

Índice general

Resumen	V
Resum	VII
Abstract	IX
Índice general	XI
Índice de tablas	XV
Índice de figuras	XIX
1 Introducción	1
1.1 Motivación y objetivos	2
1.2 Preguntas de investigación	4
1.3 Contribuciones	5
1.4 Estructura de la tesis	6
2 Estado de la cuestión	9
2.1 Detección automática de reutilización en textos	9
2.2 Detección automática de reutilización en código fuente	15
2.2.1 Reutilización a nivel monolingüe	16
2.2.2 Reutilización a nivel translingüe	28
2.3 Conclusiones	30

3 Recursos	31
3.1 Recursos académicos	32
3.1.1 Corpus SPADE	32
3.1.2 Corpus ILN	33
3.1.3 Corpus A&T++	34
3.2 Recursos en la Web	35
3.2.1 Corpus Google Code Jam	36
3.2.2 Corpus Rosettacode	38
3.3 Conclusiones	40
4 Modelos propuestos	41
4.1 SoCo-WCR: Modelo basado en el ratio de palabras	42
4.2 SoCo-NG: Modelo basado en n -gramas	42
4.3 SoCo-Sliding: Modelo basado en ventana deslizante	46
4.4 SoCo-COG: Modelo basado en cognados	49
4.5 SoCo-LSA: Modelo basado en análisis semántico latente	51
4.6 SoCo-ESA: Modelo de análisis semántico explícito	54
4.7 SoCo-ASA: Modelo basado en alineamiento de palabras	57
4.8 Conclusiones	61
5 Experimentación	63
5.1 Evaluación en escenarios monolingües	63
5.1.1 Impacto de cambios en identificadores	64
5.1.2 Detección de reutilización en un entorno masivo	68
5.1.3 Ajuste, comparación y ensamble de modelos	71
5.2 Evaluación en escenarios translingües	77
5.2.1 Experimentación preliminar	78
5.2.2 Detección de reutilización en ámbito académico	83
5.2.3 Detección de reutilización en corpus comparable y paralelo	83
5.3 Conclusiones	93
6 Evaluación en la competición internacional de detección de reutilización en código fuente	95
6.1 Competición internacional monolingüe SOCO	95
6.1.1 Tarea propuesta	96
6.1.2 Corpus	96

6.1.3 Evaluación	98
6.1.4 Sistemas participantes	99
6.1.5 Resultados	101
6.2 Comparación con la propuesta monolingüe	105
6.2.1 Entrenamiento y ajuste de los modelos para SOCO	105
6.2.2 Comparación de los modelos propuestos con la competición	107
6.3 Competición internacional translingüe CL-SOCO	109
6.3.1 Tarea propuesta	110
6.3.2 Corpus	110
6.3.3 Evaluación	112
6.3.4 Sistemas participantes	113
6.3.5 Resultados	115
6.4 Comparación con la propuesta translingüe	118
6.4.1 Entrenamiento y ajuste de los modelos para CL-SOCO	118
6.4.2 Comparación de los modelos propuestos con la competición	120
6.5 Conclusiones	123
7 Conclusiones y trabajos futuros	125
7.1 Aportaciones de la tesis	126
7.2 Respuestas de investigación	130
7.3 Líneas de investigación abiertas	134
Apéndices	139
A Herramienta DeSoCoRe	141
B Publicaciones relacionadas	145
Bibliografía	149
Índice alfabético	161