independientes (tres opiniones en el primer ejemplo, y dos en el segundo). En caso de extraer una sola opinión, la cobertura obtenida por el sistema al ser evaluado de acuerdo a la definición de la tarea se verá reducida.

Ironía y otros fenómenos pragmáticos

Si ya de por sí determinadas opiniones son inabordables debido a una excesiva complejidad semántica o a la necesidad de tener cierto conocimiento del mundo para poder decidir la polaridad de las mismas, la aparición de la ironía y de otros fenómenos del nivel pragmático del lenguaje dificultan aún más la tarea. El estudio de estos fenómenos dentro del PLN se encuentra aún en etapas muy tempranas. Debemos aceptar que nuestro sistema no será capaz de manejar dichas situaciones, lo que repercutirá en alguna medida en los resultados que serán presentados en el capítulo 7.

4.3. Resumen de la propuesta

La solución que planteamos al problema de la extracción de opiniones sobre características se apoya en dos pilares fundamentales. Por un lado, hemos desarrollado un sistema de extracción modular, basado en taxonomías y adaptable al dominio. Por el otro, para que dicho sistema genérico sea aplicable a un dominio concreto, deben proporcionársele un conjunto de recursos específicos del dominio, incluyendo entre otros una taxonomía de características. Dichos recursos son inducidos de manera automática (semi-automáticamente en el caso de la taxonomía) a partir de un conjunto de documentos

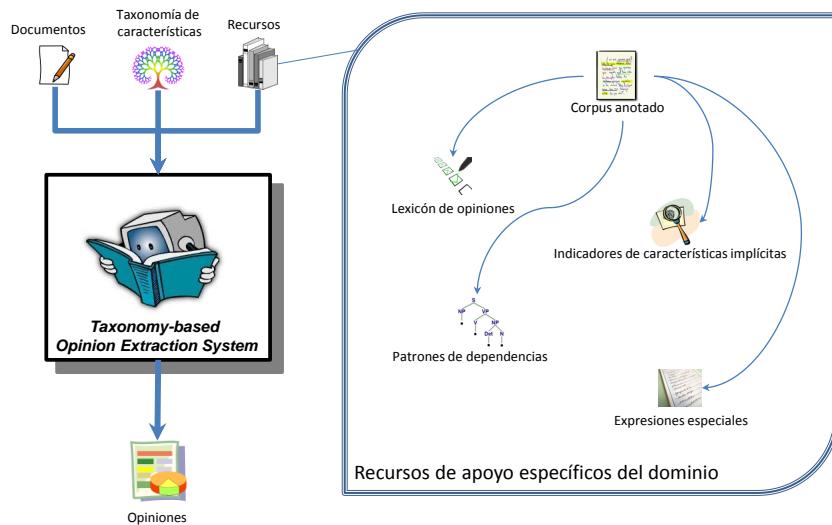


Figura 4.2: Esquema conceptual de nuestra propuesta

anotados. Además del sistema genérico de extracción, nuestra aportación incluye una metodología que describe los pasos a seguir para la obtención de los recursos necesarios para un dominio dado, además de los algoritmos de inducción de los recursos y herramientas de apoyo para las tareas manuales (la construcción de la taxonomía de características y la anotación de documentos). En la figura 4.2 se muestra una representación conceptual de nuestra propuesta.

4.3.1. Sistema de extracción genérico

Para llevar a cabo la tarea de extracción tal como la hemos definido, hemos diseñado un sistema de extracción de opiniones sobre características modular, basado en taxonomías y adaptable a dominio. El sistema está preparado para funcionar en cualquier dominio, siempre que reciba como entrada, además de los textos sobre los que se llevará a cabo la extracción, los recursos pertinentes específicos del dominio. Las principales características del sistema de extracción son las siguientes:

- **Arquitectura modular.** El sistema está compuesto de distintos componentes, cada uno de los cuales lleva a cabo una subtarea determinada, de manera que la ejecución secuencial de un conjunto de componentes lleva a cabo la tarea completa de reconocimiento y clasificación de opiniones. Hemos identificado una serie de subtareas básicas, cada una de las cuales se corresponde con un tipo de componente. Hemos llamado *componente abstracto* a cada uno de estos tipos de componentes, para

diferenciarlos de los *componentes concretos*, que representan distintas implementaciones de los subproblemas atacados por cada componente abstracto. Algunos de estos componentes concretos hacen uso de los recursos del dominio, y otros tratan de resolver el subproblema correspondiente de manera independiente del dominio. La arquitectura modular del sistema y la disponibilidad de ambos tipos de componentes concretos nos permitirá configurar distintas cadenas de ejecución, más o menos dependientes de los recursos del dominio.

- **Taxonomías de características.** El sistema lleva a cabo la extracción de opiniones sobre las características incluidas en una taxonomía. A diferencia del resto de recursos específicos del dominio, la taxonomía de características es una entrada requerida del sistema. La adaptación del sistema a la taxonomía de características concreta que reciba hace más útiles las opiniones obtenidas, de cara a agruparlas, resumirlas o representarlas gráficamente. Además, permite a quien configura el sistema de extracción para un dominio concreto especificar las características del objeto en las que está interesado.
- **Adaptabilidad al dominio.** El sistema puede ser utilizado en cualquier dominio sin necesidad de ser modificado, siempre y cuando se disponga de la taxonomía de características correspondiente al dominio. Además de la taxonomía, y según los componentes concretos utilizados en la cadena de ejecución, será necesario generar otros recursos específicos del dominio.

La descripción del sistema en profundidad, junto con un estudio experimental del mismo, puede consultarse en los capítulos 7 y 8.

4.3.2. Recursos específicos del dominio

Como ha quedado ya claramente expuesto en el resto del capítulo, los recursos específicos del dominio son una pieza fundamental de nuestra propuesta. En particular, éstos son los recursos que serán definidos en profundidad en el siguiente capítulo:

- **Taxonomía de características:** contiene el conjunto de características del dominio sobre las que se efectuará la extracción de opiniones, organizadas jerárquicamente. Además, para cada característica, se incluyen en el recurso un conjunto de palabras de característica.

- **Lexicón de opiniones:** incluye estimaciones de la orientación semántica de los términos usados como palabras de opinión en el dominio. También se incluyen estimaciones de la probabilidad de que dichos términos actúen como palabras de opinión.
- **Indicadores de características implícitas:** listado de términos usados como palabras de opinión en el dominio y cuya presencia puede estar relacionada, según cierta probabilidad, con la existencia de opiniones sobre características implícitas.
- **Patrones de dependencias:** reglas de extracción basadas en relaciones de dependencia sintáctica que permiten conectar las palabras de característica con las palabras de opinión, y las palabras de opinión entre sí.
- **Expresiones especiales:** términos participantes en las opiniones como palabras de opinión, con implicaciones singulares sobre la polaridad de las opiniones en las que intervienen. En concreto, distinguiremos tres tipo de expresiones especiales: de negación, de no negación y de polaridad dominante.

Todos los recursos son generados de manera automática (semiautomática en el caso de la taxonomía de características), a partir de un conjunto de documentos anotados. Los algoritmos de inducción de los recursos serán descritos en el capítulo 6. Además, en dicho capítulo se establecerá una metodología a seguir para generar los recursos para un nuevo dominio, incluyendo qué pasos deben seguirse y en qué orden, y aportando herramientas de apoyo para la parte manual del proceso (construcción de la taxonomía de características y anotación de los documentos).

Capítulo 5

Recursos para la extracción de opiniones

Resumen: En el presente capítulo se definen una serie de recursos de apoyo específicos del dominio, cuya intención es la captura de conocimiento útil para la extracción de opiniones sobre características. Además de la definición, mostraremos la sintaxis utilizada para representar cada uno de los recursos.

5.1. Introducción

Una de las ideas centrales de nuestra propuesta para la extracción automática de opiniones es la captura de conocimiento acerca de las particularidades del dominio para el que deseamos construir el sistema, y de la manera en que la gente expresa sus opiniones en dicho dominio. A raíz de la definición de la tarea, hemos identificado una serie de necesidades de información cuya disponibilidad debería facilitar el proceso de extracción. Las enunciamos a continuación en forma de preguntas:

- ¿Cuáles son las características opinables del objeto? O en todo caso, ¿en cuáles de ellas estamos interesados?
- ¿Qué palabras son más frecuentemente utilizadas para expresar opiniones para cada característica de las anteriores? ¿Qué polaridad de la opinión implican?
- ¿Qué palabras están correlacionadas con la existencia de opiniones sobre características implícitas?

- ¿De qué manera se relacionan sintácticamente en una opinión las palabras que hacen referencia a la característica y las palabras que expresan la opinión?
- ¿Qué expresiones con efectos concretos sobre la polaridad de las opiniones son utilizadas? Por ejemplo, ¿qué palabras ocasionan una inversión en la polaridad de las opiniones en las que participan?

Hemos concretado cada una de estas necesidades en la definición de un recurso que capture todo el conocimiento del dominio útil para la extracción automática de opiniones. La generación de estos recursos, aún siendo en buena parte automática, requiere de un cierto esfuerzo manual. Siendo un trabajo que debe realizarse una sola vez por dominio, creemos que es indispensable para ser capaces de construir un sistema de extracción que se adapte convenientemente a cualquier dominio de actuación. Nuestra filosofía será la utilización de métodos automáticos siempre que sea posible, y en aquellos casos en que sea necesaria la intervención humana, aportaremos todas las herramientas de apoyo posibles para facilitar la labor y mejorar la calidad de los recursos generados.

En este capítulo presentaremos cada uno de los recursos que utiliza nuestro sistema. Para cada uno de ellos, además de la definición, mostraremos la sintaxis utilizada para su representación. Los métodos y algoritmos para la generación de los recursos, así como la secuencia en la que dicho proceso se lleva a cabo, serán explicados en el capítulo 6.

5.2. Corpus

La pieza fundamental en la que se apoya todo el proceso de generación de recursos para un dominio dado es un conjunto de documentos de opinión de dicho dominio. Si estamos construyendo un sistema de extracción de opiniones sobre productos, podemos utilizar los *reviews* escritos por usuarios disponibles en múltiples páginas web de Internet, tales como www.epinions.com o www.ciao.com. Si el sistema tiene como objetivo la extracción de las opiniones acerca de cierta entidad o persona, podemos utilizar los comentarios dejados por los internautas en foros o blogs relacionados con dicha entidad o persona.

5.2.1. Calidad de los documentos

Dado que en la mayoría de los casos los documentos que obtendremos han sido redactados por usuarios anónimos de determinadas páginas web,

foros o blogs, la calidad de los textos puede variar desde textos bien escritos a textos completamente ilegibles. En el caso de aplicación que utilizamos como ejemplo en el capítulo 6, nos hemos encontrado con textos escritos completamente en mayúsculas, textos en los que no se hace uso alguno de los signos de puntuación, numerosas faltas de ortografía, etc. Y no sólo es un problema la calidad en la forma de los documentos: también el contenido puede ser un problema. Es fácil encontrar documentos cuyo contenido, en parte o totalmente, no hace referencia al objeto en cuestión, o bien se habla del objeto muy tangencialmente y no se vuelcan opiniones sobre el mismo.

Existen dos maneras de abordar este problema. La primera sería simplemente no hacer nada, permitiendo que todos esos documentos de baja calidad pasen a formar parte del corpus. Por contra, podríamos establecer una serie de filtros que permitieran seleccionar para el corpus aquellos documentos que presentan una mejor calidad. Para ello, podríamos utilizar diccionarios para medir la corrección ortográfica, contabilizar los signos de puntuación, estableciendo un intervalo de frecuencias esperadas, o utilizar modelos de lenguaje para descartar aquellos documentos cuyo contenido esté redactado sin coherencia gramatical. Con esto conseguiríamos un corpus de más calidad, más fácilmente procesable por los procesadores lingüísticos, y del que podremos extraer recursos también de más calidad.

Esta segunda opción parece a priori la más adecuada, pero debemos plantearnos cuál será el marco de explotación del sistema de extracción que deseemos construir. Es posible que nuestro sistema tenga que trabajar con algunos documentos de baja calidad, y por tanto puede ser deseable disponer de una cierta proporción de ese tipo de documentos en nuestro corpus, tanto para que las particularidades de dichos documentos se reflejen en los recursos como para que el corpus constituya un mejor muestreo del conjunto de documentos objetivo del sistema, permitiendo una evaluación más real de la precisión y la cobertura y un mejor ajuste de parámetros. Volveremos sobre estas consideraciones en el capítulo 6, donde llevamos a cabo un acercamiento a medio camino entre las dos formas de actuar a la hora de seleccionar los documentos que formarán el corpus.

5.2.2. Tamaño del corpus

El número de documentos es otra variable que debemos decidir. A mayor número de documentos anotados, mayor representatividad de los mismos y por tanto mayor calidad de los recursos generados. En contraposición, un mayor número de documentos exige de un mayor esfuerzo de anotación y validación de los mismos. En principio, deberíamos anotar tantos documentos como nuestros recursos nos permitan. No obstante, es previsible que exista

un cierto número de documentos anotados a partir del cuál obtengamos una ganancia mínima al añadir nuevos documentos; éste sería el valor óptimo de documentos, siempre que no estemos limitados por los recursos de que dispongamos. Independientemente del número de documentos a anotar, es conveniente recolectar el mayor número posible de documentos del dominio, ya que estos documentos, aún sin anotar, serán de utilidad para la aplicación del método de expansión automática de los lexícones de opinión (ver sección 6.8).

En el capítulo 8 llevaremos a cabo algunos experimentos variando el tamaño del corpus y midiendo las repercusiones en el rendimiento del sistema.

5.2.3. Representatividad de los documentos

La representatividad del conjunto de documentos, es decir, lo bien que representan los documentos seleccionados al conjunto de documentos existentes en la web, está relacionada con el tamaño del corpus (a priori, a más documentos, más representatividad) y con la calidad de dichos documentos (cualquier proceso que descarte los documentos con peor calidad está mermando la representatividad real del conjunto finalmente seleccionado). Una vez fijados estos factores, influyen también en la representatividad los siguientes parámetros:

- La proporción de opiniones positivas y negativas contenidas en los documentos. Es deseable que la proporción obtenida en los documentos seleccionados sea similar a la existente en el conjunto total de documentos. Dado que esto puede ser difícil de medir, al menos sería deseable que el número de opiniones positivas y negativas esté equilibrado. Conseguir un conjunto de documentos que cumpla esta condición puede ser más o menos fácil dependiendo del tipo de documentos con el que estemos trabajando. Si nuestro corpus está compuesto por *reviews* de productos, podemos utilizar la puntuación asignada por los usuarios (que suele estar comprendida en un pequeño intervalo discreto, por ejemplo de 1 a 5) para seleccionar documentos uniformemente distribuidos a lo largo de dichas puntuaciones. Si los documentos provienen de foros o blogs, conseguir el equilibrio entre opiniones positivas y negativas puede ser más complicado, e implica un cierto trabajo manual previo de lectura y clasificación de los documentos según su contenido.
- La diversidad de objetos concretos sobre los que versan las opiniones. Dado que un dominio puede englobar a múltiples objetos concretos, el conjunto de documentos seleccionados debería recoger la mayor diversidad posible de objetos concretos dentro del dominio. Por ejemplo, si

estamos construyendo un sistema de extracción de opiniones sobre hoteles, es preferible que el corpus contenga *reviews* de múltiples hoteles, a poder ser con diferentes características particulares (por ejemplo, desde hostales a hoteles de 5 estrellas), antes que seleccionar un conjunto de *reviews* que hablen sobre el mismo hotel. Esto será especialmente importante para la generación de la taxonomía de características del dominio, ya que si los documentos seleccionados no recogen una suficiente variabilidad de objetos concretos, la taxonomía generada podría no recoger todas las posibles características del dominio. Siguiendo con el ejemplo de los hoteles, la elección de *reviews* para unos pocos hoteles que, supongamos, no dispusieran de piscina, ocasionaría que dicha característica no se viera reflejada en la taxonomía generada, y por tanto, el sistema no extraería correctamente opiniones acerca de dicha característica cuando procese *reviews* de otros hoteles que sí dispongan de la misma.

- Intervalo de tiempo de los documentos seleccionados. La proporción de opiniones positivas y negativas, las características en las que se centran la mayor parte de las opiniones, o incluso el tipo de lenguaje empleado, son variables que pueden cambiar con el tiempo. Por tanto, una selección de documentos que abarquen un mayor intervalo de tiempo será más representativa. En algunos casos, dependiendo del marco de aplicación del sistema o del dominio sobre el que actúa, podría interesarnos centrarnos en documentos más recientes, si prevemos que el sistema será utilizado principalmente para extraer opiniones acerca de objetos de una cierta novedad (como suele ocurrir por ejemplo con la mayoría de productos tecnológicos).

5.2.4. Definición del recurso

El corpus R_D está formado por un conjunto de documentos $\{r_1, r_2, \dots, r_n\}$. Cada uno de los documentos está formado por un conjunto de oraciones $\{s_1, s_2, \dots, s_n\}$, cada una de las cuales está formada por una lista de *tokens* $\{t_1, t_2, \dots, t_n\}$ y una lista de sintagmas o *chunks* $\{c_1, c_2, \dots, c_n\}$. Cada *token* t_k es una tupla $(word_k, lemma_k, pos_k, depHead_k, depType_k)$, donde $word_k$ es la palabra tal como ha sido escrita por el autor del documento, $lemma_k$ es el lema de la palabra anterior, pos_k es la categoría morfosintáctica o *part-of-speech*, $depHead_k$ es una referencia a otro token de la oración, del cuál depende sintácticamente la palabra actual según nos indica el análisis de dependencias de la oración, y $depType_k$ es el tipo de dicha relación de dependencia. Así mismo, cada uno de los *chunks* es una lista de tokens conti-

guos de la oración. Todas esta información lingüística será obtenida mediante la aplicación de herramientas de análisis. En la sección 6.2 describiremos las herramientas que hemos utilizado para llevar a cabo esta tarea.

5.2.5. Sintaxis del recurso

Cada documento del corpus es almacenado en un fichero de texto plano. Cada *token* es descrito en una línea del fichero, con cada una de las informaciones asociadas al token incluidas por columnas. Cada *token* tiene asignado un número identificativo único para la oración. A continuación mostramos un par de oraciones de ejemplo, extraídas del corpus de *reviews* de auriculares que utilizaremos en la sección 6:

The	1	the	AT	B-NP	det(3)
poor	2	poor	JJ	I-NP	mod(3)
design	3	design	NN1	I-NP	subj(14)
of	4	of	IO	B-PP	mod(3)
the	5	the	AT	B-NP	det(6)
band	6	band	NN1	I-NP	subj(8)
that	7	that	DD1	I-NP	-
holds	8	hold	VVZ	B-VP	-
these	9	these	DD2	B-NP	det(10)
headphones	10	headphone	NN2	I-NP	obj(8)
in	11	in	II	B-PP	mod(10)
place	12	place	NNL1	B-NP	pcomp-n(11)
is	13	be	VBZ	B-VP	-
unacceptable	14	unacceptable	JJ	B-ADJP	pred(13)
.	15	.	YF	O	-
If	1	if	CS	B-SBAR	-
you	2	you	PPY	B-NP	subj(4)
're	3	be	VVO	O	-
willing	4	willing	JJ	O	pred(3)
to	5	to	TO	O	aux(6)
modify	6	modify	VVO	I-VP	-
the	7	the	AT	B-NP	det(8)
band	8	band	NN1	I-NP	obj(6)
,	9	,	YC	O	-
these	10	these	DD2	B-NP	det(12)
wireless	11	wireless	JJ	I-NP	mod(12)
headphones	12	headphone	NN2	I-NP	subj(17)
can	13	can	VM	B-VP	aux(14)
be	14	be	VBO	I-VP	-

a	15	1	AT1	B-NP	det(17)
good	16	good	JJ	I-NP	lex-mod(17)
deal	17	deal	NN1	I-NP	pred(14)
.	18	.	YF	O	-

La primera columna corresponde a la palabra, y la segunda al número identificativo del *token*. La tercera columna es el lema, y la cuarta la categoría morfosintáctica, según el *tagset* de la herramienta de etiquetado utilizada. La quinta columna codifica los *chunks*, según notación IOB: una cadena comenzando por B indica el comienzo de un sintagma; una cadena comenzando por I indica que el sintagma continúa. La cadena que sigue a la B o a la I codifica el tipo de sintagma. Aquellas palabras que no forman parte de ningún *chunk* están etiquetadas con la letra O. Por último, la última columna almacena la información relacionada con el análisis de dependencias. Entre paréntesis aparece el número del *token* del que depende la palabra actual, precedido por el tipo de la relación de dependencia. Para indicar el final de una oración, se utiliza una línea en blanco.

Además del corpus así codificado, una versión más legible de los documentos, en la que sólo aparecen las palabras junto al número identificativo de cada una, es almacenada usando un formato XML para su uso en la etapa de anotación de evidencias de opinión. A continuación mostramos un ejemplo del formato utilizado, correspondiente a un *review* de un hotel:

```
<?xml version="1.0" encoding="UTF-8"?>
<document id="1">
  <sentence id="1">
    Cheap(1) ,(2) inconvenient(3) location(4) ,(5) cleanliness
    (6) is(7) NOT(8) their(9) priority(10) !(11) !(12) !(13)
  </sentence>

  <sentence id="2">
    I(1) definitely(2) do(3) not(4) recommend(5) staying(6)
    there(7) due(8) to(9) the(10) discomfort(11) and(12)
    inconvenience(13) of(14) this(15) hotel(16) and(17)
    location(18) .(19)
  </sentence>

  <sentence id="3">
    I(1) decided(2) to(3) stay(4) there(5) during(6) my(7) trip
    (8) to(9) Disney_World(10) for(11) affordability(12) and
    (13) got(14) CHEAP(15) !(16)
  </sentence>

  <!--...continuan el resto de oraciones del documento.-->
</document>
```

5.3. Taxonomía de características

La taxonomía de características cumple una triple función en el sistema. En primer lugar, es un catálogo de las características opinables en las que nuestro sistema se va a centrar. Aquellas características que no estén en la taxonomía serán por tanto ignoradas. Por otro lado, el recurso contiene un conjunto de palabras de característica para cada una de las características, que será utilizado por el sistema de extracción para detectar menciones a dichas características en los documentos de entrada. Por último, la taxonomía establece una jerarquía de dependencias entre las distintas características, la cuál será utilizada por algunos de los algoritmos de inducción de recursos (además de ser útil para la agregación y visualización de las opiniones extraídas, como propondremos en la sección 8.5.7).

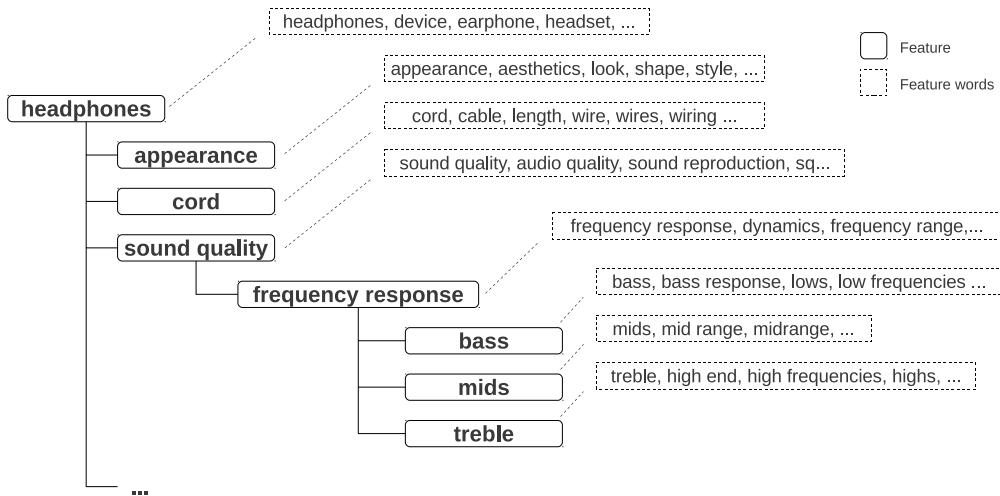
5.3.1. Definición del recurso

La taxonomía de características contiene el conjunto de características para las que deseamos que el sistema extraiga opiniones. Dicho conjunto F'_P es un subconjunto de F_P , el conjunto de todas las características opinables del objeto abstracto P . Para cada característica f_i , el recurso contiene un conjunto de palabras de característica FW'_{f_i} , un subconjunto de FW_{f_i} (idealmente, lo más cercano posible al conjunto completo). Todos estos pares (f_i, FW'_{f_i}) están organizados jerárquicamente: el objeto abstracto (que es en sí mismo una característica) es el nodo raíz de la taxonomía, con un conjunto de características colgando de él. Cada característica puede ser recursivamente descompuesta en un conjunto de subcaracterísticas. A modo de ejemplo, en la figura 5.1 se muestra una representación de una parte de la taxonomía de características correspondiente al objeto abstracto *headphones*.

No todos los nodos deben tener una lista de palabras de característica asociada: se permite incluir nodos que sirven para agrupar algunas características bajo algún criterio. Estos nodos no se corresponden con una característica opinable directamente referenciable en los documentos, y por ello no contienen palabras de característica asociadas. Su función es facilitar la legibilidad de la taxonomía, y servir como aglutinadores de opiniones en la fase de agregación y visualización.

5.3.2. Sintaxis del recurso

El formato que utilizamos para representar las taxonomías está basado en XML. A continuación mostramos un ejemplo correspondiente a la taxonomía parcial cuya representación ha sido mostrada en la figura 5.1:

Figura 5.1: Porción de la taxonomía de características para *headphones*

```

<?xml version="1.0" encoding="UTF-8"?>
<taxonomy objectClass="headphones">
    <featureRoot name="headphones" featWords="headphones , device ,
        earphone , headset">
        <feature name="appearance" featWords="appearance , aesthetics
            , look , shape , style"/>
        <feature name="cord" featWords="cord , cable , length , wire ,
            wires , wiring"/>
        <feature name="sound quality" featWords="sound quality ,
            audio quality , sound reproduction , sq"/>
        <feature name="frequency response" featWords="frequency
            response , dynamics , frequency range">
            <feature name="bass" featWords="bass , bass response ,
                lows , low frequencies"/>
            <feature name="mids" featWords="mids , mid range ,
                midrange"/>
            <feature name="treble" featWords="treble , high end , high
                frequencies , highs"/>
        </feature>
    </feature>
    <!--...continuan el resto de features de la taxonomia.-->
</featureRoot>
</taxonomy>
    
```

5.4. Corpus anotado

El corpus anotado, en el que se han identificado manualmente todas las evidencias de opinión contenidas en los documentos del corpus, es la base para la generación del resto de recursos, tal como veremos en el capítulo 6. Además, también nos permitirá evaluar la precisión y cobertura del sistema de extracción, según veremos en el desarrollo experimental propuesto en el capítulo 8.

5.4.1. Definición del recurso

El corpus anotado está compuesto por un conjunto de documentos $R'_D = \{r'_1, r'_2, \dots, r'_n\}$, donde cada documento r'_i es un par (r_i, OE_i) , siendo r_i un documento del corpus R_D y $OE_i = \{oe_{i_1}, oe_{i_2}, \dots, oe_{i_n}\}$ un conjunto de evidencias de opinión encontradas en r_i . Estas evidencias de opinión serán anotadas por uno o varios anotadores, según la metodología que explicaremos en la sección 6.5.

5.4.2. Sintaxis de las anotaciones

Las anotaciones se llevan a cabo sobre los ficheros XML generados anteriormente, añadiendo a cada elemento *sentence* un elemento *opinion* por cada evidencia de opinión observada en dicha oración, como puede verse en el siguiente ejemplo:

```
<sentence id="...">
  The(1) staff(2) is(3) rude(4) and(5) unresponsive(6) .(7)

  <opinion polarity="-" feature="staff"
    featWords="2" opWords="4" />
  <opinion polarity="-" feature="staff"
    featWords="2" opWords="6" />
</sentence>
```

Para cada opinión, se deben indicar como mínimo la polaridad, la característica y las palabras de opinión (atributos *polarity*, *feature* y *opWords*, respectivamente). La polaridad se indica utilizando el carácter “+” o “-”, según sea positiva o negativa. La característica debe coincidir con uno de los elementos de la taxonomía. Las palabras de opinión se indican mediante la lista de los números asociados a los *tokens* correspondientes, separados por comas. Si se trata de una opinión sobre característica explícita, deben indicarse también las palabras de característica, mediante el atributo *featWords*, y también mediante la lista de números de los *tokens* correspondientes.

En caso de que la característica sea referenciada mediante un pronombre, nuestro esquema de anotación contempla la inclusión de la palabra o palabras anteriores a las que referencia dicho pronombre, usando el atributo *featRef*. Por último, aquellas palabras que no participan en la polaridad de la opinión pero sí influyen en la intensidad de la misma, han sido anotadas usando el atributo *potency*. A continuación se muestra una anotación que incluye estos dos atributos:

```
<sentence id="...">
  It (1) seemed (2) really (3) nice (4) .(5)

  <opinion polarity="+" feature="hotel"
    featWords="1" featRef="hotel" opWords="4" potency="3" />
</sentence>
```

A efectos de aplicar la metodología descrita en este trabajo, no es necesario anotar estos dos últimos atributos (*featRef* y *potency*). Aún así, siendo pequeño el esfuerzo que requiere por parte del anotador, decidimos en su momento incluirlos en el proceso. En el caso del atributo *featRef*, nos permitirá llevar a cabo evaluaciones del sistema sin contar con un módulo de resolución de correferencias, simulando su aplicación mediante la sustitución de los pronombres en cuestión por sus sintagmas referenciados. En cuanto a las palabras de potencia, su inclusión en el esquema de anotación nos pareció oportuna para futuros experimentos de inducción de la intensidad de las opiniones extraídas.

5.5. Lexicón de opiniones

El lexicón de opiniones recoge información relativa a las palabras que han sido anotadas como palabras de opinión en el corpus. Por ejemplo, ¿qué polaridad tienen al ser utilizadas para calificar determinada característica del objeto? ¿O con qué probabilidad son utilizadas en el contexto de una opinión y con qué probabilidad no lo son? Este conocimiento nos será de gran ayuda a la hora de identificar y clasificar correctamente opiniones del dominio.

5.5.1. Definición del recurso

El lexicón de opiniones está compuesto por una lista de términos (potenciales palabras de opinión) y una serie de valores para cada término. El recurso recoge la siguiente información para cada término:

- *Support*: número de apariciones del término en el corpus anotado.

Cuento mayor es este valor, más precisas serán las estimaciones siguientes.

- *Feature-based opinion word probabilities*: estimación de la probabilidad de que la aparición de este término lleve pareja la existencia de una opinión sobre una característica concreta de la taxonomía (o alguna de las características que dependen de ella). Por tanto, habrá tantas estimaciones como características contenga la taxonomía. Por ejemplo, un término cuyo valor de esta medida sea 0 para una característica determinada, no ha sido anotado como palabra de opinión en ninguna opinión sobre dicha característica o sobre alguna característica hija de la misma. El valor asociado a la característica raíz de la taxonomía indica la probabilidad de que la aparición del término implique la existencia de una opinión sobre cualquier característica.
- *Feature-based semantic orientation polarities*: estimación de la polaridad de la orientación semántica del término cuando es utilizado en una opinión acerca de una determinada característica. De nuevo disponemos de tantas estimaciones como características posea la taxonomía. El rango de valores de la medida es $[-1,0; 1,0]$. Para un término que siempre haya sido usado en opiniones positivas o negativas para una determinada característica, el valor asignado será 1,0 o -1,0 respectivamente. Los valores intermedios indican un cierto grado de ambigüedad en la polaridad. Las estimaciones se calculan a partir de las opiniones positivas y negativas anotadas en el corpus acerca de la característica en cuestión y de todas las características en que aquella se descompone en la taxonomía.

Formalmente, el lexicón es una lista $\text{lexicon} = \{le_1, le_2, \dots, le_n\}$, con un elemento por término, en el que cada elemento es una tupla

$$le_i = (term, support, fbOpProbs, fbPolarities)$$

, donde $term$ es el término en cuestión, $support$ es un número entero mayor o igual que 1, y $fbOpProbs$ (*feature-based opinion word probabilities*) y $fbPolarities$ (*feature-based semantic orientation polarities*) son dos aplicaciones $F_P \rightarrow [0; 1] \in \Re$ y $F_P \rightarrow [-1; 1] \in \Re$, respectivamente.

Es importante señalar que el valor absoluto de las estimaciones de polaridad de la orientación semántica no está relacionado con la intensidad de las implicaciones positivas o negativas del término, como ocurre cuando se utiliza el concepto de orientación semántica en otros trabajos de clasificación de opiniones. En nuestro caso, el valor absoluto está relacionado con una cierta

probabilidad de que el término tenga implicaciones positivas o negativas. La mayoría de las veces un valor distinto de 1,0 o -1,0 indica la existencia de valores de la polaridad opuestos para algunos *subfeatures*. Por ejemplo, la estimación de la polaridad de “*cheap*” para el objeto abstracto “*headphones*” fue en nuestros experimentos igual a 0,4693, siendo -1,0 para la mayoría de las características en que se descompone el objeto (*appearance*, *durability*, *sound quality*,...) y 1,0 en una única pero más frecuentemente observada característica (*price*).

5.5.2. Sintaxis del recurso

El formato en el que se expresan estas medidas está basado en XML. A continuación mostramos una porción del diccionario de opiniones del dominio *headphones*:

```
<?xml version="1.0" encoding="UTF-8"?>
<opinionLexicon abstractObject="headphones">
    <entry term="acceptable" support="11">
        <fbEntry feature="headphones" opProb="0.5384"
            polarity="1.0"/>
        <fbEntry feature="sound quality" opProb="0.4615"
            polarity="1.0"/>
    </entry>
    <entry term="adjustable" support="12">
        <fbEntry feature="comfort" opProb="0.1176" polarity="1.0"/>
        <fbEntry feature="component" opProb="0.2352"
            polarity="1.0"/>
        <fbEntry feature="earcups" opProb="0.1176" polarity="1.0"/>
        <fbEntry feature="headband" opProb="0.1176" polarity="1.0"/>
        <fbEntry feature="headphones" opProb="0.3529"
            polarity="1.0"/>
    </entry>
    <entry term="light" support="64">
        <fbEntry feature="bass" opProb="0.0256" polarity="-1.0"/>
        <fbEntry feature="frequency response"
            opProb="0.0256" polarity="-1.0"/>
        <fbEntry feature="headphones" opProb="0.5641"
            polarity="0.8888"/>
        <fbEntry feature="sound quality" opProb="0.0256"
            polarity="-1.0"/>
        <fbEntry feature="weight" opProb="0.5384" polarity="1.0"/>
    </entry>
    <!-- continua ... -->
</opinionLexicon>
```

Los atributos *support* y *opProb* de los elementos de tipo *entry* se corresponden con las medidas de *support* y *opinion word probability*. Los atributos

opProb y *polarity* de los elementos de tipo *fbEntry* se corresponden con las medidas *feature-based opinion word probability* y *feature-based semantic polarity*. Para aquellas características para las que no aparezca un elemento *fbEntry*, ambas medidas se consideran nulas, de forma que no es necesario incluir sus entradas explícitamente en el recurso.

5.6. Indicadores de características implícitas

Al examinar las evidencias de opinión anotadas cuya característica está implícita (es decir, aquellas en las que no se han anotado palabras de característica), es fácil encontrar correlaciones inequívocas entre determinadas palabras de opinión y la característica sobre la que se está opinando. Por ejemplo, *comfortable* y *affordable* son palabras de opinión con polaridad positiva, cuya aparición suele ir ligada a la existencia de una opinión acerca de las características *comfort* y *price*, respectivamente. Por tanto, la presencia de estas palabras de opinión puede ser un buen indicativo de la existencia de una opinión implícita.

El recurso de indicadores de características implícitas (*implicit feature cues*) trata de capturar este tipo de información, que puede sernos muy útil para descubrir opiniones sobre características implícitas.

5.6.1. Definición del recurso

El recurso está formado por una lista de términos cuya presencia implica con cierta probabilidad la existencia de una opinión sobre determinada característica. Para cada término de la lista, se dispone de la siguiente información:

- *Support*: número de apariciones del término en el recurso. Cuanto mayor sea este valor, más fiables serán las estimaciones realizadas para el cálculo de probabilidades.
- *Feature-based implicit feature probabilities*: para cada una de las características de la taxonomía, un valor real entre 0 y 1 que indica la probabilidad de que la presencia del término en cuestión vaya unida a la existencia de una opinión sobre característica implícita en la que participe el término.
- *Feature-based implicit feature, not explicit, conditional probabilities*: se define de forma similar a la anterior, para cada una de las características de la taxonomía; se trata de la misma probabilidad anterior, pero condicionada a la no participación del término en una opinión explícita.

De manera formal, el recurso es una lista $cues = \{cue_1, cue_2, \dots, cue_n\}$, con un elemento por término, en el que cada elemento es una tupla

$$cue_i = (term, support, fbImplProbs, fbCondImplProbs)$$

, donde $term$ es el término en cuestión, $support$ es un número entero mayor o igual que 1, y $fbImplProbs$ (*feature-based implicit feature probabilities*) y $fbCondImplProbs$ (*feature-based implicit feature, not explicit, conditional probabilities*) son dos aplicaciones $F_P \rightarrow [0; 1] \in \mathfrak{R}$.

Para entender la diferencia entre las dos probabilidades incluidas en el recurso, supongamos la existencia de un par de términos t_1 y t_2 en el recurso, cuyos valores de $fbImplProbs$ y $fbCondImplProbs$ para una característica determinada f sean los siguientes:

$$fbImplProbs_{t_1}(f) = 1, fbCondImplProbs_{t_1}(f) = 1,$$

$$fbImplProbs_{t_2}(f) = 0,9, fbCondImplProbs_{t_2}(f) = 1.$$

En esta situación, el término t_1 ha sido siempre observado formando parte de una opinión implícita; si la estimación ha sido realizada sobre un número suficiente de observaciones (es decir, si el valor de $support$ es suficientemente alto), ante la presencia del término en un nuevo documento de entrada podemos inferir, con bastante seguridad, la existencia de una opinión implícita sobre f . En el caso del término t_2 , la certeza no será tan alta, puesto que la probabilidad $fbImplProb$ indica que en un pequeño número de casos la presencia del término no estuvo relacionada con la existencia de una opinión sobre f . Sin embargo, si previamente hemos llevado a cabo la extracción de las opiniones sobre características explícitas, y el término t_2 no participó en ninguna de las extraídas, entonces la presencia del mismo nos informa con bastante seguridad de la existencia de una opinión implícita sobre f (dando por supuesto que la extracción de opiniones sobre características explícitas fuese perfecta).

5.6.2. Sintaxis del recurso

Como en el resto de recursos, el formato que utilizamos está basado en XML. A continuación mostramos una porción del recurso de indicadores de características implícitas para el dominio *headphones*:

```
<?xml version="1.0" encoding="UTF-8"?>
<implicitCues abstractObject="headphones">
  <entry term="attractive" support="9">
    <fbEntry feature="appearance" prob="0.44" condProb="0.57">
  </entry>
```

```

<entry term="awesome" support="27">
  <fbEntry feature="appearance" prob="0.04" condProb="0.08">
    <fbEntry feature="headphones" prob="0.15" condProb="0.33">
  </fbEntry>
<entry term="feel great" support="5">
  <fbEntry feature="comfort" opProb="1.0" condProb="1.0">
</fbEntry>
<entry term="inexpensive" support="31">
  <fbEntry feature="price" opProb="0.77" condProb="0.77">
</fbEntry>

<!-- continua ...-->
</implicitCues>

```

Los atributos *prob* y *condProb* de los elementos de tipo *fbEntry* se corresponden con las medidas *feature-based implicit feature probabilities* y *feature-based implicit feature, not explicit, conditional probabilities*. Para aquellas características para las que no aparezca un elemento *fbEntry*, ambas medidas se consideran nulas, de forma que no es necesario incluir sus entradas explícitamente en el recurso.

Los ejemplos mostrados aquí sólo pretenden ilustrar el formato que hemos utilizado; en la sección 6.9 analizaremos los valores obtenidos experimentalmente en el recurso y sus implicaciones prácticas.

5.7. Patrones de dependencias

La taxonomía de características contiene información útil para que el sistema pueda identificar posibles palabras de característica, y el lexicón de opiniones permite encontrar y clasificar potenciales palabras de opinión. Pero el sistema de extracción también necesita enlazar correctamente unas y otras palabras, siempre que estén relacionadas. Algunas de las propuestas de otros investigadores utilizan simplemente la proximidad en el texto entre las características y las palabras de opinión para decidir que están relacionadas(Hu & Liu, 2004a; Ding et al, 2008b; Blair-Goldensohn et al, 2008). En nuestro caso, aunque también experimentaremos con aproximaciones basadas en ventanas de palabras, intentaremos explotar la estructura sintáctica de las oraciones. En concreto, utilizaremos la información proporcionada por un analizador de dependencias. La utilización de la información sintáctica ya ha sido previamente propuesta por otros autores(Popescu & Etzioni, 2005; Popescu et al, 2005; Yi & Niblack, 2005; Brody & Elhadad, 2010). Sin embargo, a diferencia de estos trabajos, en los que se dispone de un conjunto de patrones construidos a mano, en nuestro caso utilizaremos un conjunto de patrones específicos del dominio, a nivel de característica, inducidos automáticamente a partir

del corpus anotado.

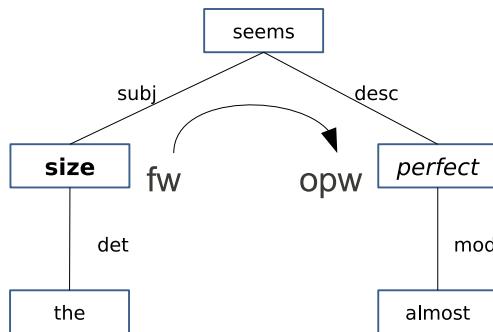
Las relaciones de dependencia conectan las palabras de las oraciones según la función sintáctica que desempeñan unas frente a otras. Se establece con ello un árbol de relaciones en el que, a diferencia del análisis sintáctico completo tradicional, los nodos del árbol son las propias palabras, no existiendo nodos intermedios (en el análisis sintáctico completo, los nodos intermedios del árbol representan unidades gramaticales a distintos niveles de abstracción, formadas por grupos de palabras, como por ejemplo sujeto, predicado, sintagma nominal, sintagma preposicional, ...). En cierto modo, el análisis de dependencias es una simplificación del análisis sintáctico completo.

Cada relación de dependencia está formada por un núcleo o *head word* y una palabra dependiente, y es etiquetada con una función sintáctica determinada (por ejemplo, *subj* para la función sujeto, o *mod* para la función modificador, entre otras¹). La idea es identificar a partir de las anotaciones del corpus aquellos patrones que se repiten más frecuentemente en los caminos que conectan las palabras de característica con las palabras de opinión relacionadas en el árbol de dependencias. Por ejemplo, para la frase “*The size seems almost perfect*” (ver figura 5.2), el patrón que se observa desde la palabra de característica “*seems*” a la palabra de opinión “*perfect*” es $N \rightarrow subj \rightarrow V \rightarrow desc \rightarrow J$, donde N , V y J representan las categorías morfosintácticas (nombre, verbo y adjetivo) de las palabras “*size*”, “*seems*” y “*perfect*”, respectivamente. Si este patrón se repite frecuentemente en nuestras anotaciones, podríamos plantearnos su utilización en el sistema de extracción para decidir qué palabras de opinión están comunicadas con ciertas palabras de característica previamente identificadas, siempre que encontremos un camino similar en el árbol de dependencias de la oración de entrada.

El ejemplo utilizado es el más simple posible, puesto que se trata de una opinión en la que participan una única palabra de característica y una única palabra de opinión. En aquellos casos en que tengamos varias palabras de cada tipo, surge el problema de cómo expresar la relación entre ambos conjuntos de palabras. Podemos buscar patrones entre las palabras dos a dos, o centrarnos en las relaciones que se establecen entre las palabras que actúan como núcleo de cada uno de los sintagmas. También podemos capturar las relaciones de dependencia existentes entre el núcleo de las palabras de opinión y el resto de palabras del mismo tipo, o entre el núcleo y las palabras correspondientes a expresiones especiales². Cada una de estas posibilidades

¹El conjunto de etiquetas utilizadas para las funciones de dependencia cambia de unos analizadores a otros; los utilizados en esta sección corresponden a las etiquetas utilizadas por el *parser* Minipar(Lin, 1998)

²Las expresiones especiales son definidas en la sección 5.8



The **size** seems almost *perfect*.

Figura 5.2: Árbol de dependencias de oración de ejemplo

está contemplada en la definición del recurso.

5.7.1. Definición del recurso

El recurso está formado por cinco listas ordenadas de patrones para cada una de las características de la taxonomía, más cinco listas con patrones aplicables a cualquier característica. Cada una de las listas contiene patrones que enlazan palabras participantes en una misma evidencia de opinión. Según los elementos que enlazan, podemos distinguir cinco tipos de patrones:

- Tipo 1: patrones para enlazar cualquier palabra de característica con cualquier palabra de opinión.
- Tipo 2: patrones para enlazar el núcleo de las palabras de característica con el núcleo de las palabras de opinión.
- Tipo 3: patrones para enlazar el núcleo de las palabras de opinión con el resto de las palabras de opinión.
- Tipo 4: patrones para enlazar el núcleo de las palabras de opinión con palabras correspondientes a las expresiones de negación.
- Tipo 5: patrones para enlazar el núcleo de las palabras de opinión con palabras correspondientes a las expresiones de polaridad dominante.

Cada una de las listas será utilizada por distintas implementaciones de los módulos abstractos que compondrán el sistema de extracción, como explicaremos en el capítulo 7. Cada uno de los patrones, que relacionan una palabra origen con una palabra destino, se define en base a una serie de restricciones que se concretan en los siguientes datos:

- **Característica sobre la que se aplica:** indica la característica de las opiniones sobre las que se aplica el patrón. Esto nos permite disponer de un juego de patrones específico para cada característica; nuestra hipótesis es que las opiniones sobre determinadas características suelen expresarse de maneras determinadas, y es por ello que incluimos la característica como restricción del patrón. No obstante, se contemplan también patrones sin esta restricción, aplicables a opiniones sobre cualquier característica.
- **Listas de restricciones morfosintácticas:** contienen la secuencia de etiquetas morfosintácticas de las palabras participantes en el camino que une las palabras origen y destino. Esta secuencia está dividida en dos listas, una correspondiente a la parte ascendente del camino, y otra a la parte descendente³. En el ejemplo anterior, la primera lista sería $\{N, V\}$ y la segunda $\{V, J\}$. Si ambas palabras, origen o destino, dependen una de la otra, directa o transitivamente, entonces una de las listas contendrá únicamente una etiqueta morfosintáctica (la correspondiente a la propia palabra origen o destino). Las cadenas utilizadas para especificar cada uno de los elementos de las listas se corresponden con prefijos de las etiquetas utilizadas por el etiquetador morfosintáctico que hayamos utilizado. Por ejemplo, N significa que se trata de un nombre en el *tagset* del etiquetador que hemos utilizado en nuestros experimentos (Penn Treebank(Marcus et al, 1994)); podríamos indicar una restricción más detallada mediante la cadena NNP , que exigiría que se tratara de un nombre propio. También permitimos la no imposición de restricción alguna en los patrones acerca de las categorías morfosintácticas de las palabras participantes. Pretendemos con ello experimentar con los distintos grados de restricción.
- **Listas de tipos de relaciones de dependencia:** contienen la secuencia de tipos de relaciones de dependencia entre cada dos palabras del camino. Al igual que en el caso anterior, se trata de un par de listas, una correspondiente a la parte ascendente del camino y otra a la descendente. En el ejemplo anterior, la primera lista sería $\{subj\}$ y la segunda $\{desc\}$. Ésta es la información principal del patrón, y la única de las restricciones que no puede ser omitida. Si ambas palabras, origen o destino, dependen una de la otra, directa o transitivamente, entonces una de las listas estará vacía.

³Dos nodos cualesquiera de un árbol están comunicados mediante un camino mínimo formado por la unión de los dos subcaminos que comunican a cada uno de los nodos con el ancestro común más cercano.

- **Palabra destino:** de manera opcional, un patrón puede indicar la palabra destino sobre la que se aplica. Si al aplicar el patrón, la palabra destino encontrada no es la especificada en esta restricción, el resultado será descartado. Esta restricción se utilizará únicamente en los patrones de tipo 4 y 5.

Además de la propia definición del patrón, en el recurso se incluye la siguiente información para cada patrón:

- *Support*: número de veces que el patrón ha sido observado en el corpus anotado.
- *Precision*: número real entre 0 y 1, que mide la precisión del patrón al enlazar palabras origen y destino del tipo adecuado, atendiendo al tipo del patrón. Por ejemplo, una precisión igual a 0,95 para un patrón del tipo 1 nos indicaría que la aplicación del patrón partiendo de una palabra de característica cualquiera correctamente anotada nos conduciría a una palabra de opinión relacionada con la anterior, en un 95 % de los casos.
- *Recall*: número real entre 0 y 1, que mide la cobertura del patrón actual. Por ejemplo, una cobertura igual a 0,01 para un patrón del tipo 1 indicaría que aplicar este patrón a todas las palabras de característica correctamente etiquetadas conduciría a capturar sólo el 1 % de las palabras de opinión relacionadas con esa *feature word*.
- *Accumulated precision and recall*: las listas de patrones de cada tipo para cada una de las características son ordenadas en orden decreciente de precisión, y a igual precisión en orden decreciente de cobertura. Respecto a la posición que ocupen en estas listas ordenadas, en el recurso se proporcionan valores de precisión y cobertura acumulados para cada patrón, que indican la precisión y la cobertura estimadas tras la aplicación del patrón actual y de todos los anteriores a éste en la lista. Estos valores, que siguen una progresión monótona decreciente en el caso de la precisión acumulada y monótona creciente en el caso de la cobertura acumulada, permiten seleccionar un subconjunto de patrones tras establecer ciertos valores umbrales para la precisión y/o la cobertura que deseemos obtener en la aplicación de los mismos.

Formalmente, el recurso de patrones de dependencias está formado por cinco listas de patrones para cada característica, más cinco listas sin restricciones de característica:

$$\text{patterns}(f_i) = \{pats_1(f_i), pats_2(f_i), pats_3(f_i), pats_4(f_i), pats_5(f_i)\},$$

$$f_i \in F_P \cup \{*\}.$$

Cada una de las listas $pats_j(f_i)$ está formada por patrones de cada uno de los tipos descritos anteriormente, de la forma:

$$pat_j = (pos_{asc}, pos_{desc}, dep_{asc}, dep_{desc}, support, p, r, accP, accR),$$

donde pos_{asc} y pos_{desc} son las listas ascendente y descendente de restricciones morfosintácticas, dep_{asc} y dep_{desc} son las listas ascendente y descendente de tipos de relaciones de dependencia, $support$ es un número entero mayor o igual que 1, y p , r , $accP$ y $accR$ son números reales entre 0 y 1 correspondientes a la precisión y cobertura individuales y acumuladas. Además, las listas están ordenadas en función de la precisión de los patrones, y a igual precisión, en función de la cobertura:

$$pat_i.p \geq pat_j.p, \forall i \in [1, n], j \in [1, n], i < j.$$

$$pat_i.p = pat_j.p \Rightarrow pat_i.r \geq pat_j.r, \forall i \in [1, n], j \in [1, n], i < j.$$

5.7.2. Sintaxis del recurso

A continuación mostramos una porción del recurso de patrones de dependencias para el dominio *headphones*:

```
<?xml version="1.0" encoding="UTF-8"?>
<patterns objectClass="headphones">
  <patternList type="1" feature="bass">
    <pattern order="1" destWord="*" posAsc="N,N,J" posDesc="J"
             depAsc="nn,subj" depDesc="" p="1.0" r="0.009"
             pAcc="1.0" rAcc="0.009"/>
    <pattern order="2" destWord="*" posAsc="N,V,V" posDesc="J,V"
             depAsc="subj,conj" depDesc="pred" p="1.0" r="0.006"
             pAcc="1.0" rAcc="0.015"/>
    <pattern order="3" destWord="*" posAsc="N,J,V" posDesc="X,V"
             depAsc="subj,pred" depDesc="neg" p="1.0" r="0.006"
             pAcc="1.0" rAcc="0.021"/>
    <pattern order="4" destWord="*" posAsc="N" posDesc="V,N"
             depAsc="" depDesc="mod" p="1.0" r="0.004"
             pAcc="1.0" rAcc="0.026"/>
    <!-- continua ... -->
  </patternList>
  <!-- continuan patrones del resto de tipos para todas las
       caracteristicas de la taxonomia ... -->
</patterns>
```

Dentro de los elementos $type1$, $type2, \dots$, $type5$ aparecen los patrones de cada tipo, ordenados según lo explicado anteriormente. El carácter “*” es utilizado en el atributo *feature* para indicar ausencia de restricción de la característica a la que son aplicables los patrones, y de igual forma en el atributo *destWord* para indicar ausencia de restricción de la palabra destino.

En la sección 6.10 analizaremos los valores obtenidos experimentalmente en el recurso y sus implicaciones prácticas.

5.8. Listas de expresiones especiales

Una de las subtareas que debe realizar nuestro sistema es la inducción de la polaridad de las palabras de opinión previamente identificadas. Esto puede implementarse de múltiples maneras, haciendo uso de distintas soluciones que nos permitan predecir la orientación semántica de las palabras y expresiones participantes (por ejemplo, haciendo uso de un lexicón de opiniones). Sin embargo, existen ciertas expresiones que pueden formar parte de las palabras de opinión y que exigen un tratamiento distinto al resto a la hora de calcular la orientación semántica de la opinión en la que participan.

Hemos identificado tres tipos de expresiones con estas características: las expresiones de negación, las expresiones de no-negación y las expresiones de polaridad dominante. En las siguientes subsecciones explicamos las particularidades de cada tipo y mostramos los ejemplos que hemos reunido hasta el momento.

5.8.1. Expresiones de negación

Todos los trabajos de clasificación basada en la opinión, ya sea a nivel de documentos, frases o sintagmas, y los trabajos de extracción de opiniones, tienen en cuenta de alguna forma el tratamiento de las negaciones. Es evidente que si t es un término con orientación semántica positiva, $not\ t$ tendrá orientación semántica negativa.

En nuestro caso, el módulo encargado de calcular la polaridad de las opiniones extraídas debe tener en cuenta la existencia de este tipo de partículas, para calcular correctamente las orientaciones semánticas de las palabras de opinión involucradas. Aunque en un primer momento decidimos incluir en la propia lógica de los módulos encargados de este cálculo el tratamiento especial de la palabra *not*, la constatación de que existen bastantes más expresiones con el mismo comportamiento, y la intención de conseguir que el sistema trabaje en otros idiomas distintos al inglés, nos llevaron a definir una lista de expresiones de negación, como un recurso más que el sistema

recibirá como entrada. La particularidad de este recurso (y del resto de listas que presentaremos a continuación) es que no es dependiente del dominio: podemos reutilizar las listas de expresiones especiales de un dominio a otro, añadiendo las nuevas expresiones que vayan apareciendo en el proceso de anotación.

La lista de expresiones de negación de la que disponemos es la siguiente:

barely, hardly, lack, least, never,
no, not, not too, scarcely

No se incluyen contracciones entre verbos auxiliares y la palabra *not*, dado que el proceso de tokenización en la etapa de preprocesado separa dichas contracciones en un par de *tokens*. Algunos ejemplos en los que participan estas expresiones son los siguientes:

This car is barely drivable.
Camera is hardly working.
The images lack of clarity.
It is the least funny movie I have ever seen.
I would never recommend it.
The plot is scarcely surprising.

El caso de la expresión *not too* es particular y será explicado cuando introduzcamos las expresiones de polaridad dominante.

5.8.2. Expresiones de no-negación

Las expresiones que hemos llamado de no-negación son aquellas que, incluyendo alguna de las expresiones de negación, no deben ser consideradas como negaciones a efectos de cambiar la polaridad de las palabras de opinión a las que acompañan, sino que simplemente deben ser ignoradas en el cálculo de la orientación semántica. Por ejemplo, en la oración “*It is not just lightweight, but also durable.*”, aún apareciendo la expresión de negación “*not*” delante de “*lightweight*”, no se invierte la polaridad de esta última. En este caso, “*not just*” es una expresión de no-negación. Sólo hemos identificado dos de estas expresiones, que son las siguientes:

not just, not only

Estas expresiones deberán ser comprobadas por el módulo de cálculo de la polaridad de las opiniones antes que las expresiones de negación.

5.8.3. Expresiones de polaridad dominante

Las expresiones de polaridad dominante son aquellas cuya presencia en un conjunto de palabras de opinión determinan de manera inequívoca la polaridad de la misma, sin importar la polaridad del resto de palabras a las que acompañen. Por ejemplo, puede ser deseable que determinado objeto sea ligero, pero si lo calificamos como *demasiado* ligero, la opinión vertida es claramente negativa, a pesar de la polaridad a priori positiva del término *ligero*. Es la palabra *demasiado* la que nos informa de que la opinión es negativa, de forma inequívoca, sin importar la polaridad del término al que acompaña.

Se dispone por tanto de dos listas de expresiones de este tipo, una con aquellas que indican una polaridad positiva, y otra con las que indican polaridad negativa. Hemos identificado las siguientes:

- Positivas:

enough, sufficient, sufficiently, reasonably

- Negativas:

insufficient, insufficiently, excessive, excessively,
overly, too, at best, too much, unnecessarily

En el caso de la palabra *too*, cuando aparece precedida por *not* no se comporta como una expresión de polaridad dominante, sino como una expresión de negación. Es por eso que la expresión *not too* aparece en la lista de expresiones de negación. El módulo encargado del cálculo de la polaridad de las opiniones deberá por tanto comprobar la existencia de expresiones de negación previamente a las expresiones de polaridad dominante.

Capítulo 6

Metodología para la generación de los recursos

Resumen: En el presente capítulo proponemos una metodología para la generación de los recursos definidos en el capítulo 5 para un dominio específico, incluyendo un flujo de trabajo y presentando una serie de herramientas y algoritmos encaminados a automatizar el proceso en la mayor medida posible. Se utilizará un dominio concreto (*headphones*) para ejemplificar cada uno de los pasos de la metodología, y para llevar a cabo algunos experimentos con la intención de ponderar las distintas opciones planteadas en algunos de los algoritmos.

6.1. Introducción

Según nuestra propuesta, la construcción de un sistema de extracción de opiniones para un determinado dominio pasa por la generación de los recursos anteriormente descritos. En comparación con un sistema independiente del dominio y completamente no supervisado, nuestro acercamiento exige un mayor esfuerzo de adaptación a un nuevo dominio. En la presente sección, mostramos una metodología para la generación de los recursos que pretende minimizar el esfuerzo necesario, mediante la definición de herramientas de apoyo y algoritmos de inducción.

En la figura 6.1 se muestra un resumen del proceso de generación de recursos, y se indica qué partes del proceso son llevadas a cabo de manera manual, automática o semiautomática. Tras reunir un corpus de documentos del dominio deseado, el primer paso consiste en la construcción de una primera versión de la taxonomía de características, para lo que se parte de una lista

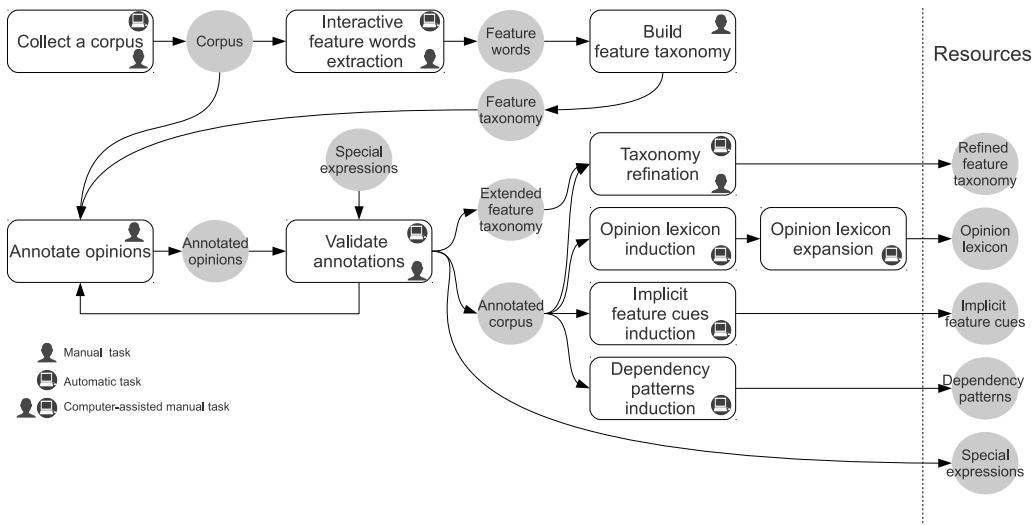


Figura 6.1: Proceso de generación de los recursos

de *feature words* generada semi-automáticamente mediante la utilización de una herramienta de extracción. Esta primera versión de la taxonomía es utilizada en la tarea de anotación de opiniones, a partir de la cual obtendremos el corpus anotado. La tarea de anotación va acompañada de un proceso de validación de las anotaciones, que permite asegurarnos de que el recurso obtenido posee la mejor calidad posible. Además de las evidencias de opinión, en este paso son identificadas nuevas expresiones de negación, no-negación y polaridad dominante que pasarán a formar parte de las listas respectivas. El proceso de validación también genera como salida una versión ampliada de la taxonomía de características, que deberá ser refinada antes de obtener la versión final de la misma. A partir del corpus anotado, los tres recursos restantes serán obtenidos mediante una serie de algoritmos de inducción.

A continuación explicamos cada uno de los pasos para la generación de los recursos, incluyendo las herramientas de apoyo y algoritmos empleados. Además, mostraremos algunos ejemplos de los recursos obtenidos al aplicar la metodología a un dominio de ejemplo (*headphones*). Este caso de aplicación también será utilizado para experimentar en algunos de los pasos con los distintos parámetros configurables de algunos de los algoritmos propuestos.

6.2. Recolección y procesado del corpus

La obtención de los documentos de opinión del dominio es el primer paso a dar para la generación de los recursos. Si estamos construyendo un sistema

ma de extracción sobre *reviews* de productos, podemos emplear una multitud de webs de *reviews* de productos existentes en Internet, tales como www.epinions.com o www.cnet.com, entre otras. Si estamos construyendo un sistema de extracción de opiniones sobre otro tipo de objetos, debemos identificar fuentes de Internet a partir de las cuales obtener textos de opinión adecuados. Por ejemplo, si el sistema en cuestión trata de extraer opiniones acerca de determinado organismo, una opción sería utilizar los textos escritos en foros públicos puestos a disposición por el propio organismo o de manera externa al mismo.

Una vez identificadas las fuentes de los documentos, es preciso extraer los documentos mediante la utilización de un *crawler*, seleccionar un subconjunto de los documentos extraídos que satisfagan ciertos criterios de calidad y representatividad, y por último procesar los documentos seleccionados mediante ciertas herramientas de análisis lingüístico, para así obtener un corpus que se adapte a la definición y sintaxis expuestas en la sección 5.2.

6.2.1. Extracción de los documentos

La extracción de los documentos una vez seleccionada la fuente o fuentes de los mismos, se puede llevar a cabo mediante algún *crawler*: herramientas que permiten configurar ciertas expresiones regulares que, ejecutadas sobre el código *html* recuperado a partir de ciertas direcciones URL, consiguen extraer los campos de texto en los que estamos interesados. En nuestros experimentos, hemos utilizado la herramienta de código abierto *Web-Harvest*¹, aunque existen muchas otras opciones disponibles. Para construir fácilmente las expresiones regulares que identifican los campos de texto (en el caso de *Web-Harvest* se trata de expresiones *XPath*), utilizamos las herramientas *XPath Checker*² y *Solvent*³, ambos disponibles en forma de *plug-ins* para el navegador *Firefox*⁴.

Para el dominio *headphones*, trajimos todos los *reviews* de auriculares disponibles a fecha de septiembre de 2008 en la web www.epinions.com. Obtuvo un total de 2591 documentos, correspondientes a *reviews* acerca de 444 modelos distintos de auriculares.

¹<http://web-harvest.sourceforge.net/>

²<https://addons.mozilla.org/en-US/firefox/addon/1095/>

³<http://simile.mit.edu/wiki/Solvent>

⁴<http://www.mozilla-europe.org/>

Reviews	599
Oraciones	8302
Palabras	142832

Cuadro 6.1: Algunos datos del corpus obtenido para el dominio *headphones*

6.2.2. Selección de los documentos

Si bien a la hora de extraer los documentos es pertinente obtener cuantos más documentos mejor, dado que los documentos que finalmente conformen nuestro corpus deberán ser manualmente anotados, debemos seleccionar un subconjunto de documentos que permitan que dicha tarea sea abordable según los recursos de los que dispongamos. En principio, cuantos más documentos seamos capaces de anotar, más calidad tendrán los recursos obtenidos a partir de ellos; algunos experimentos serán llevados a cabo en el capítulo 8 para medir las repercusiones del tamaño del corpus en el rendimiento del sistema.

Tras estimar el número de documentos a cuyo etiquetado podemos abordar con los recursos disponibles, debemos escoger los documentos que conformarán el corpus. En primer lugar, filtraremos aquellos documentos que no superen ciertos criterios de calidad: documentos demasiado concisos o demasiado largos, escritos usando exclusivamente letras mayúsculas, o en los que aparezcan caracteres extraños. Después, llevaremos a cabo una selección aleatoria a partir de los documentos restantes. En el caso de aplicación que nos ocupa, dado que disponemos de una puntuación global de 1 a 5 asignada por cada autor al producto, hicimos uso de dicha puntuación para seleccionar el mismo número de críticas con cada una de las puntuaciones posibles. Dado que el número de modelos de auriculares disponibles era alto, no fue necesario ningún control para asegurar que los documentos seleccionados abarcaran una amplia cantidad de objetos concretos diferentes, tal como es deseable desde el punto de vista de la representatividad del corpus (ver sección 5.2.3). No controlamos la densidad de opiniones de los documentos seleccionados, por lo que en el corpus existen documentos con casi ninguna opinión real; hemos preferido no filtrar este tipo de documentos, para que nuestro corpus refleje más fielmente el tipo de situaciones que se le presentarán al sistema de extracción de opiniones.

Una vez filtrados y seleccionados los documentos, obtuvimos un corpus de 599 *reviews*. Mostramos algunos datos del mismo en la tabla 6.1.

6.2.3. Procesado de los documentos

Una vez se han seleccionado los documentos del corpus, estos son analizados por una serie de procesadores lingüísticos para llevar a cabo *tokenización* y *lematización*, detección de oraciones, etiquetado morfosintáctico, análisis sintáctico superficial (*chunking*) y análisis de dependencias. En nuestros experimentos, hemos utilizado la herramienta *Freeling* (Atserias et al, 2006) para el tokenizado, la lematización y la detección de oraciones. El etiquetado morfosintáctico fue llevado a cabo utilizando *TnT* (Brants, ????), a partir de un modelo construido utilizando el corpus *Susanne* (Sampson, 1992). El *chunking* se llevó a cabo mediante la herramienta *Yamcha* (Kudo & Matsumoto, 2003), utilizando modelos pre-entrenados por Mihai Surdeanu que forman parte de la suite *BIOS*⁵. Para el análisis de dependencias empleamos *Minipar* (Lin, 1998) y *MaltParser* (Hall, 2006). Por supuesto, estas herramientas pueden intercambiarse por otras que cumplan la misma función.

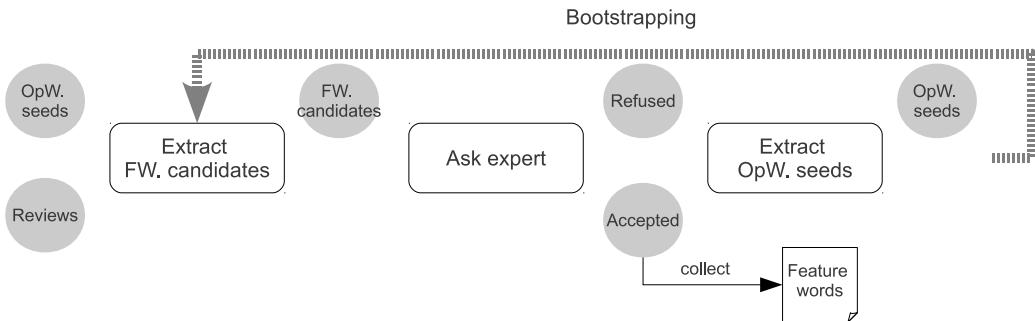
6.3. Extracción semiautomática de *feature words*

Como paso previo a la construcción de la taxonomía, se genera una lista de *feature words* del dominio. Esto se lleva a cabo mediante la ejecución de una herramienta interactiva que, a partir del corpus completo, muestra candidatos a *feature words*, que deben ser confirmados o rechazados manualmente. La herramienta utiliza la información aportada por los candidatos aceptados y rechazados para refinar a lo largo de la ejecución la lista de candidatos que irá mostrando, lo que conforma un proceso iterativo que termina en el momento en que el usuario de la herramienta lo decida (presumiblemente, cuando se rechacen un número elevado de candidatos consecutivos).

Dos son los principios en los que se basa el algoritmo de extracción de la lista de candidatos a *feature words*:

1. Las expresiones con las que la gente suele referirse a las características opinables de un objeto abstracto tienden a converger, por lo que es esperable que las mismas *feature words* se repitan a lo largo de los documentos del corpus (Hu & Liu, 2004b).
2. Dado que trabajamos con textos de opinión, las *feature words* aparecen frecuentemente en opiniones, y por tanto cerca de palabras de opinión.

⁵<http://www.surdeanu.name/mihai/bios/index.php>

Figura 6.2: Extracción interactiva de *feature words*

6.3.1. Descripción del algoritmo

Siguiendo estos dos principios, proponemos un algoritmo de *bootstrapping* para extraer candidatos a *feature words*. La idea es utilizar ciertas palabras de opinión como semillas⁶, para localizar posibles *feature words* a partir de algunos patrones morfosintácticos sencillos en los que participan dichas semillas. Por ejemplo, si observamos frecuentemente en el texto la aparición de “*excellent sound quality*”, podemos considerar que “*sound quality*” es un buen candidato a ser propuesto como *feature words*. A su vez, a medida que descubrimos nuevos *feature words*, podemos utilizar los mismos patrones morfosintácticos anteriores para localizar nuevas palabras de opinión, que a su vez pueden ser utilizadas para descubrir nuevas *feature words*.

Los candidatos a *feature words* que mediante este procedimiento se van generando, deberán ser aceptados o rechazados por la persona encargada de la construcción de la taxonomía. De esta forma, el proceso se lleva a cabo como se ilustra en la figura 6.2. En un primer momento, la extracción de candidatos a *feature words* se lleva a cabo utilizando las semillas de opinión iniciales. La lista obtenida es ordenada por número de apariciones, de manera que se comienza mostrando al usuario aquellos candidatos más frecuentes. El usuario deberá confirmar o rechazar cada uno de los candidatos. En determinado momento, se lleva a cabo una iteración para refinar la lista de candidatos, utilizando los nuevos candidatos aceptados para encontrar nuevas semillas de opinión, y empleando a su vez estas nuevas semillas para refinar la lista de candidatos. El algoritmo que busca los nuevos candidatos lo hace apoyándose en la lista de candidatos y los conteos de la frecuencia de aparición de los mismos previamente calculados, actualizando las frecuencias y añadiendo nuevos candidatos a partir de las nuevas semillas.

⁶Utilizamos cuatro semillas con muy claras implicaciones evaluativas: *excellent, good, poor* y *bad*

Utilizamos dos construcciones simples para encontrar las *feature words* a partir de las semillas de opinión (y nuevas semillas de opinión a partir de las *feature words* que ya han sido aceptadas). Por un lado, la palabra de opinión puede aparecer como modificador de las *feature words*, precediéndolas, como en el ejemplo anterior (“*excellent sound quality*”). Por otro lado, también contemplamos oraciones copulativas en las que la característica hace la función de sujeto, siendo la palabra de opinión el atributo (por ejemplo, “*The sound quality is excellent*”). Podemos añadir además distintas restricciones morfosintácticas a ambos participantes:

- Basándonos en la información proporcionada por el etiquetador de *part-of-speech*, podemos exigir que las *feature words* sean nombres, y las palabras de opinión, adjetivos.
- Basándonos en la información proporcionada por el analizador sintáctico superficial, podemos exigir que las *feature words* estén contenidas en un sintagma nominal. Así mismo, las palabras de opinión podrían formar parte de un sintagma nominal (si aparecen como modificador de las *feature words*) o de un sintagma adverbial (si aparecen como atributo en una oración copulativa).

El momento en el que la aplicación decide extraer nuevas semillas de opinión y refinar la lista de candidatos a mostrar al usuario puede ser:

- Cada vez que se acepte un nuevo candidato. En este momento, hay nueva información, con lo cual tiene sentido refinar los siguientes candidatos que se van a mostrar al usuario. En el lado negativo, a priori esto haría que el proceso fuese más lento.
- Cada vez que se acepten n candidatos. De esta manera, el proceso es algo más ágil. Parece además lógico pensar que hasta que no hay un número suficiente de candidatos validados, el proceso de refinamiento no cambiará significativamente la lista de próximos candidatos a validar.
- Cada vez que se rechace un candidato. De esta manera, mientras el usuario esté aceptando candidatos, está obteniendo una buena impresión del sistema, y por tanto no es necesario refinar los candidatos que se mostrarán.
- Igual que el anterior, pero esperando a que se rechacen n candidatos.

También podemos variar el número de nuevas semillas de opinión que se añaden en cada iteración del proceso, desde añadir una única semilla nueva en cada paso (suficiente para encontrar nuevos candidatos a *feature words*, y permitiendo una ejecución rápida) hasta añadir todas las semillas encontradas (encontrando más cantidad de candidatos, pero ocasionando una ejecución más lenta).

6.3.2. Simulación de la extracción de *feature words*

Para estudiar la influencia del uso de las restricciones morfosintácticas, el criterio de refinamiento de la lista de candidatos y el número de semillas nuevas en cada iteración, hemos implementado una simulación del proceso, en la que en lugar de la participación del experto, utilizamos una taxonomía previamente construida para decidir qué candidatos son aceptados y qué candidatos rechazados. Aplicaremos esta simulación al dominio *headphones* y analizaremos la influencia de los distintos parámetros; para ello utilizaremos algunas métricas que nos informarán de la efectividad de la herramienta.

Para construir la taxonomía de características que nos sirva de guía en la simulación, partimos de una lista ordenada por frecuencia de aparición de los nombres y parejas de nombres que aparecen en el corpus de *reviews* de *headphones*. A partir de esta lista, se construyó manualmente la taxonomía. El esfuerzo necesario para ello es grande, pues la lista de candidatos a *feature words* es enorme (7315 para el corpus de *headphones*); es precisamente esto lo que pretendemos solucionar mediante la utilización de la herramienta descrita.

Para cada uno de los experimentos, fijamos los parámetros y llevamos a cabo la simulación, en la que para cada candidato sugerido por la herramienta decidimos si debe ser aceptado o rechazado en función de si dicho candidato pertenece o no a la taxonomía-objetivo. De esta manera, podemos medir la influencia de los parámetros de la herramienta en el proceso. Para cada experimento, hemos observado las siguientes métricas:

- *Suggestion precision*: porcentaje de aceptación de los candidatos sugeridos por la herramienta en cada momento de la simulación. Es deseable que se mantenga lo más alto posible a lo largo de la ejecución, de manera que el proceso converja lo más rápidamente posible hacia la lista final de *feature words*.
- *Feature words recall*: porcentaje del total de palabras de característica incluidas en la taxonomía-objetivo que ya han sido aceptadas en cada momento de la simulación. Cuanto más rápidamente crezca este valor, mejor funciona el proceso.

- *Simulation time*: segundos transcurridos desde el comienzo del proceso en cada momento de la ejecución. El proceso de refinamiento de los candidatos puede volverse demasiado lento para determinados valores de los parámetros. Este valor deberá ser tenido por tanto en cuenta para evitar que el proceso se vuelva demasiado tedioso para el experto.

6.3.3. Influencia de las restricciones morfosintácticas

Las primeras simulaciones ejecutadas estuvieron encaminadas a medir la influencia en el proceso de las restricciones morfosintácticas impuestas a los candidatos a palabras de característica o de opinión. Hemos llevado a cabo tres experimentos:

- No aplicar restricciones morfosintácticas. En este caso, cualquier palabra o par de palabras podrá ser un candidato a *feature word*, y cualquier palabra suelta podrá ser identificada como palabra de opinión.
- Aplicar restricciones basadas en etiquetas *part-of-speech*. Las palabras de característica deben ser uno o dos nombres consecutivos, y las palabras de opinión serán adjetivos sueltos.
- Aplicar restricciones basadas en *chunks*. En este caso, empleamos la información proporcionada por el analizador sintáctico superficial, para exigir a las palabras de característica que formen un sintagma nominal, y a las palabras de opinión que formen un sintagma adjetival, o bien formen parte de un sintagma nominal.
- Aplicar ambas restricciones anteriores. Las palabras de característica deberán ser nombres que conformen un sintagma nominal, y las palabras de opinión, adjetivos que formen un sintagma adjetival o que formen parte de un sintagma nominal.

El resto de parámetros quedó fijado con los valores más conservadores: un candidato aceptado para refinar, y una nueva semilla de opinión por iteración (a priori, la opción que debe conseguir una mayor precisión en las sugerencias). En la figura 6.3 se muestra la evolución a lo largo de la ejecución de las métricas anteriormente descritas. Las gráficas muestran los valores de las métricas para las primeras 1000 sugerencias realizadas por la herramienta, salvo la gráfica correspondiente a la precisión de la sugerencias, en la que hemos optado por mostrar un intervalo menor de la ejecución, puesto que es al principio de la ejecución donde se producen variaciones más interesantes. En resumen, se observa que la aplicación de restricciones mejora la cobertura de

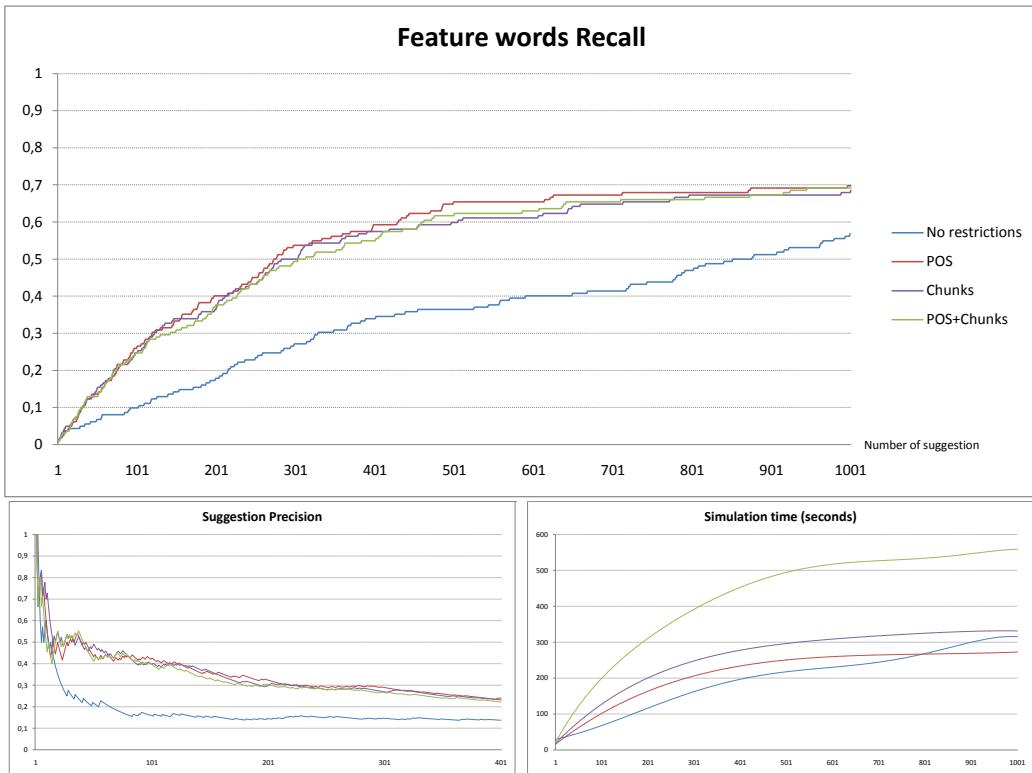


Figura 6.3: Simulación del proceso interactivo de extracción de *feature words*: influencia de las restricciones morfosintácticas.

feature words y consigue una mayor precisión en las sugerencias. En cuanto a los distintos tipos de restricciones empleadas, las diferencias observadas son pequeñas. Descartamos el empleo de restricciones basadas en *POS* y *chunks* a la vez, porque el tiempo de ejecución se ve sensiblemente incrementado en este caso. Como elección final, escogeríamos las restricciones basadas en *part-of-speech*, que consigue pequeñas mejoras en la cobertura y un tiempo de ejecución algo menor; aunque el empleo de restricciones basadas en *chunks* también sería aceptable.

Nuestra intuición previa era que la información relativa a los *chunks* representaría un valor añadido con respecto al empleo únicamente de etiquetas *POS*. Sin embargo, téngase en cuenta que la precisión obtenida por los analizadores morfosintácticos es muy superior a la obtenida por los analizadores sintácticos superficiales; la mejor caracterización que a priori debería representar la información aportada por los sintagmas queda desvirtuada por la menor precisión de dichos analizadores. Hemos observado este mismo fenómeno en otros momentos de nuestra investigación: el empleo de analiza-

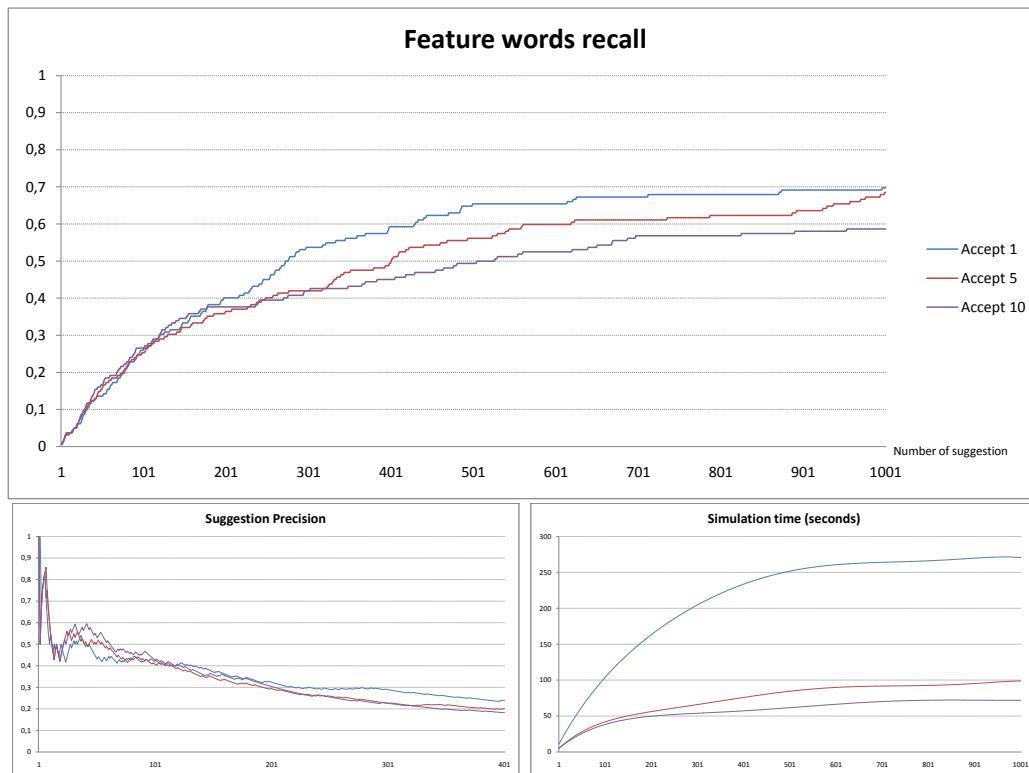


Figura 6.4: Simulación del proceso interactivo de extracción de *feature words*: influencia del número de candidatos aceptados antes de iterar.

dores lingüísticos de mayor nivel de abstracción lleva parejo un mayor número de errores, lo que puede ocasionar que soluciones a priori más sofisticadas obtengan peores resultados que las que se basan simplemente en información léxica o morfosintáctica.

6.3.4. Influencia del número de candidatos aceptados o rechazados

Pretendemos ahora decidir si es más adecuado refinar la lista de candidatos tras un número determinado de candidatos aceptados o rechazados. En las figuras 6.4 y 6.5 se muestra el comportamiento de las métricas cuando se refina la lista de candidatos tras aceptar o rechazar 1, 5 o 10 candidatos. Se utilizan únicamente restricciones basadas en *POS* y se añade una única semilla de opinión en cada iteración.

En resumen, a mayor número de candidatos aceptados o rechazados para refinar, menor cobertura de *feature words* y menor tiempo de ejecución. En

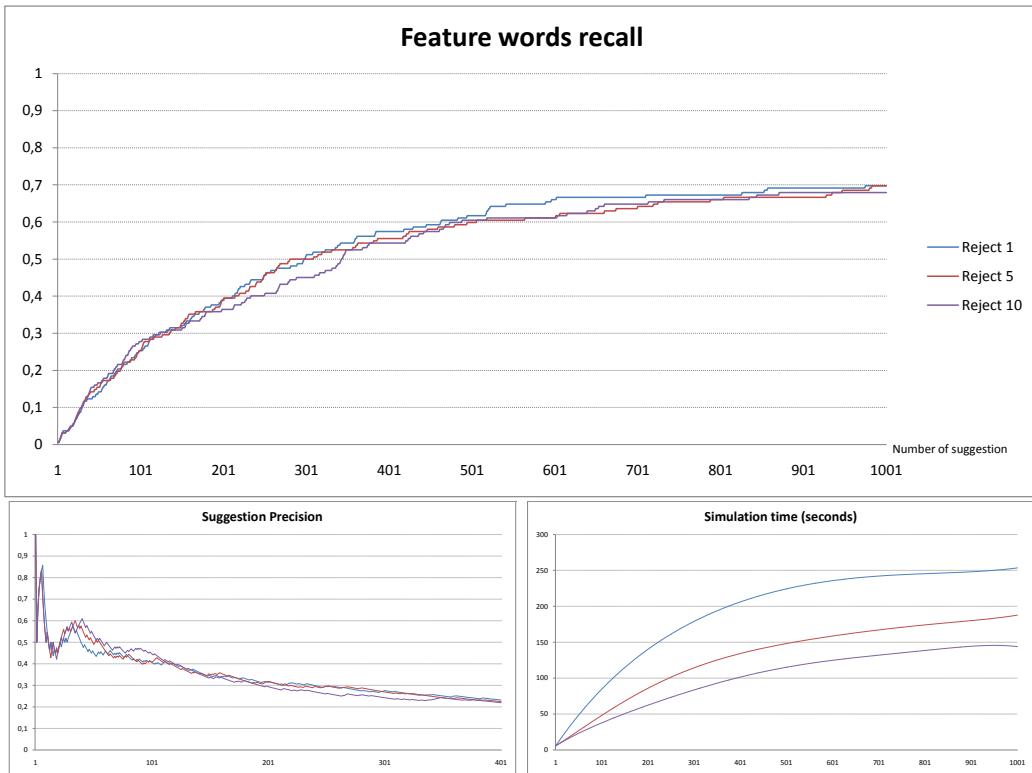


Figura 6.5: Simulación del proceso interactivo de extracción de *feature words*: influencia del número de candidatos rechazados antes de iterar.

el caso de utilizar el criterio de aceptación, las diferencias en el tiempo y especialmente en la cobertura son mayores que en el caso de utilizar el criterio de rechazo. Esto puede ser debido a que al utilizar el criterio de rechazo no se refina la lista de candidatos mientras no se haya aceptado algún candidato desde la última iteración (ya que mientras no haya candidatos aceptados nuevos, no se pueden obtener nuevas semillas de opinión); de forma que a medida que avanza la ejecución y decrece la precisión de las sugerencias, la aparición de candidatos aceptados se distancia y el sistema se comporta de forma similar a si utilizara el criterio de aceptación para refinar. Por otro lado, tanto para el criterio de aceptación como para el de rechazo, los valores mayores del parámetro conducen a una mejor precisión de las sugerencias en el periodo inicial de la simulación.

En el caso del criterio de aceptación, un valor adecuado para el parámetro sería 5, pues a pesar de conllevar un crecimiento más lento de cobertura de *feature words* a lo largo de la ejecución, con respecto al refinamiento por aceptación de un solo candidato, la cobertura final tiende al mismo valor, y

se reduce significativamente el tiempo de ejecución. En cuanto al criterio de rechazo, las menores diferencias para la cobertura de *feature words* y la precisión de la sugerencias entre los distintos valores nos hacen decantarnos por el valor más alto de los experimentados (10), lo que permite reducir el tiempo de ejecución y obtener una precisión de las sugerencias mayor en el periodo inicial de la ejecución de la herramienta. De todas formas, dependiendo del tamaño del corpus que estemos utilizando y de las especificaciones de la máquina donde ejecutemos la herramienta, podemos escoger valor mayores o menores de los parámetros para conseguir tiempos de ejecución razonables.

En la figura 6.6 se comparan las simulaciones correspondientes a refinar tras 5 candidatos aceptados y refinar tras 10 candidatos rechazados. Hemos incluido además una simulación en la que sólo se refina cuando no quedan candidatos que validar en la lista actual, que en cierto modo sirve como cota de los valores anteriores (puesto que es la situación a la que tienden los experimentos anteriores a medida que incrementamos los valores de los parámetros). En nuestro caso, preferimos decantarnos por el criterio de rechazo, ya que las métricas de cobertura y precisión tienen un mejor comportamiento en este caso, a cambio de una pequeña penalización en tiempo.

6.3.5. Influencia del número de nuevas semillas de opinión por iteración

El número de nuevas semillas de opinión añadidas en cada iteración es un parámetro que también puede influir en el comportamiento de la herramienta. Hemos experimentado con distintos valores del parámetro, y a modo de resumen mostramos el comportamiento de las métricas para tres valores (1, 3 y 6). En todos los casos, hemos utilizado únicamente restricciones basadas en *POS* y hemos refinado la lista de candidatos tras ser rechazados 10 candidatos. Añadir una única semilla de opinión por iteración permite un mejor tiempo de ejecución y una mayor precisión de las sugerencias en el periodo inicial. Aunque valores mayores del parámetro consiguen un crecimiento mayor de la cobertura, los valores que finalmente se alcanzan para la misma son iguales, y los tiempos de ejecución crecen sensiblemente. Proponemos por tanto usar valores bajos para este parámetro: una única semilla nueva por iteración sin el corpus es voluminoso o disponemos de una máquina lenta, o 3 semillas nuevas por iteración si la velocidad de ejecución no es un problema.

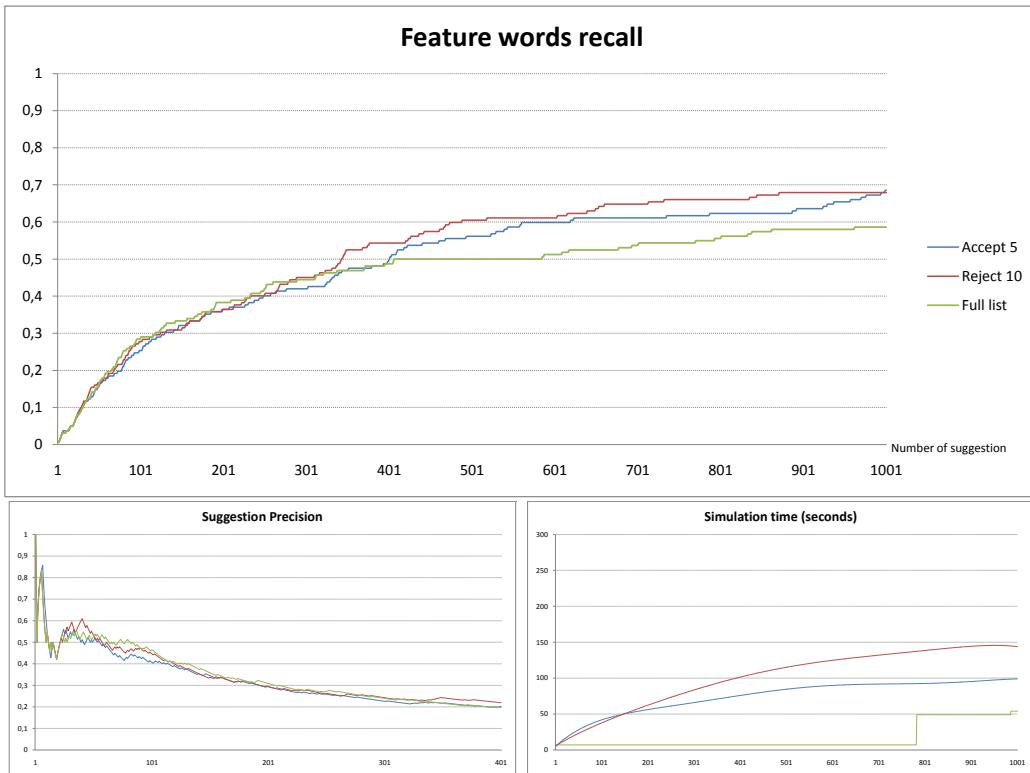


Figura 6.6: Simulación del proceso interactivo de extracción de *feature words*: comparativa entre iterar al aceptar, al rechazar o al vaciar lista de candidatos.

6.3.6. Resultados de la extracción de *feature words* para el dominio *headphones*

En la tabla 6.2 se muestran algunos datos de la ejecución de la herramienta para el dominio *headphones*: número de sugerencias realizadas, número de candidatos aceptados y rechazados, porcentaje de *feature words* de la taxonomía que han sido descubiertos y porcentaje de características de la taxonomía que quedan representadas por al menos uno de los candidatos aceptados. Esta medida es si cabe más importante que la propia cobertura de palabras de característica, puesto que la existencia de al menos un representante de cada característica es suficiente para construir una buena taxonomía; las *feature words* que falten se irán añadiendo posteriormente en el proceso de validación, a medida que vayan siendo anotadas. Aunque la herramienta permite al usuario decidir cuándo se termina la ejecución, en nuestros experimentos hemos detenido la misma en el momento en que la precisión de las sugerencias bajó hasta el 20 %.

El proceso demuestra ser efectivo, consiguiéndose descubrir al menos

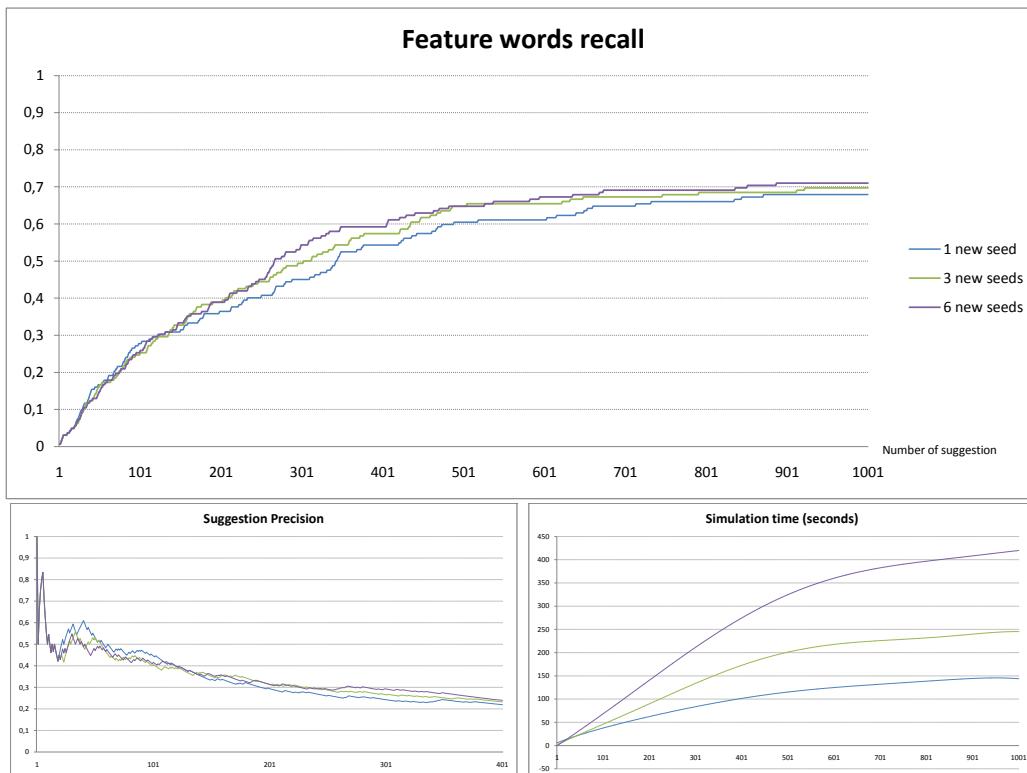


Figura 6.7: Simulación del proceso interactivo de extracción de *feature words*: influencia del número de nuevas semillas de opinión en cada iteración.

una *feature word* para más del 91 % de las características de la taxonomía-objetivo; y esto se consigue revisando tan sólo 485 candidatos (una cantidad considerablemente menor que los 7315 de los que partimos para elaborar la taxonomía-objetivo). Queda por medir la cobertura con respecto a la versión final de la taxonomía, tras el proceso de refinado que se llevará a cabo tras la anotación de evidencias de opinión en el corpus (ver sección 6.6); como se verá, al deshacernos de las características con muy poca representación en las anotaciones, la cobertura sobre *features* de la simulación anterior mejorará al ser medida frente a dicha versión refinada de la taxonomía (de hecho, en el caso de aplicación que nos ocupa, la cobertura fue del 100 % de las características de la taxonomía refinada final).

En la tabla 6.3 mostramos además la lista de *feature words* obtenida.

Nº de sugerencias	485
Nº de candidatos aceptados	97
Nº de candidatos rechazados	388
% de cobertura de <i>feature words</i>	59,87 %
% de cobertura de <i>features</i>	91,42 %

Cuadro 6.2: Resumen de la ejecución de la herramienta de extracción de palabras de característica para el dominio *headphones*

audio, audio quality, bag, bass, bass reproduction, bass response, battery, battery life, bud, buds, build, cable, cans, case, clarity, color, colors, comfort, connection, construction, cord, cost, cups, cushions, design, device, durability, ear bud, ear buds, ear cup, ear cups, ear pads, ear pieces, earbud, earbuds, earpads, earphone, earphones, earpieces, feel, fit, foam, frequency range, frequency response, head phones, headband, headphone, headphones, headset, headsets, high-end, highs, isolation, length, look, looks, lows, mids, music, music quality, noise cancellation, noise isolation, noise reduction, output, package, packaging, pads, performance, phone, phones, plug, portability, pouch, price, product, range, reduction, reproduction, response, sensitivity, shape, size, sound, sound isolation, sound quality, sound reproduction, sounds, speakers, style, transmitter, treble, value, volume, volume control, warranty, weight, wire

Cuadro 6.3: Lista de *feature words* obtenidas para el dominio *headphones*

6.4. Construcción de la taxonomía de características previa

A partir de la lista de *feature words* obtenida, se construye la taxonomía de características manualmente, agrupando aquellas *feature words* que hagan referencia a una misma característica. Una vez obtenido el conjunto de *features*, estos deben ser organizados en alguna jerarquía, de la que el propio objeto abstracto será el nodo raíz.

Por supuesto, no existe una única taxonomía correcta. Recordemos que el principal papel de la misma es indicar en qué características del objeto estamos interesados, puesto que será sobre dichas características sobre las que trabajará el sistema extractor. Por tanto, a pesar de que una determinada característica se vea reflejada en la lista de *feature words* obtenida en el paso anterior, podríamos decidir no incluirla en la taxonomía si no estamos interesados en extraer opiniones acerca de la misma.

No nos ha parecido apropiado investigar métodos automáticos para la generación de la jerarquía de la taxonomía, puesto que el esfuerzo manual que se requiere para su elaboración es pequeño, y una taxonomía bien construida es clave para el buen funcionamiento del sistema.

6.4.1. Taxonomía de características previa para *headphones*

Tras agrupar los elementos de la lista de *feature words* generada mediante la herramienta, obtuvimos el resultado mostrado en la tabla 6.4. Para cada uno de los grupos, escogemos un nombre descriptivo de la característica. Generalmente, este nombre será uno de los términos del grupo, aunque en ocasiones puede resultar más natural encontrar un nombre más descriptivo. Esto pasa por ejemplo en la características *appearance*. Se agrupan en ella *feature words* que en principio podrían haber constituido varias características: no es exactamente lo mismo el “color” que la “forma” o el “estilo”. Pero en esta ocasión hemos considerado que no es necesario que el sistema contemple ese nivel de detalle; hemos preferido agrupar todas estas características en una sola. De ahí que sea más adecuado utilizar un nombre que englobe a todas las *feature words*.

La creación de los grupos que dan lugar a las características es un proceso para el que no existe una única solución correcta. Más bien, son los encargados del sistema los que deben decidir qué nivel de detalle consideran necesario, siempre pensando en la salida deseada del sistema de extracción. Esto mismo ocurre en la creación de la jerarquía de la taxonomía.

<i>Feature</i>	<i>Feature words</i>
appearance	color, colors, look, looks, shape, style
bass	bass, bass reproduction, bass response, lows
battery	battery
battery life	battery life
case	bag, case, pouch
comfort	comfort, feel, fit
cord	cable,cord, length, wire
design	design
durability	build, construction, durability
earbuds	bud, buds, ear bud, ear buds, ear pieces, earbud, earbuds, earpieces
earcups	cans, cups, ear cup, ear cups
earpads	cushions, ear pads, foam, pads
frequency response	frequency range, frequency response, response
headband	headband
headphones	device, earphone, earphones, head phones, headphone, headphones, headset, headsets, phone, phones, product, speakers
highs	clarity, high-end, highs, treble
isolation	isolation, noise cancellation, noise isolation, noise reduction, reduction, sound isolation
jack	connection, plug
mids	mids
packaging	package, packaging
performance	performance
portability	portability
price	cost, price, value
range	range
sensitivity	sensitivity
size	size
sound quality	audio, audio quality, music, music quality, reproduction, sound, sound quality, sound reproduction, sounds
transmitter	transmitter
volume	output, volume
volume control	volume control
warranty	warranty
weight	weight

Cuadro 6.4: Características obtenidas por agrupación de las *feature words* para *headphones*

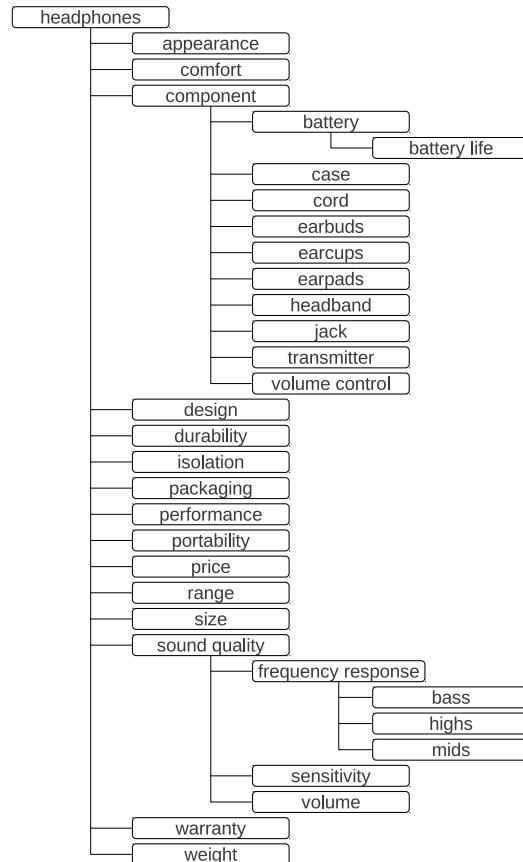


Figura 6.8: Taxonomía de características previa para *headphones*

Las características obtenidas fueron ordenadas jerárquicamente para obtener la taxonomía previa que se muestra en la figura 6.8. El propio objeto es la característica que debe aparecer como raíz de la taxonomía. Aparece una nueva característica *component*, cuya utilidad es agrupar bajo una misma categoría a todas las características que hacen referencia a partes físicas del objeto. Este tipo de características aglutinadoras no tienen *feature words* asociadas, siendo su finalidad únicamente proporcionar una jerarquía más legible y ordenada, y permitir la agregación de las opiniones extraídas sobre determinadas características relacionadas. Debemos tratar que las relaciones padre-hijo sean coherentes con la semántica de las características, puesto que en cierto punto del proceso de inducción de recursos (concretamente, en la inducción del lexicón de opiniones), las opiniones anotadas acerca de una característica hija influirán en las estimaciones realizadas para las características padre (ver sección 6.7).

6.5. Anotación y validación de evidencias de opinión

Una vez reunido el conjunto de documentos y construida la taxonomía de características, es necesario anotar todas aquellas evidencias de opinión que se encuentren en los documentos. Es importante que la persona o personas que lleven a cabo las anotaciones tengan completamente claro el tipo de opiniones que pretendemos extraer, ya que son exclusivamente este tipo de opiniones las que deben ser anotadas. Es por ello que hemos confeccionado una guía para anotadores, en la que se describen las características de las opiniones que se desean anotar. También constan en la guía el formato en que se introducirán las anotaciones y algunos ejemplos que ilustran cómo se debe proceder ante algunos casos que pueden ocasionar dudas al anotador.

El proceso de anotación de las opiniones es un paso crítico en la generación de los recursos. Cualquier fallo cometido en el mismo puede verse reflejado en éstos, y consecuentemente afectar al sistema de extracción. Tras algunas experiencias previas de anotación, hemos identificado los fallos que se producen más frecuentemente e implementado una herramienta de validación que utiliza ciertos indicadores para detectar esos posibles fallos e informar al anotador.

El proceso de anotación y validación se lleva a cabo por paquetes de documentos: cuando se completan las anotaciones para un paquete de n documentos, se ejecuta la herramienta de validación sobre dicho paquete. La herramienta informa de posibles errores al anotador; éste debe corregir lo que estime oportuno y volver a ejecutar la validación. El anotador puede informar a la herramienta de la existencia de documentos inapropiados (por ejemplo, documentos que no contienen opinión alguna, o en gran medida ininteligibles debido a la existencia de caracteres incorrectos), que serán descartados del resto del proceso. Una vez que el anotador está satisfecho con el resultado de la validación, debe informar a la herramienta, y el paquete queda registrado como *validado*. El anotador puede continuar anotando el siguiente paquete de documentos.

La herramienta lleva a cabo validaciones relacionadas con la sintaxis, y otras relacionadas con posibles incoherencias entre el contenido de cada uno de los atributos de las anotaciones a validar y otras anotaciones previamente validadas.

6.5.1. Validación de sintaxis

Consiste en comprobar que las anotaciones siguen la sintaxis propuesta, incluyendo al menos los atributos obligatorios (*polarity*, *feature* y *opWords*). Se comprueba que el atributo *polarity* contenga un carácter “+” o “-”, y que las listas de números correspondan con *tokens* de la oración correspondiente.

6.5.2. Validación de las características

La aplicación comprueba que las características utilizadas en las anotaciones se corresponden con algún elemento de la taxonomía. Si no fuese así, informa de ello al anotador. Éste debe decidir si se trata de un error de escritura, en cuyo caso debe corregir la anotación en cuestión. Si no se trata de un error de escritura, y considera que la característica especificada debería formar parte de la taxonomía, puede indicárselo a la herramienta, que se encarga de añadirla a la misma (el anotador debe seleccionar un elemento ya existente de la taxonomía del que dependerá el nuevo elemento). Esta decisión debe ser sopesada, puesto que la taxonomía de características no tiene por qué contener todas las características opinables del dominio, sino sólo aquellas en las que estemos interesados para la aplicación que estemos desarrollando.

Algo similar se lleva a cabo con las palabras de característica: aquellas que hayan sido anotadas y no aparezcan en la taxonomía, pueden corresponderse con un error en la anotación, o bien deben ser añadidas a la taxonomía en la característica indicada por la anotación. En este caso, la decisión puede tomarla directamente el anotador, puesto que es deseable que la taxonomía capture todas las palabras de característica posibles para cada uno de los *features*.

Dado que por definición las palabras de característica constituyen distintas denominaciones para una característica determinada, se exige que aquellos tokens etiquetados como palabras de característica contengan al menos una palabra cuya categoría morfosintáctica sea un nombre. Si se encuentran anotaciones que no cumplen esta condición, se informa de ello al anotador. Se permite que el anotador ignore esta advertencia, pues la no existencia de un nombre puede estar debida a un fallo del etiquetador morfosintáctico utilizado.

6.5.3. Validación de la polaridad

Uno de los errores que hemos observado más se repiten, y que tienen consecuencias graves en recursos como el lexicón de opiniones, es la elección

errónea de la polaridad de las opiniones anotadas. Al tratarse de un campo con dos posibles valores, un fallo de concentración del anotador ocasiona fácilmente errores en este sentido. Pero algunos de estos errores son fáciles de detectar de manera automática, basándonos en paquetes previamente validados.

La herramienta de validación almacena los términos que han sido utilizados en todos los paquetes validados hasta el momento, y la polaridad de dichas palabras inducida a partir de algunas reglas simples (la generación de los términos a partir de las palabras de opinión y la inducción de la polaridad se explicarán en detalle en la sección 6.7). El algoritmo de inducción de la polaridad utiliza las listas de expresiones especiales para saber interpretar correctamente la polaridad de las palabras de opinión participantes en una evidencia de opinión (ver sección 5.8 para una descripción de las mismas).

Cuando la herramienta de validación encuentra un término anotado con una polaridad contraria a la observada anteriormente, informa al anotador de la situación por si constituye un error de etiquetado. Por supuesto, no en todos los casos esta situación se corresponde con una anotación errónea; algunas palabras de opinión llevan implicaciones negativas al ser utilizadas en un contexto determinado, e implicaciones positivas en otras situaciones. Más concretamente, algunas palabras de opinión muestran este comportamiento distinto al ser utilizadas en opiniones sobre características distintas. Por ejemplo, la palabra “*cheap*” utilizada en una opinión sobre la *apariencia* de un producto, suele tener connotaciones negativas. Sin embargo, la misma palabra utilizada para calificar el *precio* de ese producto, constituye una opinión positiva. De todas formas, son pocas las palabras que presentan esta ambigüedad, por lo que el método de validación anterior es útil para detectar errores en la elección de la polaridad de las opiniones anotadas.

El anotador deberá decidir en cada caso sospechoso detectado si se trata de un error en las anotaciones, o si la palabra o palabras de opinión utilizadas presentan ambigüedad. Ante esta última situación, la herramienta almacena el término en una lista de términos ambiguos, que no serán tenidos en cuenta en sucesivas validaciones de la polaridad. Además de las dos situaciones anteriores, también es posible que el algoritmo de inducción de la polaridad de las palabras de opinión no haya obtenido correctamente la polaridad. Esto puede ser debido a la participación de nuevas expresiones especiales aún no catalogadas. En este caso, el anotador debe informar de ello a la herramienta, y ésta le solicitará que introduzca una nueva expresión especial, ya sea de negación, no-negación o polaridad dominante.

6.5.4. Validación de la cobertura

Para intentar minimizar el número de evidencias de opinión omitidas por el anotador, la herramienta de validación informa a éste de la existencia de palabras potencialmente significativas en los documentos a validar, que no hayan sido utilizadas en ninguna anotación. En concreto, se contempla la aparición de:

- *Feature words*: las apariciones de *feature words* de la taxonomía, que no hayan sido utilizadas en ninguna anotación, pueden estar relacionadas con evidencias de opinión pasadas por alto por el anotador.
- *Implicit feature cues*: se buscan apariciones de palabras utilizadas como *opinion words* en anotaciones ya validadas de opiniones cuyos *features* sean implícitos, y que no participen en ninguna anotación en el documento actual. Por ejemplo, la aparición de “*comfortable*” en un *review* sugiere una posible opinión sobre el *feature* implícito “*comfort*”, lo que se deduce de la aparición de “*comfortable*” como *opinion word* en varias anotaciones previamente validadas en las que no aparecen *feature words*.

En ambos casos, la mayoría de las veces la observación de dichas palabras no se corresponde realmente con una evidencia de opinión. Sin embargo, el esfuerzo empleado por el anotador para comprobar cada una de las sugerencias es pequeño, y a cambio nos aseguramos de que se omitirán el mínimo número posible de evidencias de opinión.

En resumen, el proceso de validación de anotaciones permite obtener documentos anotados con más precisión y omitiendo menos evidencias de opinión. Además, obtiene como salida no sólo los documentos con anotaciones validadas, sino también versiones refinadas de la taxonomía de características y las listas de expresiones de negación, no-negación y polaridad dominante, recursos que serán utilizados por el sistema de extracción. En la figura 6.9 se muestra un diagrama de flujo que resume el proceso de validación.

6.5.5. Corpus anotado para *headphones*

La anotación del corpus del dominio *headphones* fue llevada a cabo aplicando la metodología descrita, incluyendo el proceso de validación, tras lo cual quedaron 587 documentos. En la tabla 6.5 se muestran algunos datos

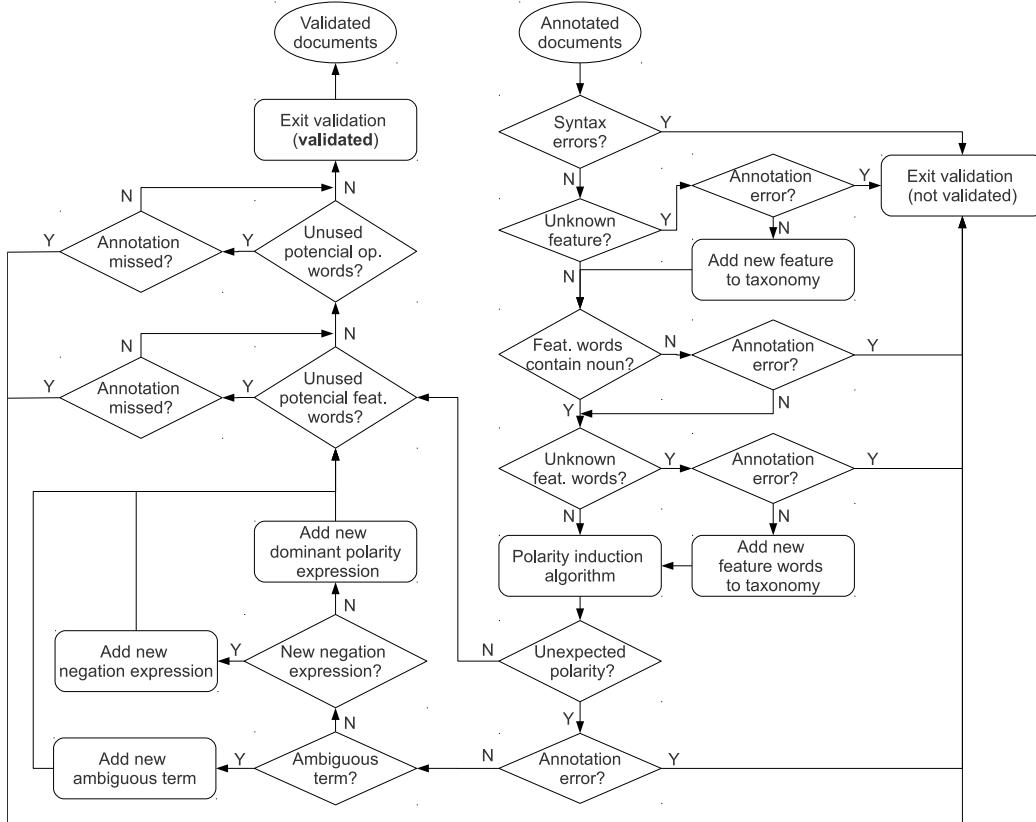


Figura 6.9: Proceso de validación de anotaciones

sobre el corpus anotado obtenido. Estos datos nos permiten conocer la naturaleza del problema; por ejemplo, el número de oraciones que contienen evidencias de opinión representa menos de un tercio del total. Será por tanto importante para el sistema discernir adecuadamente qué oraciones contienen opiniones, y cuáles no. El número de opiniones positivas y negativas anotadas está fuertemente descompensado, a pesar de que el corpus está formado uniformemente por *reviews* con todas las puntuaciones totales. El número de opiniones sobre características implícitas es más elevado de lo que podríamos esperar a priori, representando más de un tercio del total.

6.6. Refinamiento de la taxonomía de características

La herramienta de validación permite al anotador añadir nuevas características y *feature words* al texto. De esta manera, se genera una versión

Reviews	587
Palabras	139331
Oraciones	8151
Oraciones que contienen opiniones	2545
Número de características en la taxonomía	36
Evidencias de opinión	3897
Evidencias de opinión con.....	
... polaridad positiva/negativa	72,16 %/27,84 %
... característica implícita/explícita	36,56 %/63,44 %
... con una/dos/tres o más palabras de opinión	80,59 %/15,99 %/3,42 %
... con característica referenciada por pronombre	2,60 %
... con expresiones de negación	4,64 %
... con expresiones de polaridad dominante	2,50 %

Cuadro 6.5: Estadísticas del corpus anotado del dominio *headphones*

expandida de la taxonomía que incluye los nuevos elementos identificados. Además de los cambios introducidos por el proceso de validación, una vez las anotaciones han terminado, el análisis de las mismas nos sirve también para refinrar la taxonomía. Es posible que algunas de las características incluidas en la taxonomía tengan un reflejo muy pequeño o nulo en las anotaciones. Estas características, que tienen muy baja representación en los documentos, pueden ser eliminadas de la taxonomía, puesto que previsiblemente el sistema final no tendrá que extraer ninguna (o casi ninguna) opinión acerca de las mismas; de camino mejoraremos la precisión del sistema, al descartar *feature words* cuya aparición podría ser incorrectamente interpretada como una opinión.

A lo largo de la metodología de generación de los recursos, encontramos por tanto 3 versiones distintas de la taxonomía:

- Taxonomía previa: la generada a partir de la herramienta de extracción de *feature words* y la posterior agrupación y jerarquización manuales.
- Taxonomía tras anotaciones: la generada como extensión de la anterior, al añadirle las nuevas características y las *feature words* encontradas durante el proceso de anotación.
- Taxonomía final: se crea partiendo de la anterior y eliminando o agrupando aquellas características que hayan aparecido con muy poca frecuencia en las anotaciones.

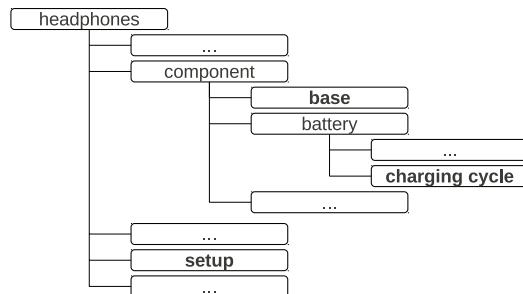


Figura 6.10: Nuevas características añadidas a la taxonomía características tras anotaciones para *headphones* (en negrita)

6.6.1. Refinamiento de la taxonomía de características para *headphones*

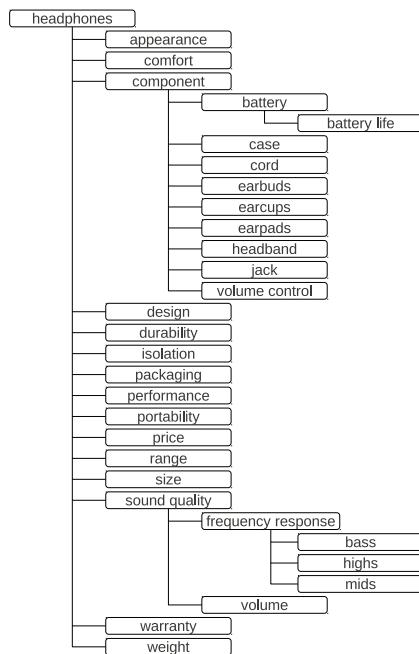
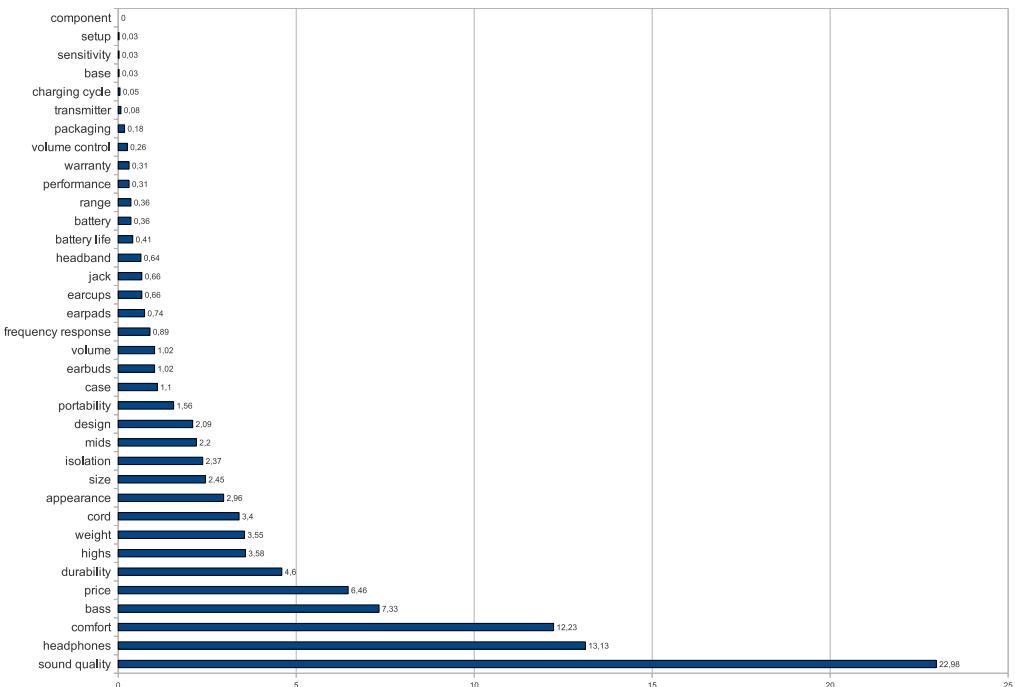
En la figura 6.10 aparecen reflejadas 3 nuevas características que fueron identificadas en las anotaciones. En la tabla 6.6 se muestran las *feature words* de cada característica. Como se observa, el número de *feature words* en la taxonomía crece considerablemente tras el proceso de anotación.

Una vez obtenida la versión tras las anotaciones de la taxonomía de características, contabilizamos el porcentaje de aparición en las anotaciones de cada una de las características (figura 6.12). Decidimos eliminar las características con una representatividad menor al 0,1 % (excluyendo a *component*, que es una característica aglutinadora, por lo que es lógico que no aparezca en las anotaciones). De las 5 características así seleccionadas, 3 se corresponden con las añadidas a la taxonomía previa tras las anotaciones (*base*, *charging cycle* y *setup*). Una vez eliminadas las 5 características poco frecuentes, obtenemos la taxonomía de características final que se muestra en la figura 6.11. Cada característica lleva asociada las *feature words* obtenidas tras las anotaciones (tabla 6.6). Las anotaciones correspondientes a las características excluidas son eliminadas del corpus.

Con respecto a esta taxonomía final, la cobertura sobre características obtenida por la herramienta de extracción de *feature words* es del 100%: la lista incluía al menos una *feature word* correspondiente a cada una de las características de la taxonomía final.

<i>Feature</i>	<i>Feature words</i>
appearance	aesthetics, appearance, color, colors, look, looks, shape, style, styling
comfort	comfort, feel, fit, fitting
component	
base	plug base
bass	bass, bass reproduction, bass response, low, low end, low end frequencies, low frequencies, low range, lows
battery	batteries, battery
battery life	battery life
charging cycle	charging, charging cycle
case	bag, carrying case, case, pouch
cord	audio cord, cable, cables, cord, cords, length, wire, wires, wiring
design	design
durability	build, construction, durability, flexibility, strength
earbuds	bud, buds, ear bud, ear buds, ear clip, ear clips, ear pieces, earbud, earbuds, earpieces
earcups	cans, cups, ear cup, ear cups, earcups
earpads	cushion, cushions, ear pads, earpads, foam, padding, pads, pillows
frequency response	dynamic range, frequency range, frequency response, response
headband	head band, headband
headphones	device, earphone, earphones, head phones, headphone, headphones, headset, headsets, phone, phones, product, set of earbuds, speakers
highs	clarity, high, high end, high frequencies, high range, high-end, highend, highs, treble
isolation	noise blocking, noise cancel, noise cancelation, noise canceling, noise cancellation, noise cancelling, noise-canceling, noise-cancellation, isolation, noise isolation, noise reduction, noise stopping, nr, reduction, shielding, sound isolation
jack	conectors, connection, connections, connector, connectors, input jack, jack, jacks, plug
mids	mid, mid range, mid-range, middle, midrange, mids
packaging	package, packaging
performance	performance
portability	portability
price	cost, price, pricing, value
range	range, wireless reception
setup	setup
size	size
sensitivity	sensitivity
sound quality	audio, audio quality, music, music quality, reproduction, sound, sound quality, sound reproduction, sounding, sounds, sounds quality, sq
transmitter	transmitter
volume	output, sound output, volume
volume control	volume control
warranty	warranty
weight	weigh, weight

Cuadro 6.6: *Feature words* para cada una de las características de la taxonomía tras anotaciones para *headphones*

Figura 6.11: Taxonomía final de características para *headphones*Figura 6.12: Porcentaje de aparición de las características en las anotaciones para *headphones*

6.7. Inducción del lexicón de opiniones

El lexicón de opiniones se genera de manera automática a partir de las anotaciones del corpus. Una vez obtenido, el lexicón es ampliado mediante un método automático, similar al explicado en la sección 2.3.2 y que fue utilizado en (Cruz et al, 2009b) para la generación de lexicones orientados al dominio.

6.7.1. Descripción del algoritmo

Explicamos a continuación el algoritmo de inducción del lexicón de opiniones. El primer paso es la obtención de la lista de términos utilizados como palabras de opinión (algoritmo 1): para cada opinión anotada, la secuencia completa de palabras de opinión participantes, excluyendo las expresiones especiales, conforman un término de la lista. Después, para cada término de esta lista, contamos el número total de apariciones en el corpus, el número de apariciones formando parte de opiniones, el número de apariciones formando parte de una opinión para cada una de las características de la taxonomía, y el número de apariciones formando parte de una opinión positiva o negativa para cada una de las características de la taxonomía. A partir de estos conteos podemos estimar las probabilidades y la polaridad de cada término y almacenarlos con el formato adecuado en el recurso (algoritmo 2). Nótese que, mientras las probabilidades son calculadas a partir de la aparición de los términos como subconjunto de las palabras de opinión, las polaridades se calculan a partir de las apariciones exactas de los términos como palabras de opinión, una vez eliminadas las expresiones especiales.

```

input :  $D'_P$ : a set of documents,  $\text{noNegExprs}$ ,  $\text{negExprs}$ ,  $\text{domExprs}^+$ ,  

        $\text{domExprs}^-$ : sets of special expressions  

output:  $\text{terms}$ : a set of terms

 $\text{terms} \leftarrow \{\}$ ;
 $\text{specialExprs} \leftarrow \text{noNegExprs} \cup \text{negExprs} \cup \text{domExprs}^+ \cup \text{domExprs}^-$ ;
foreach document  $d$  from  $D'_P$  do
    foreach sentence  $s = \{w_1, w_2, \dots, w_n\}$  from  $d$  do
        foreach opinion evidence  $oe = (\dots, \text{opW} \subseteq s)$  in  $s$  do
             $\text{opW}' \leftarrow \text{opW} - \text{specialExprs}$ ;
            if not isEmpty( $\text{opW}'$ ) then
                 $\text{terms} \leftarrow \text{terms} \cup \text{opW}'$ ;
return  $\text{terms}$ ;

```

Algoritmo 1: Extracción de lista de términos de opinión

```

input : terms: a set of terms,  $D'_P$ : a set of documents,  $F_P$ : a feature taxonomy, noNegExprs, negExprs, domExprs+, domExprs-: sets of special expressions
output: lexicon: a set of lexicon entries
 $le_i = (\text{term} \in \text{terms}, \text{support} \in \mathbb{N}^+, \text{opProb} \in [0, 1] \in \Re, \text{fbOpProbs} : F_P \rightarrow [0, 1] \in \Re, \text{fbPolarities} : F_P \rightarrow [-1, 1] \in \Re)$ 

lexicon  $\leftarrow \{\}$ ;
specialExprs  $\leftarrow \text{noNegExprs} \cup \text{negExprs} \cup \text{domExprs}^+ \cup \text{domExprs}^-$ ;
foreach term from terms do
    support  $\leftarrow 0$ ;
    opCount  $\leftarrow 0$ ;
    fbOpProbs  $\leftarrow \emptyset$ ;
    fbPolarities  $\leftarrow \emptyset$ ;
    foreach f from  $F_P$  do
        fbOpCount  $\leftarrow 0$ ;
        fbPolSum  $\leftarrow 0$ ;
        foreach document d from  $D'_P$  do
            foreach sentence s  $= \{w_1, w_2, \dots, w_n\}$  from d do
                support  $\leftarrow \text{support} + \text{countOccurrences}(\text{term}, s)$ ;
                foreach opinion evidence
                    oe  $= (\dots, f' \in F_P, \text{opW} \subseteq s, \dots)$  in s do
                        if term  $\subseteq \text{opW}$  then
                            opCount  $\leftarrow \text{opCount} + 1$ ;
                            if  $f' = f$  or isChild(f', f,  $F_P$ ) then
                                fbOpCount  $\leftarrow \text{fbOpCount} + 1$ ;
                                if term  $= \text{opW} - \text{specialExprs}$  then
                                    polarity  $\leftarrow \text{inducePolarity}(oe, \text{noNegExprs},$ 
                                    negExprs, domExprs+, domExprs-);
                                    fbPolSum  $\leftarrow \text{fbPolSum} + \text{polarity}$ ;
                            
```

fbOpProbs(f) $\leftarrow \text{fbOpCount} / \text{support}$;
 fbPolarities(f) $\leftarrow \text{fbPolSum} / \text{support}$;

opProb $\leftarrow \text{opCount} / \text{support}$;
 lexicon $\leftarrow \text{lexicon} \oplus (\text{term}, \text{support}, \text{opProb}, \text{fbOpProbs}, \text{fbPolarities})$;

return lexicon;
Algoritmo 2: Inducción del diccionario de opiniones

```

input : (polarity  $\in \{-1, 1\}, \dots, opW\}$ ): an opinion evidence, noNegExprs,
        negExprs, domExprs+, domExprs-: sets of special expressions
output: polarity'  $\in \{-1, 0, 1\}$ 

polarity'  $\leftarrow 0$ ;
opW'  $\leftarrow opW - noNegExprs$ ;
oppositePolarity  $\leftarrow$  false;
foreach expression expr from negExprs do
  if expr  $\in opW'$  then
    | opW'  $\leftarrow opW' - expr$ ;
    | oppositePolarity  $\leftarrow$  not oppositePolarity;

foreach expression expr from domExprs+  $\cup$  domExprs- do
  if expr  $\in opW'$  then
    | if expr  $\in domExprs^+$  then
      | | polarity'  $\leftarrow 1$ ;
    | | else
      | | polarity'  $\leftarrow -1$ ;
    | | break;

  if polarity' = 0 then
    | polarity'  $\leftarrow$  polarity

  if oppositePolarity then
    | | return -polarity';
  else
    | | return polarity';

```

Función inducePolarity

<i>Feature</i>	<i>Op. Word Prob.</i>	<i>S.O. Polarity</i>
appearance	0,03	-1,0
bass	0,03	-1,0
frequency response	0,23	0,1428
headphones	0,37	-0,2728
mids	0,07	-1,0
sound quality	0,33	-0,2

Cuadro 6.7: Estimaciones para el término *flat* en el lexicón inducido para *headphones* (valor de *support* del término: 30)

6.7.2. Inducción del lexicón de opiniones para *headphones*

Al aplicar el algoritmo de inducción sobre el corpus anotado de *headphones*, obtuvimos un lexicón con 796 términos, de los cuales 19 presentaron algún tipo de ambigüedad en la polaridad.

Un ejemplo de ambigüedad es el mostrado en la tabla 6.7, en la que se muestran las estimaciones de probabilidad de palabra de opinión y polaridad de orientación semántica del término *flat* para distintas características. Como puede observarse, la polaridad del adjetivo *flat* ha recibido un valor estimado igual a -1 para la mayoría de las características en las que ha sido observado, pero obtiene un valor positivo para la categoría *frequency response*, y un valor negativo pero distinto de -1 para *sound quality* y para *headphones*. ¿A qué son debidas estas estimaciones? Si observamos las evidencias de opinión en el corpus en las que participa la palabra en cuestión, encontramos entre otras las siguientes anotaciones (en las siguientes oraciones sólo señalamos las evidencias de opinión en las que participa la palabra *flat*, subrayando las palabras de característica relacionadas e indicando la polaridad etiquetada mediante un signo + o - al final de la oración):

- Small, light, good fit, some noise isolation, *flat frequency response*. (polaridad: +)
- Uncomfortable earcups, mids/lows cold and *flat*, they are guaranteed to break within months. (polaridad: -)
- The morning got the Bose headphones I put on one of my favorite cds (happens to be Paul Oakenfold) and to my disappointment the sound was incredible *flat* and lifeless. (polaridad: -)
- Not very adjustable, Flat/cheap appearance, Dual wires. (polaridad: -)

Al ser aplicado a la respuesta en frecuencias de unos auriculares, *flat* tiene connotaciones positivas (cuanto más “plana” es la respuesta de un dispositivo reproductor de audio, mayor fidelidad en la reproducción). Sin embargo, en las opiniones acerca de la calidad de sonido (*sound quality*), o acerca de los graves, medios y agudos (*bass*, *mids* y *highs*), las connotaciones son negativas (en este caso, “plano” es sinónimo de corriente, sin pegada, que no llama la atención). Igual polaridad encontramos al emplear el término para calificar la apariencia. Todo esto queda reflejado en los valores inducidos para las polaridades de cada una de las características, con la particularidad de que algunas de las características anteriores dependen unas de otras en la jerarquía. En concreto, *bass*, *mids* y *highs* son hijas de *frequency response*, que a su vez depende de *sound quality*. Por ejemplo, dado que para el cálculo de las estimaciones relacionadas con la característica *frequency response* se utilizan las anotaciones de opiniones sobre las características hijas además de sobre la propia característica, la polaridad del término *flat* para dicha característica obtiene un valor de 0,1428. El valor obtenido para la característica raíz *headphones* (-0,27) nos indica que ante una evidencia de opinión como *flat headphones*, existe una cierta ambigüedad en la polaridad de la misma, siendo la polaridad negativa la que más posibilidades de acierto tiene (al menos, de acuerdo a lo representativo que sea el corpus con respecto a la realidad). Otros términos para los que se observa ambigüedad en el lexicón son *cheap*, (*cheap durability* frente a *cheap price*), *huge* (*huge size* frente a *huge sound*), o *tight* (*tight headphones* frente a *tight bass*). La lista completa de términos con ambigüedad es la siguiente:

average, big, boomy, bright, cheap, deep, flat, flawed, high,
large, light, low, quiet, sharp, small, soft, thick, thin, tight

La utilización de la jerarquía establecida entre las características en la taxonomía para la estimación de las probabilidades y las polaridades del lexicón de opiniones consigue generalizar las observaciones, permitiendo posteriormente inducir la polaridad de un término al ser aplicado a una característica para la que no se han observado apariciones en el corpus, siempre y cuando dicha característica se descomponga en otras para las que sí fueron observadas opiniones. Pero al mismo tiempo, es preciso tener un cuidado especial a la hora de construir la jerarquía de características: debemos garantizar que las características relacionadas genealógicamente en la taxonomía son coherentes con la semántica *es-un*.

Existen algunos términos que obtienen valores de polaridad iguales a 0 para determinadas características, como es el caso del término *awful* (ver tabla 6.8). En este caso, se ha asignado el valor 0 a la polaridad del término

<i>Feature</i>	<i>Op. Word Prob.</i>	<i>S.O. Polarity</i>
appearance	0,0909	-1,0
battery	0,0909	-1,0
component	0,0909	-1,0
headphones	0,3636	-1,0
sound quality	0,1818	0,0

Cuadro 6.8: Estimaciones para el término *awful* en el lexicón inducido para *headphones* (valor de *support* del término: 11)

awful para la característica *sound quality*. Esto es debido a que no existe en el corpus anotado ninguna evidencia de opinión en la que dicho término aparezca por sí solo como palabra de opinión. Sí aparece sin embargo en compañía de otras palabras (en concreto, el término *sounds awful* aparece en varias evidencias de opinión). En este tipo de situaciones, el algoritmo de inducción no es capaz de distinguir si a cada una de las palabras participantes en una expresión se les puede atribuir la polaridad de la opinión en la que han sido utilizadas. Por ejemplo, en este caso no habría sido una mala elección atribuir la polaridad negativa de la expresión *sound awful* a la palabra *awful*, pero sí lo habría sido atribuirla a *sound*. Por otro lado, tal como se observa en el valor estimado para la probabilidad de palabra de opinión (0,18), las opiniones en las que se emplean expresiones formadas por varias palabras sí son tenidas en cuenta para el cálculo de las probabilidades de opinión para cada una de las palabras participantes en la expresión.

6.8. Ampliación del lexicón de opiniones

Una vez inducido el lexicón de opiniones, se procede a su ampliación mediante la aplicación de un método basado en el algoritmo PolarityRank(Cruz et al, 2009d) (ver sección 2.3.2 para una descripción del algoritmo). Se trata de aprovechar la información proporcionada por el resto de documentos del corpus (aquellos cuyas opiniones no han sido anotadas). En concreto, y de forma similar a la propuesta de (Hatzivassiloglou & McKeown, 1997), la aparición de dos términos de opinión coordinados mediante una conjunción “*and*” indica que dichos términos poseen probablemente orientaciones semánticas de igual signo. De manera análoga, aquellos términos que aparecen en construcciones conjuntivas conectados mediante la conjunción “*but*” suelen poseer orientaciones semánticas de signo contrario. Basándonos en este principio, construiremos un grafo a partir de todos los documentos del corpus (anotados y sin anotar). A partir de este grafo y los valores de orientación

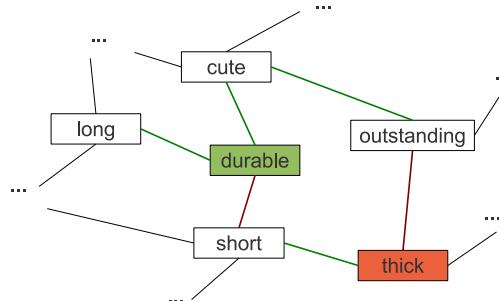


Figura 6.13: Ejemplo de grafo de orientaciones semánticas. Los términos *durable* y *thick*, con orientación semántica positiva y negativa respectivamente en el lexicón, actúan como semillas.

semántica estimados en el lexicón de opiniones, procederemos a inferir valores de orientación semántica para términos no observados en las anotaciones. El proceso es similar al explicado en la sección 2.3.2, con la salvedad de que, en lugar de utilizar como semillas un conjunto cerrado de términos positivos y negativos elegidos manualmente, se utilizarán los términos contenidos en el lexicón de opiniones inducido anteriormente.

6.8.1. Descripción del método de ampliación

El método para ampliar el lexicón de opiniones se inicia con la estimación de la orientación semántica de nuevos términos no aparecidos en el lexicón de opinión inicial. Para ello, partimos de un grafo como el descrito en la sección 2.3.2, en el que los nodos correspondientes a términos existentes en el lexicón serán utilizados como semillas positivas y negativas. La aplicación de PolarityRank nos permitirá obtener valores de *positividad* y *negatividad* de los términos, a partir de los cuales realizar estimaciones de la orientación semántica de los mismos.

A continuación se describe el método paso por paso:

1. A partir de todos los documentos de opinión disponibles para el dominio, se construye un grafo de construcciones conjuntivas de la manera explicada en la sección 2.3.2. Es posible descartar aquellas aristas referentes a términos que hayan sido observados en construcciones conjuntivas un número de veces menor a un umbral.
2. Para cada una de las características incluidas en la taxonomía del dominio, creamos una copia del grafo anterior. En cada uno de estos grafos, buscamos aquellos nodos cuyos términos asociados tengan valores de

orientación semántica no nulos para la característica en cuestión en el lexicón de opiniones a expandir. Dichos valores son asignados a las constantes e^+ o e^- de dichos nodos, en función del signo de la orientación semántica: los valores positivos serán asignados a la constante e^+ de los nodos, y los negativos a e^- . Es posible no tener en cuenta algunos términos del lexicón en este proceso, si el valor de *support* para el término en el lexicón es menor a un umbral.

3. Los valores de e^+ y e^- de los nodos de cada uno de los grafos son normalizados linealmente, para asegurar que la suma de los valores de cada una de las constantes sea igual al número de nodos en cada grafo (tal como se exige en la definición de PolarityRank).
4. Calculamos las puntuaciones PR^+ y PR^- para los nodos de cada uno de los grafos anteriores. A partir de los valores obtenidos en cada uno de los grafos, estimamos la orientación semántica de cada término asociado a cada nodo n mediante la fórmula:

$$SO(n) = \frac{PR^+(n) - PR^-(n)}{PR^+(n) + PR^-(n)} \quad (6.1)$$

Las orientaciones semánticas así calculadas están contenidas en el intervalo $[-1, 1]$.

5. Para cada característica y cada término representado en el grafo asociado, para los que el lexicón de opinión inicial no contenga estimación de la orientación semántica, añadimos la estimación realizada mediante la fórmula anterior.

Para las nuevas entradas del lexicón obtenidas mediante el método anterior, es también necesario obtener estimaciones de probabilidad de palabra de opinión (atributo *feature-based opinion word probability* en el lexicón de opiniones). Para ello, seguimos los mismos pasos anteriores, pero empleando la fórmula original de PageRank, y utilizando los valores de probabilidad de los términos del lexicón para inicializar las constantes e de los nodos en cuestión. La puntuación obtenida por cada nodo es utilizada como estimación de la probabilidad en el lexicón ampliado.

6.8.2. Ampliación del lexicón de opiniones para *headphones*

A la hora de aplicar el método descrito de ampliación al lexicón de opiniones para *headphones*, debemos decidir los valores de los siguientes umbrales:

- *minAbsWeightArcs*: el mínimo número de apariciones en los documentos de una construcción conjuntiva entre dos términos concretos, para que dicha relación se refleje en una arista del grafo.
- *minSupport*: el mínimo número de observaciones a partir de las cuáles fueron estimadas las métricas contenidas en el lexicón de opiniones original, para que dicha información sea utilizada para inicializar los valores de las constantes e^+ y e^- .

A priori, valores menores de estos umbrales permitirán una ampliación mayor del lexicón de opiniones, al permitir generar grafos con más aristas y más nodos actuando de semilla. Por otro lado, valores más altos de los umbrales quizás permitan filtrar relaciones conjuntivas erróneas o poco frecuentes, así como nodos semilla para los cuales las estimaciones contenidas en el lexicón no sean demasiado fiables.

Estudiemos la influencia de estos parámetros. Para ello, una vez escogidos unos valores concretos para los mismos, llevamos a cabo el siguiente desarrollo experimental:

1. Construimos un grafo mediante el método descrito anteriormente, a partir del conjunto completo de documentos de opinión disponibles para *headphones*.
2. Dividimos aleatoriamente el conjunto de documentos anotados en 10 partes iguales. Reservamos una de las partes como conjunto de *test*, y con el resto de partes inducimos un lexicón de opiniones.
3. Ampliamos el lexicón obtenido mediante el método descrito anteriormente.
4. Inducimos un lexicón de opiniones a partir del conjunto de *test*.
5. Obtenemos las siguientes métricas, a partir de ambos lexicones:

- *Cobertura*: porcentaje de entradas⁷ del lexicón de *test* contenidas en el lexicón ampliado, de entre el total de entradas contenidas en este último.

⁷Una entrada del lexicón está constituida por un par término/característica. Por ejemplo, si el lexicón contiene valores de probabilidad y orientación semántica para 4 características distintas de un término determinado, diremos que contiene 4 entradas para dicho término.

- *Precision*: porcentaje de entradas cuya estimación de polaridad de orientación semántica en el lexicón de opinión ampliado coincide con la estimación de polaridad en el lexicón de test, de entre el total de entradas contenidas en ambos lexicones.
 - $F_{\frac{1}{2}}$: media armónica con $\beta = \frac{1}{2}$ de los valores de cobertura y precisión anteriores.
 - *Error medio de orientación semántica*: media de los valores absolutos de las diferencias entre los valores estimados para las orientaciones semánticas en el lexicón ampliado y el lexicón de test, para aquellas entradas contenidas en ambos lexicones.
 - *Error medio de probabilidad*: media de los valores absolutos de las diferencias entre los valores estimados para las probabilidades en el lexicón ampliado y el lexicón de test, para aquellas entradas contenidas en ambos lexicones.
6. Repetimos los pasos anteriores, pero escogiendo sucesivamente distintas partes del conjunto de documentos como conjunto de *test*.
 7. Calculamos los valores medios de las métricas obtenidas para cada una de las diez iteraciones del proceso.

Aplicando el desarrollo experimental descrito, estudiamos primero la influencia del umbral *minAbsWeightArcs*. Se obtienen los resultados mostrados en la figura 6.14, para valores de 1 a 5 para el parámetro, y manteniendo en todos los casos fijo el valor del umbral *minSupport*. Como era de esperar, la precisión tiende a aumentar a medida que aumenta el valor del umbral, a costa de una disminución progresiva de la cobertura. Sin embargo, tal y como refleja la media armónica, el pequeño aumento de precisión que se consigue a medida que se aumenta el valor del umbral no compensa la disminución en la cobertura. Por otro lado, el error medio en las estimaciones de orientación semántica y de probabilidad decrece a medida que aumenta el valor del umbral, de manera especialmente pronunciada al pasar del valor 1 al 2. Es por esto que, a priori, nos inclinamos por fijar el valor del umbral como *minAbsWeightArcs=2*.

En la figura 6.15 se muestra el resultado de la ejecución del mismo desarrollo experimental, variando ahora los valores del umbral *minSupport*, y manteniendo fijo el valor de *minAbsWeightArcs*. En este caso, los mejores resultados se obtienen claramente cuando se utilizan todos los nodos semilla (*minSupport=1*), observándose pocas variaciones en la precisión, cobertura y error medio de probabilidad, y cierta tendencia a crecer el error medio de

	lexicón original		lexicón ampliado	
	términos	entradas	términos	entradas
$minAbsWeightArcs=1$	796	2492	831 (+4,4 %)	5371 (+115,53 %)
$minAbsWeightArcs=2$	796	2492	809 (+1,63 %)	3544 (+42,22 %)

Cuadro 6.9: Resultados de la ampliación del lexicón de opiniones para *headphones*. En ambos casos, se utilizó $minSupport=1$.

orientación semántica, a medida que crece el valor del umbral. Por tanto, fijaremos el valor del umbral como $minSupport=1$.

Aplicando el método de ampliación con los parámetros $minAbsWeightArcs=2$ y $minSupport=1$ al lexicón de opiniones completo para *headphones*, obtenemos una ampliación del 42,22 % en relación al número de entradas en el lexicón, y del 1,63 % en relación al número de términos (ver tabla 6.9). Los porcentajes son sensiblemente mayores en el caso de utilizar el parámetro $minAbsWeightArcs=1$.

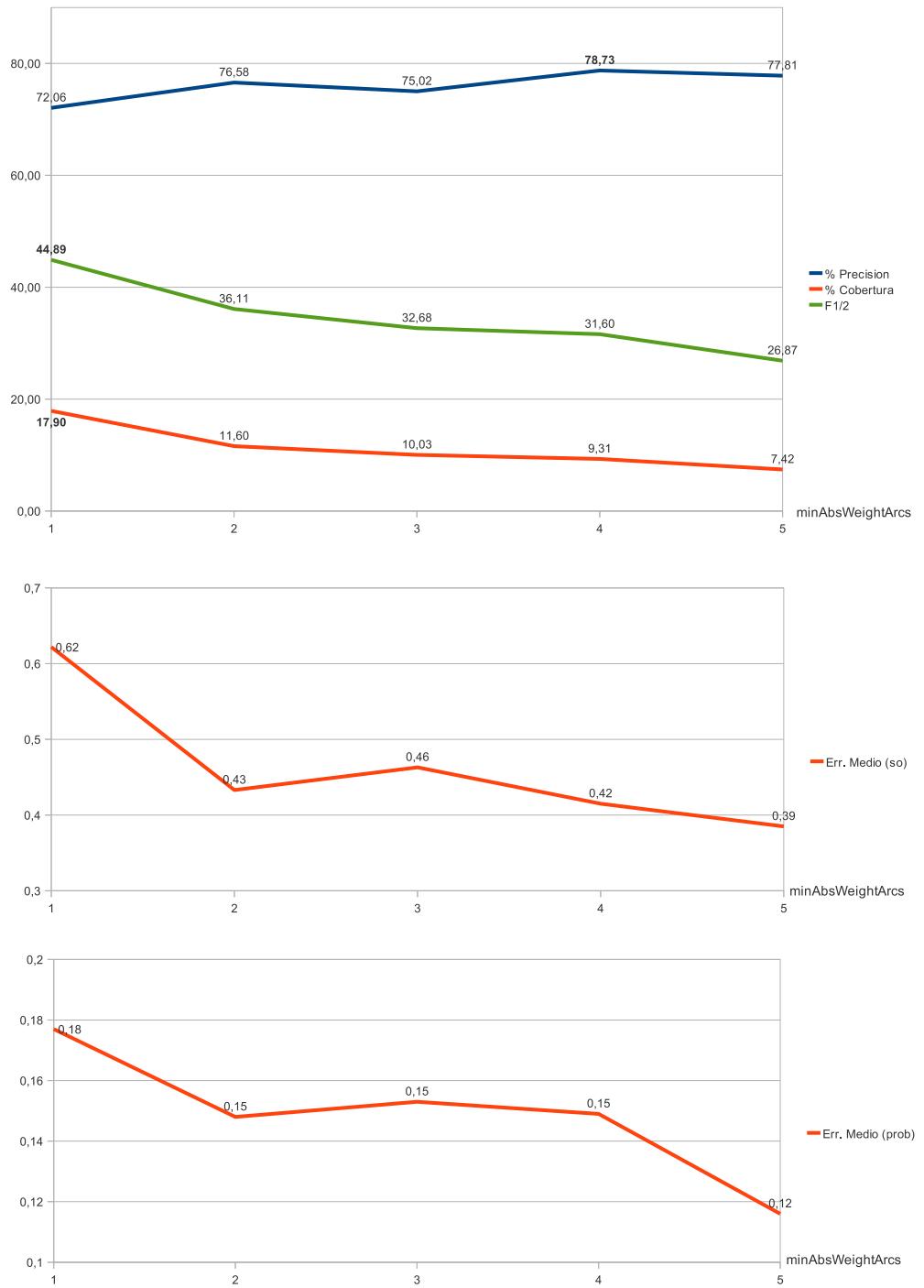


Figura 6.14: Resultados experimentos de ampliación del lexicón de opiniones para *headphones*: influencia del parámetro *minAbsWeightArcs*. El parámetro *minSupport* se mantiene con valor constante igual a 1.

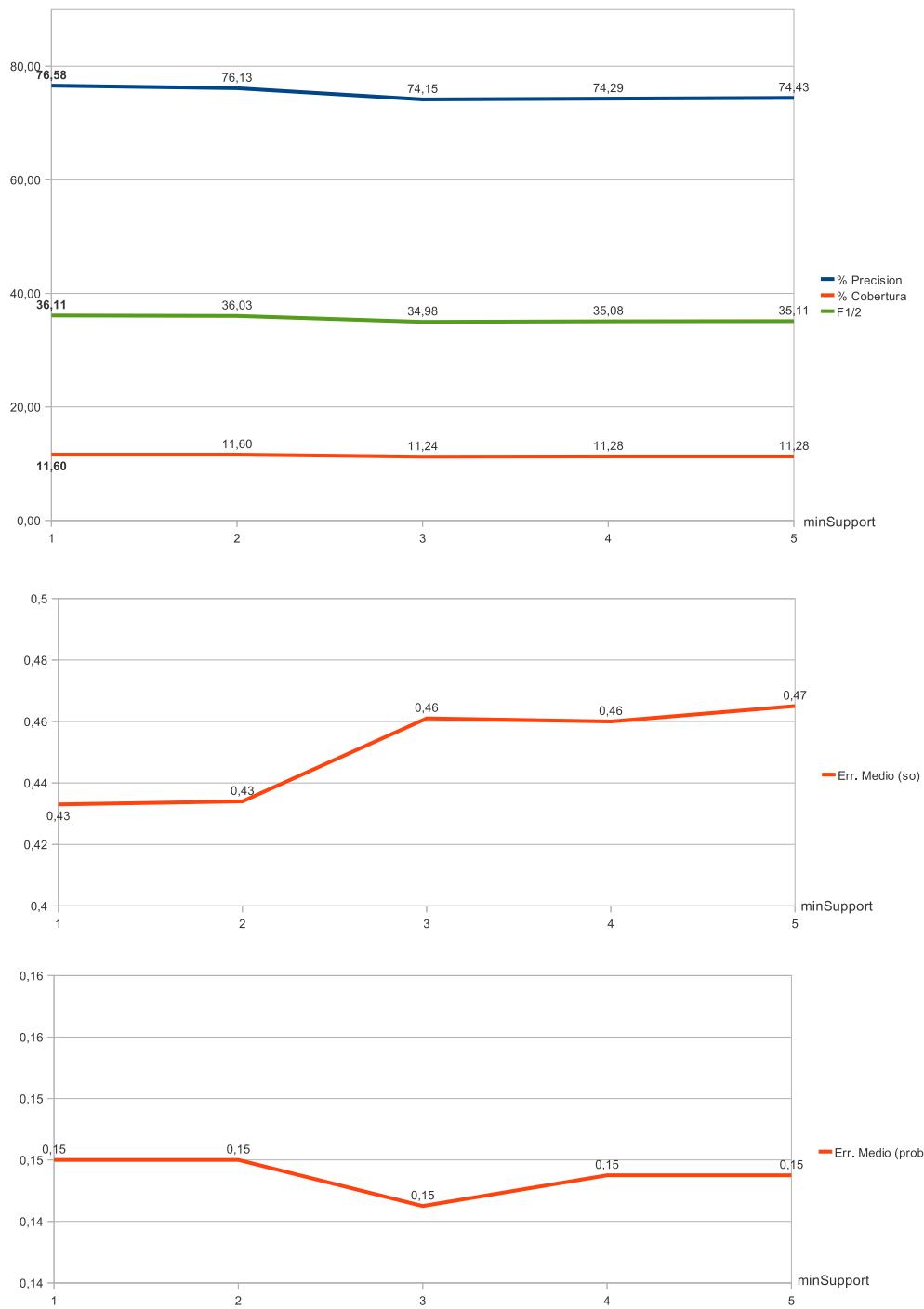


Figura 6.15: Resultados experimentos de ampliación del lexicón de opiniones para *headphones*: influencia del parámetro *minSupport*. El parámetro *minAbsWeightArcs* se mantiene con valor constante igual a 2.

6.9. Inducción de los indicadores de opiniones implícitas

El recurso se genera de manera automática a partir de las anotaciones del corpus. Describimos a continuación el algoritmo de inducción y comentamos algunos de los resultados obtenidos al aplicar dicho algoritmo al dominio *headphones*.

6.9.1. Descripción del algoritmo

De forma similar al proceso de generación del lexicón de opiniones, el primer paso es la obtención de la lista de términos utilizados como palabras de opinión, aunque en este caso se incluyen como términos cada una de las palabras sueltas que participan en opiniones con varias palabras de opinión (ver algoritmo 4). Después, para cada término de esta lista, contamos el número total de apariciones en el corpus, el número de apariciones formando parte de opiniones explícitas para cada una de las características de la taxonomía, y el número de apariciones formando parte de opiniones implícitas para cada característica. A partir de estos conteos podemos estimar las probabilidades, tal como se muestra en el algoritmo 5.

```

input :  $D'_P$ : a set of documents,  $\text{noNegExprs}$ ,  $\text{negExprs}$ ,  $\text{domExprs}^+$ ,  

        $\text{domExprs}^-$ : sets of special expressions
output: terms: a set of terms

terms  $\leftarrow \{\}$ ;
specialExprs  $\leftarrow \text{noNegExprs} \cup \text{negExprs} \cup \text{domExprs}^+ \cup \text{domExprs}^-$ ;
foreach document  $d$  from  $D'_P$  do
    foreach sentence  $s = \{w_1, w_2, \dots, w_n\}$  from  $d$  do
        foreach opinion evidence  $oe = (\dots, \text{opW} \subseteq s)$  in  $s$  do
            opW'  $\leftarrow \text{opW} - \text{specialExprs}$ ;
            if not isEmpty(opW') then
                terms  $\leftarrow \text{terms} \cup \text{opW}'$ ;
                foreach word  $w$  from opW' do
                    terms  $\leftarrow \text{terms} + w$ ;
```

return terms;

Algoritmo 4: Extracción de lista de términos para la inducción de los indicadores de opiniones implícitas

```

input : terms: a set of terms,  $D'_P$ : a set of documents,  $F_P$ : a feature taxonomy
output: cues: a set of cue entries
 $cue_i = (\text{term} \in \text{terms}, \text{support} \in \mathbb{N}^+, \text{fbImplProbs} : F_P \rightarrow [0, 1] \in \mathbb{R}, \text{fbCondImplProbs} : F_P \rightarrow [0, 1] \in \mathbb{R})$ 

cues  $\leftarrow \{\}$ ;
foreach term from terms do
    support  $\leftarrow 0$ ;
    fbImplProbs  $\leftarrow \emptyset$ ;
    fbCondImplProbs  $\leftarrow \emptyset$ ;
    foreach f from  $F_P$  do
        fbExplOpCount  $\leftarrow 0$ ;
        fbImplOpCount  $\leftarrow 0$ ;
        foreach document d from  $D'_P$  do
            foreach sentence s =  $\{w_1, w_2, \dots, w_n\}$  from d do
                support  $\leftarrow$  support + countOccurrences(term, s);
                foreach opinion evidence
                    ( $\dots, f' \in F_P, \text{featW} \subset s, \text{opW} \subseteq s, \dots$ ) in s do
                        if term  $\subseteq \text{opW}$  and f' = f then
                            if featW =  $\emptyset$  then
                                | fbImplOpCount  $\leftarrow$  fbImplOpCount + 1;
                            else
                                | fbExplOpCount  $\leftarrow$  fbExplOpCount + 1;

        fbImplProbs (f)  $\leftarrow$  fbImplOpCount / support;
        if support = fbExplOpCount then
            | fbCondImplProbs (f)  $\leftarrow 0$  ;
        else
            | fbCondImplProbs (f)  $\leftarrow$  fbImplOpCount / ( support -
                fbExplOpCount ) ;

cues  $\leftarrow$  cues  $\oplus$  (term, support, fbImplProbs, fbCondImplProbs);
return cues;

```

Algoritmo 5: Inducción de los indicadores de opinión implícita

Término	<i>Support</i>	<i>Feature</i>	<i>Prob.</i>	<i>Cond. Prob.</i>
sounds	113	sound quality	0,1327	0,1351
sounds weird	2	sound quality	1,0	1,0
sounds great	9	sound quality	0,3333	0,4285
sounds bad	3	sound quality	0,3333	0,3333
flashy	3	appearance	0,6666	1,0

Cuadro 6.10: Algunas entradas del recurso de indicadores de opinión implícita para el dominio *headphones*

6.9.2. Inducción de los indicadores de opinión implícita para *headphones*

Al aplicar el algoritmo anterior al corpus de *reviews* de *headphones*, obtuvimos 538 términos. Recordemos que estos términos se corresponden con palabras anotadas en evidencias de opinión implícitas, tanto juntas como separadas. Así por ejemplo, ante una anotación en la que aparezcan las palabras *sounds good* como palabras de opinión, tanto el término *sounds good* como los términos *sounds* y *good* aparecerán reflejados en el recurso, con sus correspondientes estimaciones de probabilidades. La inclusión en el recurso de cada una de las palabras por separado, a diferencia del lexicón de opiniones, nos permitirá detectar qué componentes de las expresiones utilizadas están correlacionadas con la aparición de opiniones implícitas, de manera que el sistema de extracción pueda determinar la existencia de opiniones implícitas ante la aparición de dichas palabras, aún cuando no aparezca literalmente la expresión que ha sido observada en el corpus anotado. Siguiendo con el ejemplo anterior, observamos en nuestro corpus múltiples construcciones similares (*sounds warms*, *sounds terrible*, *sounds weird*, ...), a partir de las cuales el algoritmo de inducción genera una entrada para el término *sounds* (ver tabla 6.10). Dado que dicha palabra ha sido también observada en diversas frases en las que no formaba parte de anotación alguna, las probabilidades estimadas son relativamente bajas. Los valores obtenidos por las expresiones completas son más altos, por lo que su observación en un nuevo documento permitirá al sistema decidir la existencia de una opinión implícita con mayor seguridad.

Algunos términos obtienen valores sensiblemente distintos para las probabilidades no condicionada y condicionada, como ocurre con el término *flashy* (ver tabla 6.10). Recordemos que el valor que aparece en la columna *Prob.* corresponde a la estimación de la probabilidad de que el término en cuestión implique la existencia de una opinión implícita sobre el *feature* indicado; por otro lado, la columna *Cond. Prob.* indica la probabilidad de un suceso idénti-

co al anterior, pero condicionado a la no participación del término en una opinión explícita. Algunos términos como *flashy* son empleados en ocasiones en opiniones explícitas, además de en implícitas, y es por ellos que obtienen un valor más alto en la probabilidad condicionada. Por ejemplo, en nuestro corpus aparecen entre otras las siguientes anotaciones en las que participa el término *flashy* (en las siguientes oraciones sólo señalamos las evidencias de opinión en las que participa la palabra *flashy*, subrayando las palabras de característica relacionadas e indicando la característica sobre la que actúa la palabra de opinión al final de la oración):

- Mediocre sound, *flashy colors*. (*explicit feature*: appearance)
- A *flashy*, unique pair of headphones. (*implicit feature*: appearance)

Por tanto, la aparición de *flashy* en un nuevo documento podría indicar al sistema la existencia de una opinión implícita, tanto más si previamente hemos descartado su participación en una opinión explícita.

Existen algunos términos para los que se han observado correlaciones con varias características (ver tabla 6.11). En algunos casos se trata de términos que pueden asociarse a opiniones de características distintas pero con cierto parecido (por ejemplo, *portability* y *size*). Esta similitud entre las características ocasiona disparidad en las anotaciones. No se trata de un problema grave, ya que el sistema podría decidirse por cualquiera de las características involucradas, y las opiniones extraídas serían correctas. En otros casos, la existencia de varias características correlacionadas con el término indica la participación del mismo en varias expresiones distintas. Por ejemplo, la palabra *awful* puede aparecer en la expresión *looks awful*, relacionada con la característica *appearance*, o en la expresión *sounds awful*, relacionada con la característica *sound quality*. Dado que ambas expresiones forman parte

Término	<i>Support</i>	<i>Feature</i>	<i>Prob.</i>	<i>Cond. Prob.</i>
compact	46	portability	0,2608	0,2608
		size	0,2391	0,25
awful	11	sound quality	0,1818	0,1818
		appearance	0,0909	0,0909
cheap	133	comfort	0,015	0,015
		price	0,2781	0,2868
		appearance	0,03	0,0303
		durability	0,015	0,016

Cuadro 6.11: Algunas estimaciones contenidas en el recurso de indicadores de característica implícita inducido para *headphones*

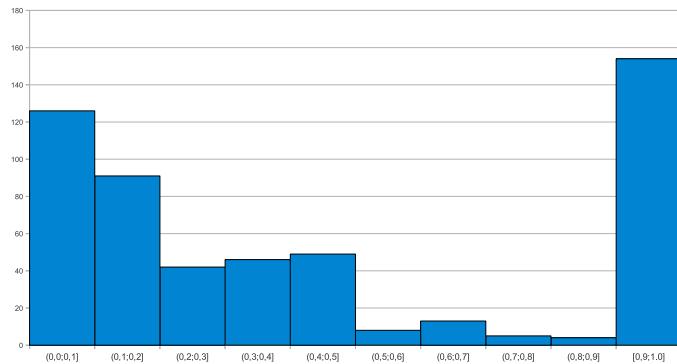


Figura 6.16: Histograma de probabilidades no condicionadas de los indicadores de opinión implícita para *headphones*

también del recurso, esto no ocasionará problema alguno al sistema. Por último, otros términos con varias características relacionadas obtienen valores significativamente mayores para determinada característica, como en el caso de *cheap* para la característica *price*; el sistema debería por tanto elegir esta característica, siempre y cuando no se observe ninguna expresión mayor conteniendo a *cheap* e incluida en el recurso.

En la tabla 6.12 se muestran los términos del recurso obtenido para los que la probabilidad no condicionada estimada es igual a 1, junto con la característica correlacionada. En la figura 6.16 se presenta un histograma que refleja el número de términos existentes en el recurso con respecto a los valores de probabilidad obtenidos por los mismos.

Característica	Indicadores de opinión implícita
appearance	look awesome, aesthetic, cheap looking, attractive looking, clunky, colorful, cool-looking, look decent, look funny, look pretty cool, look promising, look slick, look solid, looked sharp, looks fine, nicely styled, overboard, promising, retro looking, styled, translucent
bass	overbassed
comfort	chunky, feels cheap, feels terrible, fits comfortably, unsecure, comfortable, comfortable to use, cumbersome, difficult to wear, easy to wear, feel elegant, feels comfortable, fit nicely, fit ridiculously, fits ok, fits snug, fits well, fitted easily, maneuverable, nicely, sit quite uncomfortably, soft contact, super comfortable, unadjustable, uncomfortable, unobtrusive, well-padded, well-rounded
cord	tangles easy
design	worst-designed
durability	built solid, easily damaged, brittle, built cheap, dont last, durable, durable, endure much, feel substantial, look fragile, most reliable, plasticky, poorly constructed, rips, rips easily, robust quality, ruggedness, well constructed
headphones	overrated, brakes, brightest, efficient, high-quality, must buy, pro quality, recommended, very frustrating
isolation	great solution, least canceling, lowers, lowers noise
jack	easy to connect
portability	easy to carry, collapsible, nice to transport, packs, packs well, semi-portable, transport
price	expensive, over priced, spendy, too expensive, ultra-cheap, well-priced
size	gigantic, small sized, unwieldy
sound quality	clear sounding, nice sounding, sound sweet, sounded muddy, sounds weird, amazing sounding, average sounding, best-sounding, clean sounding, clean-sounding, decent-sounding, deliver beautifully, fantastic sounding, great clarity, hot-sounding, inaccurate sounding, most best-sounding, most natural sounding, outstanding musical detail, outstanding rendition, over-amplified, precise detail, preformed, preformed flawlessly, rendition, sound feeble, sound hot-sounding, sound just acceptable, sound musical, sound thin, sound way over-amplified, sound wonderfull, sounded boomy, sounded horrible, sounded ok, sounded poor, sounded terrible, sounds artificial, sounds terrible, sounds warm, sounds wonderful, thin sounding, well sounding, wonderfull
volume	enough, loud enough, plays, plays loud, super loud
weight	heavy

Cuadro 6.12: Términos del recurso de indicadores de opinión implícita para *headphones* con probabilidad igual a 1

6.10. Inducción de los patrones de dependencias

Aunque podríamos utilizar un conjunto cerrado de patrones de dependencias elaborados a mano (tal como hacen algunos artículos de la bibliografía (Popescu & Etzioni, 2005; Popescu et al, 2005)), hemos preferido aprovechar la disponibilidad de un corpus anotado para inducir automáticamente los patrones de dependencias más utilizados para cada una de las características opinables del dominio. A continuación describimos el algoritmo de inducción, y planteamos algunos experimentos en el dominio *headphones* encaminados a estudiar la influencia del uso de restricciones morfosintácticas en los patrones.

6.10.1. Descripción del algoritmo

El proceso de inducción de los patrones de dependencias está resumido en el algoritmo 6, que genera una lista ordenada de patrones del tipo especificado. Para cada evidencia de opinión anotada en el corpus, la función *getPairs* devuelve las parejas de palabras origen y destino correspondientes al tipo de patrones que se está extrayendo. Para cada una de estas parejas, se busca un camino en el árbol de dependencias que une ambas palabras (función *extractPath*). Si se encuentra dicho camino, se crean dos patrones con el mismo, uno sin restricción de característica, y otro para la característica correspondiente a la evidencia de opinión actual. Los patrones creados se van acumulando en una lista, contabilizando el número de veces que aparece cada uno. Una vez se han extraído todos los patrones, se miden la precisión y la cobertura que se obtiene al aplicar cada uno de los patrones al propio corpus anotado (función *computePrecisionAndRecall*). Estos valores son almacenados en cada patrón, y la lista de patrones se ordena por orden decreciente de precisión, y a igual precisión, por orden decreciente de cobertura (función *sortByPrecisionAndRecall*). Por último, se calculan la precisión y la cobertura acumuladas de cada una de las sublistas iniciales de patrones de la lista ordenada: en el primer patrón de la lista, estos valores coinciden con la precisión y la cobertura anteriormente calculadas; para el segundo patrón se calcula la precisión y cobertura obtenidas de la aplicación de los dos primeros patrones de la lista, y así sucesivamente (función *computeAccumulatedPrecisionAndRecall*).

```

input :  $D'_P$ : a set of documents, type  $\in [1, 5] \subset \aleph$ : type of patterns to be
       extracted, noNegExprs, negExprs, domExprs+, domExprs-: sets of
       special expressions
output: patterns = { $pat_1, pat_2, \dots, pat_n$ } : a sorted list of patterns, where
           $pat_i = (f, pos_{asc}, pos_{desc}, dep_{asc}, dep_{desc}, support, p, r, accP, accR)$ 

patterns  $\leftarrow \{\}$  ;
foreach document d from  $D'_P$  do
  foreach sentence s from d do
    foreach opinion evidence oe = ( $\dots, f \in F_P, \dots$ ) in s do
      pairs  $\leftarrow$  getPairs(oe, type, noNegExprs, negExprs, domExprs+,
                           domExprs-);
      foreach pair of words pair = (wFrom, wTo) from pairs do
        path  $\leftarrow$  extractPath(wFrom, wTo);
        if path is not null then
          if  $\exists (f, path, support) \in$  patterns then
            foreach pattern p = (feature, path, support) from
            patterns do
              patterns  $\leftarrow$  patterns - p;
              patterns  $\leftarrow$  patterns
              +(feature, path, support + 1);
            else
              patterns  $\leftarrow$  patterns +(f, path, 1);
              patterns  $\leftarrow$  patterns +(*, path, 1);

  patterns'  $\leftarrow \{\}$  ;
  foreach pattern p = (f, path, support) from patterns do
    (precision, recall)  $\leftarrow$  computePrecisionAndRecall( $D'_P$ , type, p);
    patterns'  $\leftarrow$  (f, path, support, precision, recall);
  sortByPrecisionAndRecall(patterns');
  return computeAccumulatedPrecisionAndRecall(patterns');

```

Algoritmo 6: Inducción de los patrones de dependencias

```

input : oe =(...,featW:a list of feature words, opW:a list of opinion
words), type: type of patterns to be extracted, noNegExprs,
negExprs, domExprs+, domExprs-: sets of special expressions
output: pairs: a set of pairs of source and target words

pairs  $\leftarrow \emptyset$ ;
switch type do
    Type 1: from any feature word to any opinion word.
    case 1
        foreach words wFrom from featW and wTo from opW do
             $\lfloor$  pairs  $\leftarrow$  pairs + (wFrom,wTo);
    Type 2: from the head of feature words to the head of opinion words.
    case 2
         $\lfloor$  pairs  $\leftarrow$  (getHead(opW),getHead(opW));
    Type 3: from the head of opinion words to any other opinion word.
    case 3
        foreach word wTo from opW do
             $\lfloor$  pairs  $\leftarrow$  pairs + (getHead(opW),wTo);
    Type 4: from the head of opinion words to any other opinion word from
a negation expression.
    case 4
        opW'  $\leftarrow$  opW;
        foreach expression expr from noNegExprs do
             $\lfloor$  opW'  $\leftarrow$  opW' - expr;
            foreach word wTo from opW' do
                 $\lfloor$  if wTo  $\in$  negExprs then
                     $\lfloor$  pairs  $\leftarrow$  pairs + (getHead(opW'),wTo);
    Type 5: from the head of opinion words to any other opinion word from
a dominant polarity expression.
    case 5
        opW'  $\leftarrow$  opW;
        foreach expression expr from negExprs do
             $\lfloor$  opW'  $\leftarrow$  opW' - expr;
            foreach word wTo from opW' do
                 $\lfloor$  if wTo  $\in$  domExprs+  $\cup$  domExprs- then
                     $\lfloor$  pairs  $\leftarrow$  pairs + (getHead(opW'),wTo);

return pairs;

```

Función getPairs

```

input : wFrom: source word, wTo: target word
output: path =( $pos_{asc}, pos_{desc}, dep_{asc}, dep_{desc}$ ): lists of part-of-speech tags
          and dependency relation types for the ascending and descending
          subpaths linking source and target words

ascendingPath ← {wFrom} ;
w ← wFrom;
while dependencyHead(w) is not null do
    w ← dependencyHead(w) ;
    ascendingPath ← ascendingPath + w;
descendingPath ← {wTo};
w ← wTo;
while dependencyHead(w) is not null do
    w ← dependencyHead(w) ;
    descendingPath ← descendingPath + w;
reverse(descendingPath);

foreach word w from ascendingPath do
    if w ∈ descendingPath then
        sublist from the first element to w, inclusive:
        ascendingPath ← headList(ascendingPath,w);
        sublist from w to the last element, inclusive:
        descendingPath ← tailList(descendingPath,w);
         $pos_{asc}$  ← part-of-speech tags of each word from ascendingPath;
         $pos_{desc}$  ← part-of-speech tags of each word from descendingPath;
         $dep_{asc}$  ← dependency relation types of each word from
        ascendingPath;
         $dep_{desc}$  ← dependency relation types of each word from
        descendingPath;
        path ← ( $pos_{asc}, pos_{desc}, dep_{asc}, dep_{desc}$ );
        return path;

return null;

```

Función extractPath

6.10.2. Influencia de las restricciones morfosintácticas en los patrones de dependencias

Para medir la influencia de las restricciones morfosintácticas en la precisión y la cobertura de los patrones, hemos llevado a cabo experimentos a partir del corpus de *headphones*. En cada uno de los experimentos, realizamos la inducción de los patrones utilizando un subconjunto del corpus, para posteriormente aplicar los patrones obtenidos al resto del corpus, midiendo con ello la precisión y la cobertura del conjunto de patrones obtenido. Hemos llevado a cabo validación cruzada con 10 particiones, lo que significa que cada experimento se ejecutó 10 veces sobre 10 particiones aleatorias distintas del corpus. Los resultados finales se obtienen calculando la media de los valores obtenidos en las 10 ejecuciones. Utilizamos siempre un 85 % del corpus para inducir los patrones, y el 15 % restante para medir la precisión y cobertura de los patrones obtenidos.

Hemos llevado a cabo los tres experimentos siguientes:

- Sin restricciones morfosintácticas: los patrones extraídos no incluyen restricciones.
- Con restricciones morfosintácticas simples: los patrones extraídos incluyen restricciones basadas en la categoría morfosintáctica general de los participantes. La categoría general viene codificada en la primera letra de las etiquetas *POS*.
- Con restricciones morfosintácticas completas: los patrones extraídos incluyen restricciones basadas en la categoría morfosintáctica completa de los participantes. Utilizamos la etiqueta *POS* completa.

Los experimentos se realizaron induciendo sólo los patrones de tipo 2, y sin restricciones basadas en características. La tabla 6.13 contiene los resultados obtenidos. Como puede apreciarse, a mayor nivel de restricción, el conjunto de patrones inducidos consigue mayor precisión pero menor cobertura; a pesar de la disminución de cobertura, parece preferible contemplar las restricciones completas, dado que la ganancia en precisión supera ampliamente la bajada en la cobertura. En el capítulo 8 estudiaremos la influencia final en el sistema de extracción de la elección de uno u otro tipo de restricciones.

6.10.3. Patrones de dependencias para *headphones*

En la tabla 6.14 aparecen los patrones más frecuentemente observados del tipo 1 (conectando cualquier *feature word* con cualquier *opinion word*), junto

Experimento	Núm. patrones	Precision	Recall	F_1
Sin restricciones	138,3	0,368	0,834	0,511
Restricciones simples	283,1	0,475	0,773	0,588
Restricciones completas	404,6	0,575	0,726	0,641

Cuadro 6.13: Resultados de los experimentos para medir la influencia de las restricciones morfosintácticas en los patrones de dependencias para *headphones*

a algunos ejemplos extraídos del corpus siguiendo cada uno de los patrones. Para que los ejemplos mostrados sean más generales, los patrones han sido extraídos utilizando restricciones morfosintácticas simples.

Se observan algunos patrones que contienen a otros y los amplían; por ejemplo, el sexto patrón de la tabla incluye al segundo. Estos patrones surgen a partir de opiniones en las que participan varias palabras de opinión o varias palabras de característica, y en las que el camino que conecta una de las palabras de un tipo con una de las palabras del otro contiene al camino que conecta a distintas palabras de uno de los tipos. En el ejemplo citado, el camino $J \rightarrow (\text{mod}) \rightarrow N \rightarrow (\text{subj}) \rightarrow J$ que conecta a las palabras *sound* y *bad* contiene al camino $N \rightarrow (\text{subj}) \rightarrow J$ que conecta a las palabras *quality* y *bad*; el trozo añadido corresponde al camino entre las palabras de característica *sound* y *quality*. Dado que nuestro sistema utilizará los patrones de dependencias para enlazar palabras de opinión a las *feature words* previamente identificadas con ayuda de la taxonomía de características, no parece necesario disponer de los dos patrones anteriores: únicamente con el primer patrón, aplicándolo al núcleo de las palabras de característica (*quality*) conseguimos enlazar correctamente la palabra de opinión *bad*. En cuanto a las opiniones en las que participen varias palabras de opinión, los patrones del tipo 1 se pueden descomponer en patrones del tipo 2, que conectan el núcleo de las palabras de característica con el núcleo de las palabras de opinión, y patrones del tipo 3, que conectan las palabras de opinión entre sí. En la tabla 6.15 mostramos algunos patrones de los tipos 2 y 3, junto con algunos ejemplos. En el capítulo 8 llevaremos a cabo experimentos utilizando bien patrones de tipo 1 o bien patrones de tipo 2 y 3.

El último patrón mostrado en la tabla 6.14 se ha inducido a partir de opiniones en las que participan negaciones. Se trata de un caso especial de la situación anterior, puesto que las negaciones son palabras de opinión. Los patrones del tipo 4 tratan de capturar este tipo de situaciones, capturando las relaciones entre el núcleo de las palabras de opinión y las negaciones participantes. De forma similar los patrones del tipo 5 tratan de capturar las relaciones con las palabras de polaridad dominante. En la tabla 6.15

Patrones tipo 1: <i>feature word</i> → <i>opinion word</i>				
Patrón	S	P	R	Ejemplos
N←(mod)←J	835	0,78	0,31	... good wireless headphones ... (headphones←good)
N→(subj)→J	325	0,96	0,12	The sound quality is not bad. (quality→bad)
N→(mod)→ N←(mod)←J	144	0,69	0,05	... bad quality sound ... (sound→quality←bad) ... great sound isolation ... (sound→isolation←great)
N→(nn)→ N←(mod)←J	69	0,70	0,03	... excellent build quality ... (build→quality←excellent) ... good battery life ... (battery→life←good)
N→(obj)→V	61	0,26	0,02	I love these headphones! (headphones→love) ... fatiguing bass ... (bass→fatiguing)
J→(mod)→N→(subj)→J	51	0,98	0,02	The sound quality is not bad. (sound→quality→bad) The low end and mids are good. (low→end→good)
N←(mod)←V	46	0,66	0,02	... best valued headphones ... (headphones←valued) ... coiled cord ... (cord←coiled)
N←(nn)←N	36	0,22	0,01	It comes with a leather pouch. (pouch←leather)
N→(subj)→V	32	0,27	0,01	The bass response is massively over-exaggerated. (response→over-exaggerated)
...				
N→(subj)→J→(pred)→ V←(neg)←X	16	1,00	0,01	The sound quality is not bad. quality→bad→is←not

Cuadro 6.14: Algunos patrones del tipo 1 para *headphones*. Las columnas *S*, *P* y *R* indican el número de apariciones, la precisión y la cobertura, respectivamente.

Patrones tipo 2: núcleo de las <i>feature words</i> → núcleo de las <i>opinion words</i>				
Patrón	S	P	R	Ejemplos
N←(mod)←J	808	0,77	0,34	... good wireless headphones ... (headphones←good)
N→(subj)→J	315	0,94	0,13	The sound quality is not bad. (quality→bad)
N→(obj)→V	58	0,25	0,03	I love these headphones! (headphones→love)
N←(mod)←V	45	0,64	0,02	...great sounding headphones ... (headphones←great)
Patrones tipo 3: núcleo de las <i>opinion words</i> → otras <i>opinion words</i>				
N←(mod)←J	71	0,70	0,11	... These headphones do a good job! ... (job←good)
N←(desc)←J	48	0,98	0,08	... The bass sounds a little muffled. ... (sounds←muffled)
V←(desc)←R	28	0,5	0,05	... They fit well. ... (fit←well)
V←(obj)←N	13	0,14	0,02	These headphones do a good job! (do←job)
Patrones tipo 4: núcleo de las <i>opinion words</i> → <i>negation word</i>				
J→(pred)→	28	1,00	0,16	The sound quality is not bad. bad→bad→is←not
V←(neg)←X				
V←(aux)←	23	1,00	0,13	I would not recommend them. recommend←would←not
V←(neg)←X				
J←(mod)←X	18	1,00	0,10	... not portable ... portable←not
J←(mod)←				
R←(mod)←X	11	0,92	0,06	... not very comfortable ... comfortable←very←not
Patrones tipo 5: núcleo de las <i>opinion words</i> → <i>dominant polarity word</i>				
J←(mod)←R	54	0,10	0,51	... too big ... big←too
D←(lex-mod)←R	4	1,0	0,04	... too much bass ... much←too
R←(mod)←R	4	0,27	0,04	... breaks too easily ... easily←too
...				
J←(pnmod)←R	1	0,33	0,01	... good enough for most users ... good←enough

Cuadro 6.15: Algunos patrones de tipo 2, 3, 4 y 5 para *headphones*. Las columnas *S*, *P* y *R* indican el número de apariciones, la precisión y la cobertura, respectivamente.

mostramos algunos patrones de estos tipos acompañados de algunos ejemplos extraídos del corpus.

Ni el análisis de dependencias ni el etiquetado morfosintáctico en que se basa el algoritmo de inducción de patrones están exentos de errores. Evidentemente, esto puede repercutir negativamente en el sistema. Sin embargo, algunos de los errores de las herramientas de procesamiento lingüístico tienden a repetirse para las mismas situaciones; este tipo de errores conducirán a patrones de dependencias que, si bien analíticamente parecen erróneos, al ser aplicados a textos procesados por las mismas herramientas lingüísticas anteriores (con los mismos errores en ciertas circunstancias) enlazan correctamente las palabras origen y destino. Un ejemplo de esto se observa en el segundo patrón del tipo 3 en la tabla 6.15. El etiquetador que hemos utilizado tiende a errar al etiquetar el verbo *sound*, asignándole la categoría morfosintáctica *nombre*. Dado que este error se repite frecuentemente, el algoritmo de inducción ha encontrado el patrón que se muestra, a partir del análisis de dependencias que la herramienta utilizada obtiene para construcciones como la mostrada (tomando como entrada las categorías morfosintácticas asignadas anteriormente). Este patrón demuestra ser efectivo, con una precisión cercana al 100 %, a pesar de basarse en un resultado erróneo de las herramientas de preprocesado.

A la vista de los experimentos, los patrones de dependencias parecen ser una herramienta adecuada para conseguir enlazar las palabras de característica y las palabras de opinión relacionadas con una alta precisión.

Parte IV

Un sistema de extracción de opiniones

Capítulo 7

TOES: Un Sistema Extractor de Opiniones sobre Características Adaptable al Dominio

Resumen: En el presente capítulo, presentamos el sistema de extracción de opiniones sobre características al que hemos denominado *TOES (Taxonomy-based Opinion Extraction System)*. La arquitectura modular que plantearemos permite la construcción de sistemas de extracción basados en los recursos específicos del dominio definidos en el capítulo 5, así como la de sistemas que no hagan uso de dichos recursos. Se describirán los algoritmos empleados por cada uno de los componentes concretos del sistema, así como los distintos parámetros de configuración disponibles. El sistema está completamente implementado y servirá de base al estudio experimental contenido en el capítulo 8.

7.1. Introducción

El objetivo del sistema es llevar a cabo la extracción de opiniones sobre características, según hemos definido la tarea en el capítulo 4. El sistema permite realizar la tarea para un dominio determinado, apoyándose en los recursos definidos en el capítulo 5. De manera alternativa, también contemplamos la posibilidad de llevar a cabo gran parte de la tarea de extracción

de opiniones sin hacer uso de los recursos (y por tanto, de manera independiente del dominio), posibilitando la construcción de sistemas de extracción con un menor esfuerzo (aunque, previsiblemente, con una menor precisión y cobertura).

La estructura del capítulo es la siguiente: en primer lugar, describimos conceptualmente la arquitectura del sistema. A continuación, definimos una serie de componentes abstractos participantes en el sistema, cada uno de los cuáles tiene por objeto la resolución de una subtarea determinada. Para cada una de estas subtareas, describimos los componentes concretos que hemos implementado, algunos de los cuáles están basados en distintas técnicas de la bibliografía, y otros se basan en técnicas propias que hacen uso de los recursos del dominio. Para clarificar el funcionamiento de algunos de los componentes concretos, puede consultarse el apéndice A, donde se muestran algunos ejemplos concretos de la ejecución de los mismos.

En el capítulo 8 llevaremos a cabo un desarrollo experimental sobre un dominio concreto, dirigido a evaluar cada uno de estos componentes individualmente, así como el sistema completo de extracción de opiniones, a partir de distintas combinaciones de los componentes concretos disponibles.

7.2. Descripción de la arquitectura del sistema

Hemos intentado afrontar la extracción de opiniones mediante la división del problema en distintos subproblemas atacables de manera independiente. Para cada uno de los subproblemas identificados, hemos definido un *componente abstracto* encargado de su solución. Posteriormente, hemos implementado uno o varios *componentes concretos* como instancias concretas de los anteriores, algunos basados en recursos y otros no. Finalmente, la concatenación de varios componentes concretos, junto a una serie de valores de configuración de los mismos, compondrán un sistema concreto de extracción (ver figura 7.1). Este diseño modular nos permite una gran flexibilidad para configurar distintos escenarios de experimentación, además de posibilitar futuras nuevas implementaciones de los componentes abstractos.

La entrada al sistema es un conjunto de documentos de opinión, previamente analizados mediante una serie de herramientas lingüísticas, y los recursos del dominio en cuestión en caso de participar en el sistema algún componente concreto que haga uso de los mismos. La salida es el conjunto de evidencias de opinión extraídas, con al menos los atributos *feature* y *polarity* inicializados.

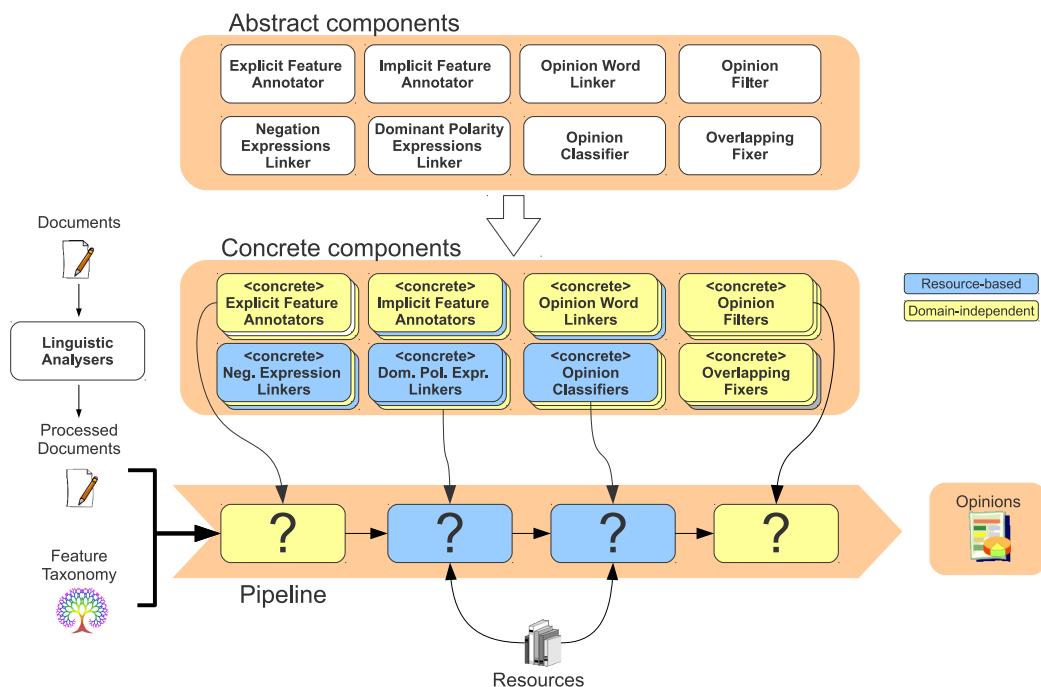


Figura 7.1: Arquitectura conceptual de TOES

El sistema no devuelve aquellas evidencias de opinión generadas en el proceso cuya polaridad haya quedado sin definir.

7.2.1. Entidades participantes

El sistema dispone de abstracciones que representan a todas las entidades participantes en el problema, que tratan de ajustarse a las definiciones contenidas en el capítulo 4. Los documentos sobre los que se va a llevar a cabo la extracción de opiniones son analizados mediante las mismas herramientas lingüísticas utilizadas para la generación de los recursos (ver sección 6.2.3), de manera que dicha información está disponible para su uso por parte de cualquier componente. También se dispone de abstracciones de los propios recursos, de forma que la información de los mismos también está disponible para su uso.

Las evidencias de opinión son representadas mediante abstracciones que disponen de los mismos atributos que hemos utilizado en su definición formal, con la diferencia de que el atributo *polarity* no es un tipo enumerado, sino un valor real. Si dicho valor real es positivo, se entiende que la polaridad de la evidencia es positiva, y viceversa. Esto permite a los componentes trabajar con valores reales de orientación semántica para las evidencias (por ejemplo,

para definir umbrales, aunque tras el proceso completo de extracción sólo será relevante la polaridad de las mismas.

7.3. Definición de componentes abstractos y concretos

En la presente sección, vamos a definir cada uno de los componentes abstractos que participan en el sistema, describiremos los subproblemas que atacan y mostraremos algún ejemplo de lo esperado de su ejecución. Además, para cada componente abstracto, repasaremos los distintos componentes concretos de los que se dispone, explicando el funcionamiento específico de cada uno.

Hemos identificado los siguientes componentes abstractos, cada uno relacionado con la solución de un subproblema distinto.

- **Anotador de palabras de característica:** se encarga de construir nuevas evidencias de opinión tentativas a partir de la observación de *feature words* en los documentos.
- **Anotador de características implícitas:** detecta posibles evidencias de opinión sobre características implícitas, a partir de la observación de palabras de opinión relacionadas con la característica.
- **Enlazador de palabras de opinión:** partiendo de evidencias de opinión con alguna palabra de característica o de opinión previamente anotada, este componente busca nuevas *opinion words* relacionadas con las anteriores y las añade a las evidencias de opinión.
- **Separador de opiniones:** divide en dos aquellas evidencias de opinión que representen a más de una opinión atómica sobre una misma característica.
- **Solucionador de opiniones superpuestas:** resuelve conflictos entre evidencias de opinión que comparten algunos de sus campos.
- **Clasificador de opiniones:** decide la polaridad de las evidencias de opinión.

Los componentes anteriores han sido enumerados según el orden lógico de aplicación para la resolución completa del reconocimiento y clasificación de opiniones. Por su importancia en dicha resolución, y también por su dificultad intrínseca, los componentes más interesantes son el anotador de palabras

de características, el anotador de características implícitas, el enlazador de palabras de opinión y el clasificador de opiniones.

A continuación detallamos el objetivo de cada componente abstracto, mostramos algunos ejemplos de los resultados esperados tras su aplicación y repasamos los componentes concretos disponibles en cada caso.

7.3.1. Anotador de palabras de característica

Este componente localiza palabras en los documentos de entrada que puedan hacer referencia a características del dominio. La aparición de dichas palabras señala la existencia de una posible opinión, por lo que una entidad de tipo opinión será añadida a la oración correspondiente. Este componente representa el punto de partida para la extracción de opiniones sobre características explícitas.

- Entradas: documento procesado.
- Salidas: evidencias de opinión con los atributos *feature* y *feature words* inicializados.

Ejemplo de aplicación

Partiendo de la siguiente oración:

These (1) headphones (2) sound (3) great (4) : (5) they (6) have
(7) clean (8) and (9) well-defined (10) highs (11) and (12)
powerful (13) low (14) frequencies (15).

, el anotador de palabras de característica generaría las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="treble" featWords="11"/>
<opinion feature="bass" featWords="14,15"/>
```

Nótese que el anotador de palabras de característica debe anotar todas las menciones a alguna característica del producto realizadas en el texto, independientemente de que dichas menciones correspondan realmente con opiniones. Posiblemente algunas de las evidencias de opinión así generadas serán eliminadas por otros componentes del sistema si no se corresponden realmente con una opinión.

Anotador de palabras de característica basado en la taxonomía *Taxonomy-based feature word annotator*

Este componente utiliza las palabras de característica contenidas en la taxonomía para anotar posibles evidencias de opinión. El funcionamiento es simple: se recorren las palabras de cada oración del documento de entrada, buscando ocurrencias de las palabras de característica de la taxonomía. Si se encuentra una secuencia de palabras que corresponde con una de las denominaciones contempladas para una característica determinada en la taxonomía, se añade a la oración una nueva evidencia de opinión, con los valores adecuados para los atributos *feature* y *feature words*. Por último, se eliminan aquellas evidencias de opinión cuyas palabras de característica sean un subconjunto de las de otra evidencia de opinión de la misma oración.

El componente dispone de un par de parámetros (*isPhraseBased* y *isNounBased*) que permiten restringir las apariciones de palabras de característica que serán anotadas, exigiendo que las mismas formen parte de un sintagma nominal, o bien que contengan algún nombre. Ambos parámetros tienen como intención evitar la anotación de palabras cuyo morfema coincide con alguna de las palabras de característica contenidas en la taxonomía, pero que no se están utilizando en el texto para referenciar a la característica en cuestión. Por ejemplo, la palabra “*sound*” aparece en la taxonomía de *headphones* asociada a la característica *sound quality* (piénsese por ejemplo en la oración “*Great sound!*”); sin embargo, “*sound*” también puede aparecer funcionando como verbo (“*They sound great!*”), en cuyo caso no se trata de una palabra de característica (en el ejemplo mostrado, se trataría de una palabra del término de opinión “*sound great*”). De esta forma, la restricción establecida por el parámetro *isPhraseBased* consiste en buscar exclusivamente palabras de característica que formen parte de sintagmas nominales, y la establecida por el parámetro *isNounBased*, que obliga a que al menos una de las palabras anotadas sea un nombre, van encaminadas a filtrar ambigüedades como la mostrada. En último término, la solución a todas las ambigüedades entre morfemas del texto y elementos de la taxonomía vendría dada por la correcta anotación de los significados concretos de las palabras de característica incluidas en la taxonomía, por ejemplo mediante el uso de los *synsets* de *WordNet*. Pero esto nos obligaría a realizar desambigüación de significados de las palabras de los documentos de entrada. Siendo ésta una tarea aún por perfeccionar, hemos considerado más conveniente la solución parcial planteada basada en la información morfosintáctica. Aún así, la información morfosintáctica que nos proporcionan los analizadores que utilizamos no está exenta de fallos, por lo que será necesario comprobar la eficacia de las restricciones planteadas (ver sección 8.3.1).

7.3.2. Anotador de características implícitas

Este componente identifica la posible presencia de opiniones sobre características implícitas, a partir de la observación de palabras de opinión relacionadas con dichas características. La aparición de estas palabras ocasionará la creación de entidades de tipo opinión que serán añadidas a la oración correspondiente. Este componente representa el punto de partida para la extracción de opiniones sobre características implícitas.

- Entradas: documento procesado.
- Salidas: evidencias de opinión con los atributos *feature* y *opinion words* inicializados.

Ejemplo de aplicación

Partiendo de la siguiente oración:

These(1) headphones(2) sound(3) great(4) : (5) they(6) have
(7) clean(8) and(9) well-defined(10) highs(11) and(12)
powerful(13) low(14) frequencies(15).

, el anotador de características implícitas generaría la siguiente evidencia de opinión:

```
<opinion feature="sound quality" opWords="3,4"/>
```

Anotador de características implícitas basado en los indicadores *Cue-based implicit feature annotator*

Este componente hace uso de los indicadores de características implícitas para anotar posibles evidencias de opinión sobre características implícitas. La definición de dicho recurso puede consultarse en la sección 5.6.1. El funcionamiento del componente es simple: se buscan apariciones de los términos incluidos en el recurso, y se crean nuevas evidencias de opinión con dichas apariciones etiquetadas como palabras de opinión. El componente dispone de parámetros para establecer umbrales de probabilidad y de soporte, de manera que no todos los términos incluidos en el recurso sean utilizados, sino sólo aquellos cuyos valores de soporte y probabilidad cumplan con dichos umbrales mínimos establecidos. Dado que algunos términos del recurso están asociados a más de una característica, el componente escogerá aquella característica cuya probabilidad sea mayor. Por último, se eliminan aquellas evidencias de opinión cuyas palabras de opinión sean un subconjunto de las palabras de opinión de otras evidencias de opinión añadidas por este mismo componente.

El componente también dispone de un parámetro que permite seleccionar si se utilizan las probabilidades *condicionadas* o *no condicionadas* (aquellas correspondientes a los atributos *feature-based implicit feature probability* y *feature-based implicit feature, not explicit, conditional probabily* del recurso, ver sección 5.6.1).

Anotador de características implícitas basado en PMI-IR *PMIIR-based implicit feature annotator*

Este componente resuelve la anotación de evidencias de opinión sobre características implícitas de manera no supervisada. Para ello, se utiliza el algoritmo *Pointwise Mutual Information - Information Retrieval*, que mide la dependencia estadística de dos palabras o expresiones, estimando ciertas probabilidades mediante el uso de búsquedas en un motor de recuperación de información web. Esta medida es utilizada en (Turney, 2002b) y (Turney & Littman, 2003b) para calcular la orientación semántica de expresiones de manera no supervisada, calculando el valor de PMIR-IR entre un término determinado y algunas palabras que actúan a modo de semillas positivas y negativas (por ejemplo, “*excellent*” como semilla positiva y “*poor*” como semilla negativa). Si un término obtiene, por ejemplo, un valor mucho mayor de dependencia estadística con las semillas positivas que con las negativas, es probable que la palabra en cuestión posea una orientación semántica positiva. El algoritmo está explicado en detalle en la sección 2.3.1.

El componente que nos ocupa utiliza la misma técnica de cálculo de dependencias estadísticas entre palabras para decidir si alguna de las palabras que aparecen en una frase determinada está relacionada con alguno de los elementos de la taxonomía de características. Para cada uno de los adjetivos encontrados en un documento de entrada, el componente calcula los valores de *PMI_{IR}* entre el adjetivo y el nombre de cada una de las características contenidas en la taxonomía. Si alguno de los valores obtenidos es lo suficientemente grande, se crea una nueva evidencia de opinión para la característica que obtuvo dicho valor y con dicho adjetivo como palabra de opinión. Se dispone de un parámetro que permite indicar el valor mínimo de *PMI_{IR}* que debe obtenerse para que el componente añada una nueva evidencia de opinión.

7.3.3. Enlazador de palabras de opinión

A partir de una evidencia de opinión previamente creada, el enlazador de palabras de opinión debe seleccionar aquellas palabras de la oración que actúen como palabras de opinión en dicha evidencia. Para llevar a cabo es-

ta tarea, el componente puede basarse en las palabras de característica o de opinión previamente anotadas en la evidencia de opinión. Las palabras de opinión añadidas pueden ocasionalmente formar parte de expresiones de negación o de polaridad dominante.

- Entradas: documento procesado, evidencias de opinión con el atributo *feature* y al menos *feature words* o *opinion words* inicializados.
- Salidas: evidencias de opinión anteriores, posiblemente con nuevas *opinion words*.

Ejemplo de aplicación

Partiendo de la siguiente oración:

These (1) headphones (2) sound (3) great (4) : (5) they (6) have
 (7) clean (8) and (9) well-defined (10) highs (11) and (12)
 powerful (13) low (14) frequencies (15).

, y las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="treble" featWords="11"/>
<opinion feature="bass" featWords="14,15"/>
<opinion feature="sound quality" opWords="3,4"/>
```

, la aplicación del enlazador de palabras de opinión debería obtener las siguientes evidencias de opinión para la oración:

```
<opinion feature="headphones" featWords="2" opWords="3,4"/>
<opinion feature="treble" featWords="11" opWords="8,10"/>
<opinion feature="bass" featWords="14,15" opWords="13"/>
<opinion feature="sound quality" opWords="3,4"/>
```

Enlazador de palabras de opinión basado en ventana

Window-based opinion word linker

Esta implementación del enlazador de palabras de opinión lleva a cabo una aproximación simple al problema de seleccionar qué palabras del contexto de las palabras de característica pueden considerarse palabras de opinión que afecten a las mismas, sin hacer uso de ningún recurso. A partir de una evidencia de opinión previamente generada y cuyo atributo *feature words* no esté vacío, se añadirán como palabras de opinión algunas de las palabras del contexto de dichas palabras de característica. Tanto el tamaño de la ventana de contexto a utilizar, como las categorías morfosintácticas que serán tenidas en cuenta como palabras de opinión, son configurables.

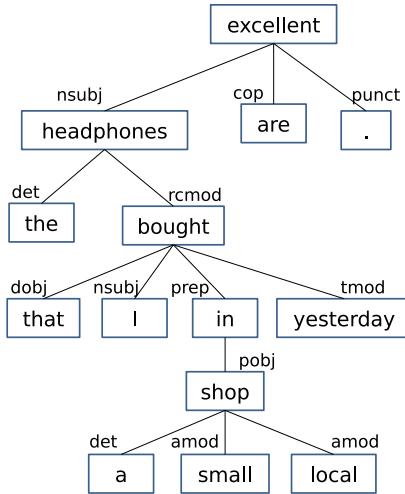


Figura 7.2: Árbol de dependencias de la oración de ejemplo

La simplicidad de este componente limita mucho su potencia. Si se utilizan ventanas pequeñas, no se capturarán correctamente aquellas evidencias de opinión en las que las palabras de característica y las palabras de opinión estén más alejadas. Si se utilizan ventanas grandes, es probable que se añadan más palabras de opinión de las que realmente están modificando a las palabras de característica.

Enlazador de palabras de opinión basado en dependencias *Dependency-based opinion word linker*

Para sobreponer las limitaciones encontradas en la aproximación basada en ventana anterior, se dispone de un enlazador de palabras de opinión que utiliza la información proporcionada por un analizador de dependencias. A partir de los patrones contenidos en el recurso de patrones de dependencias (ver definición del recurso en la sección 5.7.1), el enlazador de palabras de opinión basado en dependencias es capaz de añadir palabras de opinión con independencia de la separación entre éstas y las palabras de característica. Así, por ejemplo, en la oración “*The headphones that I bought in a small local shop yesterday are excellent.*”, cuyo árbol de dependencias se muestra en la figura 7.2, el componente será capaz de enlazar la palabra de opinión “*excellent*” con la palabra de característica “*headphones*” (que aparecen alejadas en la oración pero estrechamente relacionadas en el árbol de dependencias), siempre que el recurso de patrones de dependencias contenga el patrón adecuado.

El componente no sólo permite añadir palabras de opinión a evidencias de opinión partiendo de las palabras de característica previamente anotadas (haciendo uso de los patrones de tipo 1 o tipo 2), sino también añadir palabras de opinión relacionadas con otras palabras de opinión previamente anotadas (haciendo uso de los patrones de tipo 3)¹. Este modo de funcionamiento puede ser controlado mediante un parámetro de configuración.

Sea cual sea el tipo de patrón que se aplique, el componente buscará en las evidencias de opinión previamente generadas palabras origen válidas, y a partir de ellas buscará palabras destino dentro de la oración que se relacionen con las palabras origen de la manera expresada por el patrón utilizado (es decir, con el mismo camino ascendente y descendente de relaciones de dependencias y etiquetas morfosintácticas). Por ejemplo, si el componente está configurado para funcionar con patrones de tipo 1, tratará de aplicar cada uno de los patrones de tipo 1 disponibles en el recurso partiendo de cada una de las palabras de característica previamente anotadas en las evidencias de opinión disponibles. Aquellas palabras de la oración que estén conectadas en el árbol de dependencias con alguna de esas palabras de característica, a través de un camino cuyas relaciones de dependencias y categorías morfosintácticas coincidan con las contenidas en alguno de los patrones utilizados, serán añadidas como palabras de opinión.

En aquellos casos en que el patrón deba aplicarse a partir del núcleo de las palabras de característica o del núcleo de las palabras de opinión previamente anotadas (patrones de tipo 2 y 3), el componente selecciona como núcleo aquella palabra de la que dependen las demás. Si no se conocen relaciones de dependencia entre las palabras implicadas (algo que puede ocurrir a causa de las limitaciones de los analizadores de dependencias), se toma como núcleo a la palabra más a la derecha en la oración.

Los patrones de dependencias son aplicados por el componente en el orden en que aparecen en el recurso, esto es, comenzando por los patrones que obtuvieron una mayor precisión en la fase de generación del recurso. El componente dejará de aplicar patrones cuando la precisión de los mismos sea inferior a un umbral, o cuando se hayan seleccionado suficientes palabras de opinión. Estas características son también controladas mediante parámetros de configuración.

¹Recordemos que los patrones de tipo 1 conectan cualquier palabra de característica con cualquier palabra de opinión, los de tipo 2 conectan el núcleo de las palabras de característica con el núcleo de las palabras de opinión, y los de tipo 3 conectan el núcleo de las palabras de opinión con cualquier otra palabra de opinión.

Enlazador de expresiones especiales basado en ventana
Window-based special expression linker

Este componente concreto es una implementación del enlazador de palabras de opinión que, partiendo de evidencias de opinión con alguna palabra de opinión previamente anotada, busca expresiones especiales (de negación o de polaridad dominante) que afecten a dichas palabras de opinión, y las añade como nuevas palabras de opinión a las evidencias. La intuición nos dice que, en la mayoría de los casos, este tipo de expresiones aparecen en un contexto muy cercano de las palabras de opinión a las que afectan. Esta intuición, unida al bajo número de expresiones que consideramos, justifica la utilización de una ventana de contexto alrededor de las palabras de opinión para las que estemos buscando expresiones especiales relacionadas. Dentro de dicha ventana, el componente busca la aparición de expresiones contenidas en las listas de expresiones especiales. Se trata por tanto de un componente basado en recurso, pero independiente del dominio (puesto que dichas listas lo son).

Enlazador de expresiones especiales basado en dependencias
Dependency-based special expression linker

En un intento por superar las limitaciones del componente concreto anterior, el enlazador de expresiones especiales basado en dependencias se apoya en el recurso de patrones de dependencias para realizar la tarea de seleccionar las expresiones especiales que afectan a ciertas palabras de opinión previamente anotadas. La forma de operar del componente es similar a la del enlazador de palabras de opinión basado en dependencias: partiendo del núcleo de las palabras de opinión de la evidencia que se esté tratando, se aplican de manera secuencial los patrones de tipo 4 (para las expresiones de negación) o 5 (para las expresiones de polaridad dominante) del recurso de patrones de dependencias, añadiendo las palabras así extraídas, mientras se cumplan los valores umbrales de precisión y número máximo de palabras de opinión establecidos en el componente.

Previamente a realizar su tarea, el componente busca expresiones de no negación, marcándolas para impedir que las palabras participantes en las mismas sean extraídas por los patrones de dependencias.

7.3.4. Separador de opiniones

Según el enfoque que seguimos en nuestro trabajo, cada opinión identificable sobre una característica determinada debe ser representada y extraída

por el sistema de manera atómica, aunque se trate de opiniones contenidas en una misma oración y sobre una misma característica. El separador de opiniones tiene la finalidad de dividir en dos o más evidencias aquellas evidencias de opinión que representen a más de una opinión atómica, basándose para ello en las palabras de opinión previamente identificadas.

- Entradas: documento procesado, evidencias de opinión con al menos los atributos *feature* y *opinion words* inicializados.
- Salidas: evidencias de opinión resultantes de la división.

Ejemplo de aplicación

Partiendo de la siguiente oración:

```
These(1) headphones(2) sound(3) great(4) : (5) they(6) have  

(7) clean(8) and(9) well-defined(10) highs(11) and(12)  

powerful(13) low(14) frequencies(15).
```

, y las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2" opWords="3,4"/>  

<opinion feature="treble" featWords="11" opWords="8,10"/>  

<opinion feature="bass" featWords="14,15" opWords="13"/>  

<opinion feature="sound quality" opWords="3,4"/>
```

, la aplicación del separador de opiniones debería obtener las siguientes evidencias de opinión para la oración:

```
<opinion feature="headphones" featWords="2" opWords="3,4"/>  

<opinion feature="treble" featWords="11" opWords="8"/>  

<opinion feature="treble" featWords="11" opWords="10"/>  

<opinion feature="bass" featWords="14,15" opWords="13"/>  

<opinion feature="sound quality" opWords="3,4"/>
```

Separador de opiniones basado en conjunciones

Conjunction-based opinion splitter

El sistema sólo posee una implementación concreta del componente abstracto que nos ocupa. En concreto, el componente divide aquellas evidencias de opinión cuyas palabras de opinión estén separadas en la oración por una conjunción. El procedimiento es sencillo: se recorren las secuencias de palabras de la oración contenidas entre cada dos palabras de opinión de cada evidencia de opinión. Si entre las palabras de dicha secuencia se encuentran determinadas conjunciones (el conjunto de estas palabras es configurable), la evidencia de opinión es dividida en varias evidencias, conteniendo cada una de ellas las palabras de opinión existentes en cada una de las partes delimitadas por las conjunciones.

7.3.5. Solucionador de opiniones superpuestas

En ocasiones, algunas de las evidencias de opinión generadas por el sistema comparten algunas palabras de opinión. En algunos casos, el solucionador de opiniones superpuestas puede decidir eliminar algunas de las palabras de opinión de alguna de las opiniones involucradas, para solucionar el conflicto. En otros casos, el componente puede optar por dejar la situación tal como está, al considerar que las palabras de opinión involucradas participan realmente en varias opiniones.

- Entradas: documento procesado, evidencias de opinión con al menos los atributos *feature* y *opinion words* inicializados.
- Salidas: evidencias de opinión anteriores, posiblemente con algunas palabras de opinión menos en algunas de ellas.

Ejemplo de aplicación

Partiendo de la siguiente oración:

```
These(1) headphones(2) sound(3) great(4) : (5) they(6) have  

(7) clean(8) and(9) well-defined(10) highs(11) and(12)  

powerful(13) low(14) frequencies(15) .
```

, y las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2" opWords="3,4"/>  

<opinion feature="treble" featWords="11" opWords="8"/>  

<opinion feature="treble" featWords="11" opWords="10"/>  

<opinion feature="bass" featWords="14,15" opWords="13"/>  

<opinion feature="sound quality" opWords="3,4"/>
```

, la aplicación del solucionador de opiniones superpuestas debería obtener las siguientes evidencias de opinión para la oración:

```
<opinion feature="headphones" featWords="2" opWords="" />  

<opinion feature="treble" featWords="11" opWords="8"/>  

<opinion feature="treble" featWords="11" opWords="10"/>  

<opinion feature="bass" featWords="14,15" opWords="13"/>  

<opinion feature="sound quality" opWords="3,4"/>
```

Nótese que se han eliminado las palabras de opinión “*sound great*” asociadas a la característica *headphones* en la primera evidencia de opinión, puesto que las mismas palabras estaban asociadas a la característica más específica *sound quality*. La evidencia de opinión en cuestión se queda sin palabras de opinión; suponiendo que no se enlazaran nuevas palabras de opinión posteriormente, la evidencia sería descartada por el sistema al final del proceso, al obtener un valor nulo de polaridad.

Solucionador de opiniones explícitas superpuestas
Explicit overlapping opinion fixer

Este componente busca pares de evidencias de opinión sobre características explícitas con palabras de opinión en común, y elimina las palabras de opinión en conflicto en aquella evidencia en la que las palabras de opinión estén más distantes de las palabras de característica. De esta manera, se pretende corregir posibles errores cometidos por algún componente enlazador de palabras de opinión.

Existen sin embargo casos en los que es posible que varias evidencias de opinión sobre características explícitas compartan una o varias palabras de opinión. Esto ocurre en algunas oraciones en las que aparecen varias características referidas a través de un solo sintagma nominal, en construcciones conjuntivas. Por ejemplo, en la oración “*The highs and mids are outstanding*”, aparecen dos evidencias de opinión, una sobre la característica *highs* y otra sobre *mids*, actuando en ambas la palabra “*outstanding*” como palabra de opinión. El solucionador de opiniones explícitas superpuestas tratará de detectar estos casos, de manera que si encuentra una construcción conjuntiva entre las palabras de característica de las evidencias de opinión involucradas no se eliminarán palabras de opinión de ninguna de ellas.

Solucionador de opiniones implícita/explícita superpuestas
Implicit/Explicit overlapping opinion fixer

Este componente busca pares de evidencias de opinión, una sobre característica explícita y otra sobre característica implícita, que compartan una o varias palabras de opinión. En función de la característica de cada una de las evidencias, decide de cuál de ellas debe eliminar la palabra o palabras de opinión compartidas:

- Si la característica implícita es una especialización de la característica explícita (es decir, si aparece en la taxonomía dependiendo de la misma), las palabras de opinión en conflicto son eliminadas de la evidencia de opinión sobre característica explícita. Se entiende que en estos casos, la opinión implícita está concretando la característica con mayor precisión.
- En otro caso, se eliminan las palabras de opinión de la evidencia sobre característica implícita.

Veamos un ejemplo de cada caso. En la oración “*These headphones sound great*”, ante la existencia de dos evidencias de opinión, una explícita sobre la

característica *headphones* y otra explícita sobre *sound quality*, ambas con las palabras de opinión “*sound quality*”, el componente eliminaría las palabras de opinión de la primera evidencia, al ser la característica *headphones* un nodo padre de *sound quality* en la taxonomía de características del dominio. Sin embargo, en la oración “*The case is too big*”, y ante la existencia de dos evidencias de opinión, una explícita sobre la característica *case* y otra implícita sobre *size*, ambas con las palabras de opinión “*too big*”, el componente eliminaría las palabras de opinión de la segunda evidencia, al no ser *size* una especialización de *case* en la taxonomía.

7.3.6. Clasificador de opiniones

El clasificador de opiniones es el componente encargado de decidir la polaridad de las evidencias de opinión, a partir de la información previamente anotada en la misma (generalmente, característica y palabras de opinión, incluyendo expresiones de negación y de polaridad dominante). En algunos casos, el clasificador de opiniones puede no estar seguro de la polaridad de una opinión (o incluso decidir que no se trata de una opinión), en cuyo caso la dejará sin valor. Además, el clasificador comprobará previamente si la evidencia de opinión que se pretende clasificar ya posee un valor de orientación semántica. Si es así, no procesará la evidencia. De esta forma se permite enlazar varios clasificadores de opiniones, de manera que un clasificador posterior sólo clasificará aquellas evidencias de opinión que no hayan sido clasificadas por los componentes anteriores.

- Entradas: documento procesado, evidencias de opinión con al menos los atributos *feature* y *opinion words* inicializados.
- Salidas: evidencias de opinión anteriores, posiblemente con el atributo *polarity* inicializado.

Ejemplo de aplicación

Partiendo de la siguiente oración:

These(1) headphones(2) sound(3) great(4) : (5) they(6) have
 (7) clean(8) and(9) well-defined(10) highs(11) and(12)
 powerful(13) low(14) frequencies(15) .

, y las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2" opWords="" />
<opinion feature="treble" featWords="11" opWords="8" />
<opinion feature="treble" featWords="11" opWords="10" />
<opinion feature="bass" featWords="14,15" opWords="13" />
```

```
<opinion feature="sound quality" opWords="3,4" />
```

, la aplicación del clasificador de opiniones superpuestas debería obtener las siguientes evidencias de opinión para la oración:

```
<opinion feature="headphones" featWords="2" opWords="" />
<opinion feature="treble" featWords="11" opWords="8"
    polarity="+"/>
<opinion feature="treble" featWords="11" opWords="10"
    polarity="+"/>
<opinion feature="bass" featWords="14,15" opWords="13"
    polarity="+"/>
<opinion feature="sound quality" opWords="3,4" polarity="+"/> 
```

Clasificador de opiniones basado en el lexicón de opiniones

Lexicon-based opinion classifier

Este componente utiliza la información contenida en el lexicón de opiniones para decidir la polaridad de las evidencias de opinión previamente generadas. Además de este recurso, también se utilizan las listas de expresiones especiales.

Dada una evidencia de opinión sobre una característica determinada, el funcionamiento es el siguiente:

1. Tras comprobar que la orientación semántica aún no ha sido calculada, se buscan expresiones de no negación dentro de las palabras de opinión. En caso de encontrarse, las palabras participantes no serán tenidas en cuenta en el resto del proceso de clasificación.
2. Se buscan expresiones de negación. Si se encuentra un número impar, la polaridad finalmente calculada será invertida. Las palabras participantes en las expresiones encontradas no serán tenidas en cuenta en el resto del proceso de clasificación.
3. Se buscan expresiones de polaridad dominante. Si se encuentra alguna, se decide la polaridad de la opinión en base a la polaridad de dicha expresión, y el proceso acaba.
4. Si no se encontraron expresiones de polaridad dominante, se procede a buscar en el lexicón de opiniones el término formado por todas las palabras de opinión. Si se encuentra dicho término, y sus valores de probabilidad y orientación semántica en el lexicón para la característica de la evidencia son superiores a ciertos valores (estos umbrales son configurables mediante parámetros del componente), el componente decide

la polaridad de la evidencia de opinión en función del valor encontrado en el recurso (y de las posibles expresiones de negación encontradas).

5. Si no se obtuvo una estimación de la polaridad a partir del paso anterior, se procede a buscar en el lexicón los subtérminos formados por aquellas palabras de opinión que aparezcan de manera contigua en la oración. Si dicha búsqueda también es infructuosa (o los valores encontrados no superan los umbrales configurados), se buscan en el lexicón cada una de las palabras de opinión individualmente. En ambos casos, se estima la orientación semántica del conjunto mediante la media de las orientaciones semánticas de los términos contenidas en el lexicón para la característica de la evidencia. El componente decide la polaridad de la evidencia a partir del valor obtenido (y de las posibles expresiones de negación encontradas).

El componente permite la expansión semántica basada en WordNet de los términos a buscar en el lexicón. De esta forma, si una palabra determinada no se encuentra en el lexicón, pero sí algún sinónimo o antónimo de la misma, el componente puede utilizar los valores de probabilidad y orientación semántica contenidos en el lexicón para esa otra palabra para estimar los valores de probabilidad y orientación semántica de la palabra original. Este comportamiento puede ser desactivado mediante un parámetro.

El componente dispone también de parámetros para ajustar los valores mínimos de *support*, *feature-based opinion word probability* y *feature-based semantic orientation polarity* exigibles para que las estimaciones contenidas en el lexicón de opinión sean consideradas. Para clarificar el funcionamiento del componente, pueden consultarse ejemplos concretos de ejecución del mismo en la sección A.9.

Clasificador de opiniones basado en PMI-IR *PMIIR-based opinion classifier*

Este componente lleva a cabo la clasificación de opiniones de manera no supervisada, utilizando el algoritmo *Pointwise Mutual Information - Information Retrieval* (Turney, 2002b) (Turney & Littman, 2003b) para realizar estimaciones reales entre -1,0 y 1,0 de la orientación semántica de los términos de opinión participantes en las evidencias de opinión a clasificar. El algoritmo está explicado en la sección 2.3.1.

La forma de proceder del componente es similar a la del clasificador basado en el lexicón de opiniones, con dos diferencias fundamentales. En primer lugar, en sustitución de los valores de orientación semántica del lexicón, se utilizan las estimaciones generadas mediante la fórmula anterior, llevando

a cabo las búsquedas necesarias en AltaVista². Téngase en cuenta que las orientaciones semánticas así calculadas son independientes del dominio y de la característica. Además, no se dispone de los valores de probabilidad de que un término funcione como término de opinión.

En segundo lugar, las estimaciones de orientación semántica se llevan a cabo siempre para los términos obtenidos a partir de las palabras de opinión que aparecen de manera contigua en la oración, a diferencia del clasificador anterior, que permitía también llevar a cabo estimaciones basadas en el término formado por todas las palabras de opinión y por cada una de las palabras por separado. El resto del funcionamiento del componente es similar al del clasificador anterior, incluido el tratamiento de las expresiones de no negación, de negación y de polaridad dominante.

Clasificador de opiniones basado en WordNet

WordNet-based opinion classifier

El clasificador de opiniones basado en WordNet utiliza una medida de similitud entre adjetivos basada en distancias en un grafo construido a partir de las relaciones de sinonimia de WordNet para estimar la orientación semántica de las palabras de opinión. El método está basado en el utilizado en (Kamps et al, 2004b). Para calcular la orientación semántica de una palabra, se calcula la distancia en el grafo anterior entre la palabra y dos semillas, una representando la positividad (“good”) y otra la negatividad (“bad”). A partir de dichas distancias, y de la distancia entre ambas semillas, se estima la orientación semántica de una palabra w según la siguiente función, extraída de (Kamps et al, 2004b):

$$EVA(w) = \frac{d(w, "bad") - d(w, "good")}{d("good", "bad")}$$

Los valores así obtenidos están contenidos en el intervalo $[-1, 0; 1, 0]$. En nuestra implementación de esta función, hemos utilizado una versión modificada de la distancia definida en (Kamps et al, 2004b): además de las relaciones de sinonimia, hemos utilizado las relaciones de similitud entre adjetivos disponibles en WordNet. De esta forma, la distancia entre dos palabras w_1 y w_2 se calcula generando los conjuntos de sinónimos y adjetivos similares de la primera palabra. Si el conjunto generado contiene a la segunda palabra, la distancia es igual a 1. Si no, se genera un nuevo conjunto con todos los sinónimos y adjetivos similares de cada una de las palabras del conjunto anterior. Si el nuevo conjunto contiene a w_2 , la distancia es igual a 2. Si no, el

²Hemos implementado una caché de búsquedas para minimizar en lo posible la dependencia con el buscador web.

proceso se repite nuevamente, las veces necesarias, siendo la distancia igual al número de conjuntos generados hasta encontrar a w_2 .

El componente utiliza esta función para estimar la orientación semántica de los adjetivos participantes como palabras de opinión en las evidencias de opinión a clasificar, estimando la orientación semántica de las evidencias como la media de las orientaciones semánticas no nulas de los adjetivos participantes (e invirtiendo el signo en caso de haberse encontrado un número impar de expresiones de negación participando en las palabras de opinión). Nótese que las orientaciones semánticas son calculadas a nivel de palabra, y sólo para los adjetivos, por lo que la orientación semántica estimada para una evidencia de opinión con las palabras de opinión “*sound great*” será igual a la orientación semántica calculada mediante la función *EVA*(“*great*”). Esto puede ser una limitación con respecto a los dos clasificadores anteriormente descritos, en los que se estima la orientación semántica de grupos de palabras de manera conjunta.

Se dispone de un parámetro para controlar el valor mínimo que debe obtenerse como estimación de la orientación semántica de un término para que el resultado de la estimación sea tomado como válido. El componente lleva a cabo un tratamiento de las expresiones especiales similar a los anteriores clasificadores.

Clasificador de opiniones basado en SentiWordNet

SentiWordNet-based opinion classifier

Este clasificador de opiniones utiliza el recurso SentiWordNet 3.0 (Baccianella & Sebastiani, 2010). SentiWordNet es un recurso léxico, de generación semiautomática, en el que a cada *synset* de WordNet se le asignan tres puntuaciones: una de *positividad*, otra de *negatividad* y una tercera de *objetividad*. El clasificador de opiniones basado en SentiWordNet utiliza las puntuaciones de positividad y negatividad para estimar las orientaciones semánticas de las evidencias de opinión. Para ello, busca cada una de las palabras de opinión en el recurso; dado que nuestro sistema no incluye una etapa de desambigüación de significados, se utiliza la información contenida en SentiWordNet para todos los *synsets* a los que pertenezca la palabra (con su correspondiente categoría morfosintáctica) cuya orientación semántica se está estimando. Para cada *synset*, se estima la orientación semántica como la diferencia entre las puntuaciones de positividad y negatividad. Cada palabra de opinión recibe entonces como orientación semántica la media de las orientaciones semánticas así calculadas para todos los *synsets* a los que pertenezca la palabra con su correspondiente categoría morfosintáctica. Por último, la orientación semántica final para la evidencia de opinión será la media de las

orientaciones semánticas de aquellas palabras encontradas en SentiWordNet.

El componente dispone de un parámetro que permite establecer el valor absoluto mínimo que debe obtenerse como estimación de la orientación semántica de un término para que el resultado de la estimación sea tomado como válido. El tratamiento de las expresiones especiales es similar al del resto de clasificadores de opiniones.

Clasificador de opiniones basado en conjunciones

Conjunction-based opinion classifier

A diferencia del resto de clasificadores de opiniones, el clasificador basado en conjunciones no permite clasificar las evidencias de opinión directamente a partir de las palabras de opinión. Para llevar a cabo su tarea, otro clasificador o clasificadores de opiniones deben haber sido ejecutados previamente. La tarea del componente concreto que nos ocupa es la de clasificar un pequeño número de evidencias de opinión que no hayan conseguido ser clasificadas previamente, basándose en la polaridad de otras evidencias de opinión de la misma oración. Nos basamos para ello en el mismo principio que en el método de ampliación del lexicón de opiniones (ver sección 6.8), según el cual las orientaciones semánticas de dos adjetivos relacionados mediante una conjunción “*and*” tienden a compartir polaridad. De forma similar, las orientaciones semánticas de dos adjetivos relacionados mediante una conjunción “*but*” suelen tener polaridades contrarias (Hatzivassiloglou & McKeown, 1997).

Basándose en este principio, el funcionamiento del componente es simple. Para cada evidencia de opinión cuya polaridad sea nula, se buscan otras evidencias de opinión en la misma oración con polaridad no nula, cuyas palabras de opinión estén situadas en la oración formando una construcción conjuntiva con las palabras de opinión de la evidencia que se está intentando clasificar. Si se encuentra alguna evidencia con dichas características, se decide la polaridad en función de dicha evidencia y de la conjunción utilizada en la expresión.

7.3.7. Componentes *simulados*

Para llevar a cabo algunos de los experimentos que detallaremos en el siguiente capítulo, algunos componentes concretos que hemos llamado *simulados* fueron implementados. Estos componentes llevan a cabo su tarea de manera perfecta, utilizando para ellos las evidencias de opinión anotadas en el corpus. Mediante la utilización de estos componentes en determinados *pipelines* experimentales, podremos medir la precisión del resto de componentes concretos de manera individual.

Figura 7.3: Ejemplo de *pipeline*

Se dispone de componentes concretos simulados para los siguientes componentes abstractos (se muestra el nombre que recibe cada componente trámposo en el sistema):

- Anotador de palabras de característica (*Fake feature word annotator*)
- Anotador de características implícitas (*Fake implicit feature annotator*)
- Enlazador de palabras de opinión (*Fake opinion word linker*)
- Clasificador de opiniones (*Fake opinion classifier*)

El anotador de palabras de característica simulado, por ejemplo, generaría una evidencia de opinión por cada evidencia sobre característica explícita anotada en el corpus, completando tan solo los campos *feature* y *feature words* de las evidencias creadas. Usando este componente en lugar del anotador de palabras de característica basado en la taxonomía de características, podríamos estudiar la precisión de otros componentes concretos, sabiendo que los errores que se produzcan serán responsabilidad de esos otros componentes, y no del anotador de palabras de característica.

7.3.8. *Pipelines* de procesado

Para llevar a cabo el proceso completo de extracción de opiniones, debemos ejecutar una lista de componentes concretos en un orden determinado. Dado que en algunos casos hemos implementado varios componentes concretos que resuelven un mismo subproblema desde acercamientos distintos, y que además dichos componentes disponen de parámetros que permiten configurarlos para funcionar de maneras sensiblemente distintas, son muchas las combinaciones posibles con las que abordar el problema. Llamamos *pipeline* a cada lista de componentes concretos posible, con un orden determinado de ejecución y un conjunto de valores para los parámetros de configuración de cada uno de los componentes participantes.

El sistema permite definir un *pipeline* mediante un formato XML en el que se describen los componentes en un orden determinado y los valores de los parámetros. Por ejemplo, el *pipeline* mostrado en la figura 7.3 se describe con el siguiente código:

```

<pipeline name="explicit , dependencies , lexicon">
    <component name="Explicit"
        type="Taxonomy-based_FeatureWordAnnotator">
        <param name="isPhraseBased" value="false"/>
        <param name="isNounBased" value="false"/>
    </component>
    <component name="Linker"
        type="Dependency-based_OpinionWordLinker">
        <param name="type" value="1"/>
        <param name="minPrecision" value="0.1"/>
        <param name="minAccPrecision" value="0.7"/>
        <param name="maxOpWords" value="2"/>
    </component>
    <component name="Negation"
        type="Window-based_SpecialExpressionLinker">
        <param name="type" value="1"/>
        <param name="windowSize" value="2"/>
    </component>
    <component name="Dominant"
        type="Window-based_SpecialExpressionLinker">
        <param name="type" value="2"/>
        <param name="windowSize" value="1"/>
    </component>
    <component name="Polarity"
        type="Lexicon-based_OpinionWordLinker">
        <param name="useWordNetSynToExpand" value="false"/>
        <param name="useWordNetAntToExpand" value="false"/>
        <param name="useCompletePhrase" value="true"/>
        <param name="useContiguousWordsAsPhrase" value="true"/>
        <param name="useEachWordAsPhrase" value="true"/>
        <param name="useGlobalSO" value="false"/>
        <param name="featureBasedProbabilities" value="true"/>
        <param name="minSupport" value="4"/>
        <param name="minProb" value="0.3"/>
        <param name="minAbsSO" value="1"/>
    </component>
    <component name="PolarityFixer"
        type="Conjunction-based_OpinionClassifier">
        <param name="andWords" value="and , "/>
        <param name="butWords" value="but"/>
        <param name="allowHolesBetweenOpWords" value="false"/>
    </component>
</pipeline>

```

Nótese que se permite la inclusión de varios componentes concretos para un mismo componente abstracto. En el código anterior, por ejemplo, se incluyen dos implementaciones del clasificador de opiniones, una basada en el lexicón de opiniones y otra en conjunciones. Esto también será frecuente

cuando se empleen patrones de distintos tipos para llevar a cabo el enlazado de las palabras de opinión; en este caso, el *pipeline* en cuestión contendrá varias instancias del enlazador de palabras de opinión basados en dependencias, cada una con su propia configuración de parámetros.

Capítulo 8

TOES: Evaluación y ajuste de parámetros

Resumen: En el presente capítulo, llevamos a cabo una serie de experimentos con la intención de evaluar tanto los componentes concretos individuales definidos en el capítulo 7 como el sistema completo TOES. A partir de los resultados obtenidos en los distintos experimentos, plantearemos algunas conclusiones sobre la tarea que hemos abordado y sobre nuestro acercamiento, incluyendo una reflexión sobre la idoneidad del uso de una metodología supervisada como la propuesta en este trabajo de tesis y un ejemplo de aplicación de agregación y visualización de las opiniones extraídas por el sistema.

8.1. Introducción

En las siguientes secciones llevaremos a cabo diversos experimentos de cara a evaluar los distintos componentes concretos de manera independiente, así como algunos *pipelines* determinados, construidos con la intención de comparar acercamientos que hagan o no uso de los recursos específicos del dominio. Hemos empleado el mismo dominio que usamos en el capítulo 6 (*headphones*). Son tres los objetivos de la experimentación planteada. En primer lugar, pretendemos decidir los valores más apropiados de los parámetros disponibles en los componentes del sistema. En segundo lugar, queremos medir la efectividad de los distintos componentes concretos disponibles para cada una de las subtareas planteadas, de manera que podamos decidir qué componentes son más apropiados para la constitución de un sistema que resuelva la extracción de opiniones de la mejor manera posible. Por último,

los valores obtenidos en las distintas métricas de evaluación que utilizaremos nos servirán para determinar la eficacia del sistema en su conjunto, y para sacar algunas conclusiones a partir del análisis de dichos resultados. En el capítulo 9 se muestran resultados obtenidos para un par de dominios más (*hotels* y *cars*).

8.2. Metodología de las evaluaciones

En todas las evaluaciones utilizamos validación cruzada a partir de 10 particiones aleatorias del corpus anotado (*10-fold cross validation*); cada resultado mostrado se calcula como la media de diez repeticiones del experimento, empleando en cada repetición una de las particiones como corpus de evaluación y las nueve particiones restantes para inducir los recursos y ajustar los parámetros de los componentes.

Las métricas utilizadas serán las habituales en problemas de recuperación y extracción de información, precisión (*precision*) y cobertura (*recall*), que se definen de la siguiente manera:

$$\text{precision} = \frac{\text{correct}}{\text{correct} + \text{incorrect}} \quad \text{recall} = \frac{\text{correct}}{\text{correct} + \text{missing}} \quad ,$$

donde *correct* es el número de *entidades* correctamente reconocidas (aquellas que se corresponden con *entidades* existentes), *incorrect* es el número de *entidades* incorrectamente reconocidas (aquellas que no se corresponden con ninguna *entidad* existente), y *missing* es el número de *entidades* reales no reconocidas. Según la subtarea que estemos evaluando, una *entidad* será un objeto distinto. Por ejemplo, cuando evaluemos los enlazadores de palabras de opinión, las entidades serán las palabras a enlazar; cuando evaluemos el sistema de extracción de opiniones, las entidades serán las opiniones.

A partir de la precisión y la cobertura, calcularemos la media armónica de ambos valores, según la fórmula:

$$F_{\beta} = \frac{(1 + \beta^2) \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

De esta manera, obtenemos un solo valor que facilita la comparación de los resultados obtenidos en distintos experimentos. Si hacemos $\beta = 1$, damos la misma importancia a la precisión y a la cobertura. En nuestros experimentos, además de F_1 , también calcularemos $F_{\frac{1}{2}}$, que otorga a la precisión el doble de importancia que a la cobertura¹.

¹Tal como expondremos más adelante, en algunas ocasiones puede ser más importante

En los experimentos en que se lleve a cabo una clasificación binaria (evaluación de los clasificadores de opiniones y evaluación de la tarea *opinion classification*), utilizaremos la medida de exactitud (*accuracy*), similar a la precisión, que mide la proporción de entidades correctamente clasificadas. En estas tareas no tiene sentido hablar de cobertura, ni por tanto de media armónica.

8.3. Evaluación individual de los componentes

En esta sección, evaluaremos individualmente la eficacia de los componentes concretos a la hora de resolver la subtarea correspondiente. No se evaluarán ni el separador de opiniones basado en conjunciones ni los solucionadores de opiniones superpuestas, puesto que las tareas que resuelven no son evaluables de manera independiente.

En algunos casos, mostraremos gráficas que ilustran la influencia de algunos de los parámetros de los componentes en los resultados obtenidos por los mismos.

8.3.1. Evaluación del anotador de palabras de característica basado en la taxonomía

En la tabla 8.1 se muestran los resultados obtenidos por el anotador de palabras de característica basado en la taxonomía. Las palabras anotadas por el mismo fueron comparadas con las palabras de característica de las evidencias de opinión del corpus anotado. La baja precisión obtenida por el componente se debe a que muchas de las apariciones de las palabras de característica contenidas en la taxonomía no se corresponden con la existencia de opiniones en los documentos. En principio, esto no debe considerarse un problema, puesto que la función del anotador de palabras de característica es detectar todas las posibles evidencias de opinión sobre características explícitas, lo cual parece conseguir (se obtiene una cobertura máxima de más del 99 %), siendo responsabilidad de otros componentes (enlazadores de palabras de opinión y clasificadores de opiniones) decidir qué apariciones de palabras de característica corresponden realmente a evidencias de opinión. La alta cobertura conseguida corrobora la hipótesis inicial según la cual el conjunto de

la precisión del sistema que la cobertura. De ahí que, en dichas situaciones, sea interesante tratar de optimizar el valor de $F_{\frac{1}{2}}$

términos utilizados por los usuarios para nombrar las características de un producto tiende a converger.

El empleo de restricciones morfosintácticas (aparición de algún nombre en las palabras de característica e inclusión de las mismas en un sintagma nominal) reduce considerablemente la cobertura en favor de pequeños aumentos en la precisión. Aunque dichas restricciones deberían cumplirse teóricamente para todas las palabras de característica, los errores cometidos por los analizadores lingüísticos utilizados hacen que el anotador omita demasiadas apariciones de palabras de característica cuando se utilizan las restricciones morfosintácticas. No parece por tanto una buena opción emplear dichas restricciones.

Restricciones	p	r	F_1
Ninguna	0,2418	0,9947	0,3891
Incluyen nombre	0,2585	0,9466	0,4062
Sólo sintagmas nominales	0,2566	0,9436	0,4035
Ambas	0,2622	0,9214	0,4083

Cuadro 8.1: Evaluación individual del anotador de palabras de característica basado en la taxonomía para *headphones*.

8.3.2. Evaluación del anotador de características implícitas basado en PMI-IR

En la tabla 8.2 se muestran distintos resultados obtenidos por el anotador de características implícitas basado en PMI-IR para *headphones*, correspondientes a los distintos valores del umbral de PMI para los que se consiguen los mejores resultados de cada una de las métricas de evaluación empleadas. Para obtener los resultados mostrados, se compararon las evidencias de opinión sobre características implícitas obtenidas por el componente con las evidencias de opinión sobre características implícitas anotadas en el corpus.

Los resultados obtenidos son bastante modestos, con una precisión máxima del 0,4359 y una cobertura máxima del 0,551. Téngase en cuenta que sólo se consideran correctas aquellas evidencias de opinión extraídas cuya característica coincide con la de la evidencia correspondiente anotada en el corpus (y la taxonomía de *headphones* contiene 31 características distintas).

Optimiza...	p	r	F_1	Umbrales	
				PMI	
precision	0,4359	0,1595	0,2336	6,0	
recall	0,1183	0,551	0,1947	0,0	
F_1	0,4293	0,1611	0,2342	5,6	

Cuadro 8.2: Evaluación individual del anotador de características implícitas basado en PMI-IR para *headphones*.

8.3.3. Evaluación del anotador de características implícitas basado en los indicadores

El anotador de características implícitas basado en los indicadores obtuvo resultados mucho mejores que el componente anterior, con una precisión máxima de 0,8887 y una cobertura máxima del 0,9456 (tabla 8.3). Es previsible además que la precisión mejore en el sistema completo, puesto que algunas de las evidencias de opinión extraídas por el componente corresponden en realidad a opiniones sobre características explícitas, situación que puede ser parcialmente corregida por el solucionador de opiniones implícita-explicativa superpuestas (piénsese por ejemplo en la oración “*The case is too big*”, en la que “*big*” parece un claro indicador de la característica *size*, pero que en realidad corresponde a una palabra de opinión participante en una opinión sobre la característica explícita “*case*”).

Optimiza...	p	r	F_1	$F_{\frac{1}{2}}$	Soporte	
					Prob. mínima	Soporte mínimo
precision	0,8887	0,0435	0,0829	0,1819	0,8	20
recall	0,1452	0,9456	0,2518	0,1748	0	1
F_1	0,5224	0,6117	0,5635	0,5381	0,25	1
$F_{\frac{1}{2}}$	0,6449	0,4745	0,5467	0,6017	0,35	5

Cuadro 8.3: Evaluación individual del anotador de características implícitas basado en los indicadores para *headphones*.

En la figura 8.1 puede observarse la influencia del umbral de probabilidad del componente. A medida que se aumenta dicho umbral, la precisión del componente va creciendo (en detrimento de la cobertura) hasta cierto valor del umbral ($\approx 0,75$) a partir del cual se produce una caída repentina de la precisión. La elección de distintos valores del umbral permite obtener compromisos distintos entre la precisión y la cobertura para el componente: valores en torno a 0,25 conducen a un equilibrio entre precisión y cobertura;

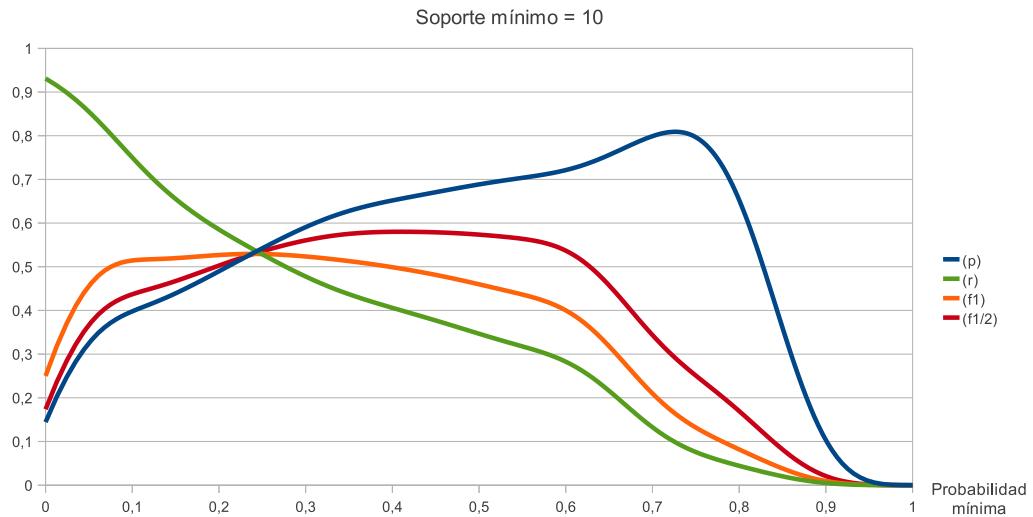


Figura 8.1: Evaluación individual del anotador de características implícitas basado en los indicadores para *headphones*: influencia del umbral de probabilidad

valores en torno a 0,4 serían más adecuados para obtener una mejor precisión sin sacrificar demasiada cobertura.

En la figura 8.2 se muestra la influencia del umbral de soporte mínimo del componente. Dicho umbral parece tener menor influencia en los resultados obtenidos, con pequeñas mejoras en la precisión a medida que el valor aumenta, a costa de mayores pérdidas en la cobertura. Parece adecuado por tanto el empleo de valores bajos para dicho umbral.

8.3.4. Evaluación del enlazador de palabras de opinión y de expresiones especiales basados en ventana

La anotación de palabras de opinión mediante ventana de palabras obtiene los mejores resultados cuando sólo se consideran como posibles palabras de opinión los adjetivos, tal como puede verse en la tabla 8.4. Los experimentos se realizaron partiendo de las palabras de característica anotadas en el corpus, y aplicando sobre las mismas el enlazador de palabras de opinión basado en ventana. Se compararon entonces las palabras de opinión extraídas con las palabras de opinión anotadas en el corpus. La inclusión de otras categorías morfosintácticas hace empeorar todas las métricas de evaluación consideradas (excepto, claro está, la cobertura). El mejor resultado para la medida F_1 se

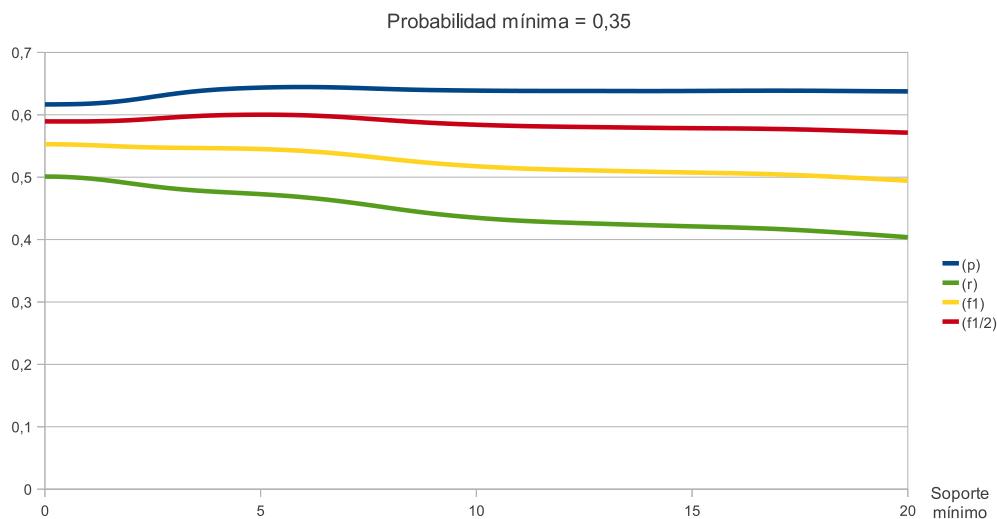


Figura 8.2: Evaluación individual del anotador de características implícitas basado en los indicadores para *headphones*: influencia del umbral de soporte

obtiene utilizando ventanas pequeñas (en torno a 3 palabras a cada lado de las palabras de característica), como se observa en la figura 8.3. Empleando una ventana pequeña, y restringiendo las palabras de opinión a enlazar a los adjetivos incluidos en dicha ventana, el resultado obtenido es mejor de lo que cabría esperar, dado lo simple de la solución aportada por el componente, con un resultado de 0, 6375 para la medida F_1 . Esto indica que en una mayoría de las evidencias de opinión sobre característica explícita, las palabras de opinión aparecen muy próximas a las palabras de característica. No obstante, existe aún un margen suficientemente grande de mejora tanto en la precisión como en la cobertura, que justifica el uso de patrones de dependencias.

El empleo adicional de enlazadores de expresiones especiales también basados en ventana permite alcanzar un valor máximo para F_1 de 0, 6479, lo que representa una mejora de más de un punto porcentual (ver tabla 8.5).

8.3.5. Evaluación del enlazador de palabras de opinión y expresiones especiales basado en dependencias

Hemos evaluado la precisión y cobertura del enlazador de palabras de opinión basado en dependencias, utilizando patrones de tipo 1, por un lado, o de tipos 2, 3, 4 y 5, por otro (en este último caso, aplicamos cuatro enlazadores de palabras de opinión consecutivos, uno para cada tipo de pa-

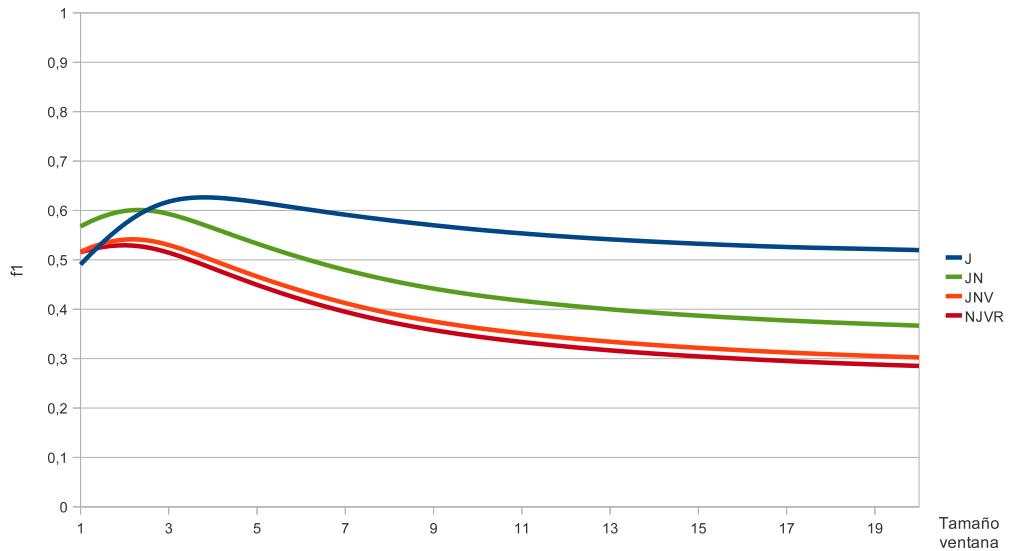


Figura 8.3: Evaluación individual del enlazador de palabras de opinión basado en ventana para *headphones*: influencia del tamaño de las ventanas y las restricciones morfosintácticas. La leyenda indica las categorías morfosintácticas permitidas en cada caso (J: Adjetivos. N: Nombres. V: Verbos. R: Adverbios)

trones). Además, hemos considerado el empleo de patrones con restricciones morfosintácticas (aquellos en los que se exigen determinadas categorías morfosintácticas a las palabras participantes en el patrón) y sin ellas (aquellos basados tan sólo en las relaciones de dependencia). En cada caso, hemos buscado los valores de los parámetros de los enlazadores que optimizan las medidas F_1 y $F_{\frac{1}{2}}$. Los experimentos se realizaron partiendo de las palabras de característica anotadas en el corpus, y aplicando sobre las mismas el enlazador o enlazadores de palabras de opinión basados en dependencias. Se compararon entonces las palabras de opinión extraídas con las palabras de

Restricciones morfosintácticas	p	r	F_1	$F_{\frac{1}{2}}$
Adjetivos	0,7079	0,5798	0,6375	0,678
Adjetivos/Nombres	0,6394	0,5932	0,6154	0,6296
Adjetivos/Nombres/Verbos	0,492	0,641	0,5567	0,516
Adjetivos/Nombres/Verbos/Adverbios	0,4542	0,6729	0,5423	0,4857

Cuadro 8.4: Evaluación individual del enlazador de palabras de opinión basado en ventana para *headphones*.

Restricciones morfosintácticas	p	r	F_1	$F_{\frac{1}{2}}$
Adjetivos	0,7034	0,6005	0,6479	0,6801
Adjetivos/Nombres	0,6371	0,6084	0,6224	0,6312
Adjetivos/Nombres/Verbos	0,4991	0,6775	0,5747	0,5268
Adjetivos/Nombres/Verbos/Adverbios	0,4548	0,6849	0,5466	0,4876

Cuadro 8.5: Evaluación individual del enlazador de palabras de opinión y el enlazador de expresiones especiales basados en ventana para *headphones*.

opinión anotadas en el corpus. Los resultados, que mostramos en la tabla 8.6, permiten concluir, en primer lugar, que es mejor utilizar cuatro enlazadores de palabras de opinión basados en patrones de dependencia de tipo 2, 3, 4 y 5, en lugar de un solo enlazador basado en patrones de tipo 1. En segundo lugar, se confirma que los patrones que incluyen restricciones morfosintácticas consiguen mejores resultados que los que no las incluyen. Esto último es coherente con los resultados que obtuvimos en los experimentos de inducción de patrones de dependencias en la sección 6.10.2.

		Sin restricciones morfosintácticas			
Patrones tipo	Optimiza...	p	r	F_1	$F_{\frac{1}{2}}$
Tipo 1	F_1	0,7330	0,7192	0,7261	0,7302
	$F_{\frac{1}{2}}$	0,7781	0,6649	0,7171	0,7525
Tipos 2, 3, 4 y 5	F_1	0,7448	0,7140	0,7291	0,7384
	$F_{\frac{1}{2}}$	0,7739	0,6972	0,7336	0,7573

		Con restricciones morfosintácticas			
Patrones tipo	Optimiza...	p	r	F_1	$F_{\frac{1}{2}}$
Tipo 1	F_1	0,7949	0,7239	0,7578	0,7796
	$F_{\frac{1}{2}}$	0,8484	0,6721	0,7500	0,8061
Tipos 2, 3, 4 y 5	F_1	0,8074	0,7262	0,7646	0,7897
	$F_{\frac{1}{2}}$	0,8490	0,6952	0,7644	0,8130

Cuadro 8.6: Evaluación individual del enlazador de palabras de opinión basado en dependencias.

Como era de esperar, los resultados obtenidos superan ampliamente a los conseguidos por el enlazador de palabras de opinión basado en ventana: 0,7646 frente a 0,6479 en la medida F_1 , y 0,8130 frente a 0,6801 en la medida $F_{\frac{1}{2}}$. En la figura 8.4 se representan los resultados de la aplicación de los enlazadores de palabras de opinión basados en dependencias de tipos 2, 3, 4 y 5, variando el parámetro de los enlazadores encargado de controlar la precisión mínima exigida a los patrones a utilizar. Variando dicho paráme-

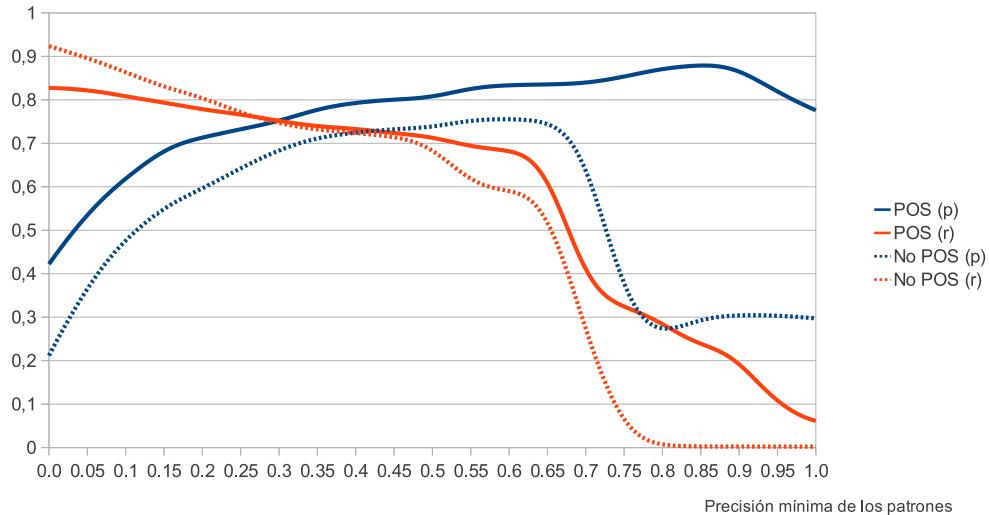


Figura 8.4: Evaluación individual del enlazador de palabras de opinión basado en patrones de dependencias de tipo 2, 3, 4 y 5 para *headphones*: influencia del umbral de precisión mínima de los patrones utilizados. Las líneas continuas representan la precisión y la cobertura obtenidas utilizando los patrones con restricciones morfosintácticas; las líneas punteadas, utilizando los patrones sin restricciones morfosintácticas.

tro, podemos aumentar progresivamente la precisión del componente, a la par que disminuimos su cobertura. Es precisamente la cobertura máxima del componente el punto más débil del mismo: en torno a 0,92, conseguida utilizando todos los patrones de dependencias disponibles en el recurso, sin restricciones morfosintácticas. Este resultado nos indica que existen palabras de opinión anotadas en el corpus que no son alcanzables utilizando ninguno de los patrones contenidos en el recurso; esto puede ser debido a errores en el analizador de dependencias, o a que aparezcan en construcciones sintácticas no observadas previamente en los documentos utilizados para inducir el recurso. Es previsible que este problema sea menor cuanto mayor sea el conjunto de documentos disponibles para la inducción de los patrones.

8.3.6. Evaluación del clasificador de opiniones basado en el lexicón

Para evaluar el clasificador de opiniones basado en lexicón (y también para el resto de clasificadores de opiniones) partimos de las evidencias de opinión

Experimento	<i>accuracy</i>
Lexicón de opinión	0,8659
Lexicón expandido (PolarityRank)	0,8717
Expansión semántica (WordNet)	0,8937
Lexicón expandido + expansión semántica	0,8966

Cuadro 8.7: Evaluación individual del clasificador de opiniones basado en el lexicón para *headphones*. En cada caso, se han escogido los valores óptimos de los parámetros para cada medida.

Clasificador	<i>accuracy</i>
Basado en WordNet	0,5513
Basado en SentiWordNet	0,6247
Basado en PMI-IR	0,6771

Cuadro 8.8: Evaluación comparativa de los clasificadores de opiniones basados en WordNet, SentiWordNet y PMI-IR para *headphones*.

del corpus anotado, comprobando en cada caso si la polaridad obtenida por el clasificador concuerda o no con la polaridad anotada. De esta manera medimos la exactitud (*accuracy*) del clasificador. En la tabla 8.7 se muestran los resultados de cuatro experimentos que pretenden medir la conveniencia o no de utilizar el lexicón ampliado mediante PolarityRank (ver sección 6.8) y las expansiones semánticas basadas en WordNet (ver sección 7.3.6). El mejor resultado se obtiene usando ambos, con un 89,66 % de las opiniones correctamente clasificadas.

8.3.7. Evaluación de los clasificadores de opiniones basados en WordNet, SentiWordNet y PMI-IR

En la tabla 8.8 se muestran los resultados conseguidos por los clasificadores de opiniones basados en WordNet, SentiWordNet y PMI-IR . El mejor resultado es conseguido por el basado en PMI-IR, seguido del basado en SentiWordNet y el basado en WordNet. En todos los casos, los valores de *accuracy* obtenidos son muy inferiores al conseguido por el clasificador basado en el lexicón (más de 20 puntos porcentuales).

Clasificador	<i>accuracy</i>
Basado en WordNet	0,5647
Basado en SentiWordNet	0,6296
Basado en PMI-IR	0,6831
Basado en lexicón	0,9009

Cuadro 8.9: Evaluación comparativa de los clasificadores de opiniones para *headphones*, al concatenar a cada clasificador un clasificador basado en conjunciones.

8.3.8. Evaluación del clasificador de opiniones basado en conjunciones

El clasificador de opiniones basado en conjunciones tiene como objetivo clasificar opiniones a partir de otras opiniones previamente clasificadas por otro clasificador (ver sección 7.3.6). Por tanto, para evaluar la eficacia de este componente, llevamos a cabo cuatro experimentos, en los que ejecutamos el clasificador basado en conjunciones tras la ejecución de cada uno de los cuatro clasificadores de opiniones de los que disponemos. Aunque modestas, se consiguen mejoras en la exactitud de todos los clasificadores iniciales (ver tabla 8.9).

8.3.9. Evaluación de la combinación de clasificadores de opiniones

En la tabla 8.10 se muestran los resultados obtenidos al utilizar un clasificador basado en lexicón combinado con los demás clasificadores. La idea es utilizar en primer lugar el clasificador basado en lexicón, y para aquellas opiniones para las que no se obtenga una predicción para la polaridad, utilizar otro clasificador basado en WordNet, SentiWordNet o PMI-IR. En todos los casos, se aplica finalmente el clasificador basado en conjunciones. Se observa que la combinación del clasificador basado en lexicón y el basado en PMI-IR obtiene los mejores resultados, con una exactitud de 0,9248, lo que significa una mejora de casi dos puntos y medio porcentuales con respecto a la aplicación individual del clasificador basado en lexicón.

8.4. Evaluación del sistema completo

Basándonos en los resultados obtenidos en las evaluaciones independientes de los componentes, en esta sección vamos a proponer y evaluar algunos

Clasificador	<i>accuracy</i>
Basado en lexicón más WordNet y conjunciones	0,9086
Basado en lexicón más SentiWordNet y conjunciones	0,9168
Basado en lexicón más PMI-IR y conjunciones	0,9248

Cuadro 8.10: Evaluación comparativa de la combinación de un clasificador basado en lexicón con cada uno de los demás clasificadores para *headphones*.

pipelines para resolver la tarea de reconocimiento y clasificación de opiniones. Empezaremos construyendo *pipelines* con componentes independientes del dominio, para posteriormente comparar los resultados con los obtenidos por otros *pipelines* que incluyen componentes basados en recursos.

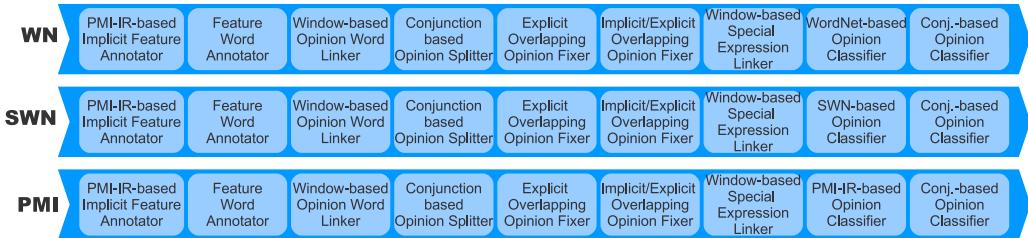
En todos los casos, mediremos la precisión y la cobertura de los *pipelines* en la tarea de reconocimiento de opiniones (*opinion recognition*). Consideraremos que una opinión extraída es correcta si se corresponde con una opinión anotada en la misma oración y sobre la misma característica². Para la tarea de clasificación de opiniones (*opinion classification*), mediremos la exactitud conseguida al clasificar las opiniones que hayan sido correctamente reconocidas

8.4.1. Evaluación de *pipelines* ligeros de recursos

Comenzamos planteando tres *pipelines* cuyos componentes no hacen uso de los recursos del dominio, salvo la taxonomía de características (ver figura 8.5). Todos ellos son iguales, a excepción del clasificador de opiniones, que está basado en WordNet en el primer *pipeline*, en SentiWordNet en el segundo y en PMI-IR en el tercero. Los componentes utilizados son, por este orden: un anotador de características implícitas basado en PMI-IR, un anotador de palabras de característica, un enlazador de palabras de opinión basado en ventana, un separador de opiniones basado en conjunciones, un solucionador de opiniones explícitas superpuestas, un solucionador de opiniones implícitas/explícitas superpuestas, un enlazador de expresiones especiales basado en ventana, un clasificador de opiniones basado en WordNet/SentiWordNet/PMI-IR (según el *pipeline*), y finalmente un clasificador de opiniones basado en conjunciones.

Las tablas 8.11 y 8.12 muestran los resultados obtenidos por cada uno de los *pipelines*, seleccionando los valores apropiados de los parámetros de los

²En ambos casos, una opinión anotada en el corpus sólo puede ser utilizada para validar una única opinión extraída. Así, por ejemplo, si se extraen dos opiniones acerca de una característica determinada, y en el corpus sólo aparece anotada una opinión sobre dicha característica, la segunda opinión extraída es considerada incorrecta

Figura 8.5: *Pipelines* ligeros de recursos

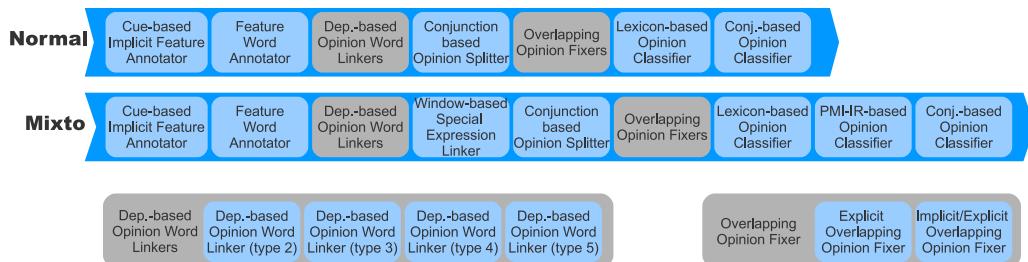
Pipeline	Opinion Recognition			Classification
	p	r	F_1	a
WN	0,5404	0,5080	0,5237	0,7785
SWN	0,5366	0,5415	0,5390	0,7817
PMI	0,5858	0,4804	0,5279	0,8345

Cuadro 8.11: Resultados obtenidos por los *pipelines* ligeros de recursos (optimizados para F_1)

componentes para optimizar los valores de F_1 y de $F_{\frac{1}{2}}$, respectivamente. Los mejores resultados los obtienen los *pipelines* basados en PMI-IR y en SentiWordNet. Los valores obtenidos son bastante modestos, lo que refuerza la idea de la necesidad de los recursos para llevar a buen término la extracción de opiniones. Aún así, un sistema basado en estos *pipelines* sería capaz de extraer correctamente entre un 30 % y un 40 % de las opiniones de un documento, ahorrándonos el esfuerzo necesario para la generación de los recursos, y a costa, eso sí, de extraer entre un 40 % y un 50 % de opiniones erróneas.

Pipeline	Opinion Recognition			Classification
	p	r	$F_{\frac{1}{2}}$	a
WN	0,6092	0,3039	0,5073	0,8706
SWN	0,6756	0,3002	0,5405	0,8940
PMI	0,6744	0,3643	0,5763	0,8688

Cuadro 8.12: Resultados obtenidos por los *pipelines* ligeros de recursos (optimizados para $F_{\frac{1}{2}}$)

Figura 8.6: *Pipelines* basados en recursos

8.4.2. Evaluación de *pipelines* basados en recursos

Hemos experimentado con *pipelines* formados por múltiples configuraciones de componentes basados en recursos, modificando el orden de los mismos, utilizando uno o dos anotadores de características implícitas, distintas combinaciones de enlazadores de palabras de opinión, etcétera. De entre todos los considerados, en la figura 8.6 se muestran los *pipelines* basados en recursos que proponemos. El primero de ellos está formado por:

- Un anotador de características implícitas basado en los indicadores de característica implícita.
- Un anotador de palabras de característica basado en la taxonomía de características.
- Cuatro enlazadores de palabras de opinión basados en los patrones de dependencias de tipo 2, 3, 4 y 5, respectivamente. En todos los casos, se utilizan patrones con restricciones morfosintácticas.
- Un separador de opiniones basado en conjunciones.
- Dos solucionadores de opiniones superpuestas (explícitas y explícita/implícita, respectivamente).
- Un clasificador de opiniones basado en el lexicón de opiniones.
- Un clasificador de opiniones basado en conjunciones.

El segundo de los *pipelines* mostrados añade un enlazador de expresiones especiales basado en ventana, y un clasificador de opiniones basado en PMI-IR (que según vimos en la sección 8.3.7, permite clasificar algunas de las opiniones para las que el lexicón no contenga información).

En las tablas 8.13 y 8.14 se muestran los resultados obtenidos por los *pipelines*, buscándose en el primer caso el mejor equilibrio posible entre la

Pipeline	Opinion Recognition			Opinion Classification
	p	r	F_1	a
Normal	0,7115	0,6617	0,6857	0,9345
Mixto	0,7043	0,6748	0,6892	0,9423

Cuadro 8.13: Resultados obtenidos por los *pipelines* basados en recursos (optimizados para F_1)

Pipeline	Opinion Recognition			Opinion Classification
	p	r	$F_{\frac{1}{2}}$	a
Normal	0,7869	0,5662	0,7300	0,9503
Mixto	0,7836	0,5736	0,7301	0,9572

Cuadro 8.14: Resultados obtenidos por los *pipelines* basados en recursos (optimizados para $F_{\frac{1}{2}}$)

precisión y la cobertura (optimización de F_1), y dándole mayor importancia a la precisión en el segundo (optimización de $F_{\frac{1}{2}}$). Los mejores resultados son obtenidos por el segundo *pipeline*, con un 0,6892 de F_1 (clasificando correctamente el 94,23 % de las opiniones extraídas) y un 0,7301 de $F_{\frac{1}{2}}$ (clasificando correctamente el 95,72 % de las opiniones). En todos los casos, se obtienen resultados sensiblemente mejores a los conseguidos por los *pipelines* ligeros de recursos.

8.5. Resumen y análisis de los resultados

En esta sección pretendemos resumir los resultados obtenidos en los experimentos anteriormente descritos. Además, expondremos algunas conclusiones y reflexiones extraídas del análisis de los resultados. De forma abreviada, las conclusiones que expondremos son las siguientes:

- Los recursos específicos del dominio constituyen un valioso mecanismo para afrontar la extracción de opiniones.
- El coste derivado de la anotación manual de los documentos del dominio para la generación de los recursos está plenamente justificado a la vista de la relación entre el esfuerzo y la mejora de los resultados.
- No todas las opiniones presentan la misma dificultad, según sean implícitas o explícitas, o afecten a unas características u otras.

Tarea	Sin recursos del dominio			Con recursos del dominio		
	p	r	F_1	p	r	F_1
Anotación palabras de caract.	-	-	-	0,2622	0,9214	0,4083
Anotación de caract. impl.	0,4293	0,1611	0,2342	0,5224	0,6117	0,5635
Enlazado palabras de opinión	0,7034	0,6005	0,6479	0,8074	0,7262	0,7646
Clasificación de opiniones	0,6831	-	-	0,9248	-	-

Cuadro 8.15: Resumen comparativo de los mejores resultados conseguidos para cada una de las subtareas en que hemos dividido la tarea de reconocimiento y clasificación de opiniones, utilizando o sin utilizar los recursos del dominio.

- Nuestro sistema permite primar la precisión o la cobertura de las opiniones extraídas, mediante el ajuste de determinados parámetros.
- Los valores obtenidos en las métricas de evaluación por nuestro sistema son buenos, teniendo en cuenta la dificultad de la tarea.
- Las opiniones extraídas por el sistema tienen la calidad suficiente para ser útiles para el cálculo de indicadores agregados.

A continuación desarrollamos cada uno de estos puntos.

8.5.1. Importancia de los recursos específicos del dominio

Los resultados obtenidos respaldan nuestra hipótesis inicial, según la cual la disponibilidad de recursos específicos del dominio permite afrontar la extracción de opiniones con mayor eficacia. En la tabla 8.15 se muestran los mejores resultados obtenidos (haciendo uso de los recursos del dominio por un lado, y sin usarlos, por el otro) en cada una de las subtareas³ llevadas a cabo por nuestro sistema para completar el reconocimiento y la clasificación de las opiniones. En todos los casos, se obtienen mejores valores de precisión y cobertura mediante el uso de los recursos del dominio. En la figura 8.7 se representan la precisión y la cobertura en la tarea de reconocimiento de opiniones y la exactitud en la tarea de clasificación de opiniones obtenidas por los distintos *pipelines* propuestos. De nuevo, el uso de recursos del dominio permite conseguir resultados sensiblemente mejores.

³En el caso de la anotación de palabras de característica no se muestran resultados sin recursos puesto que dicha subtarea no se puede resolver sin la taxonomía de características.

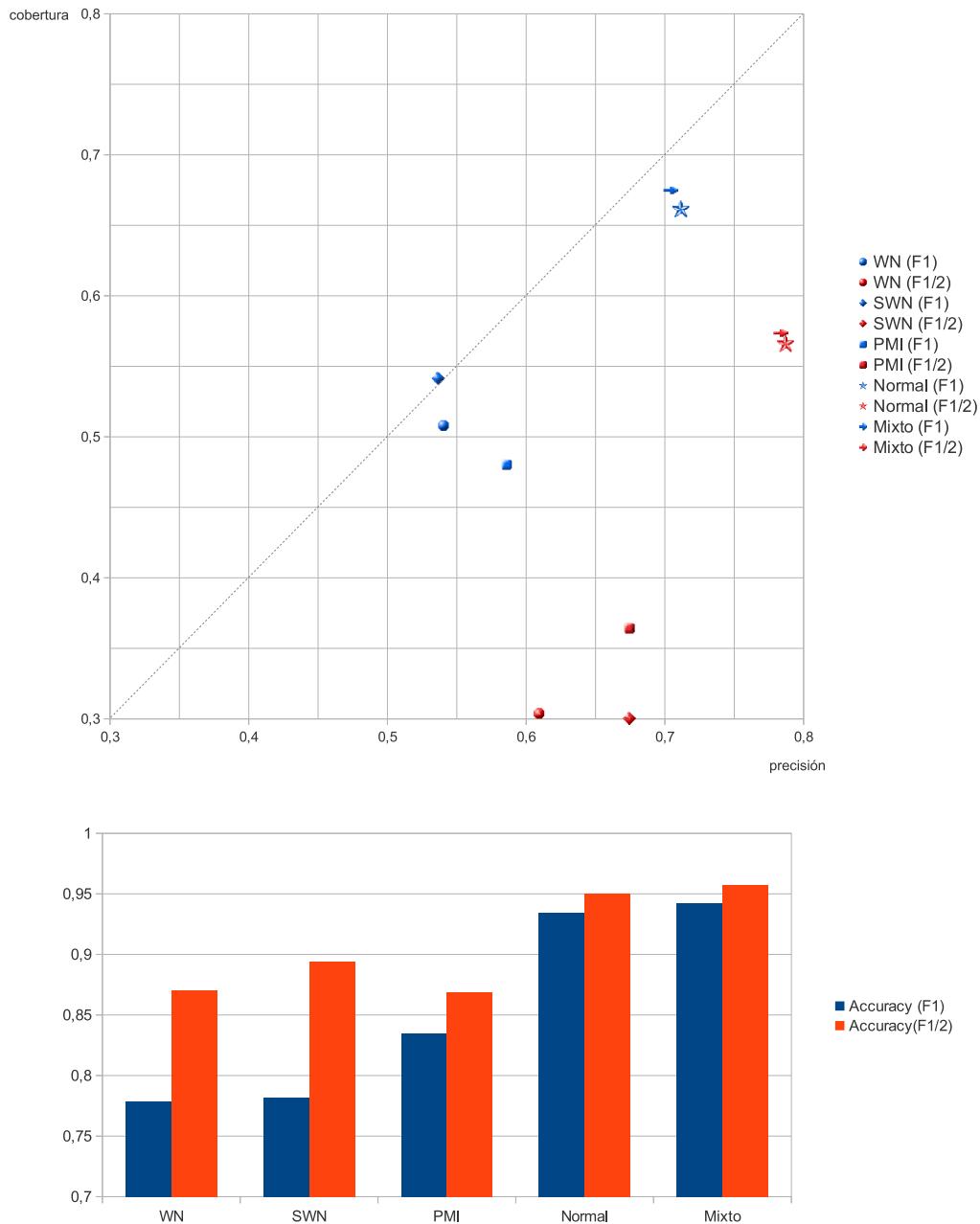


Figura 8.7: Resumen comparativo de resultados obtenidos por cada uno de los *pipelines*: precisión y cobertura (tarea *Opinion Recognition*) y *accuracy* (tarea *Opinion Classification*). Los nombres de los *pipelines* corresponden a los utilizados en la sección 8.4.1. De cada uno, se muestran los resultados obtenidos por las versiones que optimizan F_1 y $F_{\frac{1}{2}}$.

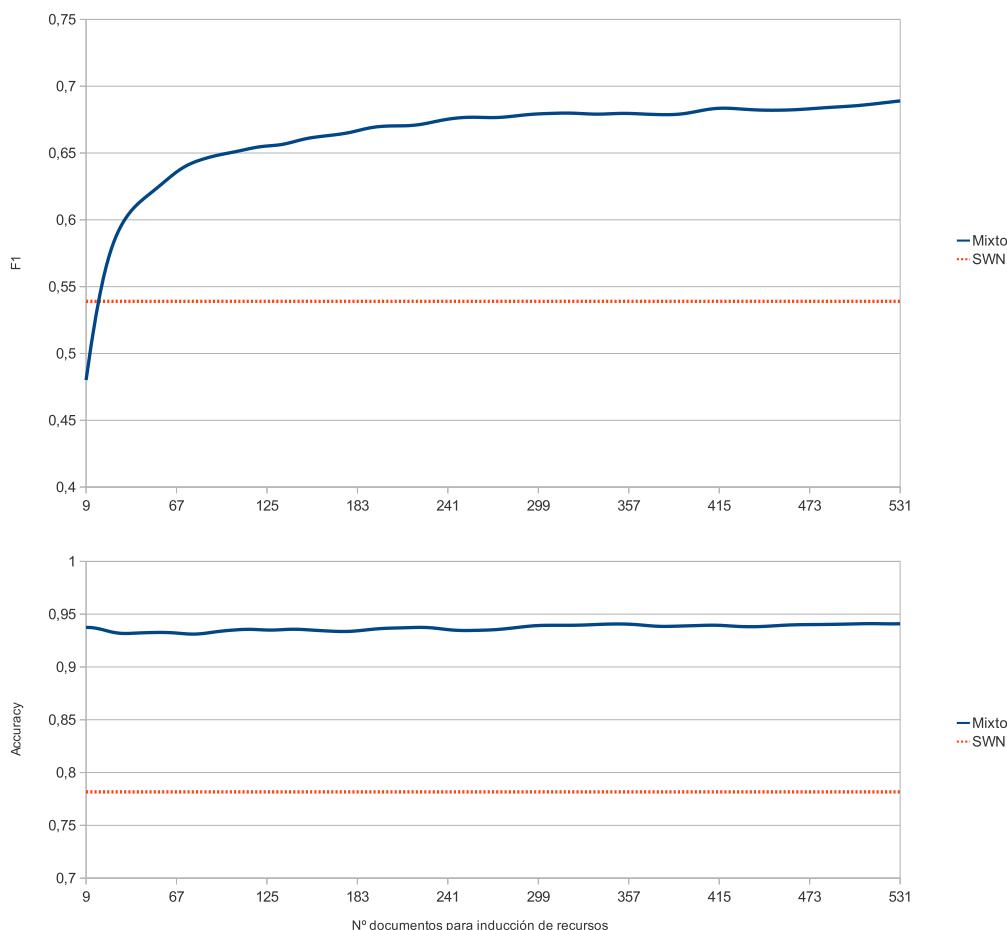


Figura 8.8: Evolución del rendimiento del sistema (*pipeline mixto*) a medida que se aumenta el número de documentos anotados utilizados en la inducción de los recursos.

Influencia del número de documentos anotados

Como ocurre en la aplicación de cualquier técnica de aprendizaje automático, a mayor número de documentos anotados, mejor calidad de los recursos generados, y consecuentemente mejores resultados del sistema de extracción. En la figura 8.8 se muestran los valores de F_1 y *accuracy* obtenidos por el *pipeline mixto*, utilizando un número creciente de documentos anotados en la etapa de inducción de los recursos. La curva obtenida parece sugerir que existe aún un margen de mejora para el sistema, a partir de

la incorporación de nuevos documentos anotados. Esto fue tenido en cuenta cuando planificamos la generación de los recursos para un nuevo dominio (hoteles), tomando la decisión de anotar un mayor número de documentos. Los resultados de la aplicación de la metodología de generación de los recursos y la evaluación del sistema de extracción de opiniones en este nuevo dominio están documentados en el capítulo 9.

En las gráficas anteriores, hemos indicado los valores obtenidos por el mejor *pipeline* ligero de recursos, poniéndose de manifiesto que, aún con pocos documentos anotados, la utilización de los componentes basados en los recursos supone una mejora considerable en los resultados.

8.5.2. Estimación de coste de la anotación y relación con la mejora del sistema

Los experimentos que hemos ido exponiendo a lo largo del capítulo demuestran que la utilización de los recursos específicos del dominio suponen una mejora considerable en los resultados obtenidos por el sistema. Sin embargo, la inducción de estos recursos sólo es posible si se dispone de un conjunto de documentos del dominio anotados. Para medir hasta qué punto este esfuerzo manual es conveniente, hemos llevado a cabo una serie de estimaciones del tiempo que requeriría la anotación de distintos conjuntos de documentos, y hemos obtenido los resultados del sistema empleando los recursos generados con esos conjuntos de documentos.

En la tabla 8.16 se muestran los resultados obtenidos por el *pipeline* mixto usando un número creciente de documentos anotados para la inducción de los recursos⁴. Teniendo en cuenta que el tiempo empleado en anotar los 599 *reviews* del dominio *headphones* fue de aproximadamente 160 horas, hemos llevado a cabo una estimación del tiempo requerido para anotar cada conjunto de documentos. Creemos que el empleo de la metodología propuesta en este trabajo de tesis está plenamente justificada, dada la mejora en los resultados obtenida por el sistema a partir de sólo unos cuantos documentos anotados: con algo más de dos horas de trabajo, se sobrepasa ampliamente la precisión en la clasificación de opiniones obtenida por cualquiera de los *pipelines* ligeros de recursos; y dedicando algunas horas más se superan sensiblemente los resultados en el reconocimiento de opiniones.

⁴Dado que usamos validación cruzada a partir de 10 particiones, hacen falta al menos diez documentos anotados para llevar a cabo los experimentos. La inducción de los recursos se lleva a cabo en este caso sobre 9 documentos en cada experimento.

Documentos anotados	Tiempo estimado de anotación	Opinion	
		Recognition	Classification
9	2,4 horas	0,48	<i>0,9348</i>
18	4,8 horas	<i>0,5561</i>	0,9377
27	7,2 horas	0,5916	0,9249
36	9,6 horas	0,6061	0,9337
45	12 horas	0,6162	0,9345
(mejor <i>pipeline</i> sin recursos)		0,5390	0,7817

Cuadro 8.16: Estimación de coste temporal y resultados del sistema para conjuntos de documentos anotados de distintos tamaños. Hemos señalado en cursiva los primeros valores de F_1 y *accuracy* que mejoran los del mejor *pipeline* sin recursos.

	Optimiza F_1			Optimiza $F_{\frac{1}{2}}$		
	p	r	F_1	p	r	$F_{\frac{1}{2}}$
Implícitas	0,6961	0,5707	0,6272	0,8173	0,4785	0,7159
Explícitas	0,7173	0,6999	0,7085	0,7759	0,6079	0,7353

Cuadro 8.17: Desglose de resultados obtenidos por el *pipeline* mixto para opiniones explícitas e implícitas (tarea *Opinion Recognition*)

8.5.3. Dificultad de las opiniones implícitas

Tal como puede observarse en la tabla 8.17, en la que se muestran los resultados de nuestro mejor *pipeline* en el reconocimiento de opiniones explícitas e implícitas medidos por separado, nuestro sistema obtiene mejores resultados en la extracción de opiniones explícitas que en la de opiniones implícitas.

Que la extracción de opiniones sobre características implícitas sea especialmente difícil no resulta sorprendente, si tenemos en cuenta que para reconocer la existencia de una opinión implícita debemos basarnos simplemente en la aparición de determinadas palabras de opinión (al menos, si seguimos el planteamiento léxico-sintáctico de nuestro sistema). Por ejemplo, la palabra “big” es un buen indicador de la característica implícita *size*, pero extraer opiniones sobre dicha característica cada vez que la observamos en un texto conduce a un gran número de opiniones incorrectas y por tanto a una baja precisión, puesto que “big” aparece en multitud de expresiones en las que no se está expresando una opinión sobre el tamaño.

8.5.4. Dificultad de determinadas características

Los resultados obtenidos por el mejor *pipeline* en el reconocimiento y clasificación de opiniones de cada una de las características de la taxonomía para *headphones* por separado (ver figura 8.9) muestran grandes diferencias entre características. Así por ejemplo, la extracción de opiniones sobre la característica *sound quality* (la característica que más aparece en las anotaciones, ver figura 6.12) obtiene un resultado de 0,7474 para F_1 y de 0,8152 para $F_{\frac{1}{2}}$, muy superiores a los resultados globales. Sin embargo, la extracción de opiniones sobre la característica *size* obtiene muy malos resultados. Creemos que estas diferencias podrían estar relacionadas con el distinto número de opiniones anotadas sobre las distintas características en el corpus, lo que da lugar a una distinta calidad de los recursos sobre según qué características. La mayor proporción de opiniones implícitas en algunas características también puede estar relacionada con las diferencias observadas (*size* es un ejemplo de característica que suele aparecer de manera implícita).

8.5.5. Precisión vs. cobertura

La arquitectura modular de nuestro sistema y la disponibilidad de parámetros de ajuste del comportamiento de los distintos componentes del mismo, permiten la construcción de sistemas de extracción de opiniones en los que prime bien la precisión de las opiniones obtenidas, bien la cobertura, o quizás un compromiso entre ambas (ver figura 8.7). Escoger una u otra opción vendrá dado por los requisitos de la aplicación concreta que se desee darle a las opiniones extraídas por el sistema: ¿se necesita una gran fiabilidad de la corrección de las opiniones obtenidas, aún cuando el número de opiniones pasadas por alto sea grande?; ¿o quizás necesitamos obtener todas las opiniones posibles, para hacer un tratamiento estadístico de las mismas, de manera que la precisión individual de las opiniones sea menos importante?.

8.5.6. Resultados finales y dificultad de la tarea

Creemos que los resultados obtenidos por el mejor *pipeline* (0,6892 de F_1 y 0,7301 de $F_{\frac{1}{2}}$ en reconocimiento de opiniones, y 0,9423 y 0,9572 de *accuracy* en la clasificación) son, cuando menos, aceptables, teniendo en cuenta la dificultad de la tarea tal como la hemos definido (ver sección 4.2.4):

- Una opinión extraída sólo es contabilizada como correcta cuando concuerda con una opinión anotada sobre la *misma característica* (téngase en cuenta el alto número de características disponibles en la taxonomía, y la posible ambigüedad entre algunas de ellas). Por ejemplo, si en el

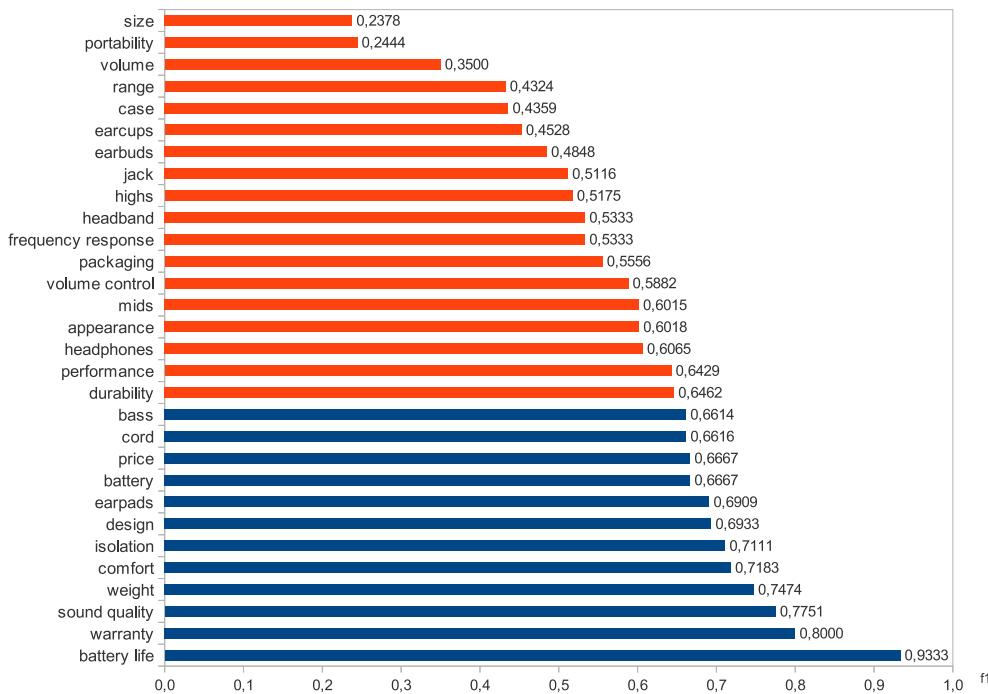


Figura 8.9: Desglose de valores de F_1 obtenidos por el *pipeline* mixto para cada característica de la taxonomía (tarea *Opinion Recognition and Classification*). En rojo se muestran aquellos resultados peores que el resultado global.

texto “*light headphones*” ha sido anotada una opinión positiva sobre la característica *weight*, y nuestro sistema extrae una opinión sobre la característica *headphones*, esta última será contabilizada como una opinión errónea (aún cuando quizás no sería así percibida por un usuario del sistema).

- Las opiniones han sido anotadas de manera atómica (ver sección 4.2.4); por ejemplo, el texto “*nice, long cord*” contiene dos opiniones anotadas sobre la misma característica. Para que la precisión y la cobertura del sistema sean iguales a 1, deben extraerse *exactamente* dos opiniones correctas. Si sólo se extrae una (en la que constarían *nice* y *long* como palabras de opinión, lo que es bastante probable), la cobertura sólo sería del 50 %.
- Muchas de las menciones a características en los documentos no corresponden a opiniones, como puede verse en el valor de precisión obtenido

en la anotación de palabras de característica (ver tabla 8.15).

- Algunas de las opiniones contenidas en los documentos no hacen referencia al objeto que está siendo analizado (se pueden referir por ejemplo a otro producto). Estas opiniones no han sido anotadas en nuestro corpus, aunque son en muchos casos extraídas por nuestro sistema.
- La calidad de los textos utilizados es bastante baja, al tratarse de textos escritos por usuarios, lo que ocasiona un gran número de errores por parte de las herramientas de análisis lingüístico en que se apoya nuestro sistema (especialmente del analizador de dependencias).

Creemos que los resultados obtenidos son buenos teniendo en cuenta el acercamiento léxico-sintáctico de nuestra propuesta : como ocurre en otras tareas del procesamiento del lenguaje natural, la utilización de herramientas lingüísticas de análisis exclusivamente léxico y sintáctico, excluyendo el uso de analizadores semánticos, razonamiento basado en conocimiento del mundo y otras tecnologías más avanzadas, establece una cota superior en los resultados alcanzables. No obstante, los resultados de precisión y cobertura son lo suficientemente altos como para que el sistema sea utilizable en la práctica, como mostramos a continuación.

8.5.7. Utilidad práctica del sistema

Las métricas de precisión y cobertura utilizadas en nuestras evaluaciones nos permiten comparar las distintas soluciones planteadas, y resumir en un valor la eficacia del sistema (por ejemplo, mediante la medida F_1). Pero, ante la visión de este único valor, cabe preguntarse: ¿son los resultados obtenidos por el sistema lo suficientemente buenos como para que las opiniones extraídas tengan utilidad práctica? La contestación a esta pregunta depende de la aplicación que vayamos a darle a la salida al sistema.

En la figura 8.10 se muestra una posible aplicación de las opiniones extraídas por nuestro sistema. Se trata de representar gráficamente un resumen de la opinión general de los usuarios de un determinado modelo de auriculares, desglosando las opiniones referidas a las distintas características del producto, y mostrando no sólo el porcentaje de opiniones positivas y negativas sino también las características que han generado más o menos comentarios. La gráfica de la figura ha sido construida contabilizando las opiniones positivas y negativas sobre cada característica, haciendo uso de las relaciones de especialización incluidas en la taxonomía; de esta forma, las opiniones sobre la característica *bass* son contabilizadas como opiniones sobre la característica *frequency response*, que a su vez son contabilizadas como opiniones sobre

la característica *sound quality*. Como resumen global de las opiniones vertidas sobre el producto, se muestra el porcentaje total de opiniones positivas y negativas.

Para responder a la pregunta sobre la utilidad práctica de nuestro sistema, hemos construido las gráficas descritas utilizando por un lado las opiniones anotadas en el corpus, y por otro las opiniones extraídas por el sistema. Aunque, evidentemente, existen diferencias, la mayor parte de ellas son irrelevantes: una persona que desee formarse una idea acerca de las bondades y defectos del producto en cuestión, sacará conclusiones parecidas ante la visión de ambas versiones del gráfico. Por ejemplo:

- La mayor parte de las opiniones sobre el producto son positivas.
- La calidad de sonido es lo más comentado por los usuarios, siendo en general positivas las opiniones al respecto. Dentro de la misma, los graves parecen ser el punto débil para algunos pocos usuarios, siendo los tonos medios y agudos buenos para todos los usuarios que han comentado acerca de ello.
- A la mayoría de usuarios les gusta la apariencia del aparato.
- El punto más débil parece ser la durabilidad o calidad de construcción: una mayoría de los usuarios tienen una mala opinión acerca de la misma.

Podemos concluir que las opiniones extraídas por el sistema son de indudable utilidad en la construcción de aplicaciones de agregación y visualización de opiniones.

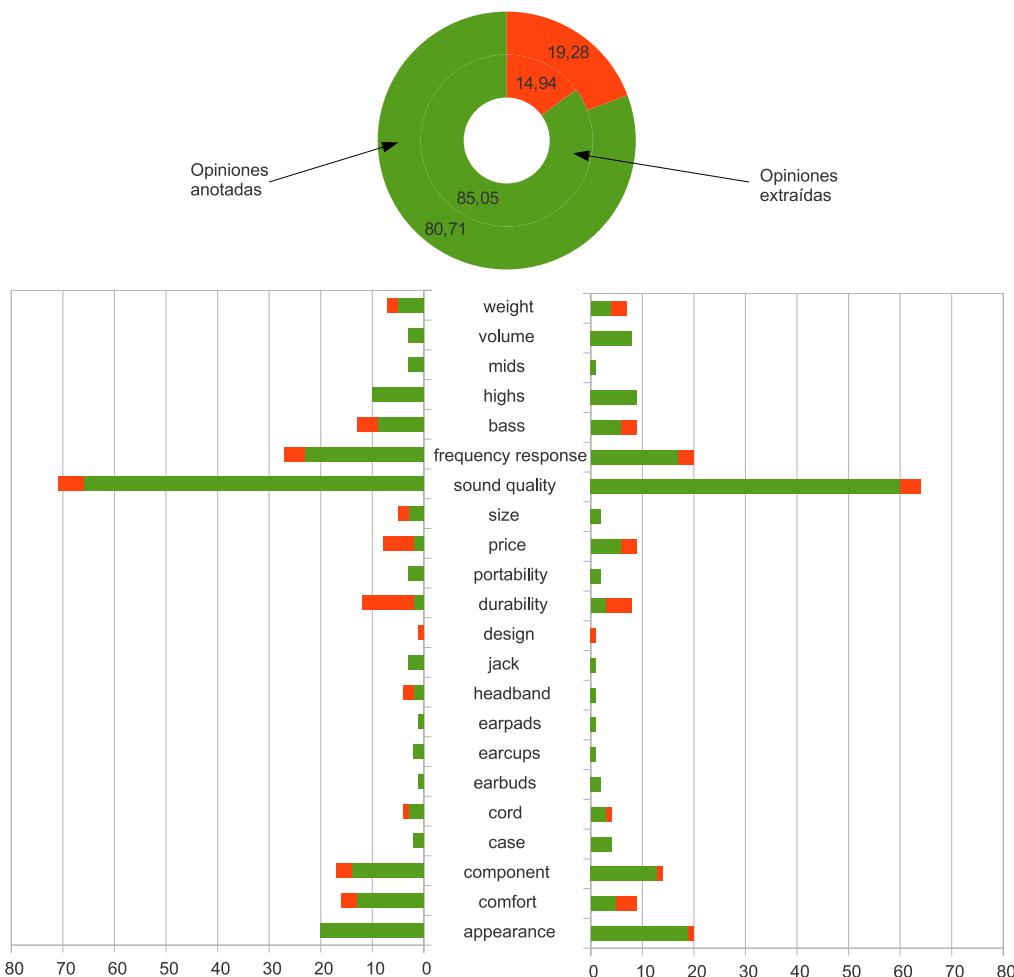


Figura 8.10: Ejemplo de agregación de opiniones para un modelo determinado de auriculares (Sony MDR-V700DJ). El gráfico de la izquierda se ha construido en base a las opiniones anotadas en el corpus; el de la derecha, a partir de las opiniones extraídas por el sistema (*pipeline mixto optimizado para F_1*). Para cada característica, se agregaron las opiniones sobre la misma o sobre alguna característica derivada (por ejemplo, las opiniones sobre *feature response* fueron utilizadas para representar la barra asociada a *sound quality*). En la parte superior, se muestran el total de opiniones positivas (en verde) y negativas (en rojo) sobre el producto anotadas en el corpus y extraídas por el sistema.

Parte V

Casos de aplicación y consideraciones finales

Capítulo 9

Casos de aplicación: hoteles y coches

Resumen: En este capítulo resumimos los resultados obtenidos de la aplicación de la metodología de generación de recursos y el sistema de extracción a dos nuevos dominios: hoteles y coches.

9.1. Introducción

Si en el caso de *headphones* escogimos dicho dominio por tratarse de un tipo de producto con una taxonomía de características no demasiado compleja y un nivel técnico de los documentos moderado, en la elección de los dos nuevos dominios ha primado un criterio práctico: se trata de dos dominios con gran potencial de aplicación, para los que existen cientos de miles de documentos en Internet. En el caso concreto de los hoteles, la extracción de opiniones se revela como una herramienta de un alto interés tanto por parte de los clientes (que están cada vez más acostumbrados a consultar las opiniones de otros usuarios antes de elegir un hotel) como por parte de las empresas (los propios hoteles, las agencias de viajes y aquellas empresas especializadas en agregar críticas de hoteles de usuarios). Al mismo tiempo, se trata de dos dominios más complejos que el empleado en los capítulos anteriores, por lo que la adaptación y evaluación del sistema en dichos dominios puede ser considerada una “prueba de fuego” sobre la capacidad de adaptación de la metodología y el sistema a nuevos dominios.

9.2. Generación de recursos

Hemos aplicado la metodología explicada en el capítulo 6 para la generación de los recursos. Comenzamos recolectando y anotando un corpus de documentos para cada uno de los dominios escogidos, junto con la definición de las taxonomías de características. Posteriormente inducimos el lexicón de opiniones, los indicadores de características implícitas y los patrones de dependencias.

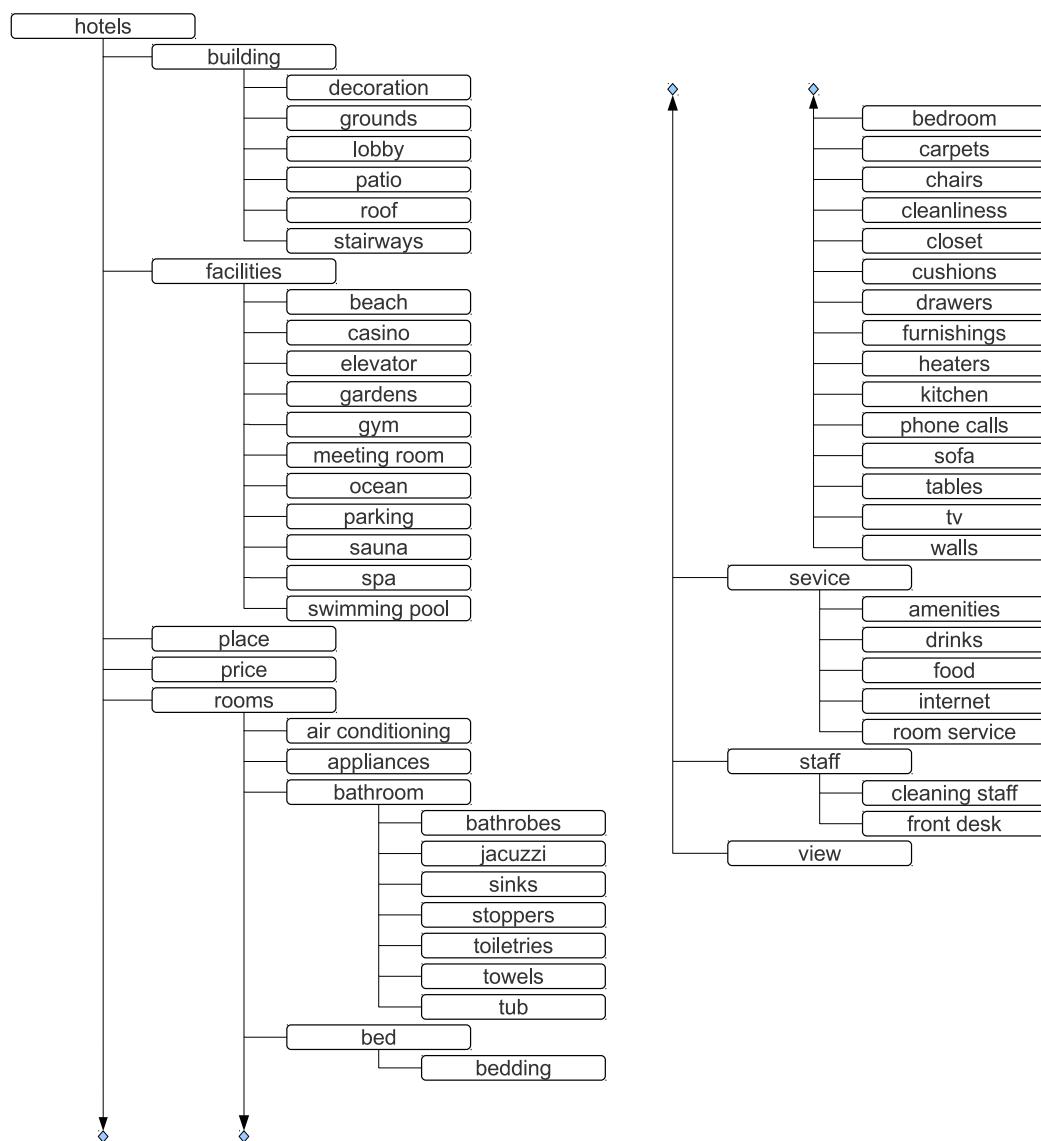
9.2.1. Corpus anotado y taxonomía de características

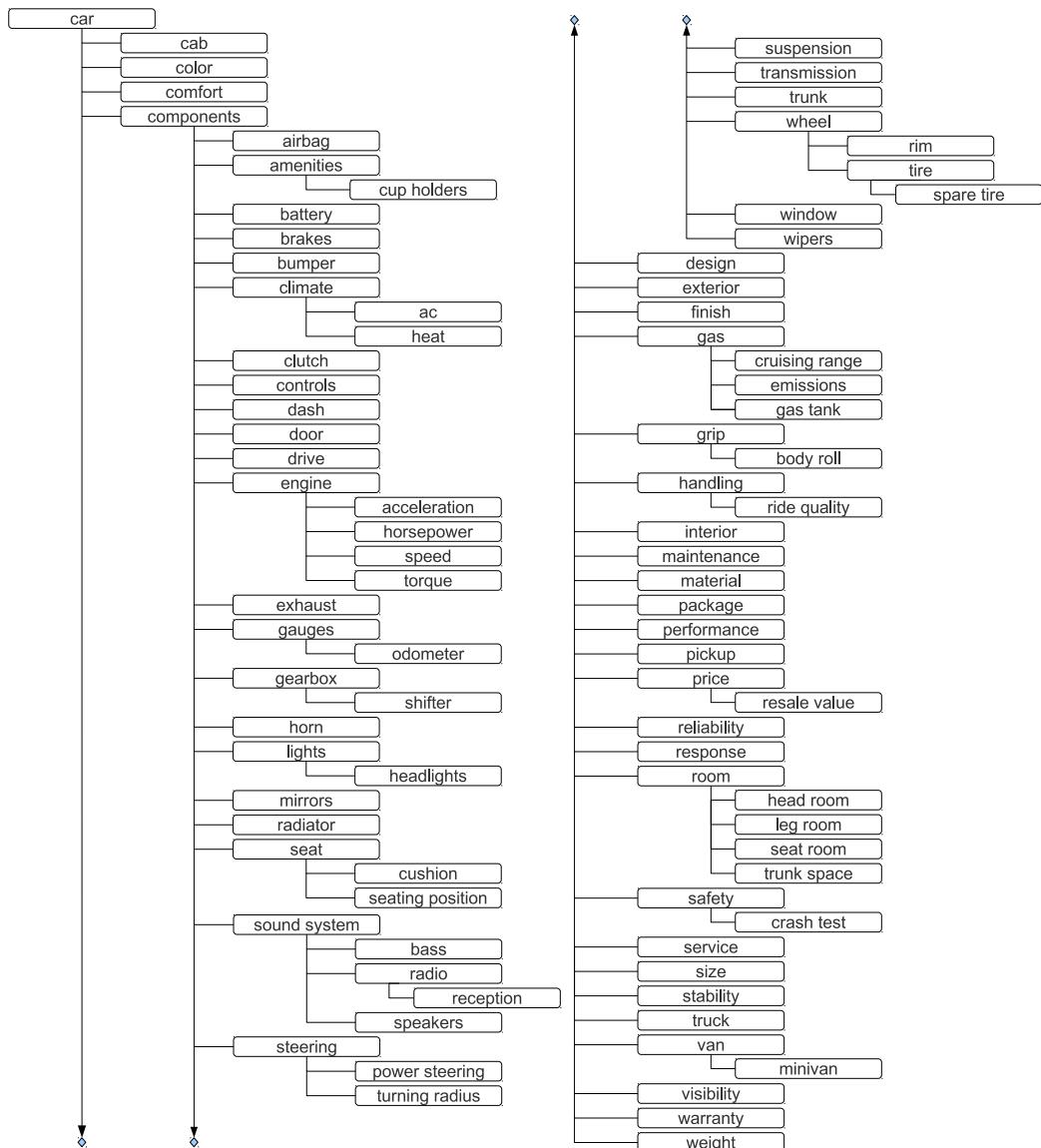
Al igual que para el dominio *headphones*, la recuperación de los documentos se llevó a cabo a partir de los contenidos de *Epinions.com*, disponibles a fecha de septiembre de 2008. Se seleccionaron 1000 documentos aleatoriamente para cada dominio, tratando de mantener una distribución uniforme entre las puntuaciones de los *reviews* seleccionados (ver sección 6.2.2). Dichos documentos fueron procesados usando las herramientas descritas en 6.2.3, y las evidencias de opinión fueron anotadas siguiendo el esquema de anotación y la herramienta de validación explicados en la sección 5.4. Con todo ello, y tras eliminar los documentos pertinentes (ver sección 6.2.2) se obtuvieron 988 documentos anotados para *hotels* y 972 para *cars*. En la tabla 9.1 se muestran algunas estadísticas relativas al corpus anotado obtenido.

	Hotels	Cars
Total de reviews	6171	23179
Reviews anotados	988	972
Palabras	631442	493459
Oraciones	33853	26307
Oraciones que contienen opiniones	7339	5989
Evidencias de opinión...	11054	8519
... con polaridad positiva/negativa	69.41 %/30.59 %	74.36 %/25.64 %
... con característica implícita/explícita	13.24 %/86.76 %	37.25 %/62.75 %

Cuadro 9.1: Estadísticas del corpus para los dominios *hotels* y *cars*

La construcción de las taxonomías de características se realizó siguiendo los pasos descritos en las secciones 6.3, 6.4 y 6.6. Las taxonomías obtenidas (figuras 9.1 y 9.2) son considerablemente más complejas que la del dominio *headphones*.

Figura 9.1: Taxonomía de características del dominio *hotels*

Figura 9.2: Taxonomía de características del dominio *cars*

9.2.2. Inducción de recursos

Una vez disponemos del corpus anotado para los nuevos dominios, aplicamos los algoritmos de inducción descritos en el capítulo 5 para obtener el lexicón de opiniones, los indicadores de característica implícita y los patrones de dependencias. En las siguientes secciones mostramos algunos datos de los recursos obtenidos a partir del conjunto total de documentos anotados disponibles; en la sección correspondiente a la evaluación del sistema, sin embargo, los recursos serán generados a partir de un subconjunto de entrenamiento y evaluados sobre el resto de documentos anotados, siguiendo el esquema de evaluación cruzada explicado en la sección 8.2.

No tratamos en este capítulo de reproducir cada uno de los experimentos llevados a cabo en el capítulo 5 referentes a las distintas opciones y parámetros de configuración de los algoritmos de inducción de los recursos. Más bien, utilizamos los algoritmos de inducción a modo de cajas negras, tal como las utilizaría alguien interesado en aplicar nuestro sistema de la manera más directa posible; el ajuste de parámetros y las decisiones acerca de las distintas opciones planteadas en el capítulo 5 se encuentran reflejadas como configuración permanente de estas herramientas.

Lexicón de opiniones

En las tablas 9.3 y 9.2 se muestran algunos datos de los lexicones inducidos para los dominios *cars* y *hotels* respectivamente, así como de las versiones expandidas mediante PolarityRank (ver sección 6.8). El número de términos hace referencia a las distintas palabras y expresiones contenidas en el lexicón; por su parte, el número de entradas se refiere a los pares (*término, característica*), para cada uno de los cuales existe una estimación de orientación semántica y de probabilidad de palabra de opinión.

El porcentaje de términos con orientación semántica ambigua¹ es mucho mayor en el dominio *hotels* que en el dominio *cars*. Esta diferencia intrínseca en la ambigüedad de las palabras de opinión utilizadas en cada dominio se verá reflejada en la precisión obtenida por los clasificadores de opiniones en ambos dominios (a favor del dominio *cars*), como se puede ver en las tablas de resultados del sistema de extracción que se muestran en la sección 9.3.2.

En cuanto al lexicón ampliado mediante PolarityRank, en ambos casos se produce un incremento notable de la ambigüedad de las estimaciones de

¹Que han recibido una estimación distinta de $-1,0$ y $1,0$, debido a que el término en cuestión ha sido observado siendo usado con polaridades opuestas en distintas evidencias de opinión, posiblemente para distintas características; ver sección 5.5.1 para una explicación más detallada.

	Original	Expandido
Número de términos ...	1186	1581 (+33.31 %)
... con orientación semántica ambigua	1.18 %	25.99 %
Número de entradas ...	4317	35610 (+824.88 %)
... con polaridad positiva/negativa	64.8 %/35.2 %	69.14 %/30.86 %
... con probabilidad igual a 1	17.07 %	2.07 %

Cuadro 9.2: Estadísticas del lexicón de opiniones para el dominio *cars*

	Original	Expandido
Número de términos ...	1257	1386 (+10.26 %)
... con orientación semántica ambigua	8.51 %	11.83 %
Número de entradas ...	5042	12065 (+139.29 %)
... con polaridad positiva/negativa	61.78 %/38.22 %	67.76 %/32.24 %
... con probabilidad igual a 1	14.48 %	6.02 %

Cuadro 9.3: Estadísticas del lexicón de opiniones para el dominio *hotels*

orientación semántica; esto es lógico por la propia naturaleza del algoritmo de *ranking* sobre grafos utilizado, ya que en la puntuación obtenida para cada nodo se reflejan en mayor o menor medida las puntuaciones de todos y cada uno de los nodos pertenecientes a la misma componente conexa del grafo. Por tanto, incluso aquellos nodos que parten con un valor no ambiguo de orientación semántica obtendrán con cierta probabilidad una estimación final ligeramente distinta a -1,0 y 1,0. A pesar de este aumento de la ambigüedad, la ampliación del lexicón permite aumentar enormemente la cobertura del mismo en ambos dominios, especialmente en el dominio *cars*. Sin duda, este mayor porcentaje de expansión se debe al mayor número de documentos no anotados disponibles para dicho dominio (23179 frente a 6171 para el dominio *hotels*).

Indicadores de opinión implícita

En las tablas 9.4 y 9.5 se muestran algunos de los indicadores de opinión implícita inducidos para los dominios *hotels* y *cars* respectivamente. En concreto, se trata de aquellos indicadores que obtuvieron una estimación de probabilidad igual a 1 (es decir, cuya aparición en el corpus de entrenamiento siempre está ligada a una opinión implícita). El alto número de indicadores del dominio *cars* está relacionado con una mayor proporción de opiniones implícitas. Es interesante el solapamiento existente en los indicadores para la característica *price*, común a ambos dominios. Esta observación da pie

a plantear la reutilización entre dominios de la información inducida para determinadas características comunes a diversos dominios.

Característica	Indicadores de opinión implícita
building	well-designed
furnishings	adequately furnished
hotel	badly decorated, cheerless, circulated, classy looking, clean-ish, comfortably appointed, disrupting, efficiently run, graceful, hi-tech, ill-kempt, imperfect, over-rated, pet-friendly, poorly staffed, puzzling, recommend, skanky, terribly managed, timid, uncared, undermaintained, wonderfully appointed
lobby	descriptive, non descriptive
place	just-central-enough
price	affordably, affordably priced, high-priced, highly-priced, low priced, mid-priced, moderately priced, over priced, priced well, reasonable priced, relatively expensive, ripoff
rooms	spotlessly kept, weird smelling
service	badly serviced
staff	short-staffed

Cuadro 9.4: Términos del recurso de indicadores de opinión implícita con probabilidad igual a 1 para el dominio *hotels*

Característica	Indicadores de opinión implícita
acceleration	accelerated easily, accelerates easily, accelerates masterfully, accelerates responsively, masterfully
brakes	brake easily, stops quickly, stops surely
car	behaved, behaves unexpectedly, cheaply built, cheaply constructed, controlable, darty, depreciate rapidly, depressing, disillusioned, disappointed, doing horrible, easily controllable, enthralling, fits nicely, incredibly disappointed, inefficient, lovable, luminous, mind-boggled, obsolete, parks easy, poorly assembled, poorly built, poorly constructed, ran fine, runs fine, runs nicely, runs quietly, runs right, runs strong, sassy, solidly-constructed, strongly built, sturdily, sturdily built, tightly constructed, trouble-prone, under-contented, under-engineered, unfaithful, v6...fun, well assembled, well behaved, well liked, well-built, well-engineered, well-equipped, wiggly
comfort	ergonomically fails, ergonomically flawed, ergonomically friendly
controls	operate easily
design	best-looking, blandly styled, butt-ugliest, clean looking, corners flat, cute looking, designed bad, horrendous looking, impressive looking, inadequately, inadequately designed, look awful, look dated, look muscular, look stunning, looked out-of-place, looks awful, looks beautiful, looks bland, looks sexy, looks stylish, looks superb, looks updated, looks weak, nicest looking, odd-shaped, plain looking, sharp looking, sophisticated looking, sportiest looking, tackiest, tackiest looking, tightly designed, worst-looking
drive	tracked well
gas	fuel-efficient, gas friendly
handling	controllably, difficult to handle, handle poorly, handled beautifully, handled controllably, handled poorly, handled predictably, handles awesome, handles beautifully, handles easily, handles excellently, handles flawlessly, handles lousy, handles poorly, handles responsively, handles smoothly, handles terribly, maneuvers well
material	well textured
performance	perform well, performed excellently, performed good, performed strongly, performed well, performs admirably, underperforming
price	affordably, affordably priced, competitively, competitively priced, competitively-priced, cost-effective, fairly priced, obtainable, outrageously, outrageously priced, pricy, reasonably-priced
reliability	all-reliable, ultra-reliable, undependable
response	extra-responsive, responds accurately, responds nicely, responds quickly
ride quality	drives decent, drives nicely, drove easily, lumpy, ride harshly, rides nicely, rides safely, rides solidly, rode quietly, rode smoothly, rough riding, undrivable
safety	crash-safe, insecure
speed	moves quick
steering	hard to steer, steers, steers great
turning radius	turns well
weight	lightweight

Cuadro 9.5: Términos del recurso de indicadores de opinión implícita con probabilidad igual a 1 para el dominio *cars*

Patrones de dependencias

En las tablas 9.6 y 9.7 se muestran el número de patrones de cada tipo², independientes de la característica, inducidos para cada dominio, junto con la precisión y cobertura acumuladas del conjunto total de patrones. Nótese la alta cobertura obtenida en ambos casos, lo que nos garantiza que el sistema podrá acceder a prácticamente todas las palabras de opinión existentes. Por otro lado, la diferencia en el número de patrones entre ambos dominios para los patrones de tipo 1, 2 y 3, nos informa de la mayor dificultad sintáctica de las opiniones en el dominio *hotels* frente al dominio *cars*. Es posible que esto se deba al carácter narrativo más acentuado en el dominio *hotels*, lo que da pie a oraciones de más longitud y con más complejidad sintáctica. Por otro lado, se observa un número similar de patrones de tipo 4 y 5 para ambos dominios (los patrones relacionados con las expresiones de negación y de polaridad dominante), lo que parece indicar que dichos patrones podrían ser reutilizados entre los distintos dominios.

Tipo patrón	Nº patrones	Acc.Prec.	Acc.Recall
1	1837	0.4732	1.0
2	1337	0.4707	1.0
3	137	0.0724	0.9689
4	62	0.8108	1.0
5	25	0.8051	1.0

Cuadro 9.6: Estadísticas de los patrones de dependencias inducidos para el dominio *hotels*

Tipo patrón	Nº patrones	Acc.Prec.	Acc.Recall
1	1221	0.4105	0.9996
2	799	0.4188	0.9996
3	209	0.1297	0.985
4	65	0.7758	1.0
5	24	0.7142	1.0

Cuadro 9.7: Estadísticas de los patrones de dependencias inducidos para el dominio *cars*

²Para una descripción de los cinco tipos de patrones inducidos, consultar la sección 5.7.1

9.3. Evaluación del sistema de extracción

En esta sección mostramos los resultados obtenidos por el sistema TOES usando los *pipelines* definidos en la sección 8.4. Hemos utilizado las configuraciones de los componentes tal cual fueron fijadas para el dominio *headphones*.

9.3.1. Evaluación de *pipelines* ligeros de recursos

En las tablas 9.8 y 9.9 se muestran los resultados obtenidos por los *pipelines* ligeros de recursos basados en WordNet, SentiWordNet y PMI (ver figura 8.5), en sus versiones optimizadas para F_1 y $F_{\frac{1}{2}}$, para los dominios *hotels* y *cars*. Los resultados son similares a los obtenidos con los mismos *pipelines* para el dominio *headphones*; es significativa la menor precisión obtenida en la clasificación de las opiniones para el dominio *hotels*, claramente relacionada con la mayor ambigüedad de los términos de opinión del dominio (ver tabla 9.3). En cuanto a la tarea de reconocimiento de opiniones, los resultados obtenidos para el dominio *cars* son peores que los de *hotels*, lo que nos indica la mayor dificultad del dominio (entre otras cosas, por el mayor número de características de la taxonomía).

Pipeline	<i>Hotels</i>				<i>Cars</i>			
	Opinion Recognition			Opinion Classif.	Opinion Recognition			Opinion Classif.
	p	r	F_1	a	p	r	F_1	a
WN	0,4811	0,6012	0,5347	0,743	0,4742	0,5433	0,5064	0,7717
SWN	0,5126	0,6137	0,5586	0,7521	0,4854	0,5536	0,5173	0,7577
PMI	0,5136	0,5392	0,5261	0,7922	0,4874	0,4919	0,4896	0,8233

Cuadro 9.8: Resultados obtenidos por los *pipelines* ligeros de recursos para los dominios *hotels* y *cars* (optimizados para F_1)

Pipeline	<i>Hotels</i>				<i>Cars</i>			
	Opinion Recognition			Opinion Classif.	Opinion Recognition			Opinion Classif.
	p	r	$F_{\frac{1}{2}}$	a	p	r	$F_{\frac{1}{2}}$	a
WN	0,5854	0,2426	0,4565	0,8503	0,6224	0,2395	0,4716	0,8853
SWN	0,7104	0,2853	0,5473	0,8731	0,6709	0,2590	0,5090	0,8972
PMI	0,5923	0,4092	0,5437	0,8323	0,5534	0,3800	0,5071	0,8611

Cuadro 9.9: Resultados obtenidos por los *pipelines* ligeros de recursos para los dominios *hotels* y *cars* (optimizados para $F_{\frac{1}{2}}$)

9.3.2. Evaluación de *pipelines* basados en recursos

En las tablas 9.10 y 9.11 se muestran los resultados obtenidos por los *pipelines* basados en recursos. Al igual que ocurrió en el dominio *headphones*, los resultados obtenidos son muy superiores a los obtenidos por los *pipelines* ligeros de recursos, lo que demuestra la significativa aportación de los recursos específicos del dominio en la solución de la tarea. Es especialmente remarcable el hecho de que el dominio *cars* consigue los mejores resultados, cuando a priori se trata de un dominio más complejo (nótese el mayor número de características en la taxonomía, o los peores resultados obtenidos por los *pipelines* ligeros de recursos). Sin duda influye en este hecho la disponibilidad de un mayor conjunto de documentos no anotados para dicho dominio, utilizados por el algoritmo de ampliación del lexicón de opiniones.

La utilización del *pipeline* mixto, que añade un segundo clasificador de opiniones basado en *PMI* (ver figura 8.6), no supone una ventaja demasiado clara en los dominios estudiados, limitándose a ligeras mejoras en algunos casos, e incluso empeorando también ligeramente en alguna ocasión.

Pipeline	<i>Hotels</i>			<i>Cars</i>			Opinion Classif.	
	Opinion Recognition			Opinion Classif.	Opinion Recognition			
	p	r	F_1		p	r	F_1	
Normal	0,6782	0,7389	0,7073	0,9116	0,7169	0,7296	0,7232	0,928
Mixto	0,6763	0,7414	0,7073	0,9122	0,7162	0,7319	0,724	0,9283

Cuadro 9.10: Resultados obtenidos por los *pipelines* basados en recursos para los dominios *hotels* y *cars* (optimizados para F_1)

Pipeline	<i>Hotels</i>			<i>Cars</i>			Opinion Classif.	
	Opinion Recognition			Opinion Classif.	Opinion Recognition			
	p	r	$F_{\frac{1}{2}}$		p	r	$F_{\frac{1}{2}}$	
Normal	0,7673	0,584	0,722	0,9366	0,7836	0,609	0,7411	0,95
Mixto	0,7668	0,5803	0,7205	0,9386	0,7821	0,6124	0,7411	0,9505

Cuadro 9.11: Resultados obtenidos por los *pipelines* basados en recursos para los dominios *hotels* y *cars* (optimizados para $F_{\frac{1}{2}}$)

9.4. Resumen de resultados

A modo de resumen, en las tablas 9.12 y 9.13 se muestran los valores de F_1 y $F_{\frac{1}{2}}$ de la tarea de reconocimiento de opiniones y los valores de *accuracy* de la tarea de clasificación de opiniones obtenidos para cada uno de los dominios con los que hemos trabajado: *headphones*, *hotels* y *cars*. Las principales conclusiones que sacamos a la vista de los resultados son las siguientes:

- La utilización de los recursos específicos del dominio supone una ventaja clara en cualquiera de los dominios. A mayor complejidad de los dominios, mayor es el aporte de los recursos.
- La disponibilidad de un gran número de documentos sin anotar supone una ventaja de cara a la aplicación del algoritmo de ampliación del lexicón de opiniones.
- Nuestro sistema permite el reconocimiento de opiniones con valores de F_1 en torno a 0,7 y de $F_{\frac{1}{2}}$ en torno a 0,73, y la correcta clasificación de entre el 91 y el 96 por ciento de las opiniones extraídas.

<i>Pipeline</i>	<i>Headphones</i>		<i>Hotels</i>		<i>Cars</i>	
	F_1	a.	F_1	a.	F_1	a.
WN	0.5237	0.7785	0.5347	0.7430	0.5064	0.7717
SWN	0.5390	0.7817	0.5586	0.7521	0.5173	0.7577
PMI	0.5279	0.8345	0.5261	0.7922	0.4896	0.8233
Normal	0.6857	0.9345	0.7073	0.9116	0.7232	0.928
Mixto	0.6892	0.9423	0.7073	0.9122	0.724	0.9283

Cuadro 9.12: Resumen de resultados del sistema TOES en los distintos dominios (*pipelines* optimizados para F_1)

<i>Pipeline</i>	<i>Headphones</i>		<i>Hotels</i>		<i>Cars</i>	
	$F_{\frac{1}{2}}$	a.	$F_{\frac{1}{2}}$	a.	$F_{\frac{1}{2}}$	a.
WN	0.5073	0.8706	0.4565	0.8503	0.4716	0.8853
SWN	0.5405	0.8940	0.5473	0.8731	0.5090	0.8972
PMI	0.5763	0.8688	0.5437	0.8323	0.5071	0.8611
Normal	0.7300	0.9503	0.7220	0.9366	0.7411	0.95
Mixto	0.7301	0.9572	0.7205	0.9386	0.7411	0.9505

Cuadro 9.13: Resumen de resultados del sistema TOES en los distintos dominios (*pipelines* optimizados para $F_{\frac{1}{2}}$)

Capítulo 10

Conclusiones

El interés en las Tecnologías del Lenguaje y el Procesamiento del Lenguaje Natural se ha venido incrementando espectacularmente en los últimos tiempos, en gran parte debido a las nuevas necesidades de tratamiento de la información derivadas de la creciente implantación de Internet en la sociedad actual. Siendo el lenguaje escrito el principal vehículo de la información contenida en este medio, tecnologías como la recuperación y la extracción de información, el análisis y representación semántica, y otras técnicas encaminadas, en fin, al tratamiento de información no estructurada, cobran especial protagonismo. En este contexto, y dado el carácter subjetivo de una gran parte de los contenidos de la Web (especialmente, de aquellos generados por usuarios a través de los nuevos servicios de la Web 2.0), numerosos investigadores han señalado la importancia del estudio de las opiniones, los sentimientos y otros fenómenos subjetivos en los textos.

El Análisis del Sentimiento o Minería de Opiniones es una nueva subdisciplina que se ocupa de este tipo de tareas. Algunos de los problemas en cuestión vienen siendo estudiados desde hace bastante tiempo (por ejemplo, la caracterización del lenguaje subjetivo), aunque han experimentado un creciente interés por parte de la comunidad en los últimos años. Muchos otros problemas han sido definidos recientemente, tales como la clasificación de documentos en base a la opinión expresada en los mismos, la clasificación de documentos según la tendencia política, la detección de emociones, y la extracción de opiniones. Esta última tarea, consistente en la extracción de representaciones estructuradas de opiniones individuales, desde una perspectiva cercana a la extracción de información, posee un gran potencial práctico, puesto que permite el tratamiento estadístico de la ingente cantidad de opiniones expresadas continuamente en diversas fuentes de Internet por los usuarios, acerca de multitud de temáticas. Las implicaciones prácticas a nivel político, económico y sociológico son evidentes.

En esta tesis, hemos afrontado la construcción de un sistema de extracción de opiniones sobre características. Esto es, para cada opinión, se identifica no sólo la polaridad positiva o negativa de la misma, sino también la característica concreta del objeto analizado sobre la que se vuelca la opinión. Son dos las principales características diferenciadoras con respecto a otras propuestas de la bibliografía. En primer lugar, en la definición de la tarea participa como pieza fundamental una representación semántica de las características del objeto, de manera que se pretende extraer opiniones sobre las características incluidas en una taxonomía, decidiendo adecuadamente a qué elemento de la taxonomía debe asociarse cada opinión extraída. En otros trabajos, la identificación de la característica se hace a nivel léxico, indicando simplemente las palabras del texto que identifican a la opinión (con los consiguientes problemas de ambigüedad y sinonimia, entre otros inconvenientes). En segundo lugar, nos centramos en la construcción de sistemas de extracción altamente adaptados a un dominio concreto, en lugar de en la construcción de sistemas genéricos de extracción de opiniones.

Con estas premisas, las principales aportaciones expuestas en la presente memoria son las siguientes:

1. La definición de un conjunto de **recursos** de apoyo a la tarea de extracción de opiniones sobre características (ver capítulo 5). Estos recursos, en su mayoría dependientes del dominio, incluyen la taxonomía de características, los indicadores de características implícitas, el lexicón de opiniones, los patrones de dependencias y las listas de expresiones especiales.
2. La definición de una **metodología** para la generación de los recursos anteriores, incluyendo un flujo de trabajo y un conjunto de herramientas y algoritmos que minimizan la participación manual en el proceso (ver capítulo 6). Hemos propuesto un algoritmo semiautomático de *bootstrapping* para la extracción de palabras de característica, como paso previo a la construcción de la taxonomía de características, y algoritmos que permiten realizar la inducción de los recursos con una mínima participación manual, a partir de un conjunto de documentos anotados. Hemos especificado el esquema de anotación necesario, y hemos diseñado una herramienta de validación de las anotaciones que permite maximizar la calidad de las mismas.
3. El diseño de un **sistema** de extracción sobre características modular y adaptable al dominio, al que hemos denominado TOES (*Taxonomy-based Opinion Extraction System*, ver capítulo 7). El carácter genérico

del mismo y la definición de una serie de componentes abstractos, encargados cada uno de ellos de un subproblema determinado, permiten la construcción bien de sistemas independientes del dominio, a partir de componentes que implementan las principales técnicas utilizadas en el estado del arte, o bien de sistemas adaptados al dominio, mediante la utilización en este último caso de los recursos de apoyo anteriores. Los componentes propuestos son parametrizables, de manera que su funcionamiento puede ajustarse para conseguir distintos compromisos entre la precisión y la cobertura del sistema.

Tanto el sistema de extracción como las herramientas descritas en la metodología han sido implementadas y utilizadas en diversos experimentos encaminados a probar la validez de nuestras hipótesis iniciales. Algunas de las conclusiones principales que hemos sacado de los resultados obtenidos son las siguientes:

1. La consideración del dominio de aplicación en la tarea de extracción de opiniones sobre características permite obtener mejores resultados de precisión y cobertura (ver sección 8.5.1). Así, el sistema construido a partir de los componentes independientes del dominio obtiene en todos los casos peores resultados que el sistema construido a partir de los componentes basados en los recursos específicos del dominio.
2. El esfuerzo manual necesario para la obtención de los recursos, que se concentra principalmente en la anotación del conjunto de documentos de opinión, es lo suficientemente pequeño como para compensar el incremento en la calidad de las salidas obtenidas por el sistema (ver sección 8.5.2).
3. Las opiniones sobre características vinculadas a la taxonomía obtenidas por nuestro sistema son de una clara utilidad práctica en aplicaciones de agregación y visualización de opiniones (ver sección 8.5.7).

Además de las aportaciones anteriormente expuestas, hemos obtenido algunos algoritmos y métodos que son aplicables en contextos distintos a la extracción de opiniones, y en algún caso incluso en contextos distintos al procesamiento del lenguaje natural:

- El algoritmo **PolarityRank**, un algoritmo de *ranking* sobre grafos, inspirado en PageRank, y con la particularidad de que trabaja sobre grafos con aristas de pesos de negativos. Al igual que PageRank permite asignar puntuaciones a los nodos de una red en función de las recomendaciones entre los mismos representadas por las aristas, PolarityRank

permite además modelar recomendaciones negativas entre los nodos, mediante aristas de pesos negativos. El algoritmo puede utilizarse en aquellos problemas de propagación de información en los que se disponga de información acerca de algunas de las entidades participantes en el problema, y se conozcan las similitudes y diferencias existentes entre todas las entidades. En nuestro grupo de investigación, hemos aplicado este algoritmo a tareas como la selección de instancias en minería de datos (Vallejo et al, 2010), el cálculo de la confianza y reputación en redes sociales (Ortega et al, 2011) y la detección de *spam* (Ortega et al, 2010). La justificación algebraica y la demostración de la convergencia del algoritmo se encuentran desarrolladas en el apéndice B.

- Un método de expansión automática de lexícones de opinión, basado en la representación en forma de grafo de los términos participantes en construcciones conjuntivas, la utilización como semilla de los valores de orientación semántica contenidos en un lexicón de opiniones que se desea expandir, y la aplicación al grafo del algoritmo de PolarityRank (ver sección 6.8). De manera independiente al trabajo aquí expuesto, hemos llevado a cabo experimentos aplicando el método de expansión a pequeños lexícones de opinión, inducidos a partir de una decena de documentos anotados, consiguiendo incrementos de la cobertura de los lexícones de entre 0,28 y 0,48, sin apenas perjuicio de la precisión (para una descripción más detallada de los experimentos y los resultados, ver (Cruz et al, 2011a)).

A partir del desarrollo de la investigación expuesta en esta memoria, hemos obtenido los siguientes resultados adicionales a las aportaciones teóricas:

- Un corpus de *reviews* de productos, anotado a nivel de características según nuestro esquema de anotación. El corpus recoge documentos de tres dominios distintos (*headphones*, *hotels* y *cars*), incluyendo para cada dominio la taxonomía de características utilizada en las anotaciones. El corpus está disponible para su utilización por parte de la comunidad investigadora¹.
- El sistema de extracción de opiniones sobre características TOES. Actualmente disponemos de una implementación completa del mismo, y estamos desarrollando una demo técnica vía web. Tenemos intención de implementar y alojar un servicio web que pueda ser utilizado públicamente para extraer opiniones de las fuentes que se especifiquen.

¹<http://www.lsi.us.es/~fermin/index.php/Datasets>

- Artículos en distintos congresos y publicaciones, algunos relacionados con la propuesta central de la tesis (Cruz et al, 2010, 2009c, 2011b,c), y otros con aportaciones relacionadas (Cruz et al, 2011a, 2009a,b, 2008; Ortega et al, 2011, 2010). Actualmente tenemos un artículo bajo revisión en la revista *Information Processing & Management*, centrado en el algoritmo PolarityRank.

Bibliografía

- Abbasi A (2007) Affect intensity analysis of dark web forums. In: Proceedings of Intelligence and Security Informatics (ISI), pp 282–288
- Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17:734–749, DOI <http://doi.ieeecomputersociety.org/10.1109/TKDE.2005.99>
- Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.7506>
- Alm CO, Roth D, Sproat R (2005) Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)
- Amatriain X, Pujol J, Oliver N (2009) I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In: Houben GJ, McCalla G, Pianesi F, Zancanaro M (eds) *User Modeling, Adaptation, and Personalization*, vol 5535, Springer Berlin Heidelberg, Berlin, Heidelberg, chap 24, pp 247–258, DOI 10.1007/978-3-642-02247-0_24, URL http://dx.doi.org/10.1007/978-3-642-02247-0_24
- Andreevskaia A, Bergler S (2006) Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)
- Atserias J, Casas B, Comelles E, González M, Padró L, Padró M (2006) Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In: Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA, Genoa, Italy, <http://www.lsi.upc.edu/nlp/freeling>

- Baccianella AES, Sebastiani F (2010) Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
- Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta
- Balahur A, Montoyo A (2008a) A feature dependent method for opinion mining and classification. In: Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on, pp 1 –7, DOI 10.1109/NLPKE.2008.4906796
- Balahur A, Montoyo A (2008b) Multilingual feature-driven opinion extraction and summarization from customer reviews. In: Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, Springer-Verlag, Berlin, Heidelberg, NLDB '08, pp 345–346, DOI http://dx.doi.org/10.1007/978-3-540-69858-6_39, URL http://dx.doi.org/10.1007/978-3-540-69858-6_39
- Bansal M, Cardie C, Lee L (2008) The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In: Proceedings of the International Conference on Computational Linguistics (COLING), poster paper
- Battistella EL (1990) Markedness : the evaluative superstructure of language / Edwin L. Battistella. State University of New York Press, Albany :
- Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis G, Reynar J (2008) Building a sentiment summarizer for local service reviews. In: NLPIX
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022
- Boldrini E, Balahur A, Martínez-Barco P, Montoyo A (2010) Emotiblog: a finer-grained and more precise learning of subjectivity expression mo-

- dels. In: Proceedings of the Fourth Linguistic Annotation Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, LAW IV '10, pp 1–10, URL <http://portal.acm.org/citation.cfm?id=1868720.1868721>
- Brants T (????) Tnt – a statistical part-of-speech tagging. URL <http://www.coli.uni-saarland.de/~{}thorsten/tnt/>
- Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '10, pp 804–812, URL <http://portal.acm.org/citation.cfm?id=1857999.1858121>
- Cardie C, Wiebe J, Wilson T, Litman D (2003) Combining low-level and summary representations of opinions for multi-perspective question answering. In: Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, pp 20–27
- Carenini G, Ng RT, Zwart E (2005) Extracting knowledge from evaluative text. In: Proceedings of the 3rd international conference on Knowledge capture, ACM, New York, NY, USA, K-CAP '05, pp 11–18, DOI <http://doi.acm.org/10.1145/1088622.1088626>, URL <http://doi.acm.org/10.1145/1088622.1088626>
- Carenini G, Ng RT, Pauls A (2006) Interactive multimedia summaries of evaluative text. In: Proceedings of Intelligent User Interfaces (IUI), ACM Press, pp 124–131
- Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G (2007) Language resources and linguistic theory: Typology, second language acquisition, English linguistics., Franco Angeli Editore, Milano, IT, chap Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. URL <http://www.unipv.it/wnop/#CITECerini07>
- Church KW, Hanks P (1989) Word association norms, mutual information, and lexicography
- Cilibrasi R, Vitanyi PMB (2005) Automatic meaning discovery using google. URL <http://homepages.cwi.nl/~paulv/papers/amdig.pdf>, v2

- Cruz F, Troyano JA, Enríquez de Salamanca F, Ortega FJ (2008) Clasificación de documentos basada en la opinión: Experimentos con un corpus de críticas de cine en español. In: Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural, vol 41, pp 73–80
- Cruz F, Troyano JA, Ortega FJ, Enríquez de Salamanca F (2009a) The *italica* system at tac 2008 opinion summarization task. In: Proceedings of the Textual Analysis Conference
- Cruz F, Troyano JA, Ortega FJ, Vallejo CG (2009b) Inducción de un lexicón de opinión orientado al dominio. In: Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural
- Cruz FL, Troyano JA, Ortega FJ, Enríquez F (2009c) Hacia una metodología para la construcción de sistemas de extracción de opiniones orientados al dominio. In: 1st Workshop on Opinion Mining And Sentiment Analysis
- Cruz FL, Troyano JA, Ortega FJ, Vallejo CG (2009d) Inducción de un Lexicón de Opinión Orientado al Dominio. Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural 43:5–12
- Cruz FL, Troyano JA, Enríquez F, Ortega J, GVallejo C (2010) A knowledge-rich approach to feature-based opinion extraction from product reviews. In: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, ACM, pp 13–20
- Cruz FL, Troyano JA, Ortega FJ, Enríquez F (2011a) Automatic expansion of feature-level opinion lexicons. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), Association for Computational Linguistics, Portland, Oregon, pp 125–131, URL <http://www.aclweb.org/anthology/W11-1716>
- Cruz FL, Troyano JA, Ortega FJ, Enríquez F (2011b) Extracción de opiniones sobre características adaptable al dominio. IV Jornadas de la Red Temática en Tratamiento de la Información Multilingüe y Multimodal URL http://sinai.ujaen.es/timm/jornadas4/timm2011_submission_6.pdf
- Cruz FL, Troyano JA, Ortega FJ, Enríquez F (2011c) TOES: A taxonomy-based opinion extraction system. In: Proceedings of the 16th Conference on Applications of Natural Language to Information Systems, Springer

- Das S, Chen M (2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)
- Dave K, Lawrence S, Pennock DM (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of WWW, pp 519–528
- Ding X, Liu B, Yu PS (2008a) A holistic lexicon-based approach to opinion mining. In: Proceedings of the Conference on Web Search and Web Data Mining (WSDM)
- Ding X, Liu B, Yu PS (2008b) A holistic lexicon-based approach to opinion mining. In: WSDM '08: Proceedings of the international conference on Web search and web data mining, ACM, New York, NY, USA, pp 231–240, DOI <http://doi.acm.org/10.1145/1341531.1341561>
- Dini L, Mazzini G (2002) Opinion classification through information extraction. In: Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining), pp 299–310
- Ekman P (1982) Emotion in the Human Face, 2nd edn. Cambridge University Press
- Esuli A, Sebastiani F (2005) Determining the semantic orientation of terms through gloss analysis. In: Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)
- Esuli A, Sebastiani F (2006a) Determining term subjectivity and term orientation for opinion mining. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)
- Esuli A, Sebastiani F (2006b) SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of Language Resources and Evaluation (LREC)
- Esuli A, Sebastiani F (2007) Pageranking wordnet synsets: An application to opinion mining. In: Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, pp 424–431, URL <http://acl.ldc.upenn.edu/P/P07/P07-1054.pdf>

- Etzioni O, Cafarella M, Downey D, Kok S, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A (2004) Web-scale information extraction in knowitall: (preliminary results). In: Proceedings of the 13th international conference on World Wide Web, ACM, New York, NY, USA, WWW '04, pp 100–110, DOI <http://doi.acm.org/10.1145/988672.988687>, URL <http://doi.acm.org/10.1145/988672.988687>
- Fagin R, Kumar R, Mahdian M, Sivakumar D, Vee E (2004) Comparing and aggregating rankings with ties. In: PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, New York, NY, USA, pp 47–58, DOI <http://doi.acm.org/10.1145/1055558.1055568>
- Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. MIT Press
- Gamon M, Aue A, Corston-Oliver S, Ringger E (2005) Pulse: Mining customer opinions from free text. In: Proceedings of the International Symposium on Intelligent Data Analysis (IDA), no. 3646 in Lecture Notes in Computer Science, pp 121–132
- Ganu G, Elhadad N, Marian A (2009) Beyond the stars: Improving rating predictions using review text content. In: WebDB, URL <http://dblp.uni-trier.de/db/conf/webdb/webdb2009.html#GanuEM09>
- Golbeck J (2006) Generating Predictive Movie Recommendations from Trust in Social Networks. In: Stølen K, Winsborough W, Martinelli F, Massacci F (eds) Trust Management, Lecture Notes in Computer Science, vol 3986, Springer Berlin / Heidelberg, Berlin, Heidelberg, chap 8, pp 93–104, DOI [10.1007/11755593_8](https://doi.org/10.1007/11755593_8), URL http://dx.doi.org/10.1007/11755593_8
- Goldberg AB, Zhu J (2006) Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In: TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing
- Hall J (2006) MaltParser – An Architecture for Inductive Labeled Dependency Parsing
- Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 174–181, DOI <http://dx.doi.org/10.3115/979617.979640>

- Hatzivassiloglou V, Wiebe J (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of the International Conference on Computational Linguistics (COLING)
- Havasi C, Speer R, Alonso J (2007) Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In: Recent Advances in Natural Language Processing, Borovets, Bulgaria
- del Hoyo R, Hupont I, Lacueva FJ, Abadía D (2009) Hybrid text affect sensing system for emotional language analysis. In: Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, ACM, New York, NY, USA, AFFINE '09, pp 3:1–3:4, DOI <http://doi.acm.org/10.1145/1655260.1655263>, URL <http://doi.acm.org/10.1145/1655260.1655263>
- Hu M, Liu B (2004a) Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp 168–177
- Hu M, Liu B (2004b) Mining opinion features in customer reviews. In: Proceedings of AAAI, pp 755–760
- Huettner A, Subasic P (2000) Fuzzy typing for document management. In: ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pp 26–27
- Hummel RA, Zucker SW (1983) On the foundations of relaxation labeling processes. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-5(3):267 –287, DOI 10.1109/TPAMI.1983.4767390
- Jin X, Li Y, Mah T, Tong J (2007) Sensitive webpage classification for content advertising. In: Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising
- Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, New York, NY, USA, WSDM '11, pp 815–824, DOI <http://doi.acm.org/10.1145/1935826.1935932>, URL <http://doi.acm.org/10.1145/1935826.1935932>
- Kamps J, Marx M, Mokken RJ, De Rijke M (2004a) Using wordnet to measure semantic orientation of adjectives. In: National Institute for, vol 26, pp 1115–1118, URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.2534>

- Kamps J, Marx M, Mokken RJ, de Rijke M (2004b) Using WordNet to measure semantic orientation of adjectives. In: LREC
- Kanayama H, Nasukawa T (2006) Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Sydney, Australia, pp 355–363
- Kim SM, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of the International Conference on Computational Linguistics (COLING)
- Kudo T, Matsumoto Y (2003) Fast methods for kernel-based text analysis. In: ACL, pp 24–31
- Lehrer Adrienne (1974) Semantic fields and lexical structure / A. Lehrer. North-Holland ; American Elsevier, Amsterdam : New York :
- Leung C, Chan S, Chung Fl, Ngai G (2011) A probabilistic rating inference framework for mining user preferences from reviews. World Wide Web 14:187–215, URL <http://dx.doi.org/10.1007/s11280-011-0117-5>, 10.1007/s11280-011-0117-5
- Lin D (1998) Dependency-based evaluation of minipar. In: Proc. Workshop on the Evaluation of Parsing Systems, Granada
- Lin WH, Wilson T, Wiebe J, Hauptmann A (2006) Which side are you on? identifying perspectives at the document and sentence levels. In: Proceedings of the Conference on Natural Language Learning (CoNLL)
- Liscombe J, Riccardi G, Hakkani-Tür D (2005) Using context to improve emotion detection in spoken dialog systems. In: Interspeech, pp 1845–1848
- Liu B, Hu M, Cheng J (2005) Opinion observer: Analyzing and comparing opinions on the web. In: Proceedings of WWW
- Liu H, Lieberman H, Selker T (2003) A model of textual affect sensing using real-world knowledge. In: Proceedings of Intelligent User Interfaces (IUI), pp 125–132
- Malouf R, Mullen T (2008) Taking sides: Graph-based user classification for informal online political discourse. Internet Research 18(2), URL <http://bulba.sdsu.edu/~malouf/papers/Takingsides.pdf>

- Marcus MP, Santorini B, Marcinkiewicz MA (1994) Building a large annotated corpus of english: The penn treebank. Computational Linguistics 19, URL <http://www.aclweb.org/anthology-new/J/J93/J93-2004.pdf>
- Mel'čuk I (1988) Dependency Syntax: Theory and Practice. State University of New York Press
- Mihalcea R, Banerjee S, Wiebe J (2007) Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the Association for Computational Linguistics (ACL), Prague, Czech Republic, pp 976–983
- Morinaga S, Yamanishi K, Tateishi K, Fukushima T (2002) Mining product reputations on the web. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp 341–349, industry track
- Mullen T, Malouf R (2006) A preliminary investigation into sentiment analysis of informal political discourse. In: AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pp 159–162
- Mullen T, Malouf R (2008) Taking sides: User classification for informal online political discourse. Internet Research 18:177–190
- Nanni DL (1980) On the surface syntax of constructions with easy-type adjectives. Language 56(3):568–581
- Nasukawa T, Yi J (2003) Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the Conference on Knowledge Capture (K-CAP)
- Ortega FJ, Macdonald C, Troyano JA, Cruz F (2010) Spam detection with a content-based random-walk algorithm. In: 2nd International Workshop on Search and Mining User-generated Contents, at CIKM 2010
- Ortega J, Troyano J, Cruz F, Enríquez de Salamanca F (2011) PolarityTrust: measuring trust and reputation in social networks. In: Fourth International Conference on Internet Technologies and Applications (ITA 11), Wrexham, North Wales, United Kingdom
- Ounis I, de Rijke M, Macdonald C, Mishne G, Soboroff I (2006) Overview of the TREC-2006 Blog Track. In: Proceedings of the 15th Text REtrieval Conference (TREC 2006)

- Page L, Brin S, Motwani R, Winograd T (1998) The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project, URL <http://citeseer.ist.psu.edu/page98pagerank.html>
- Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the Association for Computational Linguistics (ACL), pp 271–278
- Pang B, Lee L (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the Association for Computational Linguistics (ACL), pp 115–124
- Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135
- Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 79–86
- Passonneau RJ, Nenkova A, McKeown K, Sigelman S (2005) Applying the pyramid method in duc 2005. In: In Proceedings of the 2005 DUC Workshop
- Piao S, Ananiadou S, Tsuruoka Y, Sasaki Y, McNaught J (2007) Mining opinion polarity relations of citations. In: International Workshop on Computational Semantics (IWCS), pp 366–371, short paper
- Popescu AM, Etzioni O (2005) Extracting product features and opinions from reviews. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)
- Popescu AM, Nguyen B, Etzioni O (2005) Opine: extracting product features and opinions from reviews. In: Proceedings of HLT/EMNLP on Interactive Demonstrations, Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-Demo '05, pp 32–33, DOI <http://dx.doi.org/10.3115/1225733.1225750>, URL <http://dx.doi.org/10.3115/1225733.1225750>
- Qiu G, Liu B, Bu J, Chen C (2011) Opinion word expansion and target extraction through double propagation. Computational Linguistics 37(1)
- Sampson G (1992) The Susanne Corpus. School of Cognitive & Computing Sciences, University of Sussex, 1st edn

- Spertus E (1997) Smokey: Automatic recognition of hostile messages. In: Proceedings of Innovative Applications of Artificial Intelligence (IAAI), pp 1058–1065
- Stone PJ (1966) The General Inquirer: A Computer Approach to Content Analysis. The MIT Press
- Subasic P, Huettner A (2001) Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems* 9(4):483–496
- Szomszor M, Cattuto C, Alani H, O'Hara K, Baldassarri A, Loreto V, Servadio VD (2007) Folksonomies, the semantic web, and movie recommendation. In: 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, URL <http://eprints.ecs.soton.ac.uk/14007/>
- Taboada M, Gillies MA, McFetridge P (2006) Sentiment classification techniques for tracking literary reputation. In: LREC Workshop: Towards Computational Models of Literary Analysis, pp 36–43
- Takamura H, Inui T, Okumura M (2005) Extracting semantic orientation of words using spin model. In: Proceedings of the Association for Computational Linguistics (ACL), pp 133–140
- Tateishi K, Ishiguro Y, Fukushima T (2001) Opinion information retrieval from the Internet. Information Processing Society of Japan (IPSJ) SIG Notes 2001(69(20010716)):75–82, also cited as “A reputation search engine that gathers people’s opinions from the Internet”, IPSJ Technical Report NL-14411. In Japanese.
- Thomas M, Pang B, Lee L (2006) Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 327–335
- Titov I, McDonald R (2008a) A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, pp 308–316, URL <http://www.aclweb.org/anthology/P/P08/P08-1036>
- Titov I, McDonald R (2008b) Modeling online reviews with multi-grain topic models. In: Proceeding of the 17th international conference on World Wide Web, ACM, New York, NY, USA, WWW ’08, pp 111–120, DOI <http://dx.doi.org/10.1145/1367497.1367517>

- //doi.acm.org/10.1145/1367497.1367513, URL <http://doi.acm.org/10.1145/1367497.1367513>
- Tong RM (2001) An operational system for detecting and tracking opinions in on-line discussion. In: Proceedings of the Workshop on Operational Text Classification (OTC)
- Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM
- Turney P (2002a) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics (ACL), pp 417–424
- Turney PD (2002b) Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp 417–424
- Turney PD, Littman ML (2003a) Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) 21(4):315–346
- Turney PD, Littman ML (2003b) Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21:315–346
- Vallejo CG, Troyano JA, Ortega FJ (2010) Instancerank: Bringing order to datasets. Pattern Recognition Letters 31(2):133–142, URL <http://dx.doi.org/10.1016/j.patrec.2009.09.022>
- Wiebe J, Breck E, Buckley C, Cardie C, Davis P, Fraser B, Litman D, Pierce D, Riloff E, Wilson T, Day D, Maybury M (2003) Recognizing and organizing opinions expressed in the world press. In: Proceedings of the AAAI Spring Symposium on New Directions in Question Answering
- Wiebe JM, Wilson T, Bruce R, Bell M, Martin M (2004) Learning subjective language. Computational Linguistics 30(3):277–308
- Wilson T, Wiebe J, Hwa R (2004) Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI, pp 761–769, extended version in *Computational Intelligence* 22(2, Special Issue on Sentiment Analysis):73–99, 2006

- Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp 347–354
- Yi J, Niblack W (2005) Sentiment mining in WebFountain. In: Proceedings of the International Conference on Data Engineering (ICDE)
- Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the IEEE International Conference on Data Mining (ICDM)
- Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)
- Zhai Z, Liu B, Xu H, Jia P (2010) Grouping product features using semi-supervised learning with soft-constraints. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, pp 1272–1280, URL <http://www.aclweb.org/anthology/C10-1143>
- Zhao J, Liu K, Wang G (2008) Adding redundant features for crfs-based sentence sentiment classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP ’08, pp 117–126, URL <http://portal.acm.org/citation.cfm?id=1613715.1613733>
- Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a maxent-lda hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP ’10, pp 56–65, URL <http://portal.acm.org/citation.cfm?id=1870658.1870664>
- Zhu X, Ghahramani Z (2002) Learning from labeled and unlabeled data with label propagation. Tech. rep.

Parte VI

Apéndices

Apéndice A

Ejemplos de entradas y salidas de los componentes concretos de TOES

Resumen: En el presente apéndice mostramos ejemplos de aplicación de los componentes concretos del sistema definidos en el capítulo 7, que permiten aclarar el funcionamiento de los mismos, haciendo hincapié en la influencia de los distintos parámetros de configuración de los componentes.

A.1. Anotador de palabras de característica basado en la taxonomía

Taxonomy-based feature word annotator

Sea la siguiente frase del documento de entrada:

[These_DT(1) headphones_NNS(2)]_NP [have_VBP(3)]_VP [a_Z(4)
great_JJ(5) sound_NN(6) quality_NN(7)]_NP ._Fp(8)

, en la que hemos indicado la categoría morfosintáctica de cada palabra (utilizando las etiquetas del Penn Treebank), el número identificativo de cada token y los sintagmas nominales (*NP*) y verbales(*VP*). Empleando la taxonomía para el dominio *headphones* mostrada en la figura 6.11, y con independencia de los valores de los parámetros del componente, se generarían las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="sound quality" featWords="6"/>
<opinion feature="sound quality" featWords="6,7"/>
```

Dado que la segunda de las evidencias mostradas contiene como palabras de característica un subconjunto de las palabras de característica de la tercera evidencia de opinión, el componente la eliminará, quedando finalmente en la oración las evidencias siguientes:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="sound quality" featWords="6,7"/>
```

Como puede observarse, no todas las apariciones de palabras de característica se corresponderán finalmente con evidencias de opinión reales. Será responsabilidad de otros componentes determinar cuáles de las evidencias de opinión tentativas corresponden a evidencias reales y cuáles no, como veremos más adelante.

A.2. Anotador de características implícitas basado en los indicadores

Cue-based implicit feature annotator

Supongamos la siguiente oración:

These (1) headphones (2) sound (3) great (4) : (5) they (6) have
 (7) clean (8) and (9) well-defined (10) highs (11) and (12)
 powerful (13) low (14) frequencies (15).

, y las siguientes entradas en el recurso de indicadores de características implícitas:

```
<entry term="sound" support="1496">
  <fbEntry feature="sound quality" prob="0.12"
           condProb="0.12"/>
</entry>
<entry term="sound great" support="92">
  <fbEntry feature="sound quality" prob="0.29"
           condProb="0.29"/>
</entry>
<entry term="clean" support="33">
  <fbEntry feature="sound quality" opProb="0.06"
           condProb="0.11"/>
</entry>
<entry term="powerful" support="12">
  <fbEntry feature="volume" opProb="0.17" condProb="0.17"/>
</entry>
```

Suponemos que no existen entradas en el recurso para el resto de palabras de la oración anterior. La ejecución del componente con los siguientes parámetros:

- $\minProb=0.1$, $\minSupport=3$, $\useCondProbs=false$

, daría lugar a las siguientes evidencias de opinión:

```
<opinion feature="sound quality" opWords="3"/>
<opinion feature="sound quality" opWords="3,4"/>
<opinion feature="volume" opWords="13"/>
```

Dado que las palabras de opinión de la primera evidencia están incluidas en las palabras de opinión de la segunda evidencia, la primera sería eliminada, quedando finalmente:

```
<opinion feature="sound quality" opWords="3,4"/>
<opinion feature="volume" opWords="13"/>
```

Como puede observarse, la segunda de las evidencias anotadas sería incorrecta, puesto que “*powerful*” es en realidad una palabra de opinión asociada a la característica *bass* , referenciada por las palabras de característica “*low frequencies*” en la oración mostrada. Un valor mayor para el parámetro *min-Prob* (por ejemplo, 0.25) habría sorteado dicho error.

Supongamos ahora que, antes de la ejecución del componente, han sido anotadas las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2" opWords="3,4"
          polarity="+"/>
<opinion feature="treble" featWords="11" opWords="8"
          polarity="+"/>
<opinion feature="treble" featWords="11" opWords="10"
          polarity="+"/>
<opinion feature="bass" featWords="14,15" opWords="13"
          polarity="+"/>
```

, y que los valores de los parámetros del componente fueran los siguientes:

- $\minProb=0.25$, $\minSupport=3$, $\useCondProbs=true$

En este caso, no se añadirían nuevas evidencias de opinión, puesto que el único término cuya probabilidad condicionada es mayor o igual a 0.25 (“*sound great*”) ya ha sido previamente utilizado como término de opinión en una evidencia de opinión sobre característica explícita.

A.3. Anotador de características implícitas basado en PMI-IR

PMIIR-based implicit feature annotator

Supongamos la siguiente oración:

These(1) headphones(2) sound(3) great(4) : (5) they(6) have
 (7) clean(8) and(9) well-defined(10) highs(11) and(12)
 powerful(13) low(14) frequencies(15).

Sólo se consideran como posibles indicadores los adjetivos de la oración. para cada uno de estos, se estiman los valores de PMI-IR con cada una de las características de la taxonomía. Los mayores valores obtenidos para cada uno y las características correspondientes son los siguientes (los valores mostrados han sido estimados realmente haciendo uso del componente):

- “great”: 1,725 (*sound quality*)
- “clean”: 1,547 (*sound quality*)
- “well-defined”: 1,195 (*mids*)
- “powerful”: 2,123 (*sound quality*)

Suponiendo que el valor del umbral *minPMI* del componente fuese 1,5, la ejecución del componente generaría las siguientes evidencias de opinión:

```
<opinion feature="sound quality" opWords="4"/>
<opinion feature="sound quality" opWords="8"/>
<opinion feature="sound quality" opWords="13"/>
```

A.4. Enlazador de palabras de opinión basado en ventana

Window-based opinion word linker

Sea la siguiente frase del documento de entrada:

These_DT(1) headphones_NNS(2) have_VBP(3) a_Z(4) great_JJ(5)
 sound_NN(6) quality_NN(7) ._Fp(8)

, en la que hemos indicado la categoría morfosintáctica de cada palabra (utilizando las etiquetas del Penn Treebank) y el número identificativo de cada token.

Supongamos que previamente han sido generadas las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="sound quality" featWords="6,7"/>
```

La ejecución del enlazador de palabras de opinión basado en ventana con los siguientes parámetros:

- *windowSize=1, validpos="JN"*

, añadiría a las evidencias anteriores como palabras de opinión aquellos adjetivos y nombres contenidos en una ventana de contexto de una palabra alrededor de las palabras de característica. El resultado sería el siguiente:

```
<opinion feature="headphones" featWords="2"/>
<opinion feature="sound quality" featWords="6,7" opWords="5"/>
```

Aunque en el ejemplo mostrado el resultado ha sido satisfactorio, la simplicidad de este componente limita mucho su potencia. Si se utilizan ventanas pequeñas, no se capturarán correctamente aquellas evidencias de opinión en las que las palabras de característica y las palabras de opinión estén más alejadas. Si se utilizan ventanas grandes, es probable que se añadan más palabras de opinión de las que realmente están modificando a las palabras de característica. Por ejemplo, en la siguiente oración:

```
The_DT(1) headphones_NNS(2) that_WDT(3) I_PRP(4) bought_VBD
(5) in_IN(6) a_Z(7) small_JJ(8) local_JJ(9) shop_NN(10)
yesterday_NN(11) are_VBP(12) excellent_JJ(13) ._Fp(14)
```

Partiendo de la evidencia de opinión:

```
<opinion feature="headphones" featWords="2"/>
```

, si se utiliza un tamaño de ventana pequeño, no se capturaría la palabra de opinión “*excellent*”, que claramente afecta a “*headphones*”. Por el contrario, si se utilizara una ventana lo suficientemente grande como para capturar a “*excellent*”, otras palabras como “*small*” y “*local*” serían erróneamente añadidas como palabras de opinión.

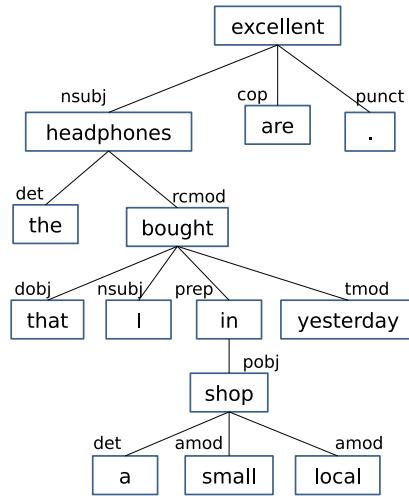
A.5. Enlazador de palabras de opinión basado en dependencias

Dependency-based opinion word linker

Supongamos la siguiente oración:

The_DT(1) headphones_NNS(2) that_WDT(3) I_PRP(4) bought_VBD
 (5) in_IN(6) a_DT(7) small_JJ(8) local_JJ(9) shop_NN(10)
 yesterday_NN(11) are_VBP(12) excellent_JJ(13) ._Fp(14)

, cuyo análisis de dependencias obtiene el siguiente resultado:



Supongamos que el recurso contiene entre otros el siguiente patrón de tipo 1:

```
<pattern order="34" feature="*" destWord="*" posAsc="N,J"  

  posDesc="J" depAsc="nsubj" depDesc="" p="0.96" r="0.12"  

  pAcc="0.97" rAcc="0.18"/>
```

Partiendo de la siguiente evidencia de opinión, previamente generada:

```
<opinion feature="headphones" featWords="2"/>
```

, ejecutamos el enlazador de palabras de opinión basado en dependencias con estos parámetros:

- *type=1, maxOpWords=2, minPrecision=0.8*
- *minAccPrecision=0.8, useFeatureBasedPatterns=false*

Siempre que con los patrones de tipo 1 anteriores al mostrado no se hayan obtenido ya dos palabras de opinión, la ejecución del componente capturaría correctamente la palabra de opinión asociada a la evidencia de opinión anterior:

```
<opinion feature="headphones" featWords="2" opWords="13" />
```

Nótese que, dado que el parámetro *useFeatureBasedPatterns* contiene el valor *false*, los patrones utilizados serán aquellos sin restricciones de característica (los que contienen un símbolo * en el atributo *feature*).

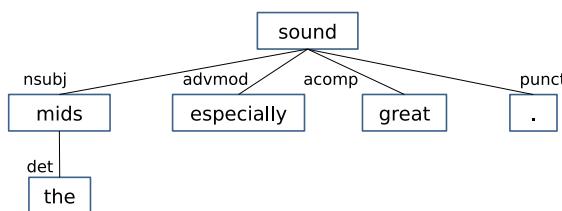
Veamos un ejemplo en el que utilizaremos dos enlazadores de opiniones basados en dependencias: uno utilizará patrones de tipo 2 y el otro de tipo 3. El primero de los enlazadores tratará de encontrar una única palabra de opinión relacionada con las palabras de característica, y posteriormente el segundo buscará otras palabras de opinión relacionadas con la anterior. Esta es la secuencia normal de aplicación (un enlazador de palabras de opinión basado en patrones de tipo 2 seguido de un enlazador de palabras de opinión basado en patrones de tipo 3), alternativa al uso de un solo enlazador basado en patrones de tipo 1. Supongamos la siguiente frase de entrada:

```
The_DT(1) mids_NNS(2) especially_RB(3) sound_VBP(4) great_JJ  
(5) ._Fp(6)
```

Partimos de la siguiente evidencia de opinión:

```
<opinion feature="mids" featWords="2" />
```

, con las relaciones de dependencias indicadas en el gráfico:



Supongamos que el recurso de patrones de dependencias contiene, entre otros, el siguiente patrón de tipo 2:

```
<pattern order="114" feature="*" destWord="*" posAsc="N,V"  
posDesc="V" depAsc="nsubj" depDesc="" p="0.33" r="0.01"  
pAcc="0.78" rAcc="0.93" />
```

y el siguiente patrón de tipo 3:

```
<pattern order="1" feature="*" destWord="*" posAsc="V"  
posDesc="J,V" depAsc="" depDesc="acomp" p="1.0" r="0.05"  
pAcc="1.0" rAcc="0.05" />
```

La aplicación de un enlazador de palabras de opinión basado en dependencias con los siguientes valores de sus parámetros:

- *type=2, maxOpWords=1, minPrecision=0.3*
- *minAccPrecision=0.7, useFeatureBasedPatterns=false*

, añadiría correctamente una palabra de opinión a la evidencia anterior:

```
<opinion feature="mids" featWords="2" opWords="4" />
```

Si ahora ejecutásemos un enlazador de palabras de opinión basado en dependencias que utilice patrones de tipo 3:

- *type=3, maxOpWords=1, minPrecision=0.9*
- *minAccPrecision=0.9, useFeatureBasedPatterns=false*

, la evidencia de opinión quedaría tal como sigue:

```
<opinion feature="mids" featWords="2" opWords="4,5" />
```

Como puede verse, hemos escogido valores adecuados de los parámetros para permitir la ejecución de los patrones indicados. La elección de unos valores adecuados para los parámetros es por tanto un aspecto crucial para que este componente funcione adecuadamente. En el capítulo 8 se ajustan los valores de los parámetros para optimizar las medidas F_1 y $F_{\frac{1}{2}}$ del sistema de extracción.

A.6. Enlazador de expresiones especiales basado en ventana

Window-based special expression linker

Sea la siguiente oración:

They_PRP(1) are_VBP(2) too_RB(3) heavy_JJ(4) . Fp(5)

, y la siguiente evidencia de opinión previamente anotada:

```
<opinion feature="weight" opWords="4" />
```

La aplicación del enlazador de expresiones especiales con los siguientes parámetros:

- *type=2, windowSize=1*

, añadiría la palabra de opinión “*too*” a la evidencia de opinión anterior:

```
<opinion feature="weight" opWords="4" />
```

No siempre las expresiones especiales aparecerán de forma contigua a las palabras de opinión a las que afectan. Por ejemplo, en la siguiente oración:

```
They.PRP(1) do_VBP(2) not_RB(3) seem_VB(4) too_RB(5)  
durable_JJ(6) _Fp(7).
```

En esta oración, la palabra de opinión “*durable*” se ve afectada por las expresiones especiales “*too*” y “*not*”. Para que el enlazador basado en ventana pueda encontrar a esta última, el valor del parámetro *windowSize* debería ser, al menos, 3. Pero el uso de ventanas de contexto grandes puede inducir a errores en otras oraciones. Por ejemplo, en la siguiente oración:

These headphones are cute and not expensive.

, la partícula “*not*”, que sólo afecta a “*expensive*”, sería incluida como palabra de opinión junto con “*cute*”, si se mantiene el tamaño de ventana anterior. Por tanto, en ciertas ocasiones este componente cometerá errores en la selección de las expresiones especiales que deben ser incluidas en determinadas evidencias.

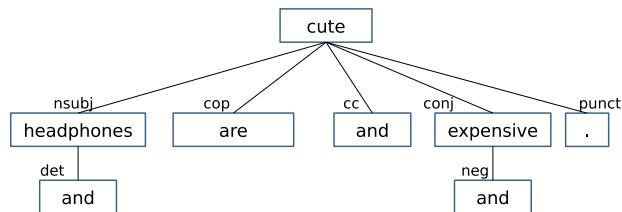
A.7. Enlazador de expresiones especiales basado en dependencias

Dependency-based special expression linker

Tomemos la siguiente oración como entrada:

```
These.DT(1) headphones.NNS(2) are_VBP(3) cute_JJ(4) and_CC  
(5) not_RB(6) expensive_JJ(7) . _Fp(8)
```

, cuyo árbol de dependencias es el siguiente:



Supongamos que el recurso contiene entre otros el siguiente patrón de tipo 4:

```
<pattern order="4" feature="*" destWord="not" posAsc="R,J"
    posDesc="" depAsc="neg" depDesc="" p="0.95" r="0.44"
    pAcc="0.96" rAcc="0.74" />
```

Partiendo de las siguientes evidencias de opinión, previamente generadas:

```
<opinion feature="appearance" opWords="4" />
<opinion feature="price" opWords="7" />
```

, ejecutamos el enlazador de expresiones especiales basado en dependencias con estos parámetros:

- $type=4$, $maxOpWords=3$, $minPrecision=0.9$
- $minAccPrecision=0.9$, $useFeatureBasedPatterns=false$

La ejecución del componente asociaría correctamente la partícula “*not*” a la segunda evidencia y no a la primera:

```
<opinion feature="appearance" opWords="4" />
<opinion feature="price" opWords="6,7" />
```

A.8. Separador de opiniones basado en conjunciones

Conjunction-based opinion splitter

Supongamos la siguiente oración:

These(1) headphones(2) sound(3) great(4) : (5) they(6) have
 (7) clean(8) and(9) well-defined(10) highs(11) and(12)
 powerful(13) low(14) frequencies(15).

, y las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2" opWords="3,4" />
<opinion feature="treble" featWords="11" opWords="8,10" />
<opinion feature="bass" featWords="14,15" opWords="13" />
<opinion feature="sound quality" opWords="3,4" />
```

La aplicación de un separador de opiniones basado en conjunciones con el siguiente valor para el parámetro de configuración:

- $splitterWords="and but ,"$

, ocasionaría la división de la evidencia de opinión anterior relativa a la característica *treble*, quedando el conjunto de evidencias de opinión como sigue:

```
<opinion feature="headphones" featWords="2" opWords="3,4"/>
<opinion feature="treble" featWords="11" opWords="8"/>
<opinion feature="treble" featWords="11" opWords="10"/>
<opinion feature="bass" featWords="14,15" opWords="13"/>
<opinion feature="sound quality" opWords="3,4"/>
```

A.9. Clasificador de opiniones basado en el lexicón de opiniones

Lexicon-based opinion classifier

Supongamos la siguiente oración de entrada:

At(1) first(2) the(3) headphones(4) were(5) not(6) comfortable
 (7) ,(8) but(9) started(10) to(11) fit(12) right(13) after
 (14) a(15) few(16) hours(17) .(18)

, y las siguientes evidencias de opinión previamente generadas:

```
<opinion feature="comfort" opWords="1,2,6,7"/>
<opinion feature="comfort" opWords="12,13"/>
```

Apliquemos un clasificador de opiniones basado en el lexicón de opiniones, con la siguiente configuración de parámetros:

- *useCompletePhrase=true, useContiguousWordsAsPhrase=true*
- *useEachWordAsPhrase=true, featureBasedProbabilities=true*
- *useGlobalSO=true, minSupport=4, minProb=0.75, minAbsSO=0.9*
- *useWordNetSynToExpand=true, useWordNetAntToExpand=true*

En primer lugar, el componente busca en las palabras de opinión apariciones de expresiones de no negación, no encontrando ninguna. Después busca apariciones de expresiones de negación, encontrándose “*not*” en la primera evidencia de opinión. El componente no considerará esta palabra en el resto del proceso de clasificación, salvo para invertir la polaridad final de la evidencia de opinión. Dado que el parámetro *useCompletePhrase* es cierto, el componente trata de localizar en el lexicón entradas para los términos “*at first comfortable*” y “*fit right*”.

Supongamos que no encuentra ninguna entrada para el primer término, pero sí para el segundo:

```
<entry term="fit right" support="11">
  <fbEntry feature="comfort" opProb="0.85" polarity="1.0"/>
  <fbEntry feature="headphones" opProb="0.85" polarity="1.0"/>
</entry>
```

Dado que tanto el valor de *support*, como los de *feature-based opinion word probability* y el valor absoluto de *feature-based semantic orientation polarity* para la característica *comfort* son superiores a los parámetros del componente *minSupport*, *minProb* y *minAbsSO* respectivamente, la evidencia de opinión es completada con el valor de la polaridad:

```
<opinion feature="comfort" opWords="12,13" polarity="1.0"/>
```

Siguiendo con la otra evidencia de opinión, al no haber sido capaz de estimar una orientación semántica para el término completo, y dado que el parámetro *useContiguousWordsAsPhrase* es cierto, el componente buscará en el lexicón los términos “*at first*” y “*comfortable*”. Para el primero no se encuentra ninguna entrada, pero sí para el segundo:

```
<entry term="comfortable" support="363">
  <fbEntry feature="comfort" opProb="0.70" polarity="1.0"/>
  <fbEntry feature="headphones" opProb="0.75" polarity="1.0"/>
</entry>
```

Sólo mostramos las entradas correspondientes a las características *comfort* y *headphones*, aunque es de suponer que existen entradas para el término “*comfortable*” al ser usado para otras características, tal como se desprende de la diferencia de valores de la medida *feature-based opinion word probability* en las entradas mostradas (ver definición de la medida en la sección 5.5.1). Dado que el valor de la probabilidad para la característica *comfort* es menor al parámetro del componente *minProb*, la estimación inicial de orientación semántica para “*comfortable*” será nula. Pero ya que el valor del parámetro *useGlobalSO* es cierto, el componente pasará a utilizar los valores de la entrada correspondiente a la característica *headphones*. En este caso, los valores de la entrada son compatibles con la configuración del componente, por lo que la orientación semántica del término se estima como 1.0. La orientación semántica final asignada será -1.0, ya que las palabras de opinión contenían una expresión de negación:

```
<opinion feature="comfort" opWords="1,2,6,7"
  polarity="-1.0"/>
```

Para ejemplificar el uso de las expansiones semánticas basadas en WordNet, supongamos la siguiente oración:

The(1) sound(2) quality(3) is(4) superb(5)

, y la siguiente evidencia de opinión previamente generada:

```
<opinion feature="sound quality" featWords="2,3"
          opWords="5"/>
```

Ejecutemos un clasificador de opinión basado en el lexicón de opiniones con los siguientes parámetros de configuración:

- *useCompletePhrase=false, useContiguousWordsAsPhrase=false*
- *useEachWordAsPhrase=true, featureBasedProbabilities=true*
- *useGlobalSO=true, minSupport=4, minProb=0.75, minAbsSO=0.9*
- *useWordNetSynToExpand=true, useWordNetAntToExpand=true*

El componente buscará una entrada en el lexicón para el término “superb”. Supongamos que encuentra la siguiente entrada:

```
<entry term="superb" support="14">
  <fbEntry feature="sound quality" opProb="0.64"
           polarity="1.0"/>
  <fbEntry feature="headphones" opProb="0.71" polarity="1.0"/>
</entry>
```

Ni el valor de probabilidad correspondiente a la característica *sound quality* ni el correspondiente a la característica raíz *headphones* cumplen con el mínimo fijado por el parámetro *minProb*. Por tanto, la orientación semántica estimada es nula. El componente procederá a realizar una expansión semántica, buscando en WordNet sinónimos y antónimos de “superb”. En el caso de los sinónimos, se encuentra la palabra “*brilliant*”, y en el de los antónimos, “*inferior*”. Supongamos las siguientes entradas en el lexicón:

```
<entry term="brilliant" support="4">
  <fbEntry feature="sound quality" opProb="0.8"
           polarity="1.0"/>
  <fbEntry feature="headphones" opProb="0.8" polarity="1.0"/>
</entry>
<entry term="inferior" support="10">
  <fbEntry feature="sound quality" opProb="0.12"
           polarity="1.0"/>
  <fbEntry feature="headphones" opProb="0.23" polarity="1.0"/>
</entry>
```

El componente realiza las estimaciones de orientación semántica de cada uno de los términos, de manera similar a los ejemplos anteriores, resultado en 1.0 para “*brilliant*” y nula para “*inferior*”. Finalmente, la orientación semántica asignada a la evidencia de opinión será 1.0:

```
<opinion feature="sound quality" featWords="2,3" opWords="5"
          polarity="1.0"/>
```

A.10. Clasificador de opiniones basado en PMI-IR

PMIIR-based opinion classifier

Supongamos la siguiente oración:

The(1) sound(2) quality(3) is(4) superb(5)

, y la siguiente evidencia de opinión previamente generada:

```
<opinion feature="sound quality" featWords="2,3" opWords="5"/>
```

Ejecutemos un clasificador de opinión basado en *PMI-IR* con los siguientes parámetros de configuración:

- *minAbsSO=0.8, useDistanceAttenuation=true*

El clasificador estimaría la orientación semántica del término *superb* mediante el algoritmo *PMI-IR*, mediante las siguientes búsquedas en AltaVista:

$$\begin{aligned} \text{hits}(\text{"excellent"}) &= 17 * 10^8 \\ \text{hits}(\text{"poor"}) &= 11,6 * 10^8 \\ \text{hits}(\text{"superb"} \text{ NEAR } \text{"excellent"}) &= 97,6 * 10^5 \\ \text{hits}(\text{"superb"} \text{ NEAR } \text{"poor"}) &= 11,9 * 10^5 \end{aligned}$$

Con lo que la estimación final de la orientación semántica de “*superb*” aplicando la fórmula anterior es:

$$SO(\text{"superb"}) = 5,596$$

Dado que el valor obtenido es mayor que el valor del parámetro *minAbsSO*, el clasificador rellena el atributo *polaridad* de la evidencia de opinión anterior con dicho valor, convenientemente ponderado por la inversa de la distancia a las palabras de característica:

```
<opinion feature="sound quality" featWords="2,3" opWords="5"
          polarity="2.798"/>
```

A.11. Clasificador de opiniones basado en WordNet

WordNet-based opinion classifier

Supongamos que a partir de la siguiente oración:

These_DT (1) headphones_NNS (2) sound_VBP (3) unclean_JJ (4)

, se ha obtenido la siguiente evidencia de opinión:

```
<opinion feature="sound quality" opWords="3,4" />
```

El componente calcula la función *EVA* para el único adjetivo participante como palabra de opinión “unclean”, calculando las distancias participantes en la misma:

$$d(\text{"unclean"}, \text{"bad"}) = 2$$

$$d(\text{"unclean"}, \text{"good"}) = 4$$

$$d(\text{"good"}, \text{"bad"}) = 3$$

Con estos valores, la orientación semántica estimada para el adjetivo, y consecuentemente para la evidencia de opinión, es de -0.667 (siempre que el valor del parámetro de configuración *minAbsSO* sea menor o igual a dicho valor):

```
<opinion feature="sound quality" opWords="3,4"
polarity="-0.667" />
```

A.12. Clasificador de opiniones basado en SentiWordNet

SentiWordNet-based opinion classifier

Partimos de la siguiente oración:

The_DT (1) sound_NN (2) quality_NN (3) is_VBZ (4) superb_JJ (5)

, y la siguiente evidencia de opinión previamente generada:

```
<opinion feature="sound quality" featWords="2,3" opWords="5" />
```

Ejecutemos un clasificador de opinión basado en SentiWordNet, con *minAbsSO*=0.8. El componente, después de comprobar que no existen expresiones especiales entre las palabras de opinión, busca en WordNet todos los synsets a los que pertenece la palabra “superb” funcionando como adjetivo.

Encuentra dos synsets, para los cuales busca en SentiWordNet los valores de positividad y negatividad asociados. Para ambos synsets, estos valores son de 0.875 y 0, respectivamente. Por tanto, la orientación semántica de cada synset es 0.875, y la orientación semántica de la palabra de opinión es igualmente 0.875. Finalmente, al ser ésta la única palabra de opinión, y siendo el valor de orientación semántica obtenido superior al parámetro *minAbsSO*, la evidencia de opinión es completada convenientemente:

```
<opinion feature="sound quality" featWords="2,3" opWords="5"
          polarity="0.875"/>
```

A.13. Clasificador de opiniones basado en conjunciones

Conjunction-based opinion classifier

Utilicemos la misma oración de ejemplo que mostramos en el ejemplo de aplicación del componente separador de opiniones basado en conjunciones:

These (1) headphones (2) sound (3) great (4) : (5) they (6) have
 (7) clean (8) and (9) well-defined (10) highs (11) and (12)
 powerful (13) low (14) frequencies (15).

Entre otras, se han generado previamente las siguientes evidencias de opinión (tras ejecutar entre otros el separador de opiniones):

```
<opinion feature="treble" featWords="11" opWords="8"
          polarity="1"/>
<opinion feature="treble" featWords="11" opWords="10"/>
```

Suponemos en este caso, como puede verse, que la ejecución previa de algún clasificador de opiniones ha clasificado la primera de las evidencias de opinión, pero no la segunda. Ejecutemos el clasificador basado en conjunciones con los siguientes valores de configuración:

- *andWords*=“and”, *butWords*=“but”
- *allowHolesBetweenOpWords*=false

En un intento por clasificar la segunda evidencia de opinión, el componente detecta otra evidencia con polaridad no nula y cuyas palabras de opinión están coordinadas con las palabras de opinión de la evidencia a clasificar, mediante el uso de la palabra “*and*”. Considerando que no hay ninguna palabra más a parte de la conjunción entre las palabras de opinión de una y otra evidencia, y que la conjunción utilizada pertenece a la lista definida

por el parámetro *andWords*, el componente propagará de manera directa la polaridad a la evidencia que está siendo clasificada.

```
<opinion feature="treble" featWords="11" opWords="8"
          polarity="1" />
<opinion feature="treble" featWords="11" opWords="10"
          polarity="1" />
```

Veamos una segunda oración de ejemplo:

The(1) cable(2) is(3) not(4) long(5) but(6) really(7)
durable(8).

, suponiendo las siguientes evidencias de opinión previamente generadas:

```
<opinion feature="cord" featWords="2" opWords="4,5" />
<opinion feature="cord" featWords="2" opWords="8"
          polarity="1" />
```

Si ejecutamos el componente con la misma configuración anterior, la primera evidencia de opinión no será clasificada, al no permitir dicha configuración la existencia de palabras distintas a la propia conjunción entre las palabras de opinión de las evidencias participantes. Si el parámetro *allowHolesBetweenOpWords* fuese cierto, entonces la evidencia de opinión sería clasificada de manera opuesta a la otra evidencia, puesto que la conjunción utilizada forma parte de la lista de palabras del parámetro *butWords*. En este caso, el resultado sería:

```
<opinion feature="cord" featWords="2" opWords="4,5"
          polarity="-1" />
<opinion feature="cord" featWords="2" opWords="8"
          polarity="1" />
```

Nótese que el clasificador no realiza ningún tratamiento con las expresiones especiales (en el ejemplo mostrado, aparece una expresión de negación), y sólo se basa en la conjunción utilizada y la polaridad de la otra evidencia participante de la expresión conjuntiva.

A.14. Solucionador de opiniones explícitas superpuestas

Explicit overlapping opinion fixer

Supongamos la siguiente oración:

The(1) highs(2) and(3) mids(4) are(5) outstanding(6) .(7)

, y las siguientes evidencias de opinión previamente generadas:

```
<opinion feature="treble" featWords="2" opWords="6"/>
<opinion feature="mids" featWords="4" opWords="6"/>
```

Dado que entre las palabras “*treble*” y “*mids*” hay una conjunción, el componente no eliminará la palabra de opinión en común. Sin embargo, en la siguiente oración:

The(1) highs(2) are(3) outstanding(4) and(5) the(6) mids(7)
are(8) ok(9) .(10)

, y ante las siguientes evidencias de opinión:

```
<opinion feature="treble" featWords="2" opWords="4"/>
<opinion feature="mids" featWords="7" opWords="4,9"/>
```

, el componente eliminará la palabra de opinión “*outstanding*” de la segunda evidencia de opinión (aquella en la que la palabra de opinión está más alejada de la palabra de característica).

A.15. Solucionador de opiniones implícita-explicita superpuestas

Implicit-Explicit overlapping opinion fixer

Supongamos la siguiente oración:

These(1) headphones(2) sound(3) great(4) .(5)

Es posible que la ejecución de algunos de los componentes anteriores haya generado las siguientes evidencias de opinión:

```
<opinion feature="headphones" featWords="2"
          opWords="3,4"/>
<opinion feature="sound quality" opWords="3,4"/>
```

La aplicación del solucionador de opiniones implícita-explicita superpuestas eliminaría en este caso las palabras de opinión en conflicto (“*sound great*”) de la evidencia de opinión sobre característica explícita, al ser *sound quality* una característica hija de *headphones* en la taxonomía de características de *headphones*:

```
<opinion feature="headphones" featWords="2" opWords="" />
<opinion feature="sound quality" opWords="3,4"/>
```

Se entiende de esta manera que, en esta situación, la segunda evidencia de opinión está afinando más a la hora de definir la característica. Este tipo de situaciones es muy común, puesto que la mayor parte de las opiniones sobre características implícitas se representan mediante oraciones en las que participa la característica raíz (por ejemplo, “*expensive headphones*” expresa

una opinión sobre la característica “*price*”; “*These headphones are too big*” expresa una opinión sobre la característica “*size*”).

Supongamos ahora la siguiente oración:

The(1) case(2) is(3) too(4) big(5) .(6)

, con las siguientes evidencias de opinión previamente generadas:

```
<opinion feature="case" featWords="2" opWords="4,5" />
<opinion feature="size" opWords="4,5" />
```

En este caso, dado que la característica “*size*” no depende de “*case*” en la taxonomía de características de *headphones*, el componente elimina las palabras de opinión de la evidencia de opinión sobre característica implícita, quedando por tanto las evidencias como sigue:

```
<opinion feature="case" featWords="2" opWords="4,5" />
<opinion feature="size" opWords="" />
```


Apéndice B

PolarityRank: justificación algebraica y convergencia^{*}

Resumen: En el presente apéndice se incluye la justificación algebraica y la demostración de la convergencia del algoritmo PolarityRank, introducido en la sección 2.3.2 y empleado en la metodología propuesta para la generación de los recursos para la extracción de opiniones, en el paso de ampliación del lexicón de opiniones (ver sección 6.8). Empezamos recordando la definición del algoritmo.

B.1. Definición

Sea un grafo dirigido $G = (V, E)$ donde V es un conjunto de nodos y E un conjunto de aristas dirigidas entre dos nodos. Cada arista de E contiene un valor real asociado o peso, distinto de cero, siendo p_{ji} el peso asociado a la arista que va del nodo v_j al v_i . Se define la operación $Out(v_i)$, que devuelve el conjunto de índices j de los nodos para los que existe una arista saliente desde v_i . Se definen las operaciones $In^+(v_i)$ y $In^-(v_i)$, que devuelven los conjuntos de índices j de los nodos para los que existe una arista entrante hacia v_i cuyo peso sea positivo o negativo, respectivamente. Definimos el PolarityRank positivo (PR^+) y negativo (PR^-) de un nodo v_i (fórmula 2.5), donde los valores de e^+ son mayores que cero para ciertos nodos que actúan como semillas positivas y cero para el resto de nodos, y los valores de e^- son

*Las demostraciones incluidas en este apéndice fueron elaboradas por el prof. D. Carlos G. Vallejo, en el marco de la redacción de un artículo conjunto sobre el algoritmo PolarityRank que fue enviado en mayo de 2010 a la revista *Information Processing & Management*. El artículo está pendiente de publicación.

mayores que cero para ciertos nodos que actúan como semillas negativas y cero para el resto de nodos.

$$\begin{aligned}
 PR^+(v_i) &= (1-d)e_i^+ + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^+(v_j) + \right. \\
 &\quad \left. + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^-(v_j) \right) \\
 PR^-(v_i) &= (1-d)e_i^- + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^-(v_j) + \right. \\
 &\quad \left. + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PR^+(v_j) \right)
 \end{aligned} \tag{B.1}$$

La constante d es un factor de amortiguación, en todos los casos positivo y menor que 1. La suma de los valores de e^+ por un lado y de e^- por otro debe ser igual al número de nodos del grafo.

B.2. Justificación algebraica

Vamos a estudiar a continuación la fundamentación algebraica del PolarityRank. Sea $n = |V|$, el número de nodos del grafo. Denotamos \mathbf{P} a la matriz de los (p_{ij}) .

Llamemos $p_j = \sum_{k \in Out(v_j)} |p_{jk}|$, esto es, la suma de los pesos de las aristas que comienzan en v_j . En la matriz \mathbf{P} , p_j es la suma de los valores absolutos de la fila j ; p_j puede escribirse como $p_j = \sum_{k=1}^n |p_{jk}|$

PolarityRank puede ahora ser expresado como

$$\begin{aligned}
 PR^+(v_i) &= (1-d)e_i^+ + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{p_j} PR^+(v_j) + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{p_j} PR^-(v_j) \right) \\
 PR^-(v_i) &= (1-d)e_i^- + \\
 &+ d \left(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{p_j} PR^-(v_j) + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{p_j} PR^+(v_j) \right)
 \end{aligned}$$

Definamos ahora la matriz $\mathbf{Q} = \mathbf{P}^t$, es decir, la traspuesta de \mathbf{P} (si la matriz \mathbf{P} describe un grafo no dirigido entonces $\mathbf{Q} = \mathbf{P}$). Llamemos $q_j = \sum_{k=1}^n |q_{kj}|$, es decir, la suma de los elementos de la columna j de \mathbf{Q} . Obviamente, $p_j = q_j$.

Definamos ahora dos matrices $\mathbf{Q}^+ = (q_{ij}^+)$ y $\mathbf{Q}^- = (q_{ij}^-)$ como

$$\begin{aligned} q_{ij}^+ &= \begin{cases} q_{ij} & \text{if } q_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \\ q_{ij}^- &= \begin{cases} -q_{ij} & \text{if } q_{ij} < 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

q_j puede verse como la suma de los elementos de la columna j de la matriz \mathbf{Q}^+ más la suma de los de la misma columna de la matriz \mathbf{Q}^- .

Ahora podemos expresar PolarityRank del siguiente modo:

$$\begin{aligned} PR^+(v_i) &= (1-d)e_i^+ + d\left(\sum_{j=1}^n \frac{q_{ij}^+}{q_j} PR^+(v_j) + \sum_{j=1}^n \frac{q_{ij}^-}{q_j} PR^-(v_j)\right) \\ PR^-(v_i) &= (1-d)e_i^- + d\left(\sum_{j=1}^n \frac{q_{ij}^-}{q_j} PR^+(v_j) + \sum_{j=1}^n \frac{q_{ij}^+}{q_j} PR^-(v_j)\right) \end{aligned}$$

Definimos a continuación la matriz $\mathbf{A}^+ = (a_{ij}^+)$ como $a_{ij}^+ = q_{ij}^+/q_j$ y la matriz $\mathbf{A}^- = (a_{ij}^-)$ como $a_{ij}^- = q_{ij}^-/q_j$, entonces

$$\begin{aligned} PR^+(v_i) &= (1-d)e_i^+ + d\left(\sum_{j=1}^n a_{ij}^+ PR^+(v_j) + \sum_{j=1}^n a_{ij}^- PR^-(v_j)\right) \\ PR^-(v_i) &= (1-d)e_i^- + d\left(\sum_{j=1}^n a_{ij}^- PR^+(v_j) + \sum_{j=1}^n a_{ij}^+ PR^-(v_j)\right) \end{aligned}$$

A continuación vamos a hacer algunas definiciones para simplificar la notación. Sea $m = 2n$. Definimos el vector \mathbf{x} de $m \times 1$ elementos como

$$\mathbf{x} = \begin{bmatrix} PR^+(v_1) \\ \vdots \\ PR^+(v_n) \\ PR^-(v_1) \\ \vdots \\ PR^-(v_n) \end{bmatrix}$$

y el vector \mathbf{e} de $m \times 1$ elementos como

$$\mathbf{e} = \begin{bmatrix} e_1^+ \\ \vdots \\ e_n^+ \\ e_1^- \\ \vdots \\ e_n^- \end{bmatrix}$$

Finalmente, definimos la matriz \mathbf{A} de $m \times m$ elementos

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}^+ & \mathbf{A}^- \\ \hline \mathbf{A}^- & \mathbf{A}^+ \end{array} \right]$$

Con los vectores y matrices auxiliares anteriores, podemos escribir las ecuaciones de PolarityRank (B.1) mucho más fácilmente como

$$\mathbf{x} = (1 - d)\mathbf{e} + d\mathbf{Ax} \quad (\text{B.2})$$

\mathbf{A} es una matriz estocástica: Todos los elementos de \mathbf{A} están entre 0 y 1 (ambos incluidos):

$$a_{ij}^+ = \frac{q_{ij}^+}{\sum_{k=1}^n q_{kj}^+ + \sum_{k=1}^n q_{kj}^-} \quad \text{y} \quad a_{ij}^- = \frac{q_{ij}^-}{\sum_{k=1}^n q_{kj}^+ + \sum_{k=1}^n q_{kj}^-}$$

y, obviamente, la suma de los elementos de cada columna es 1.

e_i^+ y e_i^- se han elegido de manera que $\sum_{i=1}^n e_i^+ = \sum_{i=1}^n e_i^- = n$ y positivos. Por tanto $\|\mathbf{e}\|_1 = m$.

Definamos ahora el vector de $m \times 1$ elementos $\mathbf{f} = \mathbf{e}/m$ (esto es, los elementos de e_i^+ y e_i^- divididos por m), y el vector \mathbf{u} como el vector de $m \times 1$ unos. Analicemos la matriz \mathbf{fu}^t :

$$\mathbf{fu}^t = \begin{bmatrix} e_1^+/m \\ \vdots \\ e_n^+/m \\ e_1^-/m \\ \vdots \\ e_n^-/m \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} e_1^+/m & \dots & e_1^+/m \\ \vdots & & \vdots \\ e_n^+/m & \dots & e_n^+/m \\ e_1^-/m & \dots & e_1^-/m \\ \vdots & & \vdots \\ e_n^-/m & \dots & e_n^-/m \end{bmatrix}$$

Sus elementos están comprendidos entre 0 y 1, y la suma de los elementos de cada columna es 1, por tanto \mathbf{fu}^t es una matriz estocástica.

Si \mathbf{x} está normalizado, de modo que $\|\mathbf{x}\|_1 = m$ entonces $\mathbf{u}^t \mathbf{x} = m$, siempre que los elementos de \mathbf{x} sean positivos; pero si comenzamos dándole valores no negativos, éstos siempre permanecerán siendo no negativos puesto que los elementos de \mathbf{A} también lo son.

La ecuación (B.2) se puede escribir entonces como:

$$\begin{aligned} \mathbf{x} &= (1 - d)\mathbf{e} + d\mathbf{Ax} = \\ &= (1 - d)\mathbf{e} \frac{1}{m} \mathbf{u}^t \mathbf{x} + d\mathbf{Ax} = \\ &= (1 - d)\mathbf{fu}^t \mathbf{x} + d\mathbf{Ax} = \\ &= ((1 - d)\mathbf{fu}^t + d\mathbf{A})\mathbf{x} \end{aligned}$$

Llamemos $\mathbf{B} = ((1-d)\mathbf{f}\mathbf{u}^t + d\mathbf{A})$. La expresión que define el PolarityRank (B.1) se puede escribir entonces como

$$\mathbf{x} = \mathbf{Bx} \quad (\text{B.3})$$

\mathbf{B} es la combinación convexa (combinación lineal donde los dos coeficientes son positivos y suman 1) de dos matrices estocásticas, por tanto \mathbf{B} es también estocástica. De aquí, y como resultado del teorema de Perron-Frobenius, la ecuación (B.3) tiene el autovalor 1, y el resto de sus autovalores tienen módulo menor que 1 (pueden ser complejos). El sistema anterior tiene una solución que es precisamente el autovector correspondiente al autovalor 1.

B.3. Convergencia

El cálculo de PolarityRank (el vector \mathbf{x}) puede realizarse a partir de la expresión (B.1) mediante un cálculo iterado que vamos a demostrar que converge.

Como hemos visto, la expresión (B.1) equivale a la expresión (B.2): $\mathbf{x} = (1-d)\mathbf{e} + d\mathbf{Ax}$. Expresemos como \mathbf{x}_k el término k -ésimo del cálculo iterado de PolarityRank. Es

$$\mathbf{x}_{k+1} = (1-d)\mathbf{e} + d\mathbf{Ax}_k$$

y

$$\mathbf{x}_k = (1-d)\mathbf{e} + d\mathbf{Ax}_{k-1}$$

Por tanto

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = d\|\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1})\| \leq d\|\mathbf{A}\| \|(\mathbf{x}_k - \mathbf{x}_{k-1})\|$$

para cualquier norma compatible con el vector y la matriz. Por ejemplo, usando la norma 1,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^m |x_i|$$

y

$$\|\mathbf{A}\|_1 = \max_{j=1\dots m} \sum_{i=1}^m |a_{ij}|$$

Es $\|\mathbf{A}\|_1 = 1$ (recordemos que es estocástica), y $d < 1$, luego $\lim_{k \rightarrow \infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \rightarrow 0$. Por tanto la convergencia está asegurada.

