

TABLA DE CONTENIDOS

	página
Dedicatoria	I
Agradecimientos	II
Tabla de Contenidos	IV
Índice de Figuras	VIII
Índice de Tablas	IX
Resumen	XII
1. Introducción	13
1.1. Marco general	13
1.2. Presentación del problema	14
1.3. Objetivos	14
1.3.1. Objetivo general	14
1.3.2. Objetivos específicos	14
1.4. Alcance	15
1.5. Descripción de capítulos	15
2. Antecedentes	17
2.1. WordNet	17
2.2. Medidas de similitud semántica entre palabras	22
2.2.1. Medidas topológicas	23
2.2.1.1. Medida de Leacock–Chodorow	23
2.2.1.2. Medida de Hirst y St–Onge	24
2.2.1.3. Medida de Wu–Palmer	24
2.2.1.4. Medida de Resnik	25
2.2.1.5. Medida de Jiang–Conrath	26
2.2.1.6. Medida de Lin	27
2.2.1.7. Medida de Mihalcea y Moldovan	27

2.2.1.8.	Medida de Lesk	29
2.2.1.9.	Medida de Banerjee–Pedersen (Lesk Adaptada) . . .	29
2.2.1.10.	Medida de Patwardhan–Pedersen (vector)	31
2.2.1.11.	Medida de Agirre y Rigau	32
2.2.1.12.	Medida de Path	32
2.2.2.	Medidas estadísticas	33
2.2.2.1.	Análisis de la Semántica Latente (LSA)	33
2.3.	Medidas de similitud sintáctica	37
2.3.1.	Distancia de Levenshtein-Damerau	38
2.4.	Noción básica de similitud entre frases	40
3.	Metodología	41
4.	Introducción al área de estudio	46
5.	Estado del arte	52
5.1.	Descripción del problema	52
5.2.	Similitud semántica	53
5.3.	Similitud semántica en contexto local	57
5.4.	Similitud semántica en contexto global	58
5.5.	Aplicaciones de similitud semántica	59
6.	Medidas de similitud entre frases	63
6.1.	Estudios preliminares	63
6.2.	Elección de medidas para implementación	69
6.3.	Propuestas de medidas para similitud de frases	70
6.3.1.	Corpus	71
6.3.1.1.	Corpus de Google	71
6.3.1.2.	Corpus de Wikipedia	76
6.3.2.	Métrica básica	77
6.3.3.	Distancia de edición orientada a palabras	78
6.3.4.	Medida secuencial	82
6.4.	Variantes sobre propuestas de medidas de similitud	85
6.4.1.	Variantes sobre Distancia de edición modificada	87
6.4.1.1.	Distancia de edición modificada con sinónimos	87

6.4.1.2.	Distancia de edición modificada con Jiang-Conrath	87
6.4.1.3.	Distancia de edición modificada con Lesk Adaptada	87
6.4.2.	Variantes sobre Medida secuencial	87
6.4.2.1.	Medida secuencial con sinónimos	90
6.4.2.2.	Medida secuencial con Jiang-Conrath	90
6.4.2.3.	Medida secuencial con Lesk Adaptada	92
7.	Pruebas	94
7.1.	Pruebas de depuración	94
7.1.1.	Pruebas de funcionalidad	94
7.1.1.1.	Pruebas de funcionalidad para Distancia de Edición adaptada	95
7.1.1.2.	Pruebas de funcionalidad para Distancia de Edición adaptada con Sinónimos	96
7.1.1.3.	Pruebas de funcionalidad para Distancia de Edición adaptada con Jiang-Conrath	97
7.1.1.4.	Pruebas de funcionalidad para Distancia de Edición adaptada con Lesk Adaptada	99
7.1.1.5.	Pruebas de funcionalidad para Medida secuencial	100
7.1.1.6.	Pruebas de funcionalidad para Medida secuencial con sinónimos	101
7.1.1.7.	Pruebas de funcionalidad para Medida secuencial con Jiang-Conrath	102
7.1.1.8.	Pruebas de funcionalidad para Medida secuencial con Lesk Adaptada	103
7.1.1.9.	Pruebas de funcionalidad para Métrica básica	104
7.1.2.	Pruebas para estudio de métricas bajo distintos casos de prueba	106
7.1.2.1.	Pruebas de estudio para Distancia de Edición adaptada	106
7.1.2.2.	Pruebas de estudio para Medida secuencial	109
7.1.2.3.	Pruebas de estudio para Métrica básica	112
7.2.	Pruebas de evaluación final	114
7.2.1.	Experimento 1	115
7.2.2.	Experimento 2	116
7.2.3.	Experimento 3	117

7.2.4. Experimento 4	118
7.2.5. Experimento 5	118
7.2.6. Experimento 6	119
7.2.7. Experimento 7	120
7.2.8. Experimento 8	120
7.3. Discusión de la evaluación experimental	121
8. Conclusiones y trabajo futuro	123
8.1. Conclusiones	123
8.2. Trabajo futuro	125
Glosario	126
Bibliografía	128
Anexos	
A: Anexo 1	144
B: Anexo 2	145

ÍNDICE DE FIGURAS

	página
2.1. Concepto de medida de similitud según Wu-Palmer.	25
2.2. Segundo paso del método Análisis de la Semántica Latente (LSA). . .	37
4.1. Estructura general de la gramática en el Procesamiento del Lenguaje Natural (PLN).	47
6.1. Métrica básica para medir similitud entre frases.	77
6.2. Distancia de edición adaptada para comparar dos frases	80
6.3. Ejemplo de matriz de costos de Distancia de edición orientada a palabras.	81
6.4. Distancia secuencial para comparar 2 frases.	83
6.5. Distancia de edición adaptada para comparar dos frases, incluyendo análisis semántico con sinónimos.	88
6.6. Distancia de edición adaptada para comparar dos frases, incluyendo análisis semántico con Jiang-Conrath.	89
6.7. Distancia de edición adaptada para comparar dos frases, incluyendo análisis semántico con Lesk Adaptada.	89
6.8. Distancia secuencial para comparar 2 frases con sinónimos.	91
6.9. Distancia secuencial para comparar 2 frases con Jiang-Conrath. . . .	92
6.10. Distancia secuencial para comparar 2 frases con Lesk Adapatada. . .	92
7.1. Precisión de las métricas con respecto al <i>ranking</i> promedio.	116
7.2. <i>Coverage</i> de las métricas con respecto al <i>ranking</i> promedio.	117
7.3. Precisión de las métricas con respecto a las frases derivadas.	117
7.4. <i>Coverage</i> de las métricas con respecto a las frases derivadas.	118
7.5. Sensibilidad de las métricas bajo la operación de intercambio.	119
7.6. Sensibilidad de las métricas bajo la operación inserción.	120
7.7. Sensibilidad de las métricas bajo la operación de borrado.	120
7.8. Eficiencia de las métricas	121

ÍNDICE DE TABLAS

	página
2.1. Sentidos para “leaf” según WordNet 3.0.	21
2.2. Cadena de hiperónimos para sustantivo “day”.	22
2.3. Ejemplo de matriz de co-ocurrencias para método de Análisis de Semántica Latente	34
2.4. Definición formal de Distancia de Levenshtein.	38
2.5. Ejemplos de modificación de cadenas aplicando Levenshtein.	38
2.6. Ejemplo de Distancia de Levenshtein.	39
2.7. Definición formal de transposición de caracteres adyacentes en Distancia de Levenshtein-Damerau.	39
2.8. Diferencia entre (de izquierda a derecha) Distancia de Levenshtein y Levenshtein-Damerau.	40
6.1. Top 10 de frases candidatas para frase: “difícil de ver el futuro es”. Error introducido: Palabras en orden incorrecto para idioma español.	66
6.2. Top 10 de frases candidatas para frase: “el perro negro”. Error introducido: Frase correcta.	66
6.3. Top 10 de frases candidatas para frase: “equivocado me he”. Error introducido: Palabras en orden incorrecto para idioma español.	68
6.4. Top 10 de frases candidatas para frase: “casa en en cordillera”. Error introducido: Palabra sobrante “en”.	68
6.5. Ejemplo de cálculo para Medida secuencial.	84
7.1. Prueba de funcionalidad sobre Distancia de Edición adaptada.	96
7.2. Prueba de funcionalidad sobre Distancia de Edición adaptada.	96
7.3. Prueba de funcionalidad sobre Distancia de Edición adaptada.	96
7.4. Prueba de funcionalidad sobre Distancia de Edición adaptada con Sinónimos.	97
7.5. Prueba de funcionalidad sobre Distancia de Edición adaptada con Sinónimos.	97
7.6. Prueba de funcionalidad sobre Distancia de Edición adaptada con Sinónimos.	97

7.7. Prueba de funcionalidad sobre Distancia de Edición adaptada con Jiang-Conrath.	98
7.8. Prueba de funcionalidad sobre Distancia de Edición adaptada con Jiang-Conrath.	98
7.9. Prueba de funcionalidad sobre Distancia de Edición adaptada con Jiang-Conrath.	99
7.10. Prueba de funcionalidad sobre Distancia de Edición adaptada con Lesk Adaptada.	99
7.11. Prueba de funcionalidad sobre Distancia de Edición adaptada con Lesk Adaptada.	100
7.12. Prueba de funcionalidad sobre Distancia de Edición adaptada con Lesk Adaptada.	100
7.13. Prueba de funcionalidad sobre Medida secuencial.	100
7.14. Prueba de funcionalidad sobre Medida secuencial.	101
7.15. Prueba de funcionalidad sobre Medida secuencial.	101
7.16. Prueba de funcionalidad sobre Medida secuencial con sinónimos.	102
7.17. Prueba de funcionalidad sobre Medida secuencial con sinónimos.	102
7.18. Prueba de funcionalidad sobre Medida secuencial con sinónimos.	102
7.19. Prueba de funcionalidad sobre Medida secuencial con Jiang-Conrath.	103
7.20. Prueba de funcionalidad sobre Medida secuencial con Jiang-Conrath.	103
7.21. Prueba de funcionalidad sobre Medida secuencial con Jiang-Conrath.	103
7.22. Prueba de funcionalidad sobre Medida secuencial con Lesk Adaptada.	104
7.23. Prueba de funcionalidad sobre Medida secuencial con Lesk Adaptada.	104
7.24. Prueba de funcionalidad sobre Medida secuencial con Lesk Adaptada.	104
7.25. Prueba de funcionalidad sobre Métrica básica.	105
7.26. Prueba de funcionalidad sobre Métrica básica.	105
7.27. Prueba de funcionalidad sobre Métrica básica.	105
7.28. Prueba de estudio sobre Distancia de Edición y sus variantes. Frase de entrada utilizada: <i>“el primer hombre que la luna pisó”</i>	107
7.29. Prueba de estudio sobre Distancia de Edición y sus variantes. Frase de entrada utilizada: <i>“la bebe dormía la en su cuna”</i>	107
7.30. Prueba de estudio sobre Distancia de Edición y sus variantes. Frase de entrada utilizada: <i>“el cigarrillo puede cáncer producir”</i>	108

7.31. Prueba de estudio sobre Distancia de Edición y sus variantes. Frase de entrada utilizada: “ <i>del río aumento del caudal</i> ”.	108
7.32. Prueba de estudio sobre Distancia de Edición y sus variantes. Frase de entrada utilizada: “ <i>en casa de palo cuchillo de herrero</i> ”.	108
7.33. Prueba de estudio sobre Medida secuencial y sus variantes. Frase de entrada utilizada: “ <i>el primer hombre que la luna pisó</i> ”.	110
7.34. Prueba de estudio sobre Medida secuencial y sus variantes. Frase de entrada utilizada: “ <i>la bebe dormía la en su cuna</i> ”.	110
7.35. Prueba de estudio sobre Medida secuencial y sus variantes. Frase de entrada utilizada: “ <i>el cigarrillo puede cáncer producir</i> ”.	111
7.36. Prueba de estudio sobre Medida secuencial y sus variantes. Frase de entrada utilizada: “ <i>del río aumento del caudal</i> ”.	111
7.37. Prueba de estudio sobre Medida secuencial y sus variantes. Frase de entrada utilizada: “ <i>en casa de palo cuchillo de herrero</i> ”.	111
7.38. Prueba de estudio sobre Métrica básica. Frase de entrada utilizada: “ <i>el primer hombre que la luna pisó</i> ”.	112
7.39. Prueba de estudio sobre Métrica básica. Frase de entrada utilizada: “ <i>la bebe dormía la en su cuna</i> ”.	112
7.40. Prueba de estudio sobre Métrica básica. Frase de entrada utilizada: “ <i>el cigarrillo puede cáncer producir</i> ”.	113
7.41. Prueba de estudio sobre Métrica básica. Frase de entrada utilizada: “ <i>del río aumento del caudal</i> ”.	113
7.42. Prueba de estudio sobre Métrica básica. Frase de entrada utilizada: “ <i>en casa de palo cuchillo de herrero</i> ”.	113