Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)

Posgrado en ciencia e ingeniería de la computación

Similitud semántica textual por medio de abstracción

T E S I S

que para obtar por el grado de:

Doctor en Ciencia e Ingeniería de la Computación

PRESENTA:

Oscar William Lithgow Serrano

DIRECTOR DE TESIS:

Dr. Luis Alberto Pineda Cortés

IIMAS, UNAM

MIEMBROS DEL COMITÉ TUTOR:

Dr. Julio Collado Vides

Centro de Ciencias Genómicas, UNAM

Dr. Fabio Rinaldi

Lingüística computacional, Universidad de Zurich

Ciudad de México, CDMX Enero, 2020

Similitud semántica textual por medio de abstracción

por

Oscar William Lithgow Serrano

M.Phil. Advanced Computer Science, University of Cambridge

Tesis presentada para obtar por el grado de

Doctor en Ciencia e Ingeniería de la Computación

en el

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
(IIMAS)

Universidad Nacional Autónoma de México

Ciudad de México, CDMX. Enero, 2020

A Debora, mi madre, por la vida, la sabiduría y su constante ejemplo de un espíritu rebelde, feliz e inquebrantable; y a mi padre por enseñarme a observar y cuestionar.

A mi hermana, mi querida simbionte, cuya agudeza, crudeza y intensidad obnubila; para muestra este posgrado.

A mi abuela Berta, a Téte, a Gina y a Raúl; familia que siempre ha estado para guiar, ayudar y escuchar, que quiero profundamente y sin quienes no sería lo que soy.

A Yolo mi esposa que con una lógica de otro mundo, su pícara ingenuidad y su ser da alegría a mis días.

A mis hijos Mina y Quillán, mis grandes motivos y cuyas risas son el significante de la vida en plenitud.

Oscar William Lithgow Serrano

Agradecimientos

Agradezco a mi asesor el Dr. Luis Alberto Pineda Cortés, quién siempre estuvo dispuesto para guiarme y cuyo tesón, cuestionamientos y profundo conocimiento fueron imprescindibles.

Agradezco profundamente al Dr. Julio Collado Vides por la guía dada en este proyecto, por confiar en mi y darme la oportunidad de colaborar en el estimulante ambiente del Programa de Genómica Computacional del Centro de Ciencias Genómicas (PGC). No quiero dejar de también agradecer su valioso ejemplo de cómo hacer ciencia con alegría.

Agradezco al Dr. Fabio Rinaldi, Dr. Ivan Vladimir Meza y Dr. Gribrán Fuentes Pineda por sus observaciones y consejos como revisores de este proyecto, asimismo agradezco al Dr. Ivan Meza y al Dr. Jorge García Flores por sus valiosos consejos durante el posgrado pero sobretodo agradezco su amistad.

Agradezco a los miembros del laboratorio del PGC, en especial a la Mtra. Scorro Gama Castro y a la Dra. Cecilia Ishida Gutiérrez quienes me permitieron y apoyaron para experimentar en sus dominios. Y por supuesto a Alberto Santos, Cesar Bonavides, Victor del Moral, Concepción Hernández y demás miembros del laboratorio.

Agradezco al Posgrado en Ciencia e Ingeniería de la computación por proveer un excelente marco para mi desarrollo. En particular agradezco al Dr. Javier Gómez Castellanos, a la Sra. Ma. de Lourdes González Lora, a la Ing. Cecilia Mandujano Gordillo y a Srta. Amalia J. Arriaga Campos por todo su apoyo.

Finalmente, quiero agradecer a la UNAM por la valiosa oportunidad de pertenecer a tan prestigiosa institución y al laboratorio de Genómica Computacional que por medio de su proyecto con el National Institutes of Health (grant number 5R01GM110597) me apoyó económicamente para la consecución de este doctorado.

Índice general

Similitud semántica textual por medio de abstracción

por

Oscar William Lithgow Serrano

Resumen

Comparar entidades u objetos para determinar qué tanto se parecen es una actividad común en los procesos computacionales que manejan información. Cuando lo que se compara es el significado de expresiones de lenguaje escrito se le conoce como medición de la similitud semántica textual (STS); y es una de las tareas básicas del área de procesamiento de lenguaje natural (PLN). Determinar qué tan similar es el significado de frases u oraciones es un paso necesario para otras tareas de PLN tales como la recuperación de información, generación de textos, resumen automático, etc. Esto ha motivado el desarrollo de un gran número de estrategias para medir la STS, las cuales abordan de formas diferentes las dificultades del lenguaje natural (lenguaje humano), como su ambigüedad. Esto se exacerba en dominios especializados en los cuales es frecuente encontrar escasez de recursos para entrenamiento; tal es el caso del área de biomédicas que se enfoca al estudio de la regulación transcripcional microbiana (RTM).

La meta de las métricas de STS no es comparar las representaciones textuales sino los significados que éstas expresan, y debido a que la finalidad de los procesos de abstracción es filtrar las características irrelevantes para la tarea, en esta tesis se propone que este tipo de procesos pueden utilizarse en una estrategia para medir la STS. En relación a esta meta se planteó la hipótesis de que si generalizamos dos expresiones en lenguaje natural hasta una representación común, existe una correlación entre la indeterminación provocada por el proceso de abstracción y la similitud semántica. Para comprobarla se aplicó la siguiente metodología: primero se generó un corpus específico al dominio RTM; posteriormente se implementaron y evaluaron las métricas del estado-del-arte que fueran compatibles con la escasez de recursos en el dominio, sus resultados constituyeron la línea base; y finalmente, se implementó y evaluó la métrica basada en abstracción que se propuso.

Tanto las métricas de la línea base como la propuesta basada en abstracción fueron evaluadas en el corpus RTM. A pesar de que la métrica basada en abstracción sólo usa *Wordnet* como fuente de conocimiento externo y que no requiere entrenamiento, obtuvo resultados superiores a los de todas las métricas evaluadas de manera individual, y sólo 2.5 % inferiores a la mejor de las estrategias de ensamble que se basa en un modelo de aprendizaje-maquina entrenado en el corpus. Además del buen desempeño, la métrica tiene la ventaja de no requerir entrenamiento y de generar paralelamente un listado de abstracciones que representan los elementos coincidentes; lo cual, favorece la explicabilidad de las calificaciones. Los resultados obtenidos comprueban la hipótesis y, además, le dan validez al enfoque de analizar el proceso de STS bajo la perspectiva de procesos de abstracción; esto abre la posibilidad de utilizar las teorías de abstracción existentes para inspirar nuevas métricas.

Capítulo 1

Introducción

Desde la invención del lenguaje la humanidad ha estado constantemente produciendo y almacenando información y en las últimas décadas ha crecido su producción y su disponibilidad de manera exponencial. Esto representa una gran oportunidad para extraer conocimiento pero también plantea retos importantes debido principalmente a dos factores: 1) que el conocimiento se encuentra inmerso en enormes cantidades de información por lo que cada vez es más importante hacer uso de procesos computacionales automatizados; 2) la información está expresada en lenguaje humano y por lo tanto los procesos computacionales deben de ser capaces de operar este tipo de lenguaje. Esto ha provocado que áreas como el *Procesamiento de Lenguaje Natural* (NLP[1]) y la *Lingüística computacional* (LC) tengan cada vez mayor auge. Dentro del área de NLP una de las tareas básicas, y paso previo para tareas más complejas, es la *medición de Similitud Semántica* (SS); es decir, determinar qué tanto se parecen los significado de dos expresiones de lenguaje natural.

Comparar el significado, de palabras, frases y oraciones (*similitud semántica*) es una tarea

[1]Por sus siglas en inglés - *Natural Language Processing*

común para los humanos y en la mayoría de los casos sencilla, por el contrario diseñar algoritmos que trabajen con el significado asociado a las representaciones textuales tiene grandes retos. Como la ambigüedad del lenguaje natural, ya que una misma palabra puede tener diferentes significados (por polisemia u homonimia), o palabras diferentes pueden significar lo mismo (sinonimia); en el caso de frases u oraciones se tienen múltiples opciones de cuáles palabras usar y de qué manera combinarlas para expresar un mismo significado. Se han propuesto muchos métodos para medir la similitud semántica entre textos (STS[2]), y aunque en la actualidad hay estrategias que dan muy buenos resultados (>90 % de precisión) no se puede considerar como una tarea resulta. Entre los principales factores están: 1) la dificultad de trasladar los modelos exitosos en un dominio controlado a otros dominios de lenguaje más reales y/o más especializados (*adaptación del dominio*) y 2) la dificultad para explicar cómo se llegó a los resultados de muchos modelos actuales (*explicabilidad*).

La presente tesis se enfoca en investigar la similitud semántica textual entre oraciones dentro de un dominio de lenguaje especializado para el cual se cuenta con pocos recursos de entrenamiento. Además se busca que la estrategia propuesta genere resultados que puedan explicarse más fácilmente. Estudios en diversas disciplinas muestran que antes de realizar comparaciones de semejanza entre dos elementos, los humanos abstraemos representaciones que facilitan la comparación semántica. Con base en esta premisa, se propone medir la similitud semántica entre dos oraciones por medio de estrategias que integren procesos explícitos de abstracción. Se plantea que si se generalizan dos expresiones de lenguaje natural hasta una representación común, existe una correlación entre el *nivel de abstracción* y la *similitud semántica* de ambas expresiones. A la vez, se encara el objetivo de la explicabilidad al generar la representación-común-abstracta en lenguaje natural.

[2]Por sus siglas en inglés - Semantic Textual Similarity

Como parte de la metodología aplicada se generó un corpus de similitud semántica entre oraciones representativo del dominio de *regulación transcripcional microbiana* (RTM). Se usó un conjunto de datos como fuente de entrenamiento y de validación, y se generó una línea base al ensamblar varias métricas de similitud semántica compatibles con el tamaño del corpus de entrenamiento.[3] Finalmente, se propuso un métrica de STS que integra procedimientos de abstracción y que produce además de una calificación de similitud, una representación abstracta en lenguaje natural que caracteriza la similitud entre las oraciones comparadas. El corpus por sí mismo representa el primero en su tipo en el dominio de RTM y el conocimiento adquirido durante su construcción puede ser un recurso valioso para proyectos similares. La métrica que se propuso e implementó tuvo resultados cuantitativos ligeramente mejores que la línea base formada a partir de las técnicas del estado-del-arte. Pero más importante, a mi mejor entender, es la primera métrica de STS que utiliza explícitamente procesos de abstracción con lo cual se abre una nueva perspectiva, bajo la cual se puede explorar cómo aplicar las muchas y diversas teorías formales de abstracción existentes a la tarea de comparar la semántica de dos expresiones de lenguaje.

1.1. Antecedentes

1.1.1. Procesamiento de Lenguaje Natural

La *Lingüística Computacional* (LC) y el *procesamiento de lenguaje natural* (NLP[4]) son las dos áreas de la computación que se enfocan al estudio del lenguaje humano. Sus inicios se dan a finales de los años 40 con las primeras investigaciones sobre traducción automática de textos [1], y a principios de los años 50 cuando Alan Turing propuso una prueba basada en conversar con

[3] Pequeño para las algunas técnicas actuales de STS, ej., basadas en deep learning.
[4] Por sus siglas en inglés - *Natural Language Processing*

la computadora para determinar si ésta "podía pensar" [2]. Se puede decir que la LC se gestó desde la lingüística y por lo tanto está más enfocada en los procesos lingüísticos, mientras que el NLP nació en los departamentos de computación y pone mayor énfasis en procesar el lenguaje para atender tareas prácticas y específicas. Sin embargo, ambas áreas tienen como fin general entender, interpretar y utilizar el lenguaje humano, por lo tanto, su línea divisoria es cada vez menos evidente.

Muchas disciplinas (ej. matemáticas, ciencias cognitivas, psicología, etc.) aportan modelos y herramientas a la LC y NLP, pero su base son las ciencias computacionales y la lingüística. Debido al tamaño y complejidad del lenguaje, su estudio se aborda desde varias perspectivas; cada una enfocada a un nivel de sistema. El estudio de los sonidos se aborda desde la *fonética y fonología*. La perspectiva *prosódica y entonativa* se enfoca a la relación entre los tonos y la intención comunicativa. El análisis de las formas de las palabras y de sus componentes con significado es visto desde la perspectiva *morfológica*. Las relaciones de estructura entre las unidades del lenguaje se aborda por la *sintáctica*. Finalmente, la *semántica* se avoca al estudio del significado puramente asociado a la expresión, mientras que la *pragmática* analiza el significado bajo un contexto más amplio en el que la meta comunicativa afecta al significado [3].

1.1.2. Similitud Semántica Textual

Dentro de la perspectiva semántica, una de las tareas básicas es *medir la similitud semántica* (SS). La SS consiste en comparar dos o más expresiones en lenguaje humano y generar una calificación que representa la similitud de sus significados. Cuando esta medición se realiza sobre lenguaje escrito (texto) se le denomina *Similitud Semántica Textual* (STS[5]) y puede aplicarse a diferentes unidades como son: palabras, frases, oraciones, párrafos, etc. Más recientemente el

[5]Por sus siglas en inglés.

estudio de la semántica de oraciones y párrafos ha tenido un auge significativo; debido principalmente a que, después de las palabras, son el siguiente nivel de composición semántica y, además, las oraciones se consideran como la unidad más pequeña con sentido completo. Por ejemplo, al evaluar la STS entre las oraciones "Ellos volaron del nido en grupos" y "Volaron hacia el nido juntos" se observa que aún cuando las dos oraciones no están expresando el mismo sentido, sí comparten parte del significado. Ambas se refieren a la acción de volar y en las dos los sujetos son un grupo; la diferencia es la dirección de vuelo en relación al nido, en una se dirigen hacia el nido y en la otra se alejan.

La STS tiene muchas aplicaciones prácticas, en ocasiones como un proceso intermedio en tareas más complejas. Entre las tareas que incorporan STS se encuentran: recuperación de información, generación de textos, resumen automático, identificación de estructuras discursivas, detección de plagio y detección de contenido duplicado [4, 5, 6, 7, 8, 9].

Otras tareas que están estrechamente relacionadas a la STS son *Detección de paráfrasis* y *detección de implicación textual* (text-entailment). La detección de paráfrasis se enfoca sólo en determinar si un texto es paráfrasis de otro por lo que se descartan grados menores de similitud; si se aplicara en los ejemplos del párrafo anterior, el resultado sería "falso". Por su parte, la detección de implicación está enfocada en determinar si dados dos textos se puede determinar si el significado de uno puede ser inferido a partir del otro; por ejemplo, "El niño ronca → el niño está durmiendo".[6] Estas tareas son diferentes a pesar de que las tres se enfocan en estudiar la relación entre los significados de los textos.

[6] Una relación de texto → hipótesis.

1.2. Planteamiento del problema y motivación

1.2.1. Similitud semántica textual en contextos especializados y con pocos datos de entrenamiento

La gran mayoría de estrategias del estado-del-arte de STS hacen uso de recursos estadísticos y probabilísticos generados de manera no supervisada a partir de grandes corpus de texto [10, 11]. Otra de las tendencias populares es el uso de modelos de aprendizaje máquina basados en redes neuronales [12, 13], que en buena parte son de tipo supervisado y requieren grandes cantidades de datos etiquetados para ser entrenados.

Sin embargo, en dominios especializados es frecuente la escasez de recursos disponibles para el entrenamiento e incluso para la validación de estrategias de NLP, incluyendo la tarea de STS. El problema es que, por un lado, compilar, depurar y etiquetar recursos específicos al dominio es una tarea compleja y costosa [14, 15], y por otro, es común que el vocabulario y los patrones específicos del dominio no se encuentren, o se encuentren escasamente, en recursos y modelos generados a partir de recursos generales [16]. Esto provoca que las estrategias del estado-del-arte frecuentemente queden excluidas o se vean menguadas en dominios especializados.

A estas limitaciones se debe el aumento de interés en desarrollar estrategias que atiendan la escasez de recursos. Uno de los enfoques más populares consiste en generar automáticamente más casos de entrenamiento a partir de recursos creados de manera manual (*aprendizaje semi-supervisado*) [17]. Otras estrategias son el *aprendizaje activo* y el *aprendizaje de supervisión débil*, cuyo objetivo es etiquetar menos ejemplares al seleccionar aquellos que aporten más información [18]. Finalmente, una estrategia cada vez más popular es la *transferencia de aprendizaje*, la cual consiste en entrenar los modelos en dominios generales y después sólo ajustarlos o refinarlos con recursos pequeños pero específicos del dominio [19, 20].

Un perspectiva diferente para abordar la escasez de recursos es desarrollar estrategias que requieran menos cantidad de datos. Esto puede lograrse combinando modelos no supervisados de representación semántica (*word-embeddings*) y métricas que no requieran aprendizaje. En esta tesis se siguió este enfoque para proponer una estrategia de STS que pueda aplicarse a dominios especializados con escasez de recursos de entrenamiento.

1.2.2. Explicabilidad

Otro problema de las estrategias del estado-del-arte, no sólo en STS sino en general en *inteligencia artificial* (IA) es que a menudo no se puede explicar su comportamiento y procesos de decisión. Modelos de aprendizaje profundo y de ensamble frecuentemente obtienen el mejor desempeño, pero también están entre los que más sufren de falta de transparencia en sus resultados [21]. Es importante que los modelos de IA sean más transparentes y permitan la explicación de sus decisiones porque facilita identificar problemas y sesgos potenciales, asegurar algoritmos con decisiones justas (no discriminación) y confirmar que funcionan como se espera [22].

Estos criterios, y no sólo los de desempeño (ej., precisión), deben ser cumplidos para que modelos de IA se puedan usar de forma segura en aplicaciones de la vida real [22, 23]. Esta necesidad ha provocado el auge del área de *inteligencia artificial explicatoria* (XAI[7]) en la que se promueve que se pueda verificar la razón de cada decisión de los modelos y que se entienda mejor el fenómeno modelado. También hay esfuerzos específicos del campo de NLP para generar explicaciones a la par de los resultados del proceso [24].

Por las razones anteriores, en este estudio se tiene como objetivo generar estrategias que favorezcan la explicabilidad de los resultados. También se desea que las representaciones semánticas se puedan reinterpretar, y que esta reinterpretación sea más fácil gracias a que las representa-

[7] Por sus siglas en inglés.

ciones estén en vocabulario y reglas gramaticales conocidas por los intérpretes [25, 26]. Por lo tanto este estudio se apoya en una línea de investigación que propone que el lenguaje natural es adecuado para representar su propia semántica [27, 28]. Su fundamento consiste en que si se usan versiones racionalizadas y restringidas del lenguaje natural es posible formular hipótesis comprobables de las intenciones expresadas a través del lenguaje [28, 29]. En resumen, un objetivo este estudio es usar representaciones semánticas en lenguaje natural para facilitar su reinterpretación y favorecer la explicabilidad de los resultados.

1.2.3. Abstracción

Debido a que el objetivo de la STS es comparar el significado, un componente clave es cómo representar la información semántica. Es común que estrategias de STS hagan uso de estructuras que representan el significado del texto (o parte de él) a través de incluir información sobre las relaciones entre los términos del texto, sus categorías gramaticales, sus roles semánticos, su distribución estadística dentro de un corpus, su posición en una taxonomía, o por la asignación arbitraria de significado a un símbolo. Al generar estas representaciones se busca separar el significado de las características propias de la expresión textual (tales como, si es voz pasiva o activa, si es predicado verbal o nominal, si es una nominalización, etc.); gracias a esto se reduce la ambigüedad y se facilita la comparación semántica. Estrategias de este tipo no sólo son aplicadas en procesos de NLP, sino que diversos estudios muestran que también los humanos realizamos esta separación e interpretación antes de comparar objetos [30, 31, 32, 33, 34]. A este mecanismo para transformar la representación original en otras que sean más adecuadas para usarlas en una tarea específica, se le conoce como *proceso de abstracción* [35].

Más allá de esta definición general e intuitiva de lo que es un proceso de abstracción, existe una amplia línea de investigación abocada a analizar la abstracción y formalizarla en teorías

[36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 34, 47, 35, 48, 49]. A pesar de que la abstracción forma parte los procesos humanos para comparar y que de manera implícita ya se usa en las estrategias de STS, este mecanismo no se aborda de manera explícita y tampoco se fundamenta aprovechando las teorías formales que existen. A mi mejor entender, la STS no se ha planteado desde esta perspectiva y, hacerlo abre la puerta para estudiar los efectos que aplicar las diferentes teorías de abstracción tendrían sobre la similitud semántica. Por estas razones, esta tesis se enfoca al problema de investigación de generar estrategias computacionales que incluyan un proceso explícito de abstracción para medir la similitud semántica entre dos oraciones.

1.2.4. El proyecto LRegulon

Un área de aplicación de la STS es en la *curación de literatura científica*, en la que personas expertas en cierto tema analizan un conjunto de publicaciones y extraen la información relevante para la tarea particular. El problema es que este proceso manual es caro y, por lo tanto, limitado en la cantidad de información que puede abarcar. Esto ha dado pie a la incorporación de estrategias de NLP para facilitar la curación.

Un caso particular es la curación de literatura para RegulonDB, la cual es una fuente de información de la *regulación transcripcional microbiana* (RTM) reconocida mundialmente. RegulonDB es una base de datos, organizada y computable, curada manualmente con la información de la regulación de la expresión de genes en E. Coli K-12 [50]. Un parte importante durante la curación de algún objeto biológico (ej., Genes, Factores de Transcripción, Interacciones de Regulación, etc.) es recabar la mayor cantidad de evidencia posible de lo que se está curando. Esto se logra buscando declaraciones, dentro del mismo artículo y en otros artículos que respalden la proposición curada. Descubrir información relevante para ser curada también se facilita cuando una proposición se ubica en contextos diferentes (artículos diferentes). En ambos casos la bús-

queda de proposiciones semejantes por medio de la STS de oraciones provee una herramienta útil para facilitar y mejorar la curación.

La colaboración con el Programa de Genómica Computacional del Centro de Ciencias Genómicas de la UNAM permitió usar las estrategias de STS analizadas y propuestas en esta tesis en una aplicación práctica y, además en un dominio especializado con escasez de recursos. Esto dio pie al proyecto LRegulon, que consiste en aplicar métodos de STS para construir redes de oraciones de la literatura de RTM e interfaces gráficas que permiten al usuario navegar en estas redes como un método diferente de lectura. En estas redes, los nodos representan las oraciones extraídas de los artículos científicos y las aristas la similitud semántica entre las oraciones. La interfaz permite navegar en diferentes conjuntos de artículos agrupados de acuerdo al criterio de análisis que se desea (ej., publicaciones específicas de RIs, publicaciones relacionadas a cierta condición de crecimiento, etc.). Al visualizar el contenido de una publicación, el usuario tiene la posibilidad de seleccionar alguna de las oraciones que sea de su interés, y la aplicación mostrará otras oraciones, de la misma publicación y de las otras publicaciones del conjunto, que estén relacionadas semánticamente. Las oraciones mostradas son hipervínculos que posicionan al usuario en el contexto en donde se encontró la oración relacionada. Debido a que esta operación puede ser repetida, el resultado es que el usuario tiene la posibilidad de cambiar una lectura vertical enfocada sólo en el tema de la publicación a una lectura horizontal que se desarrolla a través de varios artículos. Es decir, rastrear una idea de su interés a través de contextos diversos (diferentes publicaciones). En pruebas preliminares se observó que este tipo de estrategias ayuda al curador a asociar el conocimiento presente en el corpus de manera más exhaustiva, y le permite descubrir relaciones nuevas (conocimiento) al ubicarlo en proposiciones similares (significado de oraciones) pero en contextos diferentes (artículos) [51].

1.3. Hipótesis de solución

Si generalizamos dos expresiones en lenguaje natural hasta una representación común, existe una correlación entre el *nivel de abstracción* aplicado –medido a través de la entropía y la indeterminación producida durante la abstracción– y la *similitud semántica* de ambas expresiones. A mayor entropía menor cantidad de información compartida y, por lo tanto, un menor grado de similitud semántica.

1.4. Objetivos

1.4.1. Objetivo general y meta

Generar estrategias computacionales que incluyan un proceso explícito de *abstracción* para medir la similitud semántica entre dos oraciones dentro de dominios especializados y con pocos datos de entrenamiento.

1.4.2. Objetivos específicos

Bajo el contexto de un dominio especializado y para el cual se cuenta con pocos datos de entrenamiento

1. Generar corpus de entrenamiento y evaluación: Construir un corpus de pares de oraciones específicas al dominio de interés, para las cuales su similitud semántica sea evaluada.
2. Obtener un línea base con estrategias tradicionales de medición de similitud semántica: Aplicar aquellas estrategias del estado-del-arte que sean compatibles con el dominio y la disponibilidad de datos de entrenamiento.

3. Probar hipótesis: Proponer y aplicar una estrategia de similitud-semántica-entre-oraciones que esté guiada por procesos explícitos de abstracción.

1.5. Contribuciones y publicaciones

1.5.1. Contribuciones

- Una métrica para medir similitud semántica textual entre oraciones con las siguientes ventajas:
  - Aplica procesos de abstracción para generar una representación, en lenguaje natural, común a ambas oraciones. Esto favorece la explicabilidad de los resultados.
  - No requiere entrenamiento y sólo usa *Wordnet* como fuente externa de conocimiento.
- El planteamiento de la abstracción como un proceso explícito dentro del la STS; esto favorece la posibilidad de aprovechar las teorías existentes de abstracción para inspirar estrategias futuras.

1.5.2. Publicaciones

1. Rinaldi, F., Lithgow-Serrano, O., López-Fuentes, A., Gama-Castro, S., Balderas-Martínez, Y. I., Solano-Lira, H., & Collado-Vides, J. (2015). An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database. Polibits, 52, 25–31. https://doi.org/10.17562/PB-52-3

2. Lithgow Serrano, O. W., Meza Ruiz, I. V., Orozco Camacho, A. M., Garcia Flores, J., & Buscaldi, D. (2016). LIPN-IIMAS at SemEval-2016 Task 1: Random Forest Regression Experiments on Align-and-Differentiate and Word Embeddings penalizing strategies. In

Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 726–731). San Diego, California: Association for Computational Linguistics.

3. **(JCR)** Rinaldi F, Lithgow O, Gama-Castro S, Solano H, Lopez A, Rascado LJM, et al. Strategies towards digital and semi-automated curation in RegulonDB. Database [Internet]. 2017;2017:1–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467572/

   ** Incluye nota especificando misma contribución de los 2 primeros autores*

4. **(JCR)** Lithgow-Serrano O, Gama-Castro S, Ishida-Gutiérrez C, Mejía-Almonte C, Tierrafría VH, Martínez-Luna S, et al. Similarity corpus on microbial transcriptional regulation. Journal of Biomedical Semantics [Internet]. Cold Spring Harbor Laboratory; 2019 [cited 2017 Nov 22];10:8. Available from: https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-019-0200-x

5. **(JCR)** Lithgow-Serrano O, Collado-Vides J. In the pursuit of semantic similarity for literature on microbial transcriptional regulation. Pinto D, Singh V, editors. Journal of Intelligent & Fuzzy Systems [Internet]. 2019;36:4777–86. Available from: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JIFS-179026

1.6. Organización de la tesis

El resto de la tesis está organizado de la siguiente manera: en el capítulo 2 se realiza una investigación del estado-del-arte de las estrategias para medir la STS; en el capítulo 3 se exponen las teorías más relevantes que fundamentan mi propuesta para medir la STS; el capítulo 4 describe la construcción del corpus de similitud a partir de literatura del tema de regulación transcripcional microbiana; en el capítulo 5 se describen las métricas del estado-del-arte que

se implementaron y evaluaron, y que constituyen la línea base de este estudio; la propuesta y evaluación de la métrica nueva de STS que incluye procesos de abstracción se expone en el capítulo 6; y, finalmente, en el capítulo 7 se hace una recapitulación breve, se exponen las conclusiones y el trabajo futuro.

Capítulo 2

Similitud semántica textual

En este capítulo se realiza una revisión del estado-del-arte en la tarea de similitud semántica textual (STS). Primero se introduce el concepto de similitud semántica de manera general, y después como una medida. Posteriormente se analiza la similitud entre palabras y, finalmente, entre unidades mayores como frases y oraciones.

2.1. Similitud semántica

Comparar estímulos/cosas (en adelante *entidades*) es una habilidad humana. Además de objetos concretos se pueden comparar entidades abstractas como situaciones, conceptos e ideas. Esta capacidad de reconocer similitudes entre entidades es un componente central en tareas de categorización, reconocimiento de patrones, recuperación de memoria, resolución de problemas, así como en las teorías de transferencia de aprendizaje las cuales se basan en identificar habilidades conocidas para facilitar el aprendizaje de nuevas [52, 53, 54].

Esta habilidad se ha estudiado desde perspectivas cognitivas, psicológicas, neuro-cognitivas,

etc., bajo las cuales se ha llegado a cierto consenso en algunas características del proceso. Bajo la perspectiva computacional los siguientes elementos son relevantes en el proceso de comparación semántica: 1) previo a la comparación se crea una representación en un espacio psicológico propio y por lo tanto subjetivo [31, 32, 33, 30, 34]; 2) emular este espacio psicológico a través de un espacio geométrico en donde medir la distancia semántica ha dado buenos resultados empíricos [55, 56, 32, 57, 58, 59]; 3) la relación de similitud entre elementos no es invariante sino que se ajusta al objetivo y contexto [60].

La evaluación de SS también se aplica a expresiones de lenguaje natural para comparar su significado semántico. Si se toma al *significado* como la interpretación y representación mental de una entidad [61], entonces el significado asociado a los signos de lenguaje se conoce como *significado semántico* [4].[1]

Debido a la complejidad del lenguaje natural su análisis se realiza en varios niveles (fonético, morfológico, sintáctico, etc.), de los cuales el semántico es uno de los más complejos. Una de las tareas básicas en este nivel es la evaluación de la similitud semántica entre textos (STS)[62]. Esta tarea está orientada a comparar semejanzas en el significado de unidades de lenguaje de diversas granularidades, tales como palabras, oraciones, párrafos y documentos [63, 64]. La STS es una técnica clave en estrategias computacionales que buscan procesar conocimiento [65, 66]. Tareas como la recuperación de información, generación de textos, resumen automático, identificación de estructuras discursivas, detección de plagio, detección de contenido duplicado y paráfrasis, son ejemplos prácticos de su uso [6, 7, 8, 9, 67, 68].

[1] Diferente por ejemplo al significado pragmático en donde se trata la intención comunicativa tomando en cuenta también el contexto externo a la expresión.

La similitud semántica como medida

Es frecuente que las medidas de SS no puedan ceñirse a las propiedades formales (matemáticas) de medida y/o distancia que incluyen simetría, no-degeneración y desigualdad triangular. Por ejemplo, las medidas de distancia respetan la propiedad de la simetría, es decir $d(a, b) == d(b, a)$; sin embargo se ha demostrado que en ciertos casos las personas perciben de manera asimétrica la similitud dependiendo del orden de comparación, ej., una elipse es percibida más similar a un círculo que un círculo a una elipse [30, 69]. Debido a esto, la mayoría de las medidas de STS se fundamentan sólo en la intuición de que la similitud entre dos entidades debe ser mayor mientras mayor sea la preponderancia de sus semejanzas estructurales, de relación, de uso o de contexto. Es decir, cualquier función matemática, algoritmo o herramienta teórica que permita la comparación de entidades de acuerdo a su evidencia semántica puede ser considerada como una medida semántica.

En el contexto de la evaluación de SS se pueden encontrar diferentes nociones y términos que se asocian a la similitud semántica. Entre las más usadas encontramos la distancia semántica, la relación semántica y la similitud semántica. No existe una clara distinción entre estas nociones y además su definición varía entre autores, pero es frecuente que cuando se busca diferenciarlas sean interpretadas de la siguiente manera: la *distancia semántica* se asocia a la cantidad y fuerza de los vínculos semánticos entre dos entidades, es decir, mientras más vinculadas semánticamente menor la distancia; la *relación semántica* se toma como el inverso de la distancia semántica, por lo que a menor distancia mayor relación semántica; y la *similitud semántica* es una especialización de la relación semántica, es decir, también está basada en la cantidad y fuerza de los vínculos, pero sólo se toman en cuenta los de tipo taxonómico [70, 71, 72, 73].

En la práctica lo más frecuente es que como medida de SS se use cualquier herramienta matemática, función o algoritmo con la cual se pueda estimar de manera cuantitativa la fuerza

de relación semántica entre unidades del lenguaje a través de un valor numérico obtenido por la comparación de información que, de manera explícita o implícita, caracterice el significado de las entidades. Es decir, $\sigma_k : E_k \times E_k \rightarrow \mathbb{R}^+$, en donde E_k es el conjunto de entidades de lenguaje de tipo k (palabra, frase, oración, etc.) y al comparar las unidades se obtiene un valor real que representa la similitud σ_k [4].

Una de las clasificaciones más populares de medidas de STS es de acuerdo a la unidad lingüística que se compara, y consiste en dos grupos: las medidas para comparar palabras y las medidas para unidades mayores (frases y oraciones). Las medidas del primer grupo se enfrentan a la ambigüedad presente en forma de sinonimia, polisemia, homonimia, etc.; y las del segundo grupo además deben considerar la ambigüedad que se deriva de la composicionalidad del lenguaje, es decir, formas diferentes de conjuntar las palabras para expresar el mismo significado. Cabe mencionar que la evaluación de STS en textos mayores a párrafos, ej., documentos, generalmente no se aborda bajo la noción de composicionalidad. Esto se debe a que en la práctica la cantidad de evidencia semántica en textos largos permite que estrategias estadísticas y probabilísticas de tipo bolsas de palabras den buenos resultados con un menor costo computacional. En las siguientes secciones se revisan algunas estrategias representativas de STS de acuerdo a esta clasificación.

2.2. Similitud semántica entre palabras

Existen medidas que sólo se enfocan en comparar la representación textual; por ejemplo las medidas de distancia de edición que se basan en el número de cambios que se deben realizar en un texto para igualarlo a otro. Aunque es posible argumentar que este tipo de medidas consideran implícitamente la semántica –ya que las letras que conforman algunas palabras están

asociadas a su etimología que a su vez está asociada al significado– se desea que las medidas de STS aborden de manera más directa el significado. Por esta razón en la práctica se ha optado por sólo considerar como medidas de STS aquellas que usan una fuente externa[2] de información semántica.

2.2.1. De acuerdo a la fuente de conocimiento semántico

Se puede clasificar a las medidas por el tipo de recurso que usan: estructurados y no estructurados.

Basadas en recursos estructurados

Entre los recursos estructurados más comunes encontramos ontologías, taxonomías, bases de datos léxicas y diccionarios. En estos recursos, las personas codifican de manera explícita el conocimiento que tienen de las entidades a través de su descripción, su posición y cómo están relacionadas con otras entidades. En una *taxonomía* las entidades se ordenan de manera sistemática y jerarquizada formando clases de elementos con características similares. Las *ontologías* son sistemas de descripciones abstractas y comúnmente jerárquicas que representan el entendimiento común de dominios específicos de conocimiento por medio de la especificación formal de sus relaciones semánticas. Aun cuando las ontologías pueden incluir constructos lógicos complejos es frecuente que para el procesamiento computacional se usen representaciones parciales con sólo unos cuantos tipos de relaciones semánticas (ej., relaciones de tipo taxonómicas: "is-a"). De esta manera se pueden representar en grafos o redes semánticas simples que brindan buen desempeño y son más fáciles de mantener. Por su parte, en las *bases léxicas* se agrupan

[2]Se identifica como *externa* a toda fuente de conocimiento diferente al texto específico (oración, frase o palabra) que se está comparando.

"conceptos" ("Synsets") formados por conjuntos de palabras que son sinónimos. A cada uno de ellos se le adjuntan definiciones, ejemplos de uso y relaciones lógicas con otros synsets, tales como hiperónimos e hipónimos (un concepto es más o menos general que otro), y merónimos y holónimos (un concepto engloba a otro o es parte de otro).

El ejemplo más representativo de base léxica es Wordnet [74, 75], el cual es uno de los recursos estructurados más usados como fuente para STS. Entre las estrategias basadas en Wordnet encontramos las que usan la distancia más corta entre dos conceptos para estimar su similitud. Un ejemplo es la medida Leacock-Chodorow (eq. 2-1) que considera la distancia y, además, la profundidad de los conceptos en la jerarquía con el objetivo de aliviar la granularidad no-homogenea de Wordnet[73]. Otra medida es la de Wu-Palmer [76] que mide la profundidad relativa de los dos conceptos respecto al mínimo común agrupador (LCS[3]) (eq. 2-2). Estas medidas sólo consideran la estructura de la taxonomía por lo que no siempre se correlacionan bien con la similitud semántica percibida por las personas. En respuesta se han propuesto medidas que se basan en el contenido de información ($IC(w) = -logP(c)$) de los conceptos (c) representados por las palabras (w) que se desean comparar. Para calcular la IC se aproxima la probabilidad de un concepto por medio de la frecuencia con que ocurren en un corpus, esto resulta en que palabras más frecuentes sean menos informativas que las menos frecuentes. Resnik [77] propone una medida basada en el contenido de información del LCS (eq. 2-3). Lin [78] extiende la propuesta de Resnik al considerar en el cálculo la diferencia de propiedades entre los conceptos por medio de normalizar la medida con el IC de cada concepto (eq. 2-4). Una aproximación diferente es la medida de Lesk extendida [79, 80], esta usa el sobrelape de las descripciones de los conceptos ($gloss$) comparados y también la de sus relaciones directas (en donde R son las posibles relaciones) como hiperónimos e hipónimos (eq. 2-5). La medida de Pirro [81] presenta una perspectiva diferente sobre el contenido de información, usa la noción de IC

[3]Least Common Subsumer

intrínseco (iIC) que se basa en el número de subconceptos del concepto y el total de conceptos en la taxonomía. El iIC máximo se obtiene cuando el concepto no puede ser diferenciado más (hojas de la taxonomía) (eq. 2-6). Con el objetivo de considerar no solo la similitud sino también la relación semántica Pirro usa una IC-extendida (eq. 2-7) que combina la iIC del concepto con el promedio de iIC de todas las relaciones semánticas (m) que conectan con el concepto, en donde para cada relación se promedia la IC-intrínseca de los n conceptos conectados por medio de esa relación R. El resultado es una medida basada en la similitud de las propiedades de los conceptos, en donde las propiedades se extraen de sus relaciones con otros conceptos (eq. 2-8).[4]

$$sim(w_1, w_2) = -log\frac{min_length(w_1, w_2)}{2 * Depth_{max}} \tag{2-1}$$

$$sim(w_1, w_2) = 2 * \frac{depth(LCS(w_1, w_2))}{depth(w_1) + depth(w_2)} \tag{2-2}$$

$$sim(w_1, w_2) = IC(LCS(w_1, w_2)) = -logP(LCS(w_1, w_2)) \tag{2-3}$$

$$sim(w_1, w_2) = 2 * \frac{IC(LCS(w_1, w_2))}{IC(w_1) + IC(w_2)} \tag{2-4}$$

$$sim(w_1, w_2) = \sum_{r,q \in R} overlap(gloss(r(w_1)), gloss(q(w_2)) \tag{2-5}$$

[4]Similar al modelo de características [30]

$$iIC(w) = 1 - \frac{log(\#subconcepts(w) + 1)}{log(max_con)} \tag{2-6}$$

$$eIC(w) = (\alpha * iIC) + (\beta * \sum_{j=1}^{m} \frac{\sum_{k=1}^{n} iIC(w_k \in W_Rj)}{|W_Rj|}) \tag{2-7}$$

$$sim(w_1, w_2) = \frac{1}{eIC(w_1) + eIC(w_2) - 2 * eIC(LCS(w_1, w_2))} \tag{2-8}$$

Las medidas anteriores se pueden aplicar a otras taxonomías e incluso a ontologías. El uso en ontologías normalmente se hace sobre una representación parcial que sólo incluye algunos tipos de relaciones (ej., "is-a"). También es común que ontologías y taxonomías se transformen a grafos ($G = (V, E)$) en donde cada nodo (E) representa una entidad (concepto, palabra, etc.) y las aristas (V) las relaciones. Este tipo de representación permite usar teoría de grafos y de sus algoritmos para proponer medidas de similitud. Por ejemplo, medidas que estiman la relación semántica entre dos nodos usando la cantidad de interconexiones (directas o indirectas) entre ellos. También están las medidas que calculan la similitud en función del peso de la ruta más corta que conecta los nodos [82]. Otro tipo son las medidas que realizan caminatas aleatorias [83] sobre el grafo, es decir, se camina de nodo a nodo con una probabilidad de transición y una vez que se logra llegar de la entidad e_1 a la entidad e_2, se pueden usar varias medidas para estimar la similitud de los nodos. Dos ejemplos son el número promedio de pasos para conectar los nodos (hitting time) y el tiempo promedio para ir de e_1 a e_2 y de regreso a e_1 (commute time). [84]

Las medidas descritas ejemplifican diferentes enfoques sobre el tipo de información que se usa para evaluar la similitud. De acuerdo a este criterio se pueden clasificar en las que se basan

en la estructura del recurso (ej., distancia entre nodos), las que se basan en el contenido de información, las que se basan en el análisis de características de los conceptos, y los modelos híbridos [85, 86].

Basadas en recursos no estructurados

Los recursos estructurados son inexistentes o insuficientes en muchos dominios debido al costo que conlleva su creación. Esto ha motivado una creciente atención a enfoques descriptivos que extraen la información semántica a partir de una fuente no estructurada o semi-estructurada. Existen varias estrategias para generar modelos de lenguaje a través de aplicar técnicas estadísticas y probabilísticas sobre los corpora. Las estrategias se pueden agrupar por el tipo de técnica que se usa, pero de manera más general es posible clasificarlas de acuerdo al fundamento teórico sobre el que se extrae la información semántica.

El enfoque más frecuente, por mucho, es el de la *semántica distribucional* [61, 87] para el cual el significado de las palabras está asociado a su distribución en el texto. Se basa en la *hipótesis distribucional* [88] que propone que palabras que ocurren en contextos similares tienen significados similares o relacionados. Este tipo de modelos captura dos tipos de relaciones semánticas [89]: 1) las *Sintagmáticas*, que se dan entre las palabras que co-ocurren frecuentemente en el mismo contexto por lo que las relaciones de una palabra representan características dadas por su uso,[5] tanto así que las términos que más co-ocurren suelen representar propiedades prototípicas; ej., *sopa* y *caliente*. 2) las *paradigmáticas*, que se dan entre palabras que se rodean frecuentemente del mismo contexto y que por lo tanto pueden usarse como substitutos entre sí sin alterar significativamente el significado de un texto; ej.: *sopa* y *caldo*. En la práctica, la mayoría de los modelos distribucionales no diferencian entre similitud y relación semántica [63].

[5] La caracterizan de acuerdo al uso y no a una definición, por ejemplo el término tiburón es posible que esté caracterizado por palabras como *blanco*, *grande*, *dientes*, *miedo* y *mar*.

Uno de los modelos distribucionales más básicos es el VSM (Vector Space Model) [90]. En este las palabras se presentan como vectores multidimensionales dentro de los cuales cada dimensión representa a una palabra del vocabulario (V) que aparece en su contexto cercano (n palabras antes o después de la palabra objetivo) y la magnitud de la dimensión está dada por la frecuencia de co-ocurrencia. El vector resultante ($|W|$) tiene tantas dimensiones como palabras únicas en el vocabulario ($|W| = |V|$) y para medir la SS entre dos vectores se usan medidas como el coseno y la distancia cartesiana.

Una de las características de los VSM es que producen vectores muy dispersos que no cuentan con información en la mayoría de sus dimensiones pero tienen mucha evidencia en algunas (ej., dimensiones asociadas a palabras como *the*, *a*, *from*, *to*). Para normalizar el peso de los términos y privilegiar los más informativos se desarrollaron varias técnicas estadísticas entre las que se encuentran *TF-IDF* [91], basada en la frecuencia de co-ocurrencia local y la frecuencia global, y *PMI* [91], basada en la proporción de información mutua entre los términos. Otra característica es que las matrices de co-ocurrencia son muy grandes ($|V| \times |V|$) por lo que es común aplicar técnicas de análisis estadístico para reducir el número de dimensiones. Al eliminar filas y columnas muy correlacionadas se obtienen vectores más densos –frecuentemente entre 300 y 500 dimensiones– que conservan la mayor parte de la información y que, generalmente, tienen mejor desempeño en tareas de STS. [6] Entre las técnicas más usadas están *Singular Value Decomposition* (SVD), *Latent Semantic Analysis* (LSA) [92], *Hyperspace Language Analogue* (HLA) [93], *Probabilistic Latent Semantic Indexing* (PLSA) [94] y *Latent Dirichlet Allocation* (LDA) [95].

Recientemente han tenido auge estrategias que también generan vectores densos conocidos como *embeddings*, pero que en lugar de generarlos a partir de factorizar las matrices de co-

[6] Se argumenta que al generarlos se filtra parte del ruido de la información.

ocurrencia son obtenidos directamente del corpus usando *modelos de lenguaje neuronal* [96, 12], es decir, redes neuronales que aprenden la distribución de probabilidad sobre las secuencias de palabras (eq. 2-9) [12]. Los modelos de lenguaje neuronal aprenden un embedding para cada término de un vocabulario predefinido a partir de un corpus que provee los casos de entrenamiento. Durante cada paso del entrenamiento los embeddings se prueban en el corpus y se califican de acuerdo a la función objetivo del modelo; posteriormente el error resultante se propaga de regreso para actualizar el modelo y los valores de los vectores. Los embeddings resultantes tienen codificada la información implícita en el corpus de tal manera que permiten que el modelo satisfaga la función objetivo [97].

$$\hat{P}(w_t|w_1^{t-1}) \approx \hat{P}(w_t|w_{t-n+1}^{t-1}) \tag{2-9}$$

Word2Vec y Glove se encuentran entre las estrategias más exitosas de este tipo de modelos. En Word2Vec [98] se entrena una red neuronal de dos capas para reconstruir el contexto (palabras vecinas) de la palabra objetivo. La red neuronal consiste en una sola capa oculta totalmente conectada y una capa de salida con tantas dimensiones como el tamaño del vocabulario; la salida son las probabilidades de asignación para cada una de las palabras del vocabulario. Durante el entrenamiento, cada palabra del corpus se codifica en un vector *one-hot* que tiene todos sus componentes en 0 excepto el que corresponde al índice de la palabra al cual se le asigna el valor de 1. Después, el vector de entrada se multiplica en la capa oculta (la que contiene los pesos aprendidos hasta el momento), se obtiene una salida que representa la probabilidad de cada palabra del vocabulario de aparecer en la vecindad[7] de la palabra objetivo, se calcula el error con base en una función de pérdida y se propaga hacia atrás para ajustar los pesos de la capa oculta. Una vez terminado el entrenamiento se usan los pesos de la capa oculta como

[7] Generalmente 4 palabras alrededor de la palabra que se está aprendiendo.

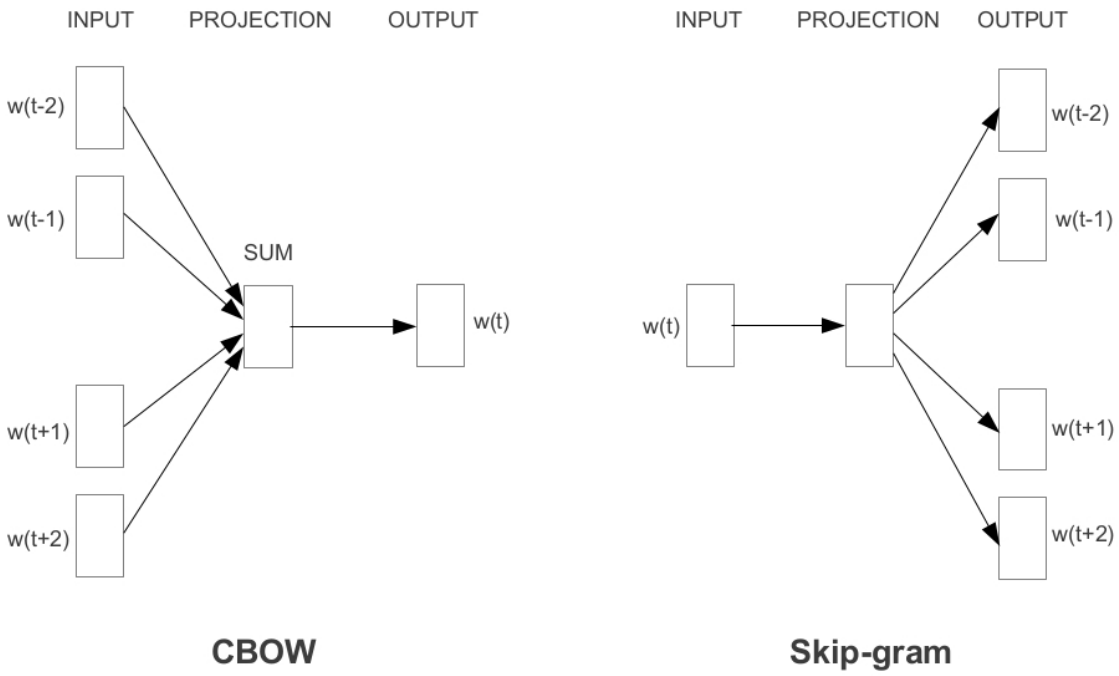

Figura 2-1: A la izquierda el modelo Word2Vec-CBOW que se entrena prediciendo la siguiente palabra a partir del contexto y a la derecha la variante Skip-Gram en la que se predice el contexto a partir de la palabra objetivo. Imagen tomada de [98].

los valores de los componentes del embedding. Por lo tanto, si se desea tener embeddings de 300 dimensiones se entrena una red con una capa oculta de 300 unidades. En Word2Vec se tienen dos variantes: CBOW en la cual se predice la palabra objetivo con base en las palabras vecinas y Skip-Gram que, por el contrario, predice el contexto (palabras vecinas) a partir de la palabra objetivo (fig. 2.2.1). En general la variantes Skip-Gram da mejores resultados y se puede representar por eq. 2-10, en donde w son palabras de un corpus, c su contexto, v_c y v_w son los vectores respectivos y se busca el parámetro θ que maximice la probabilidad del corpus.

$$argmax_\theta \prod_{w \in Text} \left[\prod_{c \in C(w)} p(c|w;\theta) \right] \quad \Big| \quad p(c|w;\theta) = \frac{e^{v_c V_w}}{\sum_{c' \in C} e^{v'_c v_w}} \tag{2-10}$$

GloVe (Global Vectors) [11], en cambio, proviene de una perspectiva más parecida a la de LSA. No sólo incorpora estadísticas locales como Word2Vec sino que también hace uso de co-ocurrencias globales de las palabras lo que le permite un mejor aprendizaje de patrones que se repiten en el corpus. En el pre-procesamiento, la matriz de co-ocurrencias se normaliza y se aplica un suavizado logarítmico. Su objetivo en el entrenamiento es aprender embeddings de manera tal que el producto punto de dos embeddings iguale el logaritmo de su probabilidad de co-ocurrencia, es decir, se asocia la distancia entre dos embeddings con la probabilidad de co-ocurrencia. La eq. 2-11 expresa la función objetivo que se busca minimizar en donde V es el vocabulario, w_i y w_j son vectores de palabras, X es la matriz de co-ocurrencia y los términos b son sesgos escalares arbitrarios ("biases") y f es una función de peso que corresponde a una función logística .

$$J = \sum_{i,j=1}^{|V|} f(x_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b_j} - log X_{ij} \right)^2 \quad \Big| \quad f(x) = \begin{cases} (x/x_{max})^\alpha & if x < x_{max} \\ 1 & else \end{cases} \tag{2-11}$$

Se puede observar que aún los modelos neuronales más recientes se basan de manera implícita en las palabras cercanas para obtener la representación semántica (embeddings). Esto se analiza en [99], en donde se muestra que es posible aproximar los modelos neuronales por medio de una factorización de matrices similar a LSA y PCA más algunos ajustes por medio de factores globales.

Si te gusta tener dinero no deberías estudiar
Deberías tener dinero si no te gusta estudiar

Tabla 2-1: Oraciones con el mismo léxico pero con significado diferente

Estrategias híbridas

Las estrategias híbridas consisten en combinar enfoques y se pueden clasificar de acuerdo a la forma de combinarlos. En las estrategias complementarias, cuando el término no se encuentra en un modelo, se recurre a otros para completar la información (ej., si la unidad no se encuentra en WordNet se usan embeddings para descubrir sinónimos que si estén presentes). En las estrategias que combinan tipos de fuentes de conocimiento se incorporan recursos estructurados durante el proceso de aprendizaje a partir de fuentes no-estructuradas; por ejemplo, se realizan caminatas aleatorias en taxonomías u ontologías para generar contextos ad hoc con los cuales se alimentan modelos neuronales para generar embeddings (ej., embeddings basados en Wordnet). En las estrategias basadas en aprendizaje máquina, usando modelos de regresión, se combinan las calificaciones de similitud obtenidas con diferentes medidas de SS.

2.3. Similitud semántica entre frases y oraciones

Las estrategias descritas hasta el momento sólo atienden la comparación de significado entre palabras y, en general, no pueden aplicarse directamente a unidades mayores como frases y oraciones. La razón es que en estas unidades, además del significado asociado a cada palabra, está involucrada la semántica derivada de la composición del lenguaje, es decir, el significado que se forma cuando se combinan las palabras en cierto orden. Esto resulta en oraciones con palabras diferentes pero que expresan el mismo significado y oraciones con las mismas palabras pero con significado diferente, como se puede ver en el ejemplo siguiente:

El muchacho bebe en tarro
El joven toma en taza

Tabla 2-2: Oraciones con léxico diferente pero con significado similar

Existen varias estrategias diseñadas para medir la SS entre frases y oraciones. Una primera clasificación consiste en distinguir entre las que generan una representación compuesta de las frases antes de compararlas y las que no.

Dentro de la última clase encontramos estrategias de *alineación* que consisten en formar los pares de palabras (una de cada oración) que más se parezcan entre sí y, una vez alineadas, ponderar la similitud de cada par y generar una calificación final. En el caso más simple, esto se traduce en alinear las palabras por su representación textual o por su lema y calcular un valor de similitud basado en la cantidad de palabras superpuestas (eq. 2-12). El problema es que esta estrategia califica las frases de la tabla 2-2 como muy poco similares aunque su significado sea prácticamente el mismo. De hecho esta medida no es semántica porque no usa ninguna fuente de conocimiento externo. Aplicando el mismo principio se han desarrollado varias medidas que sí evalúan similitud semántica. Por ejemplo, Mihalcea et al. [100] propone usar una medida de similitud palabra-palabra para buscar los mejores pares de alineación, posteriormente multiplicar la similitud del par por un factor de especificidad de la palabra (idf[8]) para evitar que palabras muy generales reciban demasiada atención y, finalmente, promediar el valor de la alineación de la S_1 respecto a la S_2 y de la de S_2 respecto a S_1 (eq. 2-13). Muchas variantes han surgido al aplicar diferentes medidas de especificidad y cambiar la medida de similitud palabra-palabra por alguna variante como las reportadas en la sección anterior (ej., PMI, LSA, medidas basadas en ontologías, etc.).

[8]Inverse Document Frequency, basado en la frecuencia del término en un corpus.

$$sim(S_1, S_2) = \frac{2 * |Aligned|}{|S_1| + |S_2|} \tag{2-12}$$

$$sim(S_1, S_2) = \frac{1}{2}\left(\frac{\sum_{w \in S_1}(maxSim(w, S_2) * idf(w))}{\sum_{w \in S_1} idf(w)} + \frac{\sum_{w \in S_2}(maxSim(w, S_1) * idf(w))}{\sum_{w \in S_2} idf(w)}\right) \tag{2-13}$$

Por otro lado están las estrategias que dependen de una representación compuesta de toda la frase, generalmente embeddings, sobre la cual se realiza la medición de similitud. Es importante aclarar que una representación compuesta no necesariamente implica que se aplique el principio de composicionalidad del lenguaje, es decir es posible generar un embedding que represente a toda una oración sin tomar en cuenta el orden o relación sintáctica entre sus partes. Tal es el caso de la estrategia más básica de esta clase, la cual consiste en ponderar y promediar los embeddings correspondientes a todas las palabras de la oración para obtener un sólo vector [101]. El resultado es un embedding con la misma dimensionalidad que los de las palabras, independientemente de cuantas palabras tenga la oración. La similitud se mide entre los vectores de las oraciones de la misma manera que se mediría entre los de las palabras, es decir, calculando el coseno o el producto punto (eq. 2-14).

$$sim(S_1, S_2) = \cos\left(\frac{\sum_{i=1}^{|S_1|} w_i}{|S_1|}, \frac{\sum_{j=1}^{|S_2|} w_j}{|S_2|}\right) \tag{2-14}$$

Las estrategias que sí toman en cuenta la composición semántica se han estudiado de manera amplia. En embeddings, un trabajo representativo es el de Mitchell y Lapata [102], en donde se investigan varias funciones de composición sobre frases cortas (de dos o tres palabras). Variantes de esta estrategia proponen operaciones diferentes (ej., sumas ponderadas [102], multiplicaciones

de matrices [103], contracciones tensoriales [104], etc.) dependiendo de las categorías gramaticales que se componen. Por ejemplo, la composición de tipo adjetivo-sustantivo se representa como $p = Ab$ en donde A es la matriz de la operación adjetivo y b el vector que representa al sustantivo. De manera general se plantea que la composición de p a partir de a y b se puede representar por una función dada por $p = f(a, b, R, K)$, en donde R es la relación sintáctica conocida a-priori y K es el conocimiento previo.

Una estrategia de composición basada en sintaxis que puede aplicarse a frases más grandes y a oraciones es la de Socher [105]. El proceso inicia con el árbol de parseo de la oración en donde a cada nodo se le asigna un vector y una matriz. El vector captura el significado de la unidad y la matriz captura su efecto sobre unidades vecinas al realizar la composición. Los vectores que se asignan a cada palabra son embeddings provenientes de co-ocurrencias similares a los descritos en la sección anterior. Si los embeddings son $x \in \mathbb{R}^n$ entonces las matrices son $X \in \mathbb{R}^{n \times n}$ y se inicializan como matrices identidad más un poco de ruido Gaussiano ($X = I + \epsilon$). Entonces, la oración se representa como los pares de vectores y matrices asociados a cada palabra $S = ((a, A), ..., (m, M))$. Debido a que el la composición se realiza sobre el árbol de parseo de abajo hacia arriba, empezando por las hojas, las funciones de composición se restringen a aquellas que resulten en vectores con la misma dimensionalidad que los embeddings de las palabras. De esta manera es posible realizar un proceso recursivo de composición (fig. 2.3). Como en esta estrategia se exploran modelos que no requieren conocimiento previo, se descarta K, y como se busca que las relaciones sintácticas queden capturadas automáticamente en las matrices durante el entrenamiento, también se descarta R. De modo que la función para componer el embedding padre p está dada por eq. 2-15, en donde $W \in \mathbb{R}^{n \times 2n}$ es una matriz global para mantener el espacio n-dimensional (es decir $P \in \mathbb{R}^{n \times n}$), y g es la función *sigmoide* o *tanh*. Por otro lado, la matriz padre P se compone por medio de la eq. 2-16. El entrenamiento se realiza como una red neuronal recurrente (RNN) en la cual primero se calculan todos los nodos del árbol de abajo

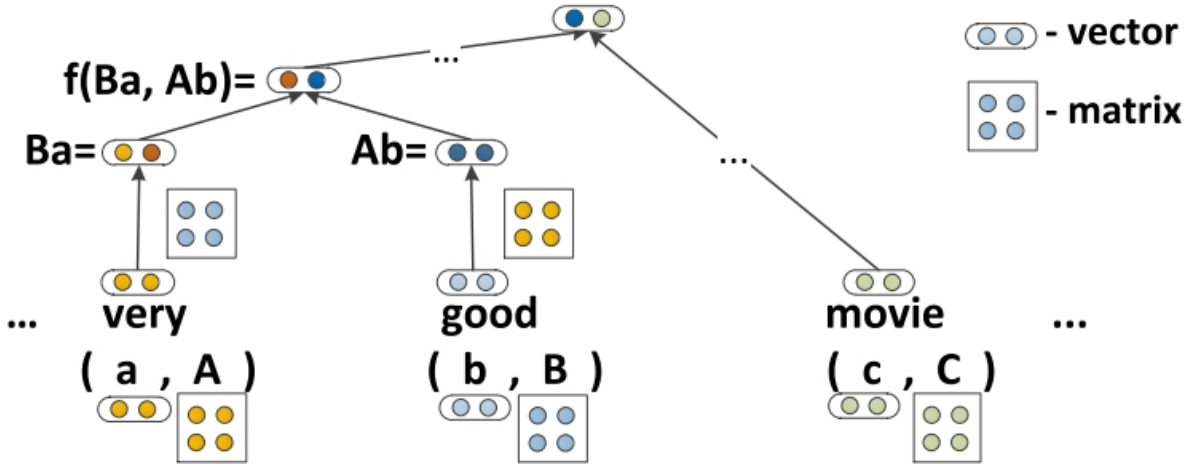

Figura 2-2: Modelo composicional basado en el árbol de parseo y embeddings de palabras. Imagen tomada de [105].

hacia arriba y posteriormente, de arriba hacia abajo se realiza una propagación de regreso de los errores de una clasificación (softmax) de cada nodo p respecto a una etiqueta (puede ser de sentimiento, de función lógica, de función semántica, etc.). Más recientemente, siguiendo la misma línea, se han propuesto modelos similares basados en otros tipos de redes neuronales; tal es el caso de [106] en donde se usa una LSTM.[9]

$$p = f_{A,B}(a, b) = f(Ba, Ab) = g\left(W \begin{bmatrix} Ba \\ Ab \end{bmatrix}\right) \tag{2-15}$$

$$P = f_M(A, B) = W_M \begin{bmatrix} A \\ B \end{bmatrix} \tag{2-16}$$

Otra perspectiva para la composición es aprender directamente del corpus una representación distribuida de la frase u oración. Uno de los métodos más representativos es *Paragraph Vector*

[9] *Long Short-Term Memory.*

[107]. Este método se basa en el mismo principio de Word2Vec-CBOW que consiste en que a partir de n palabras vecinas se predice la siguiente palabra, sólo que en este caso también se incluye un vector que corresponde al párrafo y que es concatenado a cada palabra antes de realizar la predicción. Los vectores de las palabras son globales a todos los párrafos del corpus pero los de los párrafos son independientes. Durante el entrenamiento, igual al descrito en Word2Vec, se aprende el vector del párrafo que intuitivamente captura el contexto faltante entre el sentido del párrafo y el de las palabras individuales (también se suele interpretar como el tema del párrafo).

Una forma más de abordar la composicionalidad es de manera simbólica. En esta se estipula que el significado de la estructura compleja está dado por el significado de sus partes y que existen funciones de composición semántica que están equiparadas a las reglas de combinación sintáctica del lenguaje [108]. Esta composición se realiza de manera independiente al significado de las unidades lo que facilita el proceso. El problema es que las unidades (palabras) se representan a través de símbolos desprovistos de significado por lo que la composición resultante, y su base simbólica, debe ser interpretada para poder obtener la semántica. Además, esto implica que la composición no es sensible al contexto –la base simbólica permanece invariante en todas las combinaciones de composición– [102]. Debido a que el enfoque simbólico es composicional pero cualitativo y el distribucional [109] es cuantitativo pero no-composicional, la combinación de ambos parece ser una alternativa prometedora [110]. En este sentido se han desarrollado varias estrategias que combinan los dos enfoques [103, 111, 112, 90, 113]. Por ejemplo, Clark [90] propone una teoría basada en gramática categórica y vectores de palabras para generar composicionalmente un embedding de toda la oración. En ella, los significados de las palabras son representados por vectores en un VSM, los roles gramaticales son tipos en un pregrupo, y se usa un producto tensorial para la composición de significados y tipos.

2.4. Discusión

En este capítulo se ha presentado un breve análisis de la evaluación semántica en textos. Se inició con medidas orientadas a comparar palabras y con base en ellas se revisaron estrategias diseñadas para frases y oraciones. Existe un gran número de estrategias de STS por lo que más que un análisis exhaustivo el objetivo fue exponer los diversos enfoques con que se aborda la tarea, y mencionar medidas que los ejemplifican.

En la literatura se observa que todas las estrategias tienen límites o desventajas. Las medidas basadas en grafos y en conteo de vértices son fáciles de implementar y eficientes, pero su desempeño es de los más bajos; las basadas en características tienen un mayor fundamento semántico pero dependen de que estas características estén representadas en una ontología, por lo que tienen un alcance muy limitado; las basadas en el contenido de la información derivada de corpora requieren que éste sea lo suficientemente grande y heterogéneo. Finalmente, las medidas distribucionales tienen dificultades para distinguir entre relación y similitud semántica; además tratan con palabras y no con conceptos, por lo que su evidencia semántica se ve afectada por la polisemia y sinonimia presente en el corpora. Esto afecta tanto a las medidas para palabras como a las de frases y oraciones.

En relación a las estrategias de composición es interesante observar que los enfoques con los que de se aborda el significado pueden ser alineados con la visión de Fregue. En particular se pueden diferenciar dos principios que se le atribuyen: el de contextualidad y el de composicionalidad [114, 115, 116, 117, 118, 119, 120]. En el *principio de contextualidad* [121, §60 p. 71] el significado de una palabra dentro de la oración es la contribución que aporta al significado de dicha oración;[10] es decir, la palabra no posee un significado independiente; un ejemplo que aplica este fundamento es la estrategia de Le & Mikolov [107]. Por el contrario, en el *principio*

[10] "Only in the context of a proposition do words mean something" [121, §62 p. 73]

de composicionalidad, las palabras tienen un significado y el de la oración está dado en función del significado de sus partes y del modo de composición gramatical; por ejemplo, la estrategia de Landauer [101] se alinea a este principio. La co-existencia de ambos se contrapone; sin embargo, han habido varias propuestas de unificación como la presentada por Dummet [122]. En las teorías modernas de semántica se aceptan e integran ambos principios; para esto, el de contextualidad no se toma literalmente, es decir, las palabras sí tienen un significado por sí mismas (significado semilla), el cual es "refinado" o re-definido al ser aplicadas a un contexto. En suma, la semántica de una expresión compuesta está dada en función de sus partes; es decir, las frases y oraciones constituyen el contexto local en donde se aplica la unidad léxica, bajo el cual su significado se vuelve específico, y que a su vez se combina semánticamente con las otras unidades para dar lugar al significado compuesto de la oración (o frase). Un ejemplo de estrategia que sigue esta línea es la presentada en Socher [105].

De manera general se observa que las estrategias se pueden caracterizar por: 1) el enfoque o premisa sobre la que se define qué evidencia semántica está expresada en los corpora (ej., significado dado por el contexto), 2) el tipo de representación de la unidad lingüística (ej., geométrico, probabilístico, por características, etc.), y 3) la forma (tipo de algoritmo o función) de comparar las representaciones para calcular su similitud (ej., distintivos, espaciales, estructurales y transformacionales). Por lo tanto el tratamiento semántico de una medida depende de la orientación que se tome en cada uno de estos tres aspectos. Es decir, no es sólo la medida de similitud ni solo la representación la que captura la semántica, es la estrategia en conjunto la que define cuáles rasgos del significado son evaluados. Finalmente, a pesar de la gran diversidad de medidas hay un espacio amplio para propuestas, sobre todo aquellas que privilegien la comparación de oraciones, la medición de similitud (no sólo de relación), el uso de recursos limitados y que faciliten la explicación de sus resultados.

Capítulo 3

Marco teórico

En este capítulo se abordan los conceptos y teorías que se usan como sustento teórico de la medida de STS que se propone en el capítulo 6. Se analiza el uso de lenguaje natural como medio de representación semántica y se justifica su uso en la propuesta con base en los beneficios que se han observado. Se plantea a la abstracción como una herramienta que puede facilitar la tarea de STS. Se describe la existencia de varios enfoques respecto a qué es y cómo se realiza la abstracción, y se revisan algunas teorías.

3.1. Representación semántica

Como se vio en la revisión de las medidas de SS, una característica clave de las estrategias es el enfoque con que se aborda la representación semántica. Entre la variedad de enfoques están el formal y composicional, que se basan lógica de predicados, y en la perspectiva de composición del significado. Bajo este enfoque el significado de una oración está dado por la combinación de sus partes a través de operaciones lógicas y una posterior interpretación bajo algún sistema

semántico: condiciones de verdad, teorías de modelos, semántica de mundos posibles, semántica situacional, etc [27, 29]. Sin embargo, uno de los principales argumentos en contra es que tiene dificultades para representar la variabilidad y sutileza del lenguaje humano[123]. Esto, sumado a la dificultad para definir qué es el significado de manera consensuada, motivó la búsqueda de representaciones cuya reinterpretación y verificación fuera más práctica.

Una de las teorías que surgieron como respuesta fue *Natural Semantic Metalanguage* (NSM) [124, 125]. En esta teoría, se aborda la representación de arriba hacia abajo a partir de la descomposición de una unidad compleja (ej. oración) en una más simple, y una posterior reformulación-reducción usando un metalenguaje (es decir, paráfrasis reductiva). El metalenguaje que se usa es un subconjunto estandarizado del lenguaje natural formado por unidades semánticas básicas (primitivas), obtenidas de manera empírica a través de la reducción sistemática de explicaciones al conjunto más pequeño posible y flexible de términos. Estas primitivas son básicas puesto que no se pueden definir por medio de otras unidades; además se plantea que no varían en su significado y uso en todos los lenguajes, es decir, son universales. El resultado es que la representación en NSM se da a través de la explicación semántica, que consiste en reformular con primitivas semánticas la expresión original. Es importante aclarar que las primitivas no son necesariamente simples en el sentido morfológico; a veces corresponden a unidades fraseológicas o a morfemas [126, 125], además, en las representaciones de NSM también se reúsan estructuras que expresan significados léxicos complejos (*moléculas semánticas*) [127].

Una de las características fundamentales y principales aportes de esta teoría es la *descomposición semántica*, que consiste en representar el significado de unidades complejas, como frases u oraciones, en términos de unidades de significado más simples organizadas de manera estructurada [128]. Este mismo enfoque se presenta en la teoría "Meaning-Text Theory" (MTT) [129] que se basa en establecer una correspondencia entre el conjunto de todos los significados po-

sibles (representados de manera simbólica) y el de todos los textos posibles (representados de manera fonética). Para tal propósito, la información léxica se organiza usando un diccionario combinatorio explicativo (ECD[1]) en el cual se describen relaciones semánticas y de colocación entre lexemas a través de funciones léxicas. Estas funciones permiten una descripción general y sistemática de la semántica del texto. Un postulado de esta teoría es que cada lenguaje se puede definir por medio de un modelo simbólico Meaning-Text-Model (MTM) que incluye un conjunto finito de reglas que definen la correspondencia entre los significados posibles y los textos posibles. Este postulado está estrechamente vinculado a la visión de Chomsky de la existencia de un conjunto de oraciones base (alias "kernel sentences") a partir de las cuales por medio de reglas de transformación se genera el conjunto de todas las expresiones textuales válidas [130]. Idealmente el MTM debe ser capaz de a partir de un significado generar todas las representaciones textuales posibles, y a partir de un texto generar todos los significados válidamente expresados por él.[131] La similitud entre NSM y MTT es que en ambos casos se sugiere la existencia de ciertas unidades y expresiones primitivas de las cuales se derivan las demás posibles representaciones textuales, y que tanto las funciones léxicas de MTT como las unidades del metalenguaje de NSM se presumen universales. Sin embargo, la diferencia es que en la MTT las primitivas se representan por símbolos externos al lenguaje y las reglas son diseñadas, mientras que en MSN la representación se hace con elementos de LN depurados a través de procesos empíricos en varios idiomas [128].

Estas teorías de tipo cognitivo-descomposicional se han desarrollado y aplicado por más de cuatro décadas. Entre los argumentos a favor se expone que toda representación de significado al final está ligada al lenguaje natural porque para leer e interpretar cualquier metalenguaje artificial lo que el intérprete realmente hace es trasladarlo de nuevo a su lenguaje nativo [25]. Incluso un lenguaje de descripción semántica técnica (simbólico, de propiedades, etc.) debe ser

[1] Explanatory Combinatorial Dictionary.

interpretable por los usuarios. Otra de las ventajas es que para la cognición común provee un mejor medio de representación que otros metalenguajes más técnicos porque intrínsecamente tiene la misma capacidad de representar la sutilezas de la expresión original.

Este tipo de teorías o marcos explicativos pueden aportar mucho valor a los enfoques basados en la semántica distribucional, que está presente en la mayoría de las propuestas actuales de representación. Los modelos distribucionales se basan en encontrar un gran número de características que individualmente son débiles pero en conjunto resumen el contexto de la unidad léxica y han probado tener un buen poder de representación. Sin embargo, contrario a los modelos basados en primitivas, los distribucionales no buscan características más fundamentales o abstractas que las unidades léxicas que buscan representar; tampoco tienen como objetivo diferenciar propiedades que promuevan la inferencia [132].

Más allá de aplicar la NSM o MTM de manera formal, estas teorías tienen características de especial interés para la presente investigación. La primera es la posibilidad de aplicar el enfoque descomposicional a oraciones y la posterior reducción de sus partes a moléculas más abstractas que representen su significado. La segunda es adoptar la convicción fundamental del NSM de que el lenguaje natural es adecuado para representar su semántica usando su propio léxico y sintaxis [28].

3.2. La abstracción como herramienta para simplificar representaciones

Decidir que tipo de representación semántica se usará es sólo una parte de una estrategia de STS, también se debe de establecer cómo se generarán las representaciones. En esta tesis se parte de la premisa de que toda representación derivada de la representación original con el propósito

de que sea más simple[2] es una abstracción. Saeger [35] propone que una abstracción representa un optimo local de representación dentro de un universo discursivo continuo de representaciones más detalladas. Bajo este enfoque la abstracción se entiende como un balance entre descripciones ricas y exhaustivas, pero complejas y costosas, y descripciones más específicas y menos descriptivas, pero también más simples y fáciles de usar. Una abstracción se puede ver como un vista parcial de la representación original. La relevancia de esta premisa yace en que permite plantear que si toda representación es una abstracción, entonces los procesos para generarlas son procesos de abstracción y, por lo tanto, se pueden usar como herramientas para generar representaciones semánticas.

En la revisión de estrategias-estado-del-arte (capítulo 2) se observó que existen más variantes de procesos para generar representaciones que de tipos de representaciones. Considerar a estos procesos como de abstracción permite que a pesar de la gran diversidad de propuestas éstas se puedan analizar como variantes de un mismo mecanismo. Esto abre la posibilidad de aprovechar las teorías de abstracción existentes para explicar y categorizar los métodos actuales para generar representaciones, pero también para proponer nuevos.

Es importante resaltar que para el propósito de esta tesis, el interés recae en el proceso de abstracción como una herramienta para generar representaciones que faciliten la tarea de STS.

¿Ya se usan las abstracciones?

El uso de abstracciones es común en disciplinas relacionadas con la cognición y manejo de información. Por ejemplo, en la ciencias computacionales es frecuente el uso de estructuras organizadas, ontologías y taxonomías, generadas a través de procesos de abstracción. Una ontología es una especificación explícita de la conceptualización [133]; es decir, es una teoría lógica com-

[2] En el contexto de la tarea de STS.

prometida con una conceptualización particular de la realidad que está fundamentada en una serie de reglas que definen una representación parcial. Las ontologías son una abstracción en la que se definen las entidades de manera intensional (por medio de etiquetas o asociaciones a otras clases de la ontología) o de manera extensional (asociando la entidad a sus ejemplares) [29]. Uno de los principales retos al generar una ontología es identificar las restricciones más específicas que definan las características de las abstracciones.

El lenguaje natural por sí mismo está constituido por abstracciones, esto es, la asignación de unidades simbólicas (palabras o descripciones) a conceptos. Gardenfors [134] plantea dos tipos de representaciones como parte evolutiva del lenguaje. Las que están ligadas a una entidad externa presente en el ambiente del organismo que realiza la referencia; a estas representaciones las llama *señales*; y las que apuntan a entidades no presentes en el contexto actual o reciente del individuo, a las cuales nombra como *signos*. En su enfoque los signos representan abstracciones; y de la misma manera que la teoría de *niveles de interpretación* [49], también hace uso de la distancia entre el individuo y el referente como indicador del nivel de abstracción.

Chomsky [130] usó un enfoque diferente de abstracción al proponer la existencia de un conjunto de oraciones que constituyen el núcleo del lenguaje (*kernel sentences*) a partir de las cuales se derivan las demás oraciones posibles. El *kernel* de un lenguaje $\mathfrak{L}$ con una gramática G se plantea como un conjunto de oraciones producidas a partir de la aplicación de lo que llama transformaciones obligatorias a las cadenas terminales de la gramática; y todas las demás oraciones son derivadas a partir del *kernel*. Este núcleo corresponde al conjunto de abstracciones que comprenden a todas las posibles derivaciones válidas (oraciones) de ese lenguaje. Chomsky cambió este planteamiento [135] por la noción de estructuras superficiales y estructuras profundas (*deep structures*). En su nuevo planteamiento abandona la hipótesis de que existen abstracciones generales a todo el lenguaje (*kernels*), pero conserva la idea de abstracciones locales (*deep structure*);

las cuales plantea como representaciones que facilitan el análisis e interpretación semántica de la oración. En su propuesta las *deep structures* son derivadas de la expresión original aplicando reglas sintácticas de transformación de frases.

¿Existen métodos para generar abstracciones?

Entre los primeros acercamientos metódicos a la abstracción se encuentra el desarrollo del análisis ontológico que se fundamenta en la creación de niveles de especialización o generalización [136]. Se basa en encontrar las propiedades diferenciadoras a partir de las cuales se construyen categorías generales que derivan en varias subcategorías más especializadas, por ejemplo, taxonomías. En estos primeros acercamientos se identifican dos procesos de abstracción: 1) ignorar las características físicas o sensoriales de la entidad y de esta manera tratar de llegar a la esencia de las mismas, y 2) a partir de múltiples instancias omitir las características individualizadoras y de esta manera llegar a las características universales que definan al tipo de instancias [136].

Hegel retoma la idea de reducir la información que define a los ejemplares de una entidad hasta que dicha reducción permita a múltiples instancias tener una definición común. Define al proceso de abstracción como uno iterativo en el que las partes se van aislando del todo [137].[3] Lo mismo se observa en los postulados de Locke [138], en los que la abstracción está formada por conjuntos de ideas concretas a las cuales se les omiten detalles distintivos. La conjunción de objetos determinados individualmente pero unidos colectivamente dan lugar a lo que Husserl [139] llamó el concepto de pluralidad, en el cual resalta que la naturaleza individual no tiene relevancia al momento de generar los conceptos. No es a la suma de contenidos individuales a los que Husserl reconoce como abstracción, sino a "aquello extra a los contenidos individuales" y que es lo que conjunta la unión de los individuos en un todo, en el concepto. Por lo tanto,

[3] En la visión de Hegel dicho "todo" se integra tanto por el mundo sensorial como por el conceptual.

es posible que dos *todos* sean iguales aun cuando sus partes constitutivas sean diferentes. Es decir, la formación de conceptos y abstracciones está fundamentada en las uniones por similitud. Por ejemplo, al poner atención a las conexiones de puntos en una linea o a los instantes en un periodo de tiempo se observa el concepto de unión continua, y se llega al concepto de continuo (*continuum*); aun cuando este concepto no está incluido en la representación de cada elemento particular [139].

Una acercamiento más formal del proceso de abstracción es la propuesta de Wright y Hale [36, 37]. [4] En ella se presupone la existencia y entendimiento de una relación de equivalencia R sobre el dominio de una función f que a su vez representa la función de abstracción.

$$f(a) = f(b) \iff R(a, b) \tag{3-1}$$

A esta ecuación le llamaron *principio de abstracción*. Esta ecuación no define la expresión funcional del lado izquierdo (f), sino que le impone una condición de existencia que al respetarse determina a la función; es decir, la relación R se requiere antes de determinar el lado izquierdo de la ecuación. Si dicho principio de abstracción existe, entonces existe un concepto K_f que agrupa los valores del rango de f:

$$x = instancia(K_f) \iff \exists y | x = f(y) \tag{3-2}$$

Por lo tanto, si x es una instancia de un concepto K_f que está expresado por f, y f sigue el principio de abstracción, entonces x es considerada una entidad abstracta. Esto quiere decir que bajo esta propuesta una entidad abstracta es una representación parcial generada al aplicar

[4] Inspirados en la observación de Fregue de que los términos que simbolizan entidades abstractas se forman frecuentemente por medio de expresiones funcionales (*functors*)[140].

una expresión funcional.

Un ejemplo de esta relación está dado por la eq. 3-3 en donde se establece que para entender el concepto *dirección* ($K_{dirección}$) se requiere saber que la *dirección de a* y la *dirección de b* se refieren a la misma entidad si y sólo si a y b son paralelas; es decir, se necesita saber el concepto de *paralelos* antes de entender el de *dirección* [141].

$$dirección(a) = dirección(b) \text{ si sólo si } a \text{ y } b \text{ son paralelas} \tag{3-3}$$

Barsalou et al. [142] proponen la abstracción bajo un enfoque situacional; refieren que el sistema conceptual obtiene información de conceptos abstractos cuando una situación está dada, es decir se requiere un enfoque para orientar la *abstracción*. Se plantea que cuando un evento activa diferentes entidades concretas el proceso de abstracción consiste en combinarlas con el objetivo de procesar y dar respuesta al evento [143]. Se combinan las entidades que se consideran similares de acuerdo a la situación (evento o contexto) que se atiende.

Floridi [144] también se basa en una abstracción dirigida (situacional) en su propuesta de *niveles de abstracción*. Define a un nivel de abstracción (LoA[5]) como un conjunto finito y no vacío de *observables*, los cuales a veces pueden ser moderados por *reglas* que ayudan a definir las relaciones entre los observables y sus valores válidos. A los *observables* los identifica como variables tipificadas interpretadas que reflejan un enfoque particular de la entidad; es un enfoque epistemológico en el que se tienen diferentes niveles de observación e interpretación de un sistema. Se considera exitosa la abstracción si para la situación dada se puede representar la entidad al conjuntar las simplificaciones generadas. Este mecanismo de generalización es diferente al que consiste en reemplazar características reales pero complejas del sistema, con versiones

[5] *Level of Abstraction.*

simplificadas e idealizadas (modelos). La diferencia es que el mecanismo de *idealización* describe la realidad de un sistema de una manera diferente a la que és, aproximándolo a un modelo; por su parte la generalización por abstracción, no describe una realidad aproximada, sólo la limita [145], [146].

Existe un número considerable de teorías del proceso de abstracción de las cuales sólo se mencionaron algunas, pero se revisaron muchas más. A continuación una lista de las teorías revisadas:

- Teorías basadas en Fuzzy Sets [45, 46]
- Teorías basadas en Prototipos [32, 45, 31, 147, 148]
- Teoría de Wright y Hale [36, 37]
- Teoría de Hobbs [38]
- Teoría de Lowry [40, 149]
- Teoría de Tenenberg [39]
- Teoría de Subramanian [41]
- Teoría de Giunchiglia y Walsh [42, 150]
- Teoría de Levy y Nayak [43]
- Teoría de Fine [44]
- Teoría de Ghidini y Giunchiglia [47]
- Teoría de Saeger - Teoría de canal [35]
- Teoría de Floridi - Teoría de niveles de abstracción [48]

- Teoría de Soderberg - Teoría de Nivel de Interpretación [49]

Las teorías de abstracción pueden ser clasificadas de acuerdo a dos características: el cambio de representación y al mecanismo de abstracción.

Se identifican cuatro tipos de cambio de representación [141]:

1. Mapeo entre señales sensoriales y objetos → *perceptivo/ontológico*
2. Mapeo entre predicados → *sintáctico*
3. Mapeo entre las interpretaciones semánticas de un lenguaje lógico → *semántico*
4. Mapeo entre teorías lógicas → *axiomático*

Y tres tipos de mecanismos de abstracción [151]:

1. Abstracción[6] - ignorar algunos detalles no relevantes a la tarea.
2. Reformulación - hacer un cambio en la ontología de conceptualización.
3. Aproximación - representaciones que se aproximan a la semántica o sintaxis original por medio de observaciones parciales.

En la figura 3-1 se muestran las teorías revisadas de acuerdo a estas dos clasificaciones. Una breve descripción de cada una de estas teorías se puede encontrar en el apéndice A.

Otra opción para clasificar a los mecanismos de abstracción es de acuerdo a su vaguedad. Daniliuc [152] usa el concepto de *vaguedad* como medio para el cambio de representación. En su acepción la vaguedad no está relacionada a la incapacidad de determinar de manera precisa

[6] El nombrarlo como *abstracción* no indica que las demás clases no sea consideradas formas abstracción

la extensión o valor de un concepto, sino a una estrategia de representación que aporta ganancia funcional dentro de un contexto o problema dado. Se identifican tres tipo de vaguedad, parcialidad, perspectiva o aproximación [153]. A pesar de que la teoría cognitiva de *enfoque de atención* [134] se asocia más con procesos de generalización, en realidad es compatible con los tipos de vaguedad mencionados. Es decir, la *parcialidad* es un enfoque de atención a características relevantes; la *perspectiva* es un enfoque de atención a observaciones (o sus combinaciones) relevantes; y la *aproximación* es un enfoque de atención a características relevantes dentro de un entorno de escalas (ej. tiempo, tamaño, umbral de error), en donde lo *relevante* está determinado de acuerdo al contexto.

No existe un consenso de qué es una entidad abstracta y tampoco de un único proceso para generarlas, pero independientemente del cambio de representación y del mecanismo, existe un punto común; *los procesos de abstracción buscan simplificar la representación y facilitar el manejo de una entidad para una tarea objetivo.*

3.3. Discusión

En este capítulo se propuso que las representaciones que se usan en las tareas de STS son abstracciones y que por lo tanto las estrategias para generarlas se pueden considerar procesos de abstracción. Al revisar la diversidad de teorías de abstracción se encontró que no hay un consenso en la definición de qué es una abstracción ni tampoco en el proceso de abstraer. Lo que sí comparten las teorías es que el propósito de abstraer es generar representaciones más simples.

El interés de esta tesis se enfoca en la abstracción como proceso y en particular como herramienta; por lo tanto, la definición que se adopta es la de Saeger [35]: “un mecanismo para el manejo de la complejidad de una representación permitiendo que se deriven diferentes represen-

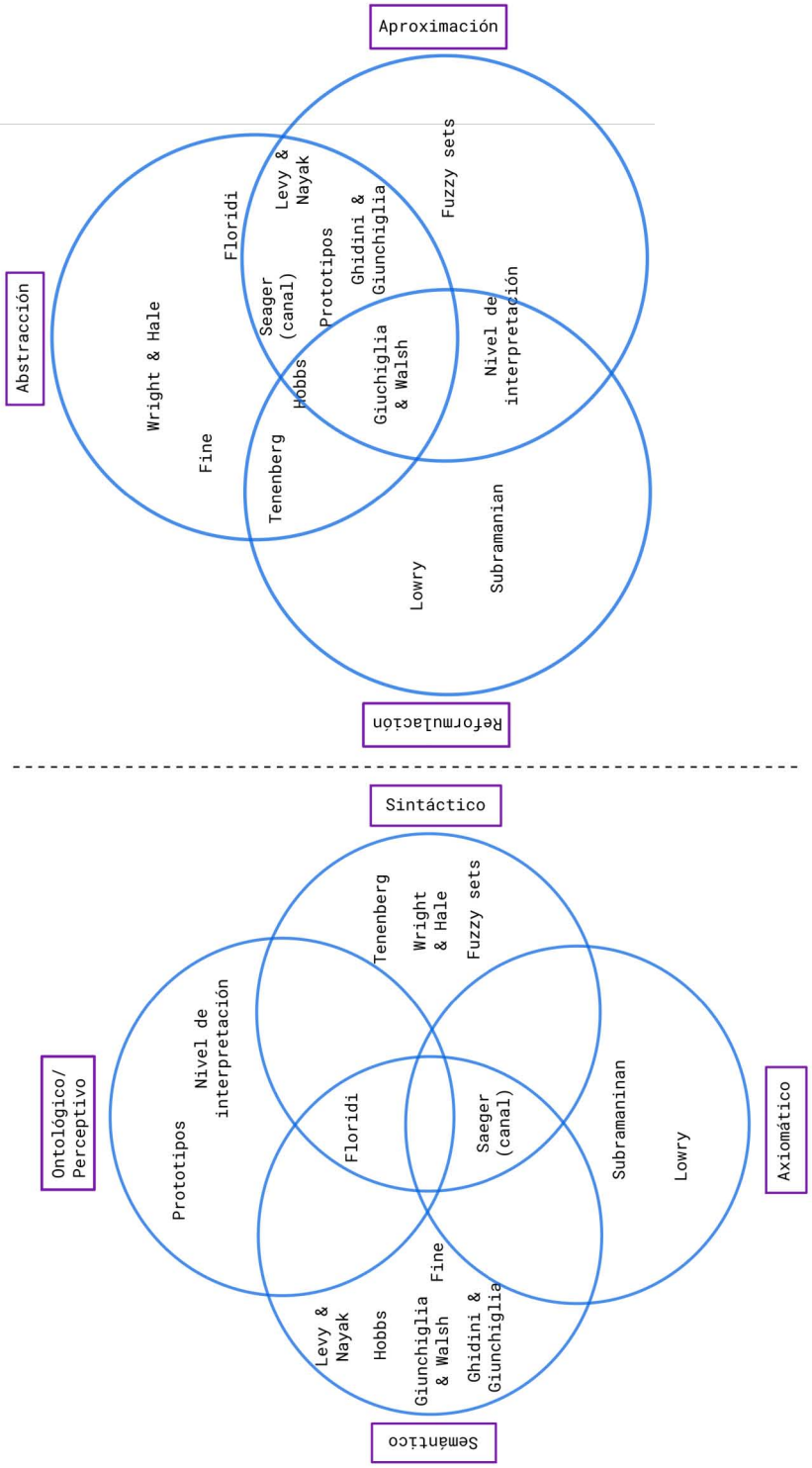

Figura 3-1: Clasificación de teorías

taciones abstractas dependiendo del uso que se les quiera dar". Estas derivaciones corresponden a un proceso de filtrado de aquellos aspectos de la representación original que no sean relevantes para el problema o contexto dado (incluyendo un instante específico en el tiempo).

Una característica distintiva entre los tipos de abstracción es si hay pérdida de información o no. Al analizar las teorías desde esta perspectiva, se observó que en la mayoría de los procesos hay pérdida de información y que por lo tanto no son reversibles; es decir, no es posible recuperar la representación original partiendo sólo de la representación abstracta. Las excepciones son las teorías de Lowry [40] y Subramanian [41]. Estas se basan en la reformulación y en consecuencia producen isomorfismos que presentan un cambio en el formato, pero no el contenido de la información. En la reformulación no hay una pérdida de información sólo un ocultamiento a través de la encapsulación tal como sucede en el cálculo lambda. La teoría de Subramanian incluso considera explícitamente una conceptualización previa Δ sobre la cual está construida la abstracción.

La perdida de información es central en el método que se propone en el capítulo 6 una vez que se asocia directamente con la disimilitud entre las expresiones originales a partir de las cuales se genera una abstracción. Otras estrategias de abstracción ya han usado antes la pérdida de información como elemento informativo durante el proceso. Un ejemplo en donde su uso es claro es la propuesta de memorias asociativas de Pineda [154], que se basa en un principio implícito similar al principio de irrelevancia de Subramanian. Al igual que en [41] en donde se busca minimizar los estados de irrelevancia en la reformulación, en las memorias asociativas se busca minimizar el aumento de la entropía causada después cada iteración de aprendizaje, tratando de que las características recordadas sean lo más representativas (relevantes) al común de las instancias del concepto que representa la memoria. Esta relación entre detalles diferenciadores de descripciones y la entropía de la conceptualización, también se ve reflejada en la observación

de Soderberg [49], al hacer más concretas las representaciones se pierde (se vuelve más opaca) la estructura que permite identificar a un ejemplar como parte de un concepto. Estos ejemplos y las teorías revisadas respaldan la validez de los procesos de abstracción con pérdida de información.

El otro objetivo de este capítulo fue motivar el uso del lenguaje natural como medio de representación semántica. Se revisaron teorías que respaldan este enfoque y que proponen reformular las expresiones originales bajo el mismo dominio simbólico (LN), y así adecuarlas para el procesamiento semántico. Esto en conjunto con la abstracción como herramienta constituyen la base teórica de la propuesta de STS que se presenta en el capítulo 6. Específicamente se plantean dos tipos de abstracción, una léxica basada en ontologías y una sintáctica que consiste en conservación de patrones de categorías gramaticales. A partir de las expresiones en LN estos procesos generan otras más abstractas pero también en lenguaje natural; la información perdida al generarlas se asocia inversamente a la similitud semántica.

Finalmente, se expone el siguiente ejemplo para reforzar el planteamiento de que los procesos de abstracción pueden usarse en el ámbito de NLP, y que de hecho ya se aplican de manera implícita. El principio subyacente de GloVe es que la proporción de co-ocurrencia entre dos palabras (i y j) está fuertemente relacionada con el significado, lo cual se plantea por medio de $F(w_i, w_j, w_k) \approx \frac{P_{ik}}{P_{jk}}$. Para el aprendizaje se establece que los embeddings que representen significados similares deben estar cercanos en el espacio geométrico; esto se implementa poniendo la restricción de que si son similares, la resta de los embeddings se debe acercar a la proporción de sus co-ocurrencias con una tercera palabra k (palabra prueba) y que esta proporción sea cercana a 1. Esto se plantea con la ecuación $F(w_i - w_j, w_k) = \frac{P_{ik}}{P_{jk}}$, en donde la relación fundamental es:

$$\boxed{w_i, w_j \text{ son semánticamente equivalentes si sólo si } w_i - w_j = \frac{P_{ik}}{P_{jk}} = 1}$$

la cual resulta ser precisamente la aplicación del principio de abstracción de Wright y Hale. Se presupone la existencia y entendimiento de una relación de equivalencia R (en el espacio geométrico de los embeddings) bajo la cual el significado las palabras i y j es el mismo si y sólo si las abstracciones ($embedding(i)$ y $embedding(j)$) son iguales.

$$significado(i) = significado(j) \iff w_i - w_j = \frac{P_{ik}}{P_{jk}} = 1$$

es decir

$$embedding(i) = embedding(j) \iff w_i - w_j = \frac{P_{ik}}{P_{jk}} = 1$$

Capítulo 4

Corpus de Similitud Semántica en RTM

La similitud semántica depende del dominio y de la tarea objetivo, es decir, de características particulares de los textos, de las relaciones en el dominio y de la perspectiva bajo la que se realiza la comparación. Este tipo de matices quedan implícitamente capturados dentro de un corpus; más aún si se construye por expertos. Por estas razones un corpus de similitud es un instrumento de suma utilidad para la tarea de STS.

Existen varios corpus de similitud semántica tanto para dominios específicos como generales. Sin embargo, cuando no existe corpora específico al dominio o el existente no se ajusta a la tarea objetivo es necesario construir uno a la medida. Tal es el caso del dominio específico de literatura biomédica referente a la regulación del inicio de la transcripción (RTM[1]) en bacterias.

En la tabla 4 se muestran un par de oraciones que provienen de publicaciones diferentes,

[1] Regulación Transcripcional Microbiana

que expresan un significado muy parecido, y que sirven de evidencia de soporte una de la otra. Estas oraciones ejemplifican lo que se desea anotar en este corpus, el cual servirá para entrenar y evaluar las estrategias de STS y, por lo tanto, será la cota superior de nuestras estrategias de STS.

Oración	Título de publicación
There is, however, some evidence that increased rob expression occurs in glucose—and phosphate—limited media in the stationary phase of cell growth, attributable to activation by factor rpoS.	MarA-mediated transcriptional repression of the rob promoter. (PMID: 16478729)
A similar rpoS dependency was observed for glucose-limited or phosphate-limited growth in which rob::lacZ transcription increased 5-fold.	Posttranscriptional activation of the transcriptional activator Rob by dipyridyl in Escherichia coli. (PMID: 11844771)

Tabla 4-1: Ejemplos de oraciones de diferentes publicaciones que expresan el mismo significado

4.1. Antecedentes

El resultado de la medición de STS va desde una coincidencia exacta, hasta significados totalmente ajenos, pasando por varios grados de semejanza. Cada nivel de semejanza en este tipo de escalas representa coincidencias de diferentes elementos de las expresiones [7]; es decir, pueden ser equivalentes, diferir en detalles menores, compartir detalles importantes o el tema general, compartir sólo el dominio y contexto ó, ser totalmente ajenas.

Entre los esfuerzos recientes más representativos tanto en la tarea de STS como en generación de corpora para la misma encontramos el *Semantic Evaluation Workshop* (SEMEVAL) [7]. El corpus de *SEMEVAL* consiste en aproximadamente 15,000 pares de oraciones anotadas a través de colaboración pública[2] con una escala entre 0 y 5, en donde 0 indica que no existe relación

[2] crowdsourcing

de significado entre las expresiones, y 5 que significan lo mismo. Sus oraciones se extrajeron de diferentes fuentes, entre las que encontramos a *Microsoft Research Paraphrase* (MSRP) y PASCAL VOC [155].

Otro es el corpus *User Language Paraphrase* (ULPC) [156], que se construyó a partir de pedirle a estudiantes que parafrasearan las oraciones originales. Consiste en 1,998 pares de oraciones anotadas en una escala de 1-6 para cada una de las dimensiones consideradas: similitud léxica, similitud semántica y consecuencia lógica.

Uno de los corpora públicos más representativos es el *MSRP* corpus [157]. Aún cuando en este no se asigna un grado de similitud sino una calificación binaria respecto a si ambas oraciones son paráfrasis una de la otra. Su importancia reside en que fue uno de los primeros y de mayor tamaño; consiste de 5,801 pares de oraciones calificadas por dos evaluadores humanos en donde el 67 % se evaluaron como "Semánticamente equivalente".

El corpus *SIMILAR* [158] se generó al evaluar cualitativamente 700 pares de oraciones extraídas del corpus *MSRP*. En este se anotaron las relaciones semánticas entre las palabras de cada par de oraciones y se asignó una calificación cualitativa de la similitud entre las oraciones: idéntica, relacionada, contextual, de dominio ó ninguna.

Finalmente, el *The Question Paraphrase* corpus [159], relacionado a los sistemas de pregunta-respuesta, consiste en 7,434 oraciones formadas a partir de 1,000 preguntas diferentes y sus paráfrasis extraídas de *WikiAnswers*.

Al haberse compilado de fuentes como noticias, textos legales, anuncios, etc., todos los corpora mencionados abarcan dominios generales por lo que se dificulta su uso en campos especializados como el de RTM. Un corpus más especializado y cuyo dominio es cercano al nuestro es BIOSSES [160]. Está constituido por 100 pares de oraciones del dominio biomédico cuya

similitud fue calificada usando la escala de SEMEVAL. Para extraer los pares de oraciones se formó un conjunto candidato de artículos que citaban alguno de los artículos de referencia (20 artículos). A partir del conjunto candidato de artículos se formaron los pares de oraciones que citaban al mismo artículo de referencia. Debido a la extensión del área de biomédicas y a lo reducido de BIOSSES, queda en evidencia que este corpus es muy general para representar el dominio de RTM; lo cual se aprecia en la ausencia de léxico esencial.

Por estas razones, decidimos crear un corpus de pares de oraciones que de manera natural se presentan en literatura de RTM. Esta literatura se obtuvo de la compilación especializada que se lleva a cabo en el proyecto RegulonDB. La similitud semántica de cada par de oraciones fue evaluada por expertos en el campo de la RTM quienes asignaron una calificación de 0-5.

4.2. Metodología

Diseño del corpus

Un corpus textual es "una colección de piezas de lenguaje escrito en un formato electrónico, seleccionado de acuerdo a criterios externos para que represente, de la mejor manera posible, una variedad del lenguaje que será usada como fuente de datos para investigaciones lingüísticas" [161].

El diseño del corpus implica definir la fuente textual, el tamaño del corpus, las reglas de evaluación, entre otras características. Que el corpus cumpla con los objetivos para los que se construyó depende en buena medida de un diseño fundamentado e informado. Por tal motivo, para diseñar este corpus se consideraron características propuestas por Sinclair [161] que se describen a continuación:

La política de muestreo define la fuente de donde se toman los textos candidatos y cómo se seleccionan. Esta política se forma por tres criterios principales: *orientación*, *criterio de selección* y *muestreo*. En relación a estos criterios el diseño del corpus se especificó como sigue: 1) en relación al criterio de *orientación*, se definió como de contraste, es decir, tiene el objetivo es mostrar las variedades de lenguaje que expresan el mismo significado; 2) con respecto al criterio de *selección*, se definió que los candidatos estén restringidos a oraciones de lenguaje escrito (granularidad y origen), en idioma inglés, tomados de artículos científicos (tipo), específicamente en el tema de regulación genética (dominio), y en donde el tipo y actitud de las oraciones es irrelevante; y 3) en relación al criterio de *muestreo*, se determinó que para la extracción inicial de oraciones candidatas se usaría una métrica básica de STS cuyas calificaciones se usarían para seleccionar un conjunto balanceado, esto es, que tenga el mismo número de pares de oraciones para cada nivel de la escala de similitud.

Representatividad y balance se refiere al tipo de características consideradas y cómo se encuentran distribuidas en los ejemplares del corpus. En nuestro caso, se privilegia que las oraciones incluyan elementos biológicos o conocimiento, y que todos los grados de similitud estén representados en más o menos igual proporción. Es importante resaltar que el objetivo principal de análisis es la similitud semántica de las oraciones, más no el tema, ni el nivel técnico, ni la especificidad o completitud de lo expresado.

El Tema del corpus impacta directamente en su variedad y tamaño. En nuestro caso el tema es la regulación genética transcripcional en bacterias. Si bien es cierto que enfocarse en un tema tan específico limita las posibilidades de uso de un corpus, en STS abarcar más temas incrementa significativamente la ambigüedad léxica y disminuye las posibilidades de representación de léxico especializado. Esto no quiere decir que sólo aquellas oraciones que explícitamente expresen temas de RTM sean incorporadas, sino que las oraciones se

tomarán de artículos científicos en este dominio específico.

La Homgenidad se abordó descartando oraciones consideradas muy cortas (menos de 10 palabras) [162] y aquellas oraciones que no formaran parte del cuerpo principal del artículo.[3]

El Tamaño del corpus debería definirse a partir del tipo de pregunta que se desea responder y las tareas en las que se usará. Sin embargo, la demanda de recursos y el costo que implica su construcción son los elementos que, en la práctica, definen el tamaño. En el caso de este corpus se debería considerar tanto el tamaño necesario para la evaluación de las estrategias, como el número de ejemplares necesarios para entrenar cada tipo de modelo que pudiera ser utilizado. Por ejemplo, si se deseara que este corpus pudiera ser fuente de entrenamiento para modelos con alta demanda de datos, como lo son las redes neuronales, una práctica común es tener al menos P^2 casos de entrenamiento en donde P representa el número de parámetros del modelo lo que se traduciría en miles de pares de oraciones, lo cual está fuera de nuestro alcance. Por esta razón, se definió el tamaño de acuerdo a lo necesario para evaluar las estrategias de STS por medio de una correlación de Pearson. Cohen [163] determinó que 85 muestras son suficientes considerando un efecto medio de $r = 0.30$, un nivel de significancia de 0.05 y un poder de 80 %. Sin embargo Moinester [164] y Chuan [165] sugieren un mínimo de 120 muestras para hacer posible un análisis de correlación y uno de regresión En consecuencia se optó por un tamaño de corpus un poco superior a los umbrales sugeridos, de 170 pares de oraciones; .

Al comparar las decisiones de diseño contra las de otros corpora notamos que en algunos casos las pautas de diseño limitan el tipo de similitud que el corpus puede representar. Por ejemplo, en el caso de MSRP se aplicaron varias restricciones para delimitar las paráfrasis candidatas. Una de las restricciones consistió en considerar sólo pares de oraciones en los que existieran por

[3] Partiendo de una clasificación estilográfica de las oraciones al ser extraídas directamente de los PDFs por medio de una herramienta que se implementó para este estudio.

lo menos 3 palabras en común y que la distancia de edición Levenshtein estuviera dentro de un umbral; sin embargo, estos criterios sesgan el corpus hacia representar una similitud textual. Otra restricción consistió en descartar pares en donde la longitud de sus oraciones difiriera en más de un tercio, limitando la posibilidad de representar similitud semántica cruzada.[4]

Compilación del corpus

De acuerdo a nuestro criterio de muestreo los candidatos se seleccionan aplicando una *estrategia básica de STS*. El proceso asigna una calificación continua entre 0 (significados no relacionados) y 1 (significados equivalentes) a los pares de oraciones candidatas. Esto se obtiene promediando los *embeddings* de las palabras contenidas en cada oración para generar uno que representa a la oración completa. Después se calcula el coseno entre las representaciones vectoriales de ambas oraciones y el resultado (entre 0 y 1) se considera como la similitud semántica. Estrategias de este tipo son consideradas como una buena línea base en tareas de STS. Posteriormente, se seleccionan el mismo numero de ejemplares para cada nivel de similitud de los pares candidatos de manera que el conjunto resultante sea lo más balanceado posible en relación a la representación de cada nivel de similitud. Es importante resaltar que los embeddings utilizados fueron entrenados en literatura de RTM [167].

El proceso de muestreo que se acaba de describir fue aplicado a los subconjuntos de datos *general* y *anaerobiosis FNR*. El subconjunto *anaerobiosis FNR*[5] se compiló manualmente por un curador experto quién, a partir de artículos relacionados a la anaerobiosis, seleccionó oraciones que consideró relevantes dentro del tema. El subconjunto *general* está constituido por oraciones tomadas de manera aleatoria a partir de la literatura de RegulonDB. Los pasos para compilar este último subconjunto fueron: 1) se extrajeron las oraciones de las 5,963 publicaciones (en

[4] Cross-level semantic similartity [166].
[5] Fumarate and Nitrate Reductase regulatory protein

formato PDF) que constituyen la literatura de RegulonDB; 2) de cada artículo se descartaron el primer 30 % y el último 30 % de las oraciones como estrategia muy básica para enfocarse en el contenido principal del artículo (ej. métodos, discusión y resultados); 3) se seleccionaron dos oraciones de cada publicación.

El 40 % de las oraciones del corpus resultante proviene del subconjunto anaerobiosis-FNR y el 60 % del subconjunto general. La figura 4.2 muestra el proceso descrito para la compilación del corpus.

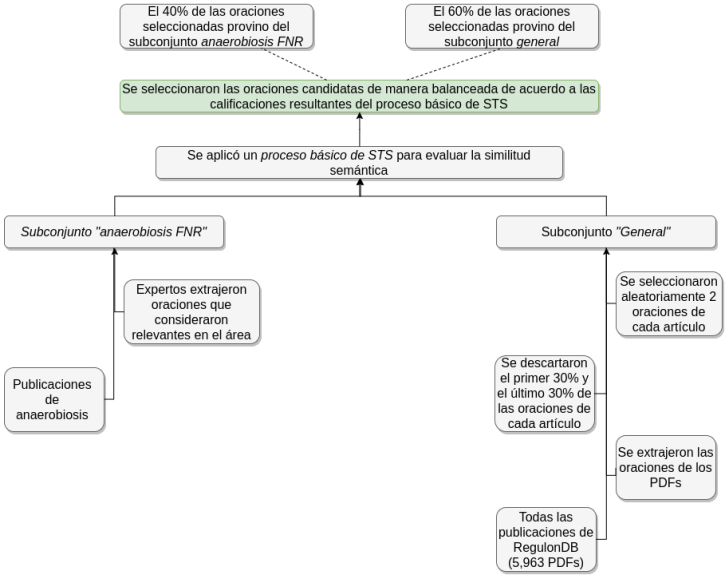

Figura 4-1: De abajo hacia arriba, se muestran los pasos ejecutados para compilar el corpus. Primero se recolectaron dos subconjuntos de oraciones: anaerobiosis-FNR y general. Posteriormente un proceso básico de STS calificó la similitud de los pares y basado en esta calificación se seleccionó un número balanceado de pares para cada nivel de similitud. Esto dio como resultado un corpus de 171 pares de oraciones de las cuales 40 % provino del subconjunto anaerobiosis-FNR y 60 % del subconjunto general. Imagen tomada de [15].

Diseño de la anotación

Aunado a diseñar el proceso de compilación se tuvo que difinir como se evaluaría la similitud semántica. Se adoptó una escala similar a la del SEMEVAL, que va de 0 a 4. Una *calificación*

de similitud del par (SPSS[6]) de 0 indica que no hay relación semántica entre las dos oraciones, mientras que una calificación de 4, que ambas oraciones expresan el mismo significado. En medio encontramos otros tres grados de similitud, como se muestra en la tabla 4.2.

SPSS	Descripción
4	Las dos oraciones son prácticamente equivalente, expresan el mismo significado.
3	Las dos oraciones son más o menos equivalentes pero algunos elementos de información importantes difieren o faltan.
2	Las oraciones no son equivalentes pero comparten algunos detalles.
1	Las oraciones no son equivalentes pero tratan el mismo tema.
0	Las oraciones no comparten similitud.

Tabla 4-2: Escala de calificaciones

La evaluación de similitud se llevó a cabo por 7 anotadores expertos en el área de RTM. Se aplicó un estudio de tipo no-totalmente-cruzado; esto es, no todos los anotadores evaluaron todos los pares, sino que cada ejemplar fue evaluado por 3 anotadores seleccionados aleatoriamente del grupo de 7 expertos.

Deleger [168] plantea que 2 evaluaciones por ejemplar es lo mínimo necesario. Con el objetivo de tener un número impar de evaluaciones para facilitar la obtención de la calificación mediana, se decidió que 3 anotadores evaluarían cada par de oraciones.

Debido a que lo que se considera "semanticamente equivalente" es subjetivo y propenso a estar sesgado por consideraciones personales, fue necesario homogenizar el proceso de anotación entre los evaluadores. Esto se logró a través de una serie de sesiones de entrenamiento durante las cuales los evaluadores se familiarizaron con el corpus, con las reglas de anotación y, además, las refinaron. En cada sesión se evaluó un subconjunto pequeño, se discutieron y resolvieron las diferencias, y se ajustaron aquellas reglas que así lo requirieron. El periodo de entrenamiento fue dado por terminado cuando se consideró que los evaluadores habían asumido las reglas y

[6]Por sus siglas en inglés - *Sentence Pair Similarity Score*

habían alcanzado un consenso sobre las mismas. En la práctica, el entrenamiento se terminó cuando el nivel de interacuerdo entre las evaluaciones de todos los anotadores sobrepasó un umbral mínimo preestablecido.

Reglas generales

Con el fin de facilitar el proceso y hacerlo menos subjetivo, a los evaluadores se les dieron algunas directrices iniciales para la anotación de similitud. Con base en los experimentos de construcción de corpus de Torres-Moreno [169], y en las observaciones de los expertos que harían la anotación, se generaron las siguientes guías generales:

- *Orden.* La diferencia de orden de las cláusulas en una oración compuesta no necesariamente afecta el significado; pueden estar dispuestas en orden diferente y seguir siendo equivalentes.
- *Cláusulas faltantes.* En el contexto de una oración compuesta o compleja, si una de las cláusulas está presente en una oración pero no en la otra esto no necesariamente indica una similitud de cero; depende de la importancia de la información faltante y la compartida.
- *Adjetivos.* La falta de algun(os) adjetivo(s) en una de las oraciones, en principio, no afecta la similitud.
- *Enumeraciones.* Si hay elementos de enumeraciones que faltan en una de las oraciones, el descenso de calificación de similitud debe ser mínimo, a menos que la enumeración exprese el significado principal de la oración. El reorden en enumeraciones no afecta la similitud.
- *Abreviaciones.* Las abreviaciones son consideradas equivalentes a la unidad no abreviada, ej. “vs” y “versus.”

- *Hipónimos e Hiperónimos* . Léxico que comparte una relación de este tipo comparte cierto grado de similitud, ej., "sugar substance" comparado con "honey" comparado con "bee honey."

- *Palabras compuestas.* Algunos términos son semánticamente equivalentes a expresiones multi-palabra, ej., "anaerobiosis" y"in the absence of oxygen," "oxidative and nitrosative stress transcriptional regulator" y "oxyR,", y "hemorrhage" y "blood loss."

- *Generalizaciones o abstracciones.* Dos textos comparten cierto grado de similitud si uno es una generalización o abstracción del otro, ej., 8 y "one-digit number."

Refinamiento consensuado

Las directrices se refinaron y enriquecieron durante las sesiones de consenso.

Con el fin de hacer más específica la escala de calificaciones en el dominio de RTM, se decidió usar las clases de objetos modeladas en RegulonDB como marcadores del tema de las oraciones. RegulonDB contiene objetos de las siguientes clases: Gen, Producto genético, Proteína, Motivo, Promotor, Unidad de transcripción (TU[7]), Interacción regulatoria (RI[7]), Reacción, Factor de transcripción (TF[7]) y Condición de crecimiento (GC[7]). Para facilitar la interpretación de la escala de similitud, se suministraron casos de ejemplo para cada nivel a los anotadores.

SPSS de 4. Ambas oraciones mencionan los mismos objetos y expresan el mismo significado, es decir, una es paráfrasis de la otra. Los siguientes son ejemplos representativos de un SPSS de 4:

- *This would mean that the IS5 element is able to provide FNR regulatory sites if inserted*

[7] Por sus siglas en inglés

at appropriate positions.

- *In any case, insertion of an IS5 element is able to increase FNR-dependent expression or to place genes under FNR control.*

SPSS de 3. Ambas oraciones comparten los mismos objetos y algunos elementos de su significado. Sin embargo, en una de las oraciones faltan elementos relevantes, no se hace referencia a los mismos objetos o se expresan conclusiones diferentes. Un caso ejemplo es cuando ambas oraciones se refieren al mismo gen y también comparten el resto de la información, pero en uno de los casos el gen es activado y en el otro reprimido. Entre los casos posibles también podemos encontrar cuando ambas oraciones tratan de la misma RI pero difieren en las condiciones en las que se presenta, o cuando ambas oraciones son paráfrasis pero una de ellas presenta más detalles que la otra.

El siguiente par de oraciones ejemplifica el último de los casos anteriores:

- *These results confirm that the N-terminal domain of NikR is responsible for DNA recognition.*
- *In preliminary experiments, we have also found that a subset of mutations within the DNA region protected by the N-terminal domain reduce the affinity of NikR for the operator—data not shown.*

SPSS of 2. Ambas oraciones comparten al menos un objeto específico y algunas otras similitudes; por ejemplo, las dos oraciones se refieren al mismo TF (ver ejemplo (a)).

En las anotaciones de los expertos se observó que las condiciones "aeróbico" y "anaeróbico" fueron consideradas como relacionadas por referirse a la disponibilidad de oxígeno. En consecuen-

cia, para este corpus, se evalúa que condiciones contrastantes como las mencionadas comparten cierto grado de similitud semántica (ver ejemplos (a) y (b)).

- Ejemplo (a)
  - *The fnr mutant was thus deficient in the anaerobic induction of fumarate reductase expression.*
  - *Aerobic regulation of the sucABCD genes of Escherichia coli, which encode K-ketoglutarate dehydrogenase and succinyl coenzyme A synthetase: roles of ArcA, Fnr, and the upstream sdhCDAB promoter.*
- Ejemplo (b)
  - *Aerobic regulation of the sucABCD genes of Escherichia coli, which encode K-ketoglutarate dehydrogenase and succinyl coenzyme A synthetase: roles of ArcA, Fnr, and the upstream sdhCDAB promoter.*
  - *Transcription of the fdnGHI and narGHJI operons is induced during anaerobic-growth in the presence of nitrate.*

SPSS of 1. Ambas oraciones tienen en común la misma clase de objetos, pero difieren en los objetos específicos.

Debido a que en la literatura de RegulonDB objetos del tipo Gen y GC son muy comunes se decidió que compartir este tipo de clase entre ambas oraciones no era condición suficiente para atribuirles una calificación SPSS de 1.

Se debe asignar un SPSS de 1 cuando se compara una oración que menciona un TF con otra que menciona un objeto de otra clase pero que hace referencia a el mismo proceso en el que el TF está involucrado

Un SPSS de 1 también debe ser considerado en aquellos casos en los que ambas oraciones hagan referencia a secuencias y genes, incluso cuando ni las secuencias ni los genes sean los mismos en ambas oraciones.

El par de oraciones siguientes es un ejemplo de SPSS 1:

- *The fnr mutant was thus deficient in the anaerobic induction of fumarate reductase expression.*

- *To test whether the formate induction of the cyx promoter could be mediated by the fhlA gene product, the expression of the cyx-lacZ fusion was examined in an fhlA deletion strain in the presence and in the absence of formate.*

SPSS of 0. Oraciones que no hacen referencia a la misma clase de objetos. Esto incluye aquellas oraciones que comparten la clase gen y GC (excepciones del nivel SPSS 1), pero no los mismos objetos específicos; esto se ejemplifica en las siguientes oraciones:

- *Carbon metabolism regulates expression of the pfl (pyruvate formate-lyase) gene in Escherichia coli.*

- *Later work showed that most mutants lacking either ACDH or ADH activities of the AdhE protein mapped in the adhE gene at 27.9 min [1,4].*

Se estipuló que la anotación tenía como objetivo evaluar la similitud semántica entre cada par de oraciones, independientemente de su tema. Asimismo, que no era necesario que las oraciones incluyeran contenido biológico o que se refirieran a objetos de RegulonDB.

Proceso de anotación

Para facilitar el proceso de anotación se proporcionó una plantilla a los anotadores (ver figura 4.2). Esta plantilla se diseñó para tener toda la información necesaria y que los anotadores no tuvieran que consultar otros archivos durante la evaluación.

Se generó una plantilla para cada anotador que incluía sólo los pares de oraciones que le correspondía evaluar. Los pares de oraciones asignados a cada anotador se seleccionaron de manera aleatoria. Cada par de oraciones se dispuso en un renglón y se incluyeron columnas, una por sesión de anotación, para que ahí se anotara la calificación de similitud. También se incluyó un resumen de la escala de similitud acordada en las sesiones de consenso.

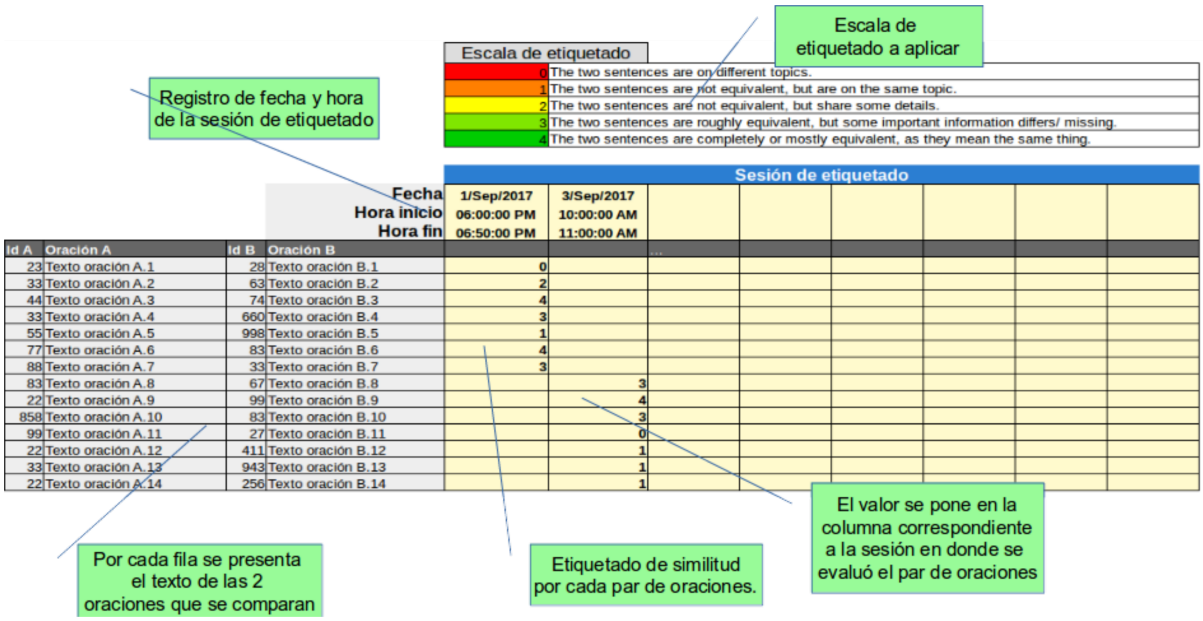

Figura 4-2: Se muestra la plantilla proporcionada a cada anotador. Cada renglón contiene el par de oraciones que se deben comparar y las celdas a la derecha están reservadas para que el anotador registre la calificación de similitud; cada columna representa una sesión diferente de anotación. En la parte superior se incluyó un resumen de la escala de similitud consensuada. Imagen tomada de [15].

Para el proceso de anotación se establecieron ciertas reglas: la evaluación de todos los pares

de oraciones debía realizarse en una semana a partir de ser entregada la plantilla; cada anotador podía realizar tantas sesiones de anotación como quisiera durante esa semana, pero se solicitó que los resultados fueran anotados en una columna diferente correspondiente a la fecha y hora de la sesión; se pidió que durante las sesiones de anotación no se realizaran otras tareas en paralelo.

Métodos para evaluación del corpus

Para evaluar la calidad de un corpus de este tipo se recomienda usar la medida de acuerdo-entre-evaluadores o confiabilidad-inter-evaluadores (IRR[8]) [170, 171, 172, 168, 173, 174]. Esta medida indica qué tanto acuerdo hay entre dos o más anotadores al calificar elementos bajo una escala nominal, ordinal, de intervalo o proporcional. Las medidas IRR se basan en la idea de que las calificaciones obtenidas O (calificaciones observadas) están dadas por $O = T + E$, en donde T es la calificación que se obtendría si no hubiera errores de medición (es decir, calificación verdadera) y E representa los errores de medición [170]. Una de las posibles fuentes de errores de medición es la inestabilidad de los instrumentos de medición cuando hay varios anotadores involucrados. Las medidas de IRR se enfocan a excluir el error de medición entre los anotadores (E) y, de esta manera, mostrar qué tanto de la variabilidad de las calificaciones observadas (O) corresponde a la varianza en las calificaciones verdaderas (T). Los coeficientes de IRR representan qué tanto se acercan las calificaciones dadas por múltiples anotadores a las calificaciones que se esperarían si todos los anotadores hubieran usado el mismo instrumento de medición. Mientras más alto resulte el coeficiente las mediciones serán más confiables.

Existen múltiples medidas estadísticas de IRR de las cuales unas se adaptan mejor que otras dependiendo de las características del estudio. Para elegir la estadística más adecuada se deben considerar: el tipo de variable que se está midiendo (ej., nominal, ordinal, etc.), si se trata de

[8]Por sus siglas en inglés

un diseño totalmente-cruzado o no, y si se desea medir la confiabilidad del anotador o de las anotaciones. De acuerdo a nuestro diseño (ver sección Diseño de la anotación), se trató de una evaluación no-totalmente-cruzada, con una variable ordinal y se deseaba medir la confiabilidad de las calificaciones. Considerando estas características se determinaron las medidas más adecuadas: *Fleiss' Kappa* (Fleiss) [175], *Krippendorff's Alpha* (Kripp), *Intra Class Correlation* (ICC) [176], *Kendall* (Kendall) [177], and *Gwet's AC1* (Gwet) [171].

Una de las IRR más usadas es el análisis de Cohen's Kappa (k) (4-5) [178]. Esta medida se representa como $k = (\mathfrak{p}_o - \mathfrak{p}_c)/(1 - \mathfrak{p}_c)$ y se trata de una relación entre la proporción de unidades en las cuales concuerdan los anotadores ($\mathfrak{p}_o$) y la proporción de unidades en las que se esperaría acuerdo por azar ($\mathfrak{p}_c$). Una variante de esta medida es *Fleiss' Kappa* (4-1); una medida no-ponderada que considera categorías sin orden. Fue diseñada para casos en los que m evaluadores son tomados de manera aleatoria de un conjunto de M evaluadores (en donde $M > m$) y cada elemento es evaluado por un grupo diferente de evaluadores.

En la ecuación 4-1 p_a representa el acuerdo entre anotadores acerca de la calificación para el elemento, y p_ϵ la proporción de elementos asignados a las categorías.

$$k = \frac{p_a - p_\epsilon}{1 - p_\epsilon} \tag{4-1}$$

La medida *Krippendorff's Alpha* (4-2) se basa en calcular el desacuerdo. Sus ventajas son que puede calcularse aun cuando falten algunos datos, que tiene buen desempeño en conjuntos de diferentes tamaños y que soporta variables de tipo ordinal, categóricas, por intervalo o de proporción. En 4-2, D_o representa el desacuerdo observado y D_ϵ el desacuerdo que se esperaría si las calificaciones fueran al azar, esto es, la proporción entre el desacuerdo observado y el

desacuerdo esperado.

$$\alpha = 1 - \frac{D_o}{D_\epsilon} \tag{4-2}$$

Intra-class correlation (4-3) es una medida de consistencia. Sirve para evaluar la confiabilidad de las calificaciones al comparar la variabilidad de las calificaciones de cada elemento con la variabilidad de todas las calificaciones de todos los elementos. Es apropiada para estudios totalmente-cruzados y para los no-totalmente-cruzados, y para estudios con dos o más anotadores por elemento. Una de sus ventajas es que al igual que Weighted-Kappa esta medida considera la magnitud del desacuerdo.

En 4-3, $var(\beta)$ representa la variabilidad debido a la diferencia entre las evaluaciones de los elementos, $var(\alpha)$ la diferencia en las reevaluaciones de los elementos, y $var(\epsilon)$ la variabilidad debido a diferencias en las escalas usadas por los evaluadores. Para este estudio, debido a que los anotadores de cada elemento son seleccionados al azar, se seleccionó la variante ICC "one-way" para evitar el sesgo sistemático entre evaluadores. Debido a que cada elemento será evaluado por el mismo número de anotadores (tres) la unidad de análisis usada para calcular la ICC fue el promedio de las calificaciones.

$$ICC = \frac{var(\beta)}{var(\alpha) + var(\beta) + var(\epsilon)} \tag{4-3}$$

La medida *Kendall's coefficient* es una medida de asociación que utiliza la posición de los elementos dentro de un ranking dado por sus calificaciones para cuantificar el grado de acuerdo entre los anotadores. Es decir, si se ordenan los elementos de acuerdo a la calificación que cada anotador le haya asignado, el coeficiente de Kendall será mayor mientras más se parezcan las

posiciones de cada elemento entre las diferentes listas (uno por cada anotador). Esta métrica tiene la ventaja de ser más resiliente a sesgos sistemáticos de los anotadores porque en lugar de utilizar las calificaciones directamente se basa en la distancia normalizada entre la posiciones. En 4-4, para una muestra de n elementos, n_c se refiere al número de elementos con la misma posición y n_d al número de elementos con posiciones diferentes.

$$W = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4-4}$$

En [179] se demostró que el coeficiente Kappa se ve seriamente influenciado por la prevalencia de una o más categorías.[9] En nuestro caso si se tiene un corpus no balanceado, debido a tener más ejemplares de uno de los niveles de similitud que de los demás, el coeficiente Kappa daría resultados sesgados. La estadística *Gwet's coefficient* 4-6 atiende estas deficiencias. Al igual que la medida Kappa, este coeficiente toma en cuenta el acuerdo debido al azar para evitar aumentar el IRR debido a acuerdos no intencionales.[10] Sin embargo, a diferencia de Kappa, este coeficiente no asume la independencia de los anotadores para definir el nivel de acuerdo al azar, sino que lo define a través de la proporción de desacuerdo (esto es, proporción entre el desacuerdo observado y el esperado al azar). La diferencia entre las medidas Gwet y Kappa radica en la forma de calcular la probabilidad de acuerdo-al-azar. Sus principales ventajas son que es más estable al ser menos sensible a la homogenidad marginal y a la prevalencia de una clase, que se puede usar en estudios con múltiples anotadores por elemento, que se puede aplicar a variables de tipo categóricas, ordinales, por intervalo y proporcionales; e incluso, que se puede aplicar cuando falten algunas calificaciones. Además, contrario a la variante Weighted-Kappa,

[9] Si la variable medida es categórica, aplica a otros tipos de variables.
[10] Cuando los anotadores asignan la misma calificación pero por razones diferentes.

no es necesario proveer pesos arbitrarios para aplicarla a variables ordinales.

$$Kappa = \frac{p - e(\kappa)}{1 - e(\kappa)} \tag{4-5}$$

$$AC = \frac{p - e(\gamma)}{1 - e(\gamma)} \tag{4-6}$$

$$e(\kappa) = \left(\frac{A1}{N}\right)\left(\frac{B1}{N}\right) + \left(\frac{A2}{N}\right)\left(\frac{B2}{N}\right) \tag{4-7}$$

$$e(\gamma) = 2P_1(1 - P_1) = 2\left(\frac{(A1 + B1)/2}{N}\right)\left(1 - \left(\frac{(A1 + B1)/2}{N}\right)\right) \tag{4-8}$$

Existen dos variantes de la estadística Gwet: AC1 y AC2. AC2 es una versión ponderada de AC1 que considera que ciertas diferencias en la clasificación son más relevantes que otras. AC2 fue diseñada para estudios con cualquier número de anotadores por elemento y en los que se aplica una escala categórica ordenada por lo que es la más adecuada para evaluar este corpus.

4.3. Resultados

Periodo de entrenamiento

Durante el periodo de entrenamiento se realizaron cuatro sesiones en las cuales todos los anotadores anotaron el mismo conjunto de pares de oraciones. Después, se revisaron todas las anotaciones y se debatieron causas y razones para los pares en donde no había consenso. Con

base en estas discusiones se ajustaron las guías de evaluación (sección refinamiento consensuado, p. 62).

Las sesiones de entrenamiento se dieron por terminadas cuando se alcanzó un coeficiente de IRR lo suficientemente alto para considerar que ya existía un consenso sistemático entre los anotadores. En la gráfica 4.3 y en la tabla 4-3 se puede observar que el acuerdo fue incrementando en cada sesión. La métrica con peores resultados al concluir el entrenamiento fue Fleiss' Kappa con un valor de 0.546, que de acuerdo a Landis [180] puede ser considerado como un IRR *moderado*. Sin embargo, para otras métricas como ICC y Gwet's AC2 se obtuvieron valores mucho más altos de 0.964 y 0.910 respectivamente. Cabe resaltar que de acuerdo a lo expuesto en la sección de métodos (p. 67) Gwet's AC2 es la estadística más adecuada para este corpus.

La gráfica 4.3 se usó para analizar si existía un sesgo sistemático de alguno de los anotadores. En ella se observó que el anotador 4 tenía consistentemente menor inter-acuerdo con el resto de los anotadores y se determinó que requería más explicación de las guías de evaluación.

Session	Kendall	Fleiss	ICC	Kripp	Gwet
1	0.216	0.024	0.454	0.116	0.545
2	0.208	0.267	0.728	0.268	0.565
3	0.430	0.390	0.813	0.439	0.826
4	0.727	0.546	0.964	0.766	0.910

Tabla 4-3: Progreso del inter-acuerdo (IRR) durante las sesiones de consenso.

Corpus

Después del periodo de entrenamiento se construyó el corpus siguiendo los lineamientos descritos en el diseño de la anotación (p. 59). El resultado fue un corpus de 171 pares de oraciones en donde cada par fue anotado por 3 anotadores seleccionados al azar del grupo de 7. Es importante mencionar que los pares anotados durante las sesiones de entrenamiento no

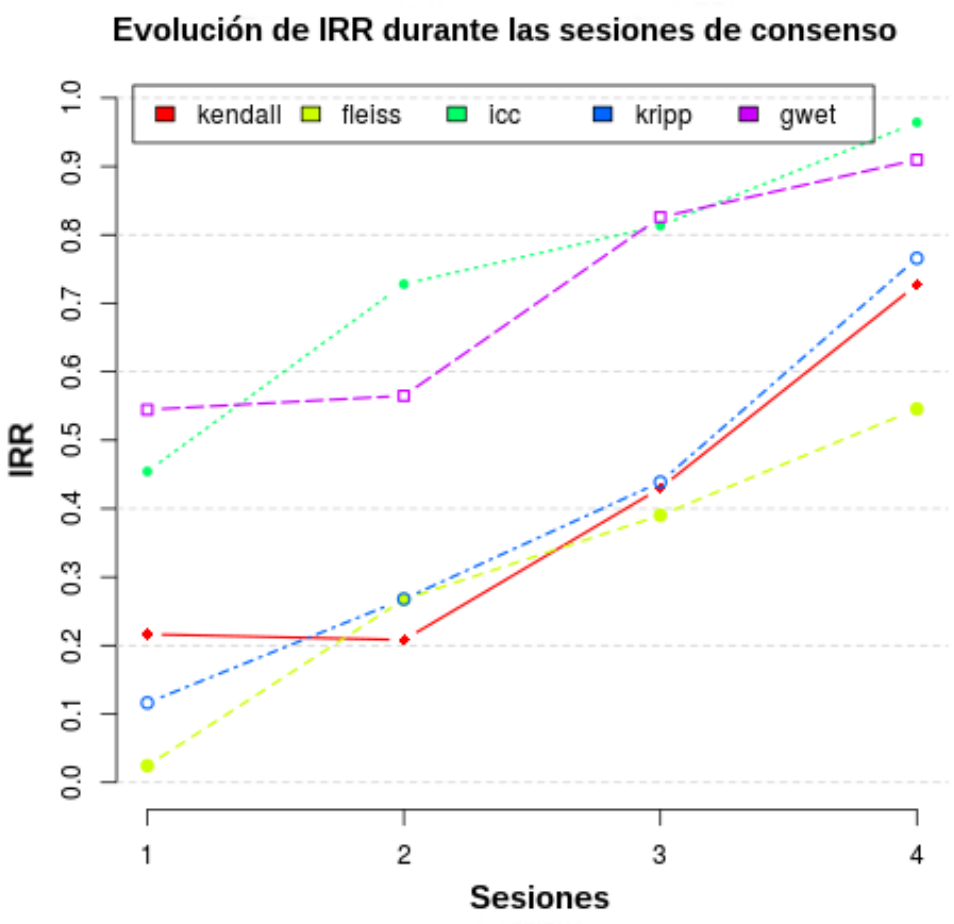

Figura 4-3: Muestra el coeficiente de IRR medido con cinco métricas diferentes. El eje x muestra el número de la sesión de entrenamiento en orden cronológico (1 representa la primera sesión y 4 la última) y el eje y representa el valor de IRR. Imagen tomada de [15].

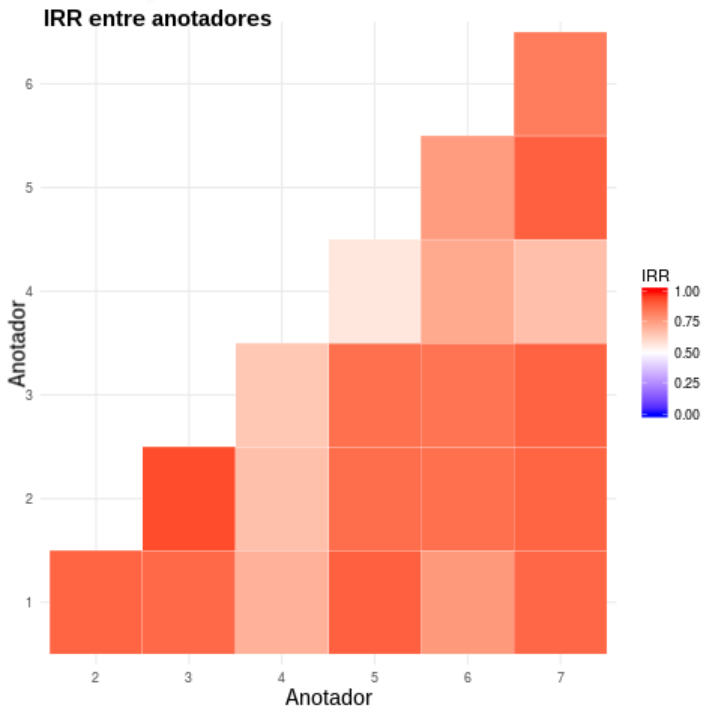

Figura 4-4: Muestra el IRR, medido con la métrica ICC, entre cada anotador y los demás. Tanto el eje x como el eje y representan a los 7 anotadores y la intersección muestra el IRR entre el anotador del eje x y el del eje y. A mayor nivel de acuerdo mayor la intensidad del color rojo. Por ejemplo, entre los anotadores 2 y 3 ($x = 3$, $y = 2$) hay mucho mayor nivel de acuerdo que entre los anotadores 4 y 5 ($x = 5$, $y = 4$). Imagen tomada de [15].

fueron incluidos en el corpus final.

Se evaluó el IRR del corpus con las cinco métricas seleccionadas y se obtuvieron buenos niveles de inter-acuerdo (ver tabla 4.3). Las distribuciones marginales de los niveles de similitud no mostraron sesgos significativos entre los anotadores (figura 4.3), pero sí una prevalencia de niveles bajos de similitud (figura 4.3). Usando la estadística Gwet's AC2 se obtuvo un coeficiente $AC2 = 0.8696$ con un intervalo de confianza de 95 % [0.8399, 0.8993]. Esto se puede considerar como *muy buen nivel de acuerdo* de acuerdo a Wongpakaran [181].

Con el propósito de descartar que el diseño no-totalmente-cruzado pudiera haber incrementado incorrectamente el IRR se efectuó el siguiente análisis: se formaron grupos de pares de oraciones que tuvieran los mismos 3 evaluadores; cada grupo se consideró como un estudio totalmente-cruzado;[11] se calculó el IRR de cada grupo y la media aritmética de todos los grupos. Estos promedios (ver tabla 4.3) resultaron muy parecidos a los coeficientes de IRR de todo el corpus con lo que se confirmó su confiabilidad.

En cuanto a la distribución de niveles de similitud (figura 4.3), se observa que existe una prevalencia de niveles bajos. Sin embargo, más del 50 % de los pares de oraciones están calificados con una similitud media (entre 1-3), con lo que se obtiene una representación aceptable del fenómeno.

Estadística	Variable	Valor	p-valor
Fleiss' Kappa	Kappa	0.443	0
Krippendorff's Alpha	Alpha	0.745	
Kendall's coefficient	W	0.741	7.86e-18
Intraclass Correlation	ICC	0.919	6.7e-83
Gwet's coefficient	AC2	0.870	0

Tabla 4-4: Inter-acuerdo del corpus medido con diferentes estadísticas

[11] Debido a que dentro de cada grupo todos los pares fueron evaluados por todos los anotadores.

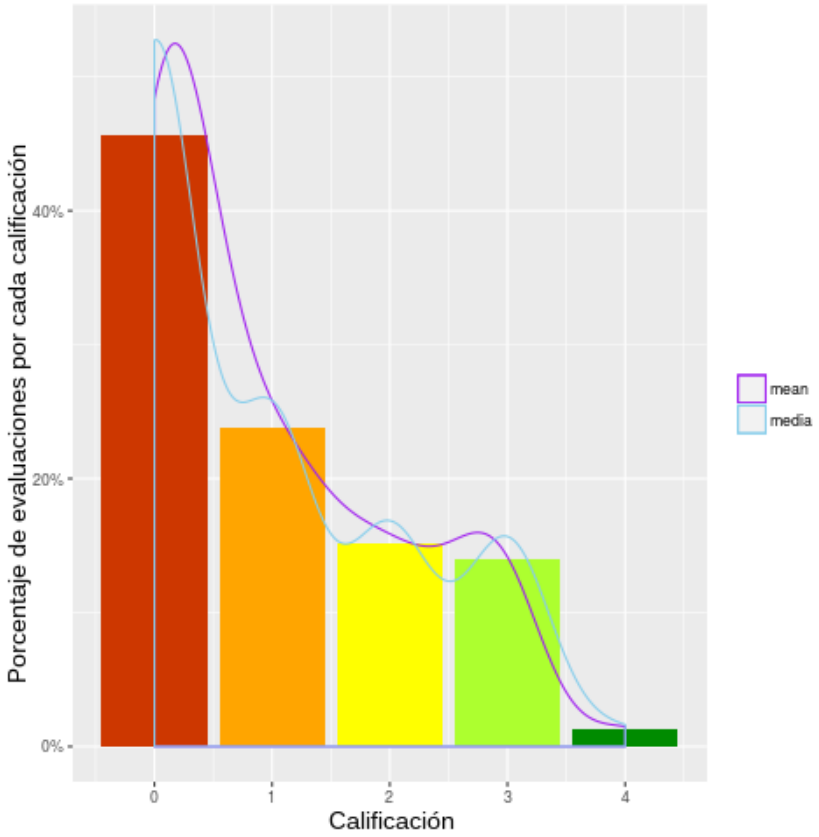

Figura 4-5: Se muestra la distribución de los pares anotados de acuerdo a su nivel de similitud. El eje x representa la escala de similitud (5 niveles) y el eje y, el porcentaje del corpus en cada nivel. No se incluyen los ejemplares del periodo de entrenamiento. Imagen tomada de [15].

Anotador	Kendall	Fleiss	ICC	Kripp	Gwet
1, 2, 3	0.782	0.597	0.941	0.814	0.864
1, 2, 7	0.641	0.512	0.926	0.705	0.894
1, 6, 7	0.788	0.358	0.912	0.756	0.686
2, 3, 4	0.669	0.442	0.916	0.691	0.907
3, 4, 5	0.712	0.310	0.894	0.708	0.802
4, 5, 6	0.593	0.268	0.753	0.602	0.818
5, 6, 7	0.833	0.409	0.913	0.772	0.784
Promedio	0.717	0.414	0.894	0.721	0.822

Tabla 4-5: IRR por grupo de anotadores

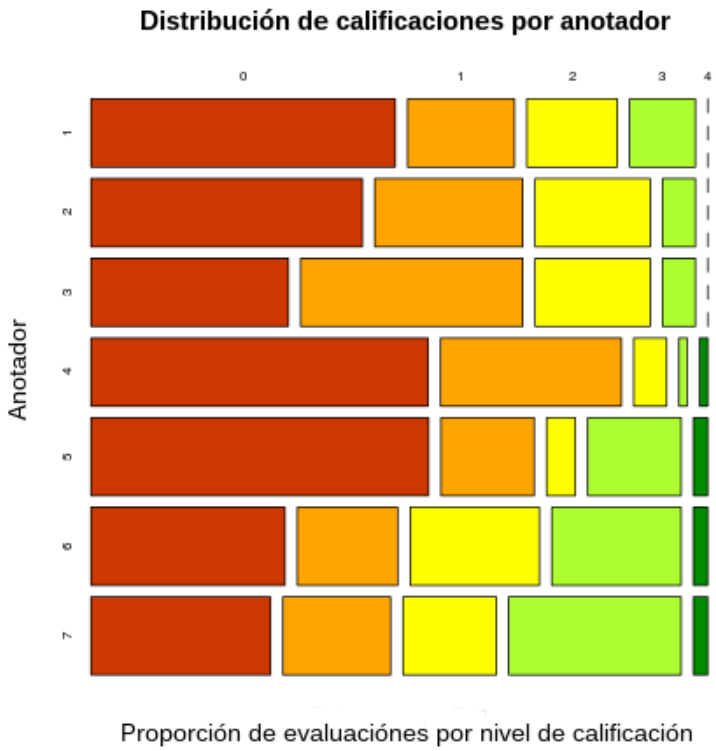

Figura 4-6: Se muestran los 7 anotadores y para cada uno la distribución de sus evaluaciones de acuerdo al nivel de similitud. El eje y representa a los 7 anotadores y en el eje x se muestra la proporción de pares de oraciones que el anotador clasificó en cada nivel de similitud. En el extremo izquierdo se muestran las anotaciones de nivel 0 de similitud (color rojo) e incrementan hasta el nivel 4 de similitud (color verde), que se encuentra en el extremo derecho. Imagen tomada de [15].

4.4. Discusión

Se observó que el IRR incrementó de manera significativa después de la tercera sesión de entrenamiento. Esto se puede explicar por dos motivos: que los anotadores ya estaban más familiarizados con el objetivo de la tarea –en la primera sesión hubo una marcada tendencia a evaluar la similitud de los objetos biológicos y no de la semejanza de significado de la oración completa– y que para la tercera sesión los anotadores ya habían recolectado un conjunto de ejemplos consensuados que les servían de referencia en caso de dudas. Esto mostró la importancia de las guías de evaluación y de las sesiones de entrenamiento y de consenso. Estas observaciones muestran que la evaluación de similitud es un proceso intuitivo, que depende del contexto personal y para el cual no hay un consenso perfecto, sobre todo en lo que se refiere a los grados de similitud [182, 183, 184, 87]. En general, los anotadores consideraron que fue más difícil evaluar aquellas oraciones que no incluían algún objeto biológico.

Como prueba final de la pertinencia de nuestro corpus (RTM) se comparó con BIOSSES; el único otro corpus de similitud que es específico al área biomédicas. Primero, en referencia a la distribución en los niveles de similitud, BIOSSES es un corpus más balanceado con 15 %, 12 %, 27 %, 35 % y 11 % respectivamente para los niveles de similitud (de 0 a 4). En cambio la distribución del corpus RTM es de 48 %, 22 %, 15 %, 14 % y 1 % para los mismos niveles. En relación a su tamaño, a pesar de sólo tener 171 ejemplares, el corpus RTM es 70 % más grande que BIOSSES que sólo cuenta con 100 pares de oraciones. Finalmente, se analizó la especialización de ambos corpora. A pesar de que el corpus BIOSSES es específico al área biomédica, es aún muy general para un dominio como el de RTM. Esto se comprueba al comparar los 50 léxicos más frecuentes de ambos corpora. En el caso de BIOSSES encontramos *célula*, *tumor*, *cáncer*, *estudio*, *reporte*, *humano*, *gen*, *pulmón*, *leucemia*, etc.; y en nuestro caso encontramos *sitio*, *expresión*, *activación*, *gen*, *proteína*, *cepa*, *regulación*, *DNA*, *región*, *downstream* y *upstream*. Los expertos

(anotadores) concordaron que el léxico del dominio RTM está muy vagamente representado en el corpus BIOSSES.

Este corpus es, al mejor de mi entendimiento, el primero en su tipo para un dominio tan específico y especializado como el de RTM. A pesar de que no se obtuvo un corpus tan balanceado como se hubiera deseado, se obtuvo una representación aceptable de los niveles de similitud. El resultado es una fuente ad hoc de entrenamiento y de prueba para las estrategias de similitud que se implementarán en este dominio y una fuente de conocimiento para otros corpora en dominios especializados.

Capítulo 5

Experimentos con métricas tradicionales

En este capítulo se reportan los experimentos realizados con métodos tradicionales para medir STS. De acuerdo a lo revisado en el capítulo 2, una de las estrategias que mejor resultados ha dado cuando no se tiene un corpus masivo de entrenamiento consiste en combinar varias métricas de similitud, las cuales individualmente no requieran entrenamiento o que puedan ser entrenadas con pocos datos, para posteriormente combinarlas por medio de algoritmos de regresión. En estos experimentos se usaron métricas textuales, distribucionales y basadas en recursos estructurados de conocimiento, y se combinaron usando regresión lineal, *Random Forest*, *Multilayer Perceptron*, *Bagging* y *Voting*. En el resto del capítulo se describen las métricas utilizadas, las estrategias de combinación, su evaluación y una discusión.

5.1. Metodología

Se usó el corpus ad hoc de similitud semántica que se construyó específicamente sobre la literatura de regulación transcripcional (RTM) [15] (capítulo 4).

El proceso consistió en preprocesar los pares de oraciones, aplicar las medidas de similitud semántica y combinar las calificaciones individuales de estas medidas.

En el paso de *preprocesamiento* se usó una lista de objetos de RegulonDB con nombres multipalabra; se ordenaron los nombres de acuerdo a su número de palabras (en orden descendente), y se identificaron en el texto buscando la coincidencia más larga de izquierda a derecha. Con este paso se identificaron casi 400 conceptos multipalabra como por ejemplo: *isoosmotic condition*, *transcriptional dual regulator*, *DNA-binding protein*, *KdpE-phosphorylated*, *acid-responsive* y *cold-shock*. Posteriormente, se lematizaron e identificaron las categorías gramaticales (POS tagging) de las unidades léxicas. Finalmente, se obtuvieron los arboles de constituyentes de las oraciones.[1]

El paso de *medición de similitud* consistió en aplicar tres tipos de medidas que consistieron en medidas textuales, distribucionales y basadas en recursos estructurados (ej., ontologías).

Finalmente, en el paso de *combinación de medidas* se usaron algoritmos de regresión y de ensamble para combinar las mediciones individuales del paso anterior. Se experimentó con los algoritmos de *Linear Regression*, *Multilayer-Perceptron* y *Random Forest*. Además se aplicaron los métodos de ensamble *Bagging* y *Voting* para investigar si proveían alguna mejora.

A continuación se describen las medidas usadas.

[1] Se usó la herramienta *Stanford CoreNLP toolkit*.

5.1.1. Medidas de similitud textual

Levenshtein

Esta métrica mide la similitud entre dos textos (s y t) comparando el número de inserciones, eliminaciones y sustituciones necesarias para transformar el texto fuente (s) en el texto objetivo (t). La calificación obtenida representa la distancia por lo que mientras mayor sea la calificación, más diferentes serán los textos. La distancia entre S_1 y S_2 está dada por $LD_{S_1,S_2}(|S_1|,|S_2|)$, en donde:

$$LD(i,j) = \begin{cases} max(i,j) \\ \quad [\text{if } min(i,j) = 0] \\ min \begin{cases} ld(i-1,j)+1 \\ ld(i,j-1)+1 \\ ld(i-1,j-1)+[S_{1_i} \neq S_{2_j}] \end{cases} \\ \quad [\text{en-otro-caso}] \end{cases} \tag{5-1}$$

Esta es una medida de distancia; para poderse usar como medida de similitud se debe normalizar y complementar. Si los textos son idénticos su calificación es de 1, y tiende a 0 al decrecer sus similitud.

$$LS(S_1,S_2) = 1 - \frac{LD(S_1,S_2)}{max(|S_1|,|S_2|)} \tag{5-2}$$

Jaccard

Esta métrica mide la similitud entre dos conjuntos finitos dividiendo la cardinalidad de su intersección entre la cardinalidad de su unión. Por lo tanto la similitud entre dos textos representados por $S_1 = w_1, w_2, .., w_n$ and $S_2 = w_1, w_2, ..., w_m$, está dada por:

$$Jaccard(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{5-3}$$

N-gramas

Esta medida es una generalización de *la secuencia común más larga* (LCS[2]) que se basa en el número de elementos de tamaño N que son compartidos por dos textos [185]. En cadenas representadas por listas de n-gramas –en donde un n-grama es un grupo de n elementos adyacentes– su similitud está dada por la proporción entre los n-gramas compartidos y el mayor de los n-gramas. Los coeficientes de *Dice* y *Overlap* son variantes de la medida de similitud por n-gramas.

$$s_n(\Gamma_{k,l}) = \quad max(s_n(\Gamma_{k-1,l}), s_n(\Gamma_{k,l-1}), s_n(\Gamma_{k-1,l-1}) + s_n(\Gamma^n_{k-n,l-n})) \tag{5-4}$$

5.1.2. Medidas distribucionales

Como ya se ha dicho, en los modelos distribucionales las palabras, oraciones e incluso párrrafos se representan como vectores multidimensionales cuya proximidad geométrica se usa para medir su similitud semántica. En estos experimentos se usaron los *embeddings* GloVe. Sin embar-

[2] *Longest Common Subsequence.*

go, estos *word-embeddings* representan unidades léxicas (palabras), por lo que antes de usarlos para comparar frases u oraciones es necesario combinar los *embeddings* correspondientes a las palabras de cada oración. A continuación se describen los métodos de combinación que se usaron.

Promedio de *word-embeddings*

Uno de los métodos más simples de composición consiste en promediar los embeddings correspondientes a las palabras de cada oración [102]. El resultado es un embedding para representar a cada oración por lo que la medida de similitud se obtiene calculando el coseno entre los vectores.

Esta técnica se usó para calcular dos valores de similitud, cada uno usando una colección diferente de embeddings :

1. Los embeddings pre-entrenados de GloVe; consisten en vectores de 300 dimensiones entrenados en el corpus *Common Crawl* (42B tokens, 1.9M vocabulario, no considera mayúsculas y minúsculas).

2. Los embeddings ad hoc, consisten en vectores de 300 dimensiones entrenados en ∼6k publicaciones del tema *Transcriptional-Regulation* (31M tokens, 108k vocabulario, no considera mayúsculas y minúsculas).

Los embeddings ad hoc se entrenaron en aproximadamente seis mil artículos científicos del tema *regulación transcripcional microbiana*. Primero se procesaron los artículos en formato PDF usando una herramienta que se desarrolló para dicho propósito. Esta herramienta permitió la extracción de oraciones completas del contenido principal de la publicación, es decir, sin tomar en cuenta pies de figura, tablas, notas al pie de página, etc. Además, durante la extracción se consideró el orden de lectura para unir oraciones segmentadas por cambios de página y de

columna. El resultado fue un archivo de texto plano por artículo con una oración completa por línea. El siguiente paso consistió en separar por palabras cada archivo y unir los archivos individuales en uno sólo. En el archivo resultante, cada línea contiene las palabras de todas las oraciones de una publicación, separadas por un espacio. Finalmente, se usaron los scripts[3] de GloVe que están disponibles al público para procesar el archivo generado en el paso anterior. Durante el procesamiento se construyó el vocabulario que incluye la frecuencia de cada palabra (uni-gramas) y en donde al final se excluyeron aquellas que tenían una cuenta menor que cinco. Después, se obtuvo una matriz de co-ocurrencia palabra-palabra para lo cual se usó una ventana de 15 palabras. Sobre estas estadísticas de co-ocurrencia se entrenaron los embeddings de 300 dimensiones.

Embedding de oración a partir de un vector común

Esta estrategia de composición consiste en:

1. Generar un diccionario común D a partir de comparar dos oraciones (A y B).
2. Para cada oración O_j
    - Crear un vector V del tamaño del diccionario D, esto es, $|V| = |D|$.
    - Calcular la similitud de cada palabra de D (D_i) con cada palabra de la oración O_j (usando el coseno entre los vectores)
        - Seleccionar el valor más alto y usarlo como la magnitud de la dimensión i de V.

Una vez que se tiene un embedding para cada oración se mide la similitud entre ellas calculando el coseno. Esta técnica se aplicó a la colección de embeddings ad hoc entrenados en la literatura de regulación transcripcional.

[3] https://github.com/stanfordnlp/GloVe

5.1.3. Medidas basadas en recursos estructurados

Como se analizó en el capítulo 2, medidas de este tipo permiten aprovechar el conocimiento que expertos han expresado en recursos estructurados; ej. ontologías, taxonomías, tesauros, etc. Este tipo de recursos con frecuencia modelan dominios generales lo que favorece la comparación semántica en diferentes ámbitos, pero la limita seriamente en aquellos que son especializados. Cuando en el texto se encuentran términos especializados que no están definidos en un recurso general se requiere de un recurso específico del dominio; tal es el caso de la literatura de la *regulación de la transcripción en genes*. Por estas razones se decidió usar un recurso específico en donde estuvieran definidos los términos especializados del dominio y uno general que modelara los demás elementos del lenguaje.

Como recurso general se usó Wordnet [186]. Aunque no es una ontología, sus relaciones de tipo *is-a* (hiperónimos e hipónimos) se pueden usar como una ontología simple. Como recurso especializado se usó una ontología basada en los objetos RegulonDB que fue desarrollada en el Centro de Ciencias Genómicas para modelar el dominio de regulación transcripcional microbiana.

Para medir la similitud entre términos se usó la métrica *Lin98* (5-5). Si para dos conceptos S_1 y S_2 se definen:

- depth(s) = profundidad de s en la ontología.
- $LCS(S_1, S_2)$ = es el padre común más cercano de S_1 and S_2, es decir, el padre común que está más abajo en la ontología.
- $IC(s) = -log\frac{1}{P(s)}$, Contenido de información de s de acuerdo al corpus.

La métrica Lin98 está dada por:

$$Lin98(S_1, S_2) = \frac{2 \cdot IC(LCS(S_1, S_2))}{IC(S_1) + IC(S_2)} \tag{5-5}$$

5.2. Resultados

Las medidas usadas se reportan en el orden siguiente:

1. *Wordnet*; medida basada en ontología usando la estrategia de vector común, Wordnet como fuente de conocimiento y Lin98 como métrica.
2. *Levenshtein*; medida normalizada de Levenshtein.
3. *Jaccard*; medida de Jaccard.
4. *Ngram*; medida basada en N-Gramas.
5. *Glove (averaged)*; promedio de *word-embedding* usando la colección de vectores GloVe pre-entrenados.
6. *AdHoc-Glove (averaged)*; promedio de *word-embedding* usando la colección de vectores GloVe ad hoc entrenados en literatura de regulación transcripcional.
7. *AdHoc-Glove (common-vector)*; embedding de oración a partir de un vector común usando la colección de vectores ad hoc entrenados en literatura de regulación transcripcional.
8. *AdHoc-ontology*; medida basada en ontología usando la estrategia de vector común, la ontología ad hoc como fuente de conocimiento y Lin98 como métrica.

Estas medidas se entrenaron y evaluaron usando el corpus RTM. Para cada par de oraciones

se usó el promedio de las calificaciones de los anotadores como medida de referencia (GS[4]). Ninguna de las medidas, incluso el GS presentaron una distribución normal (ver tabla 5-1).

Medida	p-valor
1) Wordnet	0.00009
2) Levenshtein	0.00000
3) Jaccard	0.00000
4) NGRAM	0.00000
5) Glove (averaged)	0.00003
6) AdHoc-Glove (averaged)	0.00000
7) AdHoc-Glove (common-vector)	0.00799
8) AdHoc-ontology	0.00084
GS	0.00000

Tabla 5-1: Prueba de normalidad de las medidas

5.2.1. De medidas individuales

Se realizo una prueba de correlación de Pearson (ρ) y Spearman (r_s) entre cada medida y el GS (ver tabla 6-3). La correlación más alta la obtuvo la medida *AdHoc-Glove (common-vector)* (7) con $\rho = 0.604$; y la más baja fue *Levenshtein* (2) con $\rho = 0.235$. También se hizo una comparación entre ambas correlaciones para tener una mejor perspectiva del comportamiento de las medidas. Una correlación alta de Pearson indica una relación lineal, mientras que un valor alto en Spearman indica una relación monotónica pero no necesariamente lineal. El valor de Spearman es mayor para todas las medidas pero la diferencia es muy pequeña, con excepción de la medida *AdHoc-Glove (averaged)* (6) para la cual es 12 % mayor que la de Pearson. Esto indica que una regresión lineal sería desfavorable para esta medida, que de hecho es una de las más altas.

[4] *gold standard.*

	Correlación		
Medida	ρ	r_s	diferencia
1) Wordnet	0.26059	0.27707	0.01648
2) Levenshtein	0.23564	0.26872	0.03308
3) Jaccard	0.48431	0.52356	0.03925
4) NGRAM	0.24115	0.25664	0.01549
5) Glove (averaged)	0.39060	0.43494	0.04434
6) AdHoc-Glove (averaged)	0.42398	0.54041	0.11643
7) AdHoc-Glove (common-vector)	0.60406	0.63418	0.03012
8) AdHoc-ontology	0.48449	0.48410	0.00039

Tabla 5-2: Correlación por medida

Algoritmo	Correlación	Desviación estándar
Linear Regression	0.622	0.15
Random Forest	• 0.683	0.14
Perceptron	0.642	0.18

• mejora estadísticamente significativa ($\alpha = 0.05$)

Tabla 5-3: Algoritmos de regresión

5.2.2. De la combinación de medidas

Para combinar las medidas se experimentó con los siguientes algoritmos: *Linear Regression*, *Random Forest* (con 100 % de substitución, 100 iteraciones, un tamaño de batch de 100 y sin restricción de profundidad) y *Multilayer Perceptron* (tasa de aprendizaje de 0.3, momentum de 0.2 y 6 unidades ocultas). Cada experimento se realizó aplicando una validación cruzada de 10 iteraciones (10-fold cross-validation) en el corpus RTM. En la tabla 5-3 se muestra la comparación de correlaciones de estos algoritmos. En ella se puede observar que con una correlación de Pearson de 0.683 *Random Forest* tiene el resultado más alto con una ganancia estadísticamente significativa sobre el segundo mejor (*Perceptron*).

También se experimentó con dos algoritmos de ensamble: *Bootstrap Aggregation (Bagging)* y *Voting*. En *Bagging* se obtienen varios modelos entrenando un mismo algoritmo con diferentes subconjuntos extraídos (con reemplazo) del corpus de entrenamiento, y la predicción final se obtiene promediando las predicciones de los modelos. Se usó un árbol de decisión (*Classification*

Algoritmo	Correlación	Desviación estándar
Random Forest	0.683	0.14
Bagging	0.691	0.14
Voting	0.693	0.13

Tabla 5-4: Desempeño de ensambles

and Decision Tree [CART]) con un tamaño de bolsa de 100 % y 10 clasificadores. En *Voting* se combinan las predicciones de varios modelos diferentes. Se usaron los modelos promedio aritmético, *Linear Regression*, *Random Forest* y *Perceptron*, y se promediaron las predicciones. El desempeño de estas estrategias se muestra en la tabla 6-4 en donde también se incluye *Random Forest* como elemento de contraste. Ambas estrategias de ensamble obtuvieron mejores resultados que *Random Forest* pero sus diferencias no fueron estadísticamente significativas. Por lo tanto para continuar con los experimentos se seleccionaron *Voting*, por obtener el mejor resultado, y *Random Forest* por que es más sencillo y tuvo resultados casi igual de buenos.

5.2.3. Pruebas de ablación

El último experimento fue una prueba de ablación para analizar el impacto que tiene cada medida de similitud al ser combinada con las demás. Esta prueba se aplicó a los algoritmos *Random Forest* y *Voting* (ver tabla 5-5). La medida *Glove (averaged)* (5) provoca la pérdida más grande (3 %) en *Random Forest* cuando se omite, mientras que para *Voting* la medida con mayor aporte es la *AdHoc-Glove (common-vector)* (7) que al ser omitida provoca una pérdida de 4.1 %. La medida que de acuerdo a la prueba tiene un impacto negativo es *Wordnet* (1). Cuando se omitió esta medida tanto *Voting* como *Random Forest* presentaron ganancias; en el caso de *Voting* fue de 0.7 % para llegar a la correlación de Pearson más alta de todas (0.700).

	Correlation	
Medida omitida	*Voting*	*Random Forest*
1) Wordnet	• 0.700	• 0.695
2) Levenshtein	0.689	0.687
3) Jaccard	0.686	0.678
4) NGRAM	0.686	0.683
5) Glove (averaged)	0.672	○ 0.653
6) AdHoc-Glove (averaged)	0.675	0.684
7) AdHoc-Glove (common-vector)	○ 0.652	0.678
8) AdHoc-ontology	0.693	0.679

○ mayor pérdida, • mayor ganancia

Tabla 5-5: Prueba de ablación

5.3. Discusión

En estos experimentos se hizo evidente la dificultad de la tarea de STS en dominios especializados como el de RTM. A continuación se muestra un par de oraciones del dominio que ejemplifican esta dificultad:

- *The dcuB gene is strongly activated anaerobically by FNR, repressed in the presence of nitrate by NarL, and subject to cyclic AMP receptor protein (CRP)-mediated catabolite repression.*

- *The results show that the dcuB gene is expressed exclusively under anaerobic conditions in a manner that is FNR dependent and that it is repressed by NarL in the presence of nitrate and is subject to CRP-mediated catabolite repression.*

Estos ejemplos muestran que las cláusulas subordinadas, las coreferencias lejanas, el uso de sinónimos e incluso las generalizaciones conceptuales (ej. *activated* - *expressed*) presentan retos significativos. Los no expertos calificaron la similitud de estas oraciones con un valor de 3 mientras que los expertos les asignaron una similitud de 4. Esto demuestra el efecto del vocabulario especializado en la evaluación de similitud y la necesidad de corpora específica al

dominio para que métodos de STS del estado-del-arte puedan aprender lo que los expertos toman en cuenta al calificar la similitud semántica.

Es importante resaltar que los resultados de este tipo de estrategias son bastante mejores en dominios generales; por ejemplo, Sultan aplicó una estrategia similar en SEMEVAL 2015 ([187] corrida S1) y sus resultados son significativamente superiores ($\rho = 0.8015$). Si bien este tipo de estrategias y las métricas que se consideraron en estos experimentos no requieren o requieren pocos datos de entrenamiento etiquetados, sí se necesitó construir un corpus etiquetado por expertos que fuera lo suficientemente grande para entrenar los algoritmos de regresión. También fue necesario tener una cantidad relevante de oraciones del dominio (31 millones de tokens), relativamente depuradas, para poder entrenar de manera no supervisada los embeddings de GloVe. Además se aprovechó una ontología generada específicamente para modelar la literatura de RTM. La realidad es que aun cuando esta estrategia es más fácil de implementar en dominios especializados que una basada en *deep learning*, no siempre se puede contar con los recursos que se requirieron.

Estas razones dan fundamento a proponer otras estrategias orientadas a dominios muy especializados y con pocos datos de entrenamiento y, si además pueden proveer mayor facilidad para explicar los resultados, mejor.

Capítulo 6

Métrica de STS por medio de abstracción

En los dos capítulos anteriores se abordaron los primeros objetivos de esta tesis: 1) Generar un corpus de entrenamiento y evaluación, y 2) Obtener un línea base con estrategias tradicionales de medición de similitud semántica. En este capítulo se atiende el último de los objetivos que consiste en proponer y aplicar una estrategia de similitud-semántica-entre-oraciones que haga uso de procesos explícitos de abstracción. A continuación se describe la propuesta metodológica; posteriormente se exponen los resultados de aplicar esta estrategia al corpus generado en el capítulo 4 y se comparan con los resultados de métodos tradicionales que se obtuvieron en el capítulo 5, finalmente se plantea una discusión breve de esta propuesta y algunas ideas que en un futuro se podrían explorar para continuar con el desarrollo en esta misma línea de investigación.

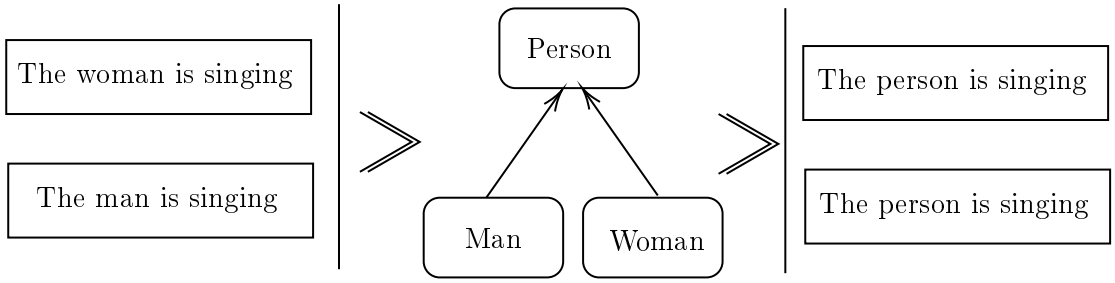

Figura 6-1: Ejemplo de abstracción léxica

6.1. Método

La propuesta se basa en comparar sub-arboles de constituyentes del mismo tipo a partir de los cuales (uno de cada oración) se genera una expresión más abstracta que representa a ambos. A los pares de sub-árboles se les aplican dos mecanismos de abstracción: una léxica y una sintáctica. La composición ordenada de ambos mecanismos genera un fragmento de oración que representa una abstracción de los fragmentos originales. Considerando que mientras más abstractas son menos informativas, se buscan aquellos fragmentos que sean lo suficientemente abstractos para representar a ambas sub-frases pero que a su vez sean lo más específicos posibles. Esta restricción sirve como guía para buscar la generalización menos general (*Least General Geralization*) [188] en el espacio de candidatos generado a partir del producto cartesiano de los sub-arboles de ambas oraciones. Por ende, para cada sub-árbol original se conserva sólo el fragmento más específico de todas las comparaciones.

El objetivo de la abstracción léxica es construir un léxico común a las frases que se comparan. Esto incluye a los términos que se presentan en ambas frases y, cuando no coinciden, a términos más abstractos que los subsumen. El resultado son las mismas frases comparadas pero expresadas con un vocabulario común entre ellas (fig. 6-1). Se trata de un mecanismo de tipo *abstracción*[1] en el cual se generalizan las representaciones hasta que son equivalentes.

[1] Ver capítulo 3 página 46

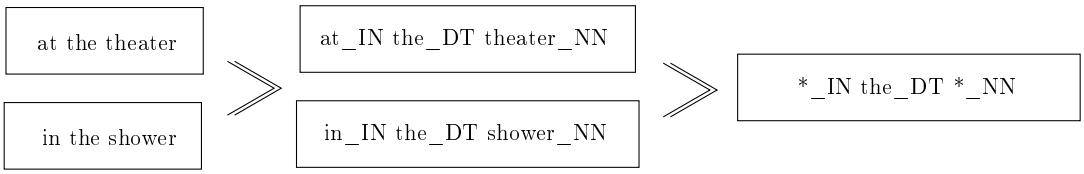

Figura 6-2: Ejemplo de abstracción sintáctica
Las unidades léxicas se transforman a unidades sintácticas al agregar el POS a cada token, finalmente se abstraen estas representaciones sintácticas.

La abstracción sintáctica parte del resultado de la abstracción léxica y tiene como objetivo generar una representación común de dos o más frases. Los términos se representan junto con su categoría gramatical (POS[2]) como una unidad léxica. Se alinean las unidades de ambas frases y aquellas que no coinciden en léxico o categoría se substituyen por una abstracción del tipo <*_POS> o <t_*> correspondientemente; es decir, se reemplaza el elemento que difiere por un comodín (ver fig. 6-2). Cabe resaltar que la alineación que se realiza no es global, sino que está acotada a sub-árboles del mismo tipo y considera las POS de las unidades. El mecanismo de abstracción que se usa es del tipo *aproximación* que se basa en la sustitución por representaciones parciales.

El producto de estas manipulaciones es una lista de fragmentos comunes (abstracciones) que representa las semejanzas léxicas y sintácticas entre las oraciones. Los mecanismos de abstracción que se usan implican una pérdida de información y de la capacidad de determinación de las representaciones nuevas (abstracciones) hacia los referentes originales. Por lo tanto, el costo de la transformación está asociado a la disminución del contenido de la información en la nueva representación; es decir, a la indeterminación causada al generar las abstracciones. Es importante recalcar que esta calificación de STS se basa en dos principios de igual importancia: 1) las abstracciones dependen de las semejanzas entre los fragmentos de ambas oraciones, y 2) se cuantifican las semejanzas midiendo la indeterminación de las abstracciones. En síntesis, si

[2]Siguiendo las directrices del esquema de anotación *Penn Treebank* [189]

existen abstracciones es porque las oraciones comparten similitudes, y lo que se mide a través de la cantidad de información perdida es el grado de similitud.

A grandes rasgos el método consiste en los siguientes pasos:

1. Obtener el árbol de constituyentes de ambas oraciones.
2. Extraer las subfrases NP, VP y PP de los árboles.
3. Comparar subfrases de oración A con subfrases de oración B del mismo tipo (ej., A:NP-1 vs B:NP-1, A:NP-1 vs B:NP-2, etc.)
4. A cada par de subfrases comparadas aplicarles un proceso de abstracción que consiste en la composición de:
   - Abstracción léxica
   - Abstracción sintáctica

   En cada paso llevar registro de la entropía introducida en el sistema debido a la indeterminación provocada.
5. Conservar la menos general de las abstracciones para cada sub-árbol (LGG). Las abstracciones más específicas son aquellas cuya entropía es menor.
6. Calcular la STS como la suma del valor de cada abstracción menos la entropía provocada por la misma.

A continuación se describe más a detalle cada uno de los pasos. Para facilitar la descripción, un ejemplo de la medición de STS entre dos oraciones se trabaja a la par en cada paso. Las oraciones que se comparan en el ejemplo son :

T_1) In May 2010 the troops attempted to invade Kabul.

T_2) The US army invaded Kabul on May 7th last year, 2010.

Primer paso

Se obtienen los árboles de constituyentes de ambas oraciones (T_1 y T_2). Un árbol de constituyentes es una estructura recursiva de representación que consiste en sub-conjuntos de una o más unidades lingüísticas (palabras) de manera tal que en cada nivel los constituyentes actúan como una unidad sintáctica. En este tipo de representación la oración original está segmentada en frases, las cuales pueden clasificarse de acuerdo al elemento que las encabeza en: frases nominales (NP), frases verbales (VP), frases preposicionales (PP), etc. Se ha demostrado que el uso de árboles de constituyentes da mejores resultados que los árboles de dependencias en tareas de alineación sintáctica [190]. Además, las unidades lingüísticas intermedias (frases) que se obtienen son útiles para generar abstracciones comunes a dos textos por las siguientes razones: 1) permiten encontrar coincidencias en granularidades menores a los textos originales pero también, al tratar con frases y no sólo con palabras, permite considerar parte de la sintaxis original. 2) las clausulas de tipo adjuntos y complementos (argumentos) son identificadas como constituyentes y ubicadas en niveles diferentes del árbol (normalmente, dentro de una NP, un complemento estará ubicado en un nivel superior al adjunto); esto permite que al comparar dos frases se pueda considerar su diferencia de niveles como un factor en el cálculo de similitud; la intuición sugiere que cláusulas similares léxica y sintácticamente pero ubicadas en diferentes profundidades del árbol de constituyentes tienen diferente relevancia semántica en el contexto de la oración. El resultado del análisis de constituyentes puede verse en la figura 6-3.

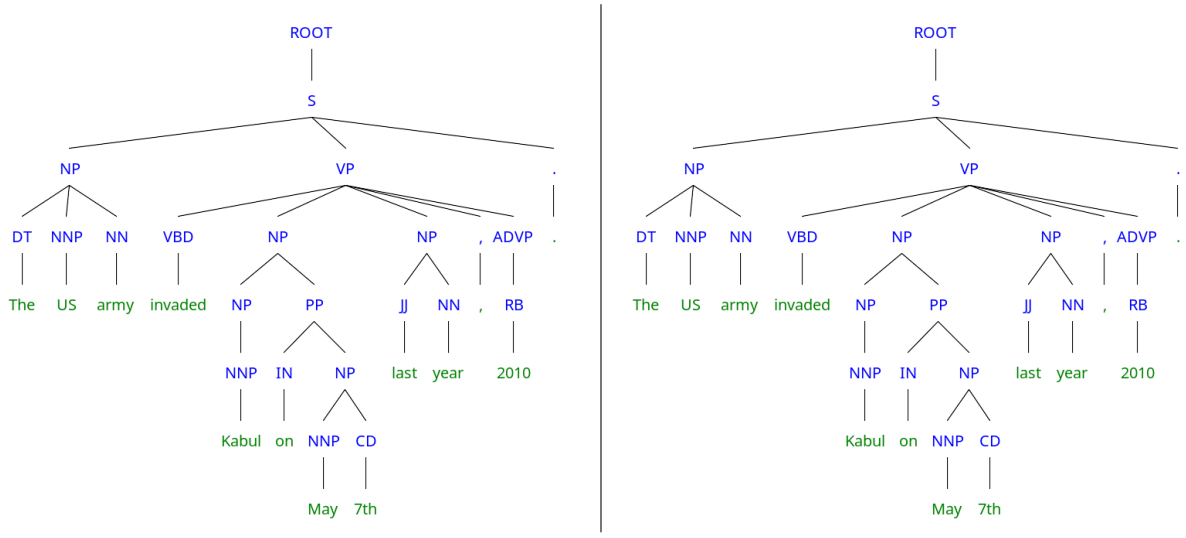

Figura 6-3: Arboles de constituyentes de oraciones de ejemplo

Segundo paso

Se obtienen las sub-frases de ambas oraciones (T_1, T_2) a partir de los constituyentes (eq. 6-1). Se exploran todos los constituyentes y, en caso de tener anidaciones se toman con traslape; es decir, la sub-frase de un constituyente incluye a los constituyentes subsumidos hasta llegar a los nodos terminales del árbol (alg. 1). Por ejemplo, si un constituyente de tipo NNP (frase nominal) contiene a un PPP (frase preposicional), se genera una sub-frase que contiene la frase nominal y la frase preposicional, y otra sub-frase con sólo el contenido de la preposicional (ver figura 6-4).

$$\forall T_i \in \{T_1, T_2\} \begin{cases} PHR_{T_i} = \{NNP_i, VBP_i, PPP_i, OTH_i \end{cases} \begin{vmatrix} NNP = \{t_1, ...t_m \mid isNNP(t) \forall\, t \in T\} \\ VBP = \{t_1, ...t_m \mid isVBP(t) \forall\, t \in T\} \\ PPP = \{t_1, ...t_m \mid isPPP(t) \forall\, t \in T\} \\ OTH = \{t_1, ...t_m \mid !isNNP(t) \\ \wedge !isVBP(t) \wedge !isPPP(t) \forall\, t \in T\} \end{vmatrix} \quad (6\text{-}1)$$

Un ejemplo de la extracción de las frases NP con traslape se puede ver en la figura 6-4.

Algoritmo 1: Paso 2 - Obtención de subfrases

Input: Arboles de constituyentes C_1 y C_2
Result: Conjunto de frases por tipo para cada árbol

```
def obtenerSubfrase(c):
    tokens = {}
    if esTerminal(c) then
        return c
    else
        foreach subConstituyente ∈ c do
            tokens ← tokens ∪ obtenerSubfrase(subConstituyente)
foreach árbol C_i do
    NNP = {} VBP = {} PPP = {} OTH = {}
    foreach constituyente c do
        switch categora(c) do
            case NNP do NNP ← NNP ∪ {obtenerSubfrase(c)}
            case VBP do NNP ← VBP ∪ {obtenerSubfrase(c)}
            case PPP do NNP ← PPP ∪ {obtenerSubfrase(c)}
            otherwise do OTH ← OTH ∪ {obtenerSubfrase(c)}
    PHR_i = {NNP, VBP, PPP, OTH}
return {PHR_1, PHR_2}
```

Tercer paso

En el tercer paso se forma el conjunto Ψ de los pares de sub-frases a comparar (eq. 6-2). Ψ está constituido por el producto cartesiano entre sub-frases del mismo tipo de la oración T_1 y de la oración T_2 (alg. 2). Por ejemplo, el sub-conjunto de las comparaciones de sub-frases nominales (N) está formado por todas las combinaciones de las *NPs* de la oración T_1 (NNP_1) con los *NPs* de la oración T_2 (NNP_2).

$$\Psi = \left\{ \begin{array}{l} N = (NNP_1 \times NNP_2) = \{(a,b) | a \in NNP_1 \wedge b \in NNP_2\} \\ V = (VBP_1 \times VBP_2) = \{(a,b) | a \in VBP_1 \wedge b \in VBP_2\} \\ P = (PPP_1 \times PPP_2) = \{(a,b) | a \in PPP_1 \wedge b \in PPP_2\} \\ O = (OTH_1 \times OTH_2) = \{(a,b) | a \in OTH_1 \wedge b \in OTH_2\} \end{array} \right\} \tag{6-2}$$

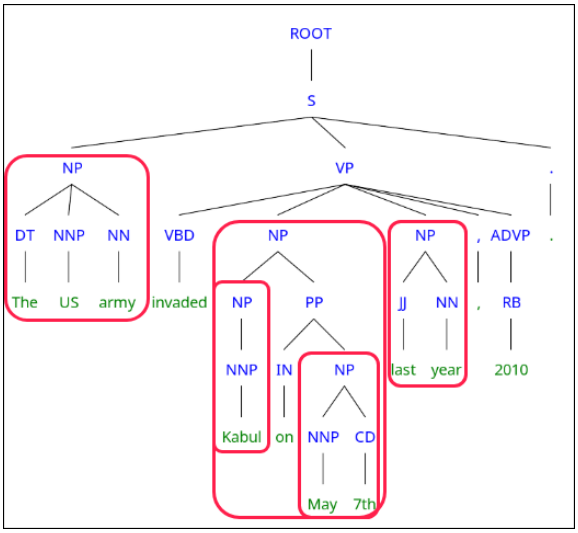

Figura 6-4: Constituyentes de tipo NP con traslape, de la oración T_2

T_1	T_2
▪ .[NNP May] [CD 2010] ▪ .[DT the] [NNS troops] ▪ .[NNP Kabul] ▪ .[NNS troops]	▪ .[DT The] [NNP US] [NN army] ▪ .[NNP Kabul] [IN on] [NNP May] [CD 7th] ▪ .[NNP Kabul] ▪ .[NNP May] [CD 7th] ▪ .[JJ last] [NN year]

Figura 6-5: Sub-frases de tipo NNP de ambas oraciones. Por simplicidad del ejemplo, sólo se muestran las frases de tipo nominal.

Algoritmo 2: Paso 3 - Formar pares de subfrases candidatas para abstracción

Input: 1 subconjunto de subfrases por cada categoría por cada oración ($\{PHR_{T1}, PHR_{T2}\}$)

Result: Pares de subfrases para abstraer (Ψ)

foreach $categ \in \{NNP, VBP, PPP, OTH\}$ **do**
  foreach $ph_1 \in PHR_{T1}[categ]$ **do**
    foreach $ph_2 \in PHR_{T2}[categ]$ **do**
      switch $categ$ **do**
        case NNP **do** $N \leftarrow N \cup \{ph_1, ph_2\}$
        case VBP **do** $V \leftarrow V \cup \{ph_1, ph_2\}$
        case PPP **do** $P \leftarrow P \cup \{ph_1, ph_2\}$
        otherwise do $O \leftarrow O \cup \{ph_1, ph_2\}$
$\Psi \leftarrow (N \cup V \cup P \cup O)$ **return** Ψ

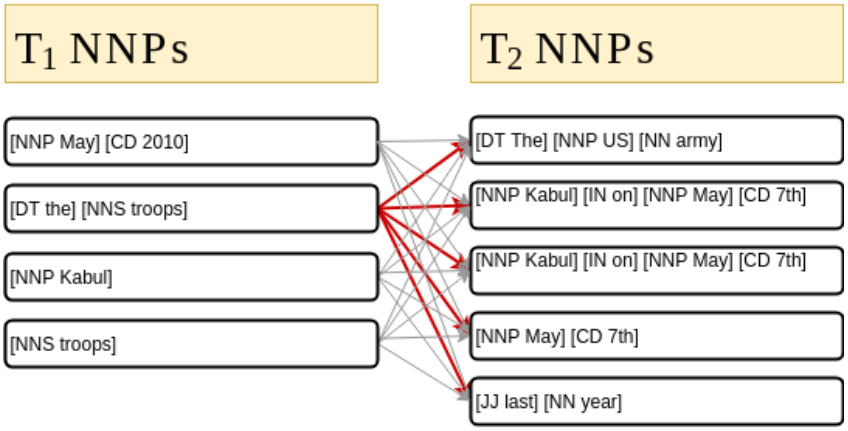

Figura 6-6: Ejemplo de comparación de sub-frases tipo NNP de T_1 con las del mismo tipo de T_2. En la figura se resalta la comparación de la segunda sub-frase de T_1 con todas las de T_2.

Cuarto paso

El cuarto paso consiste en aplicar los procesos de abstracción a cada par de sub-frases generadas en el paso anterior. Se propone una composición de abstracciones ϑ (eq. 6-3): una abstracción léxica y una abstracción sintáctica.

$$\vartheta = sabs \circ labs : (a,b) \rightarrow c \left| \begin{array}{l} labs : (a,b) \rightarrow (a',b') \\ sabs : (a',b') \rightarrow c \end{array} \right. \tag{6-3}$$

Durante la *abstracción léxica* (*labs*) (eq. 6-4), se genera una representación más abstracta común al par de unidades léxicas comparadas. Para este efecto se recurre a Wordnet.[3] El término más abstracto se obtiene buscando el hiperónimo más cercano a las unidades léxicas (t_a, t_b) que se desean abstraer (fig. 6-7). La función de abstracción léxica es sobreyectiva y probablemente no inyectiva; esto es, ambas sub-frases a y b son mapeadas a las representaciones abstractas a' y b', las cuales no necesariamente son iguales. Al sustituir por abstracciones léxicas estamos incluyendo interpretaciones compartidas y, por lo tanto, aumentando la indeterminación de la expresión. El nivel de indeterminación de un término dentro de una taxonomía tipo IS-a ha sido abordado por Resnik [191, 71], quién propone una medida del contenido de información (IC) basada en la posición del término (t) dentro de la taxonomía; su posición está implícita en la probabilidad de encontrarlo en un corpus (eq. 6-5). Sin embargo, esta medida de IC depende de un corpus externo a Wordnet en el que algunos términos pueden estar subrepresentados. Por esta razón se decidió utilizar la medida de de Seco [192] que se basa explícitamente en la posición del término para calcular su IC y no requiere de un corpus externo; asume que los conceptos más informativos son aquellos que no requieren ser diferenciados (las hojas) y el menos informativo es el nodo raíz con un valor de IC de 0. La IC de Seco se representa por la ecuación 6-6, en donde $hypo(t)$ es el número de hipónimos del término t en la ontología, y wn_{max} es una constante que representa el número máximo de conceptos que existen en la ontología. Términos más generales tienen mayor probabilidad y por lo tanto aportan menor cantidad de información; esto es, al momento de aplicarlos compiten con otras interpretaciones aumentando su grado de indeterminación. Para medir el incremento de indeterminación al abstraer dos unidades léxicas (t_a y t_b) se promedia la suma de las diferencias entre el IC de cada unidad y el de su abstracción (t_{abs}); como se muestra en la ecuación 6-7. La entropía total introducida por este proceso de

[3] Además se podría utilizar una ontología de dominio específico

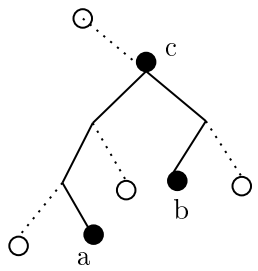

Figura 6-7: Se muestra gráficamente el hiperónimo común más cercano.

---

1 **def** `labs`(t_a, t_b)**:**
  `/* en donde` $\delta(t_a, t_b)$ `es el número de vértices en la ruta más corta entre` t_a `y` t_b `*/`
2 **return** $lcs(t_a, t_b) = \underset{c \in N}{argmin}(\delta(t_a, t_c) + \delta(t_b, t_c) + \delta(root, t_c))$

---

abstracción es la suma de las entropías de las abstracciones léxicas.

$$labs : (t_a, t_b)_{\forall (t_a, t_b) \in (a,b)} \rightarrow (t_{abs}) \left| \begin{array}{l} common\_subsumer = lso(t_a, t_b) \\ = \{\exists t \in taxonomy | parent(t_a) = t \wedge parent(t_b) = t\} \end{array} \right. \tag{6-4}$$

$$IC(t) = -log(p(t)) \tag{6-5}$$

$$IC_{wn}(t) = \frac{\frac{log(hypo(t)+1)}{max_{wn}}}{log\frac{1}{max_{wn}}} = 1 - \frac{log(hypo(t)+1)}{log(max_{wn})} \tag{6-6}$$

$$\Delta H(labs(t_a, t_b)) = \frac{(IC(t_a) - IC(t_{abs})) + (IC(t_b) - IC(t_{abs}))}{2} \tag{6-7}$$

Por su parte, la *abstracción sintáctica* (*sabs*) (eq. 6-8) mapea las sub-frases a' y b' generadas por la abstracción léxica a una sola representación abstracta c. El proceso consiste en los siguientes pasos: 1) las unidades léxicas (t) se transforman a unidades sintácticas al adjuntar-

les su categoría gramatical (POS), quedando de la forma <t_POS>. 2) se alinean las unidades sintácticas de ambas sub-frases respetando las siguientes restricciones: las cabezas se alinean primero (ej., los sustantivos en caso de NNPs); la alineación es uno-a-uno; algunas unidades pueden quedarse sin ser alineadas (está permitido saltarse unidades); la alineación es incremental de izquierda a derecha por lo que no se permite alinear unidades que ya se procesaron. 3) se abstraen las unidades alineadas para formar la representación común c al aplicar la heurística siguiente: si tanto el lema (t) como la categoría gramatical (POS) coinciden, se reduce a una sola representación de la forma <t_POS>; si coincide la categoría gramatical pero no el lema, entonces se genera una representación de la forma <*_POS>; si coincide el lema pero no la categoría gramatical, se reduce a una representación de la forma <t_*>; y finalmente, si ni el lema ni el POS coinciden, se omite el término (eq. 6-8). Al aplicar la abstracción sintáctica también se aumenta el grado de indeterminación. Su medición se basa en el trabajo de Galitsky [193, 194], en el cual, dentro del marco de una tarea de similitud semántica, se determinaron los pesos asociados al aporte semántico de una unidad léxica de acuerdo a su categoría gramatical. En este caso, el valor de la unidad es mayor mientras sea menos abstracta.

$$sabs : (t_a, t_b)_{\forall (t_a,t_b) \in (a,b)} \rightarrow (t_{abs}) = \begin{cases} t - POS & if\ lemma(t_a) == lemma(t_b) \wedge POS(t_a) == POS(t_b) \\ * - POS & if\ POS(t_a) == POS(t_b) \\ t - * & if\ lemma(t_a) == lemma(t_b) \\ descartar & else \end{cases} \tag{6-8}$$

Al aplicar el proceso de abstracción ϑ (eq. 6-3) a cada par de sub-frases de Ψ (conjunto generado en el paso 3) se obtiene el conjunto de representaciones abstractas candidatas Θ (eq.

```
def sabs(t_a, t_b):
    if lemma(t_a) == lemma(t_b) ∧ POS(t_a) == POS(t_b) then
        return lemma(t_a)_POS(t_a)
    else if POS(t_a) == POS(t_b) then
        return *_POS(t_a)
    else if lemma(t_a) == lemma(t_b) then
        return lemma(t_a)_*
    else
        return null
```

6-9).

$$\Theta = \vartheta(a,b) \forall (a,b) \in \Psi \tag{6-9}$$

El algoritmo 3 muestra el proceso descrito. La alineación que se realiza está basada en el algoritmo *Needleman-Wunsch* [195], con un costo de salto de 0.0, y un valor de coincidencia igual al valor de la abstracción sintáctica menos la entropía generada por la abstracción léxica.

Quinto paso

El quinto paso consiste en reducir el espacio de abstracciones candidatas Θ a un conjunto Φ, formado al conservar para cada sub-frase sólo la representación que preserva la mayor cantidad

Algoritmo 3: Paso 4 - Generar abstracciones candidatas

Input: Pares de subfrases para abstraer (Ψ)
Result: Pares de subfrases candidatas (Θ)

$GAP = 0.0, BAD = -1.0, MATCH = 1.0$
$DIAG = 0, LEFT = 1, UP = 2, DONE = -1$
$\Theta \leftarrow \{\}$
foreach $(F_a, F_b) \in \Psi$ **do**
    `/* Calcula la matriz de costos (Forward step) */`
    foreach $t_a \in F_a; i \leftarrow index(t_a \mid F_a)$ **do**
        foreach $t_b \in F_b; j \leftarrow index(t_a \mid F_b)$ **do**
            $t_c \leftarrow sabs(labs(t_a, t_b))$
            $mov[DIAG] \leftarrow cost[i-1][j-1] + (SAB_{weight} \mid POS(t_c), lemma(t_c) - \Delta H(labs(t_a, t_b)))$
            $mov[LEFT] \leftarrow cost[i-1][j] + GAP$
            $mov[UP] \leftarrow cost[i][j-1] + GAP$
            $cost[i][j] \leftarrow max(mov)$
            $traceback[i][j] \leftarrow index(max(mov) \mid mov)$
            $tabs[i][j] \leftarrow t_c$
    `/* Recupera la alineación óptima (Backward step) */`
    $abs \leftarrow \{\}$
    while $i > 0 \vee j > 0 \vee traceback \neq DONE$ **do**
        $abs \leftarrow abs \cup tabs[i][j]$
        switch *traceback[i][j]* **do**
            case $DIAG$ **do** $i--, j--$
            case $LEFT$ **do** $j--$
            case UP **do** $i--$
    $\Theta \leftarrow \Theta \cup [F_a, F_b, abs]$
return Θ

de la información (eq. 6-10).

$$\Phi = \left\{ \begin{array}{c} \bigcup_{i=1}^{|NNP_1|} \{\gamma | (\exists (a_i, b_j) \in N)[\gamma = max_{j=1}^{|N|}(score(\vartheta(a_i, b_j)))]\} \\ \bigcup \\ \bigcup_{j=1}^{|NNP_2|} \{\gamma | (\exists (a_i, b_j) \in N)[\gamma = max_{i=1}^{|N|}(score(\vartheta(a_i, b_j)))]\} \\ \bigcup \\ \bigcup_{i=1}^{|VBP_1|} \{\gamma | (\exists (a_i, b_j) \in V)[\gamma = max_{j=1}^{|V|}(score(\vartheta(a_i, b_j)))]\} \\ \vdots \end{array} \right\} \quad (6\text{-}10)$$

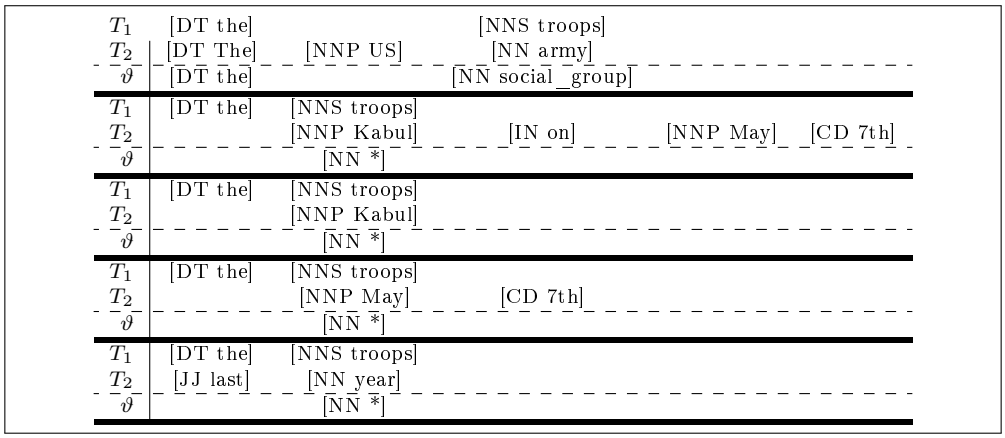

Figura 6-8: Se muestran las 5 abstracciones generadas a partir del ejemplo del paso anterior; se compara una sub-frase de tipo NNP de T_1 con todas las sub-frases de tipo NNP de T_2. Cada comparación se representa en 3 líneas, las 2 primeras son las sub-frases comparadas y la 3era (ϑ) es la abstracción generada.

Para evaluar la cantidad de información que se preserva en la representación abstracta, se asigna una calificación a cada unidad sintáctica y se resta la entropía debido a la abstracción léxica (eq. 6-11). La calificación por unidad se toma de la tabla 6-1, la cual fue calculada en [196] por medio de determinar los pesos óptimos de cada POS de manera tal que maximizaran la relevancia de resultados correctos en una tarea de búsqueda. En esta tabla se muestran los pesos asignados a una unidad léxica dependiendo de su POS, y de si en la representación abstracta se preserva sólo el POS (<POS, *>), sólo el lema (<*, t>) ó ambos. Por ejemplo, si una unidad sintáctica etiquetada como verbo se preserva con la misma categoría gramatical y las palabras originales comparten el mismo lema, entonces se asigna un valor de 0.83, pero si en la representación abstracta sólo se preservó la POS, entonces se asignará un valor de 0.2.

$$Score(C) = \sum_{i=0}^{|C|} W_{POS} * (WordGeneralization(w_i) - lso\_entropy) \tag{6-11}$$

El siguiente algoritmo (alg. 4) muestra el proceso descrito.

POS	lema	Peso
NN	t	1.0
JJ	t	0.32
RB	t	0.71
CD	t	0.64
VB	t	0.83
PRP	t	0.35
<POS>	*	0.2
*	t	0.3

Tabla 6-1: Pesos de las unidades sintácticas de acuerdo a su POS y abstracción (SAB_{weight}).

Algoritmo 4: Paso 5 - Reducir abstracciones candidatas

Input: Abstracciones candidatas (Θ)
Result: Abstracciones depuradas (Φ)

def `score(`abs`):`
  $valorAbs \leftarrow 0.0$
  foreach $tab \in abs$ **do**
    $valorAbs \leftarrow valorAbs + (SAB_{weight} \mid POS(tab), lemma(tab))$
  return $valorAbs$
$\Phi \leftarrow \{\}$
foreach $ph \in PHR_{T1}$ **do**
  $maxScore \leftarrow 0.0$, $maxAbs \leftarrow \{\}$
  for $abs \in \Theta$ **do**
    if $abs[0] == ph \wedge score(abs) > maxAbs$ **then**
      $\{maxAbs, maxScore\} \leftarrow \{abs, score(abs)\}$
  $\Phi \leftarrow \Phi \cup \{maxAbs, maxScore\}$
foreach $ph \in PHR_{T2}$ **do**
  $maxScore \leftarrow 0.0$, $maxAbs \leftarrow \{\}$
  for $abs \in \Theta$ **do**
    if $abs[1] == ph \wedge score(abs) > maxAbs$ **then**
      $\{maxAbs, maxScore\} \leftarrow \{abs, score(abs)\}$
  $\Phi \leftarrow \Phi \cup \{maxAbs, maxScore\}$
return Φ

Sexto paso

Se calcula un valor numérico asociado a la similitud semántica textual entre las oraciones T_1 y T_2. La calificación de cada abstracción perteneciente a Φ corresponde a la suma del valor

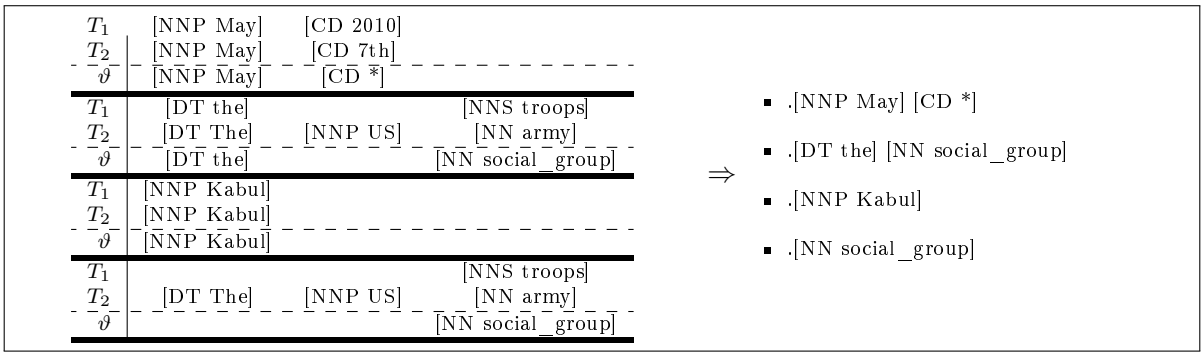

Figura 6-9: Del lado izquierdo se muestran las sub-frases de tipo NNP de T_1 y la sub-frase de T_2 con la cual se generó la mejor abstracción; del lado derecho se muestran las abstracciones resultantes de tipo NNP para T_1.

de sus unidades conservadas durante la abstracción sintáctica, menos la parte proporcional de la entropía provocada por la abstracción léxica (eq. 6-12). Finalmente, la STS es la suma de la calificación de cada abstracción multiplicada por su factor de especificidad, y normalizada por el número de abstracciones (eq. 6-13).

$$scoreSTS(\gamma) = \frac{\sum_{i=1}^{|\gamma|}(SAB_{weight} \mid tab_i) * (1 - \Delta H(tab_i))}{|\gamma|} \tag{6-12}$$

$$STS(T_1, T_2) = \frac{\sum_{i=1}^{|\Phi|} scoreSTS(\gamma) * specificity.factor(\gamma)}{|\Phi|} \tag{6-13}$$

El algoritmo 5 muestra el proceso para calcular el valor de similitud.

En el cuadro anterior sólo se mostró un subconjunto de las abstracciones producto de la comparación de ambas oraciones. El resultado considerando todos los tipos de sub-frases son las abstracciones que se muestran en el siguiente cuadro:

Algoritmo 5: Paso 6 - Calcular STS

Input: Abstracciones (Φ)
Result: Valor de STS

```
def scoreSTS(abs):
    ss ← 0.0
    foreach tab ∈ abs do
        ss ← ss + ((SAB_weight | POS(tab), lemma(tab)) * (1 − ΔH(tab)))
    return ss / |abs|
STS ← 0.0
foreach abs ∈ Φ do
    STS ← STS + (scoreSTS(abs) * specificity.factor(abs))
return STS / |Φ|
```

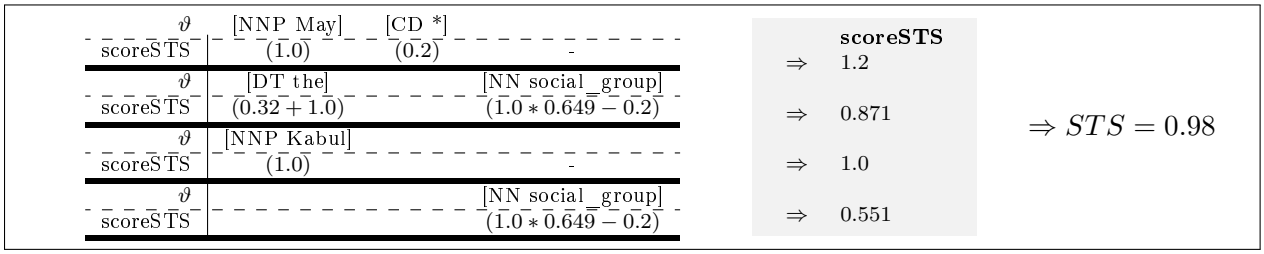

Figura 6-10: Calificación de las abstracciones de sub-frases de tipo NNP.

Ejemplo con oraciones del dominio RTM

Las siguientes son oraciones tomadas del corpus RTM y por lo tanto son ejemplos reales de la literatura de regulación transcripcional. La tabla 6-2 muestra las sub-frases que se pudieron generalizar y la abstracción común correspondiente. Con base en estas abstracciones la métrica calculó un valor de STS de 2.84 mientras que la media asignada por los expertos fue de 3.0.

A The fnr mutant , lacking the O2-responsive regulator FNR , was completely devoid of dcuC9-9lacZ expression during aerobic and anaerobic growth.

B The arcA mutant , which is deficient in the O2-responsive regulator ArcA (9) , showed only a twofold decrease in dcuC9-9lacZ expression under anaerobic conditions .

NP
==========
[NNP May] [CD *]==> (1.0 + 0.2) = 1.2
[DT the] [NN social_group]:0.649==> (0.32 + 1.0) - (1.0*0.649-0.2) = 0.871
[NNP Kabul]==> (1.0) = 1.0
[NN social_group]:0.649==> (1.0) - (1.0*0.649-0.2) = 0.551

PP
==========
[IN *] [NNP May] [CD *]==> (0.2 + 1.0 + 0.2) = 1.4

VP
==========
[VB invade] [NNP Kabul]==> (1.0 + 1.0) = 2.0

Figura 6-11: Abstracciones producidas al comparar todos los tipos de sub-frases de las oraciones T_1 y T_2. El valor de STS se calculó usando estas sub-frases.

6.1.1. Factor de especificidad en las abstracciones

La investigación de un factor de especificidad está motivada por la intuición de que la profundidad (nivel de anidamiento) de una sub-frase en el árbol de constituyentes está relacionada con su relevancia en la semántica de la oración.

Se propone usar el nivel de anidamiento de dos formas: 1) al tomar el nivel de un constituyente individual, como indicador de que tan relevante es su significado para la semántica de la oración 2) al tomar la diferencia de niveles entre dos constituyentes, como factor diferenciador de frases similares pero cuya prominencia es diferente en sus respectivas oraciones.

Inicialmente se planteo que los niveles superiores del árbol representaban contenido más relevantes en la semántica de la oración; sin embargo, en los experimentos se observó una disminución de correlación al aplicar este factor. En respuesta se experimentó dando mayor peso a frases más cercanas a los nodos terminales (más profundas en el árbol) con lo cual se obtuvo una mejora.

Empíricamente se determinó un factor de especificidad que se basa en la profundidad del constituyente a y en la diferencia de profundidad entre los constituyentes a y b (fig. 6-12). Al au-

frase A	**frase B**	**abs**	**score**
during aerobic and anaerobic growth	under anaerobic conditions	IN* JJanaerobic	0.52
dcuC99lacZ expression	dcuC99lacZ expression	IN* NNdcuc99lacz NNexpression	2.20
The fnr mutant the O2responsive regulator FNR	The arcA mutant the O2responsive regulator ArcA	DT* NN* NNmutant DT* JJo2responsive NNregulator NN*	1.40 1.72
was completely devoid of dcuC99lacZ expression during aerobic and anaerobic growth	showed only a two fold decrease in dcuC99lacZ expression under anaerobic conditions	VBD* RB* JJ* IN* NNdcuc99lacz NNexpression IN* JJanaerobic	3.32
was completely devoid of dcuC99lacZ expression during aerobic and anaerobic growth	is deficient in the O2responsive regulator ArcA (9)	VBDexist JJ* IN* NNentity NNentity	3.17

2.84 / 3.0

Tabla 6-2: Las primeras dos columnas muestran las frases de ambas oraciones y la tercera la abstracción común; en la última columna se muestra la calificación de cada abstracción, y en la parte inferior de la tabla el valor de izquierda es la calificación STS generada por la métrica y el de la derecha la asignada por los expertos.

mentar la profundidad de a este factor tiende a 1 con un coeficiente de incremento inversamente proporcional a la diferencia de profundidad entre a y b (eq. 6-14).

$$specificity.factor(a_{\in T1}, b_{\in T2} \mid \gamma) = log_{maxdepth(T_1)}(depth(a))^{\frac{depth.diff(a,b)}{maxdepth(T_1)*5}} \tag{6-14}$$

6.2. Resultados

El etiquetado POS y la generación del árbol de constituyentes se hicieron utilizando el *Stanford Core NLP.*[4] La segmentación de frases así como los métodos de abstracción se implementaron como una métrica adicional dentro del mismo paquete en donde se implementaron los algoritmos de la linea base. La nueva estrategia fue evaluada sobre el corpus de RTM.

[4] https://stanfordnlp.github.io/CoreNLP/

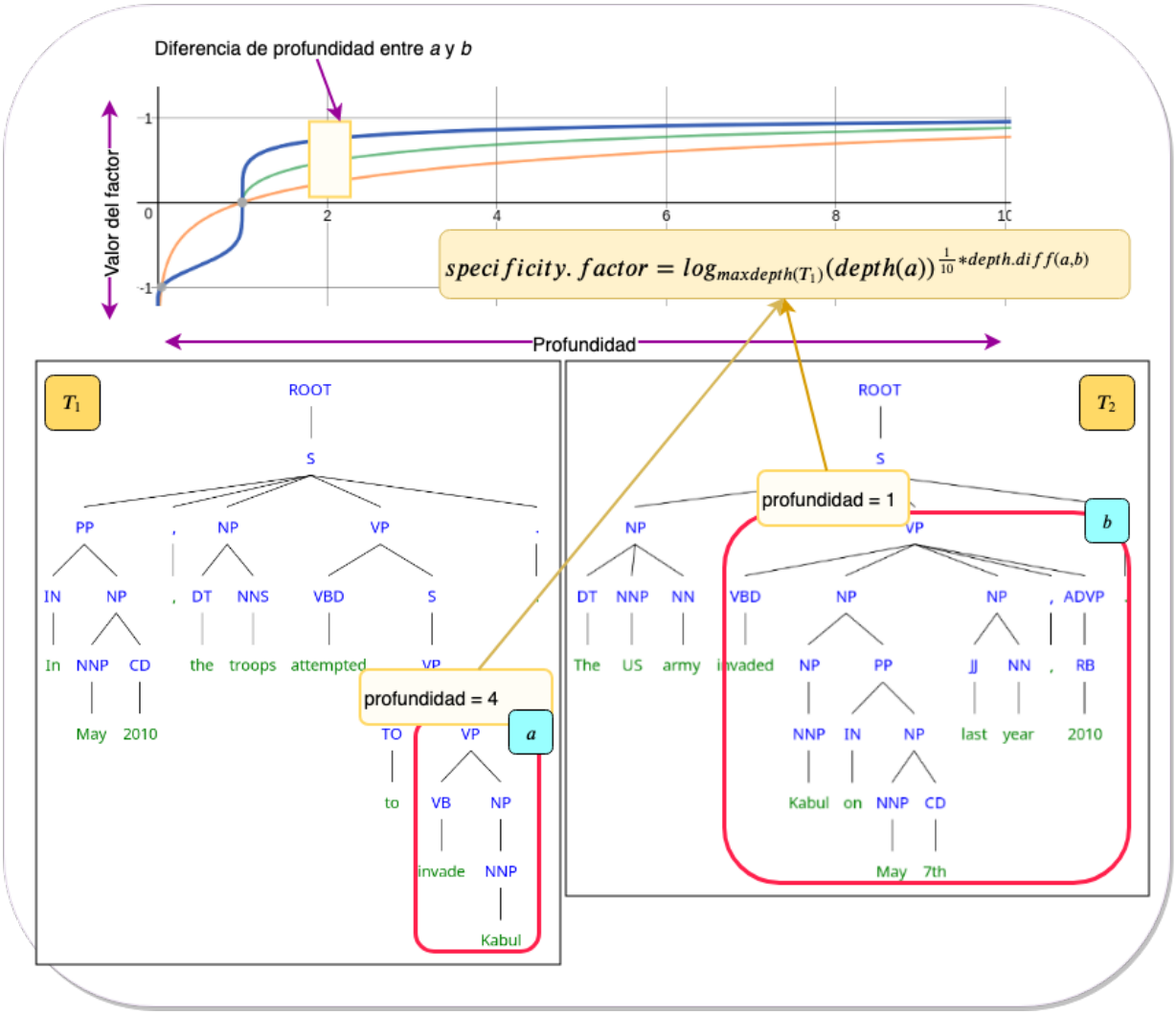

Figura 6-12: Factor de especificidad
En la gráfica se muestran 3 curvas de ejemplo del factor. Se aprecia que el valor del factor (eje Y) tiende a 1 conforme aumenta la profundidad (eje X); sin embargo, la tendencia es más lenta mientras mayor es la diferencia de profundidades entre los árboles. En las figuras de los árboles de constituyentes se muestra gráficamente la profundidad de las frases.

Los resultados de esta métrica son superiores a cada una de las métricas individuales de los experimentos de línea base descritos en el capítulo 5. Como se muestra en la tabla 6-3, la métrica propuesta es también superior a la basada en *Word Embeddings* entrenados sobre la literatura de regulación transcripcional (métrica 7). Este resultado es relevante porque muestra que la métrica basada en abstracción, que sólo usa Wordnet como fuente de conocimiento, supera a la basada en *embeddings* entrenados con los más de 6,000 artículos científicos específicos al dominio. Incluso supera a las métricas de línea base combinadas por medio de un modelo de regresión lineal entrenado en el corpus RTM (tabla 6-4).

Los resultados de los algoritmos de ensamble, *Random Forest*, *Bagging* y *Voting*, son ligeramente mejores como se muestra en la tabla 6-4. El mejor de los resultados de los experimentos de línea base *Voting* tuvo un $\rho = 0.693$ que es 2.5 % superior a la métrica de abstracción.

Además de tener resultados competitivos, la métrica de abstracción tiene la ventaja de poderse utilizar más fácilmente en dominios para los cuales se cuenta con pocos datos de entrenamiento (etiquetados y no). Hay que considerar que esta estrategia sólo usa Wordnet como recurso externo y no se realiza ningún entrenamiento sobre el corpus, lo que tienen un contraste importante con las estrategias de la línea base, en donde además de Wordnet, se utilizó Word Embeddings entrenados en literatura específica al dominio, una ontología de Regulación-Transcripcional-Microbiana y algoritmos de regresión y ensamble entrenados en el corpus de RTM.

6.3. Discusión

El análisis de constituyentes se percibe como un elemento clave porque permite tomar en cuenta la sintaxis, la cual cobra mayor importancia cuanto más complejas son las oraciones,

	Correlación	
Medida	ρ	r_s
1) Wordnet	0.26059	0.27707
2) Levenshtein	0.23564	0.26872
3) Jaccard	0.48431	0.52356
4) NGRAM	0.24115	0.25664
5) Glove (averaged)	0.39060	0.43494
6) AdHoc-Glove (averaged)	0.42398	0.54041
7) AdHoc-Glove (common-vector)	0.60406	0.63418
8) AdHoc-ontology	0.48449	0.48410
Abstracción	0.62861	0.67600

Tabla 6-3: Comparación con métricas individuales en experimentos de línea base

Algoritmo	Correlación	Desviación estándar
Linear Regression	0.622	0.15
Random Forest	0.683	0.14
Bagging	0.691	0.14
Voting	0.693	0.13
Abstracción	0.676	0.12

Tabla 6-4: Comparación con combinación de métricas en experimentos de línea base

como es el caso de las encontradas en la literatura de RTM. Se observa que las medidas que extraen el significado de la oración a partir sólo del de las palabras tienen un desempeño inferior a pesar de usar fuentes de conocimiento especializadas. Tal es el caso de las métricas 6, 7 y 8 que hacen uso de embeddings entrenados en literatura del dominio y de una ontología propia de RTM. Esto va acorde a las hipótesis de que las semántica de las oraciones está más asociada al significado de sus frases que al significado individual de sus palabras. Esto se debe a que las frases constituyen no sólo unidades sintácticas sino que representan partes coherentes del significado de las oraciones.

El tipo de alineación que se usa en las sub-frases descarta que se puedan alinear construcciones con los mismos elementos pero con orden diferente; por ejemplo, entre una expresión en voz pasiva y otra en voz activa. Esto podría verse como una limitación de la estrategia, sin embargo este tipo de variaciones sintácticas quedan cubiertas debido a que se usan tanto los constituyen-

tes superiores como todos los sub-constituyentes anidados hasta llegar a los nodos terminales. El resultado es que en caso de que no se encuentren coincidencias en sub-frases superiores se pueden encontrar en sub-frases más pequeñas. Además de la voz pasiva algunas otras variaciones que pueden generarse a partir del reposicionamiento de constituyentes, y que por ende están cubiertas por esta estrategia, son: preposición de constituyentes negativos, substitución de frase preposicional, preposición del complemento, dislocaciones izquierdas y derechas, etcétera.

El análisis de estructuras anidadas también motivó proponer un factor asociado al nivel de anidamiento al que se llamó *factor especificidad*. La evidencia sugiere que este factor es un indicio de la prominencia del constituyente para el significado de la oración. Por lo tanto es una característica útil para calificar sub-frases muy similares semánticamente pero con diferente relevancia en sus respectivas oraciones. Los experimentos sugieren que a mayor anidamiento mayor prominencia del constituyente para la semántica de la oración; es decir, la coincidencia de elementos más profundos en el árbol de dependencias es más importante. Esto insinúa que al evaluar la similitud los expertos dan mayor valor a los detalles más específicos. Es interesante que desde la perspectiva prosódica son precisamente los constituyentes más cercanos a los nodos terminales los que tienen mayor resistencia a interrupciones, lo cual se asocia a una mayor cohesión semántica de sus elementos.

El otro componente clave son los procesos de abstracción. Estos procesos se usan para generalizar las sub-frases de ambas oraciones a un conjunto de expresiones comunes. Estas abstracciones representan una simplificación de las coincidencias sintácticas y léxicas que existen entre ambas oraciones; es decir, los mecanismos de abstracción permiten descartar las características irrelevantes para la evaluación de la similitud semántica. Esto no quiere decir que la información descartada no sea importante, simplemente que no forma parte de la semántica compartida. Al aplicar los mecanismos de abstracción se produce un aumento en la indeterminación y se propu-

so que ésta se podía asociar al grado de similitud. Los resultados al evaluar la métrica validaron esta correlación y, por lo tanto, la hipótesis planteada al inicio de esta tesis.

Finalmente, otra característica relevante de esta estrategia es que además de producir una calificación de similitud se genera un conjunto de textos (abstracciones) que representan los puntos de coincidencia entre las oraciones. La interpretabilidad y explicabilidad de los resultados se facilita debido a que la califciación de STS se calcula a partir del conjunto de abstracciones comunes y que éstas están expresadas en léxico y sintaxis de lenguaje natural. Por ejemplo, para un experto en el dominio de RTM le es más fácil entender por qué la métrica asignó un STS de 2.84 al comparar las oraciones A y B si cuenta con la lista de abstracciones generadas (figura 6-2).

A The fnr mutant , lacking the O2-responsive regulator FNR , was completely devoid of dcuC9-9lacZ expression during aerobic and anaerobic growth.

B The arcA mutant , which is deficient in the O2-responsive regulator ArcA (9) , showed only a twofold decrease in dcuC9-9lacZ expression under anaerobic conditions .

Capítulo 7

Conclusión

En este capítulo se presenta una recapitulación del trabajo realizado en esta tesis así como las contribuciones generadas. Posteriormente se discuten las conclusiones y, finalmente, se exponen ideas para trabajo futuro.

7.1. Recapitulación y aportaciones

En el capítulo 1 se analizó la tarea de medición similitud semántica en el contexto de un dominio altamente especializado y con pocos datos de entrenamiento. Se motivó la búsqueda de nuevas estrategias de STS y se propuso como meta desarrollar una medida que se apoyara en procesos explícitos de abstracción. Finalmente, se plantearon los objetivos para llegar a la meta.

En el capítulo 2 se analizaron estrategias del estado-del-arte de similitud semántica; éstas se clasificaron en 2 categorías principales de acuerdo al tipo de recurso de conocimiento que usan: basadas en recursos no estructurados y basadas en recursos estructurados. Dentro del primer grupo se encuentran las estrategias que se basan en el uso de las estadísticas de concurrencia de

las palabras dentro de un corpus. En los últimos años las medidas basadas en redes neuronales son las que han tenido más auge debido a sus buenos resultados. Sin embargo, también presentan desventajas, tales como la necesidad de grandes volúmenes de datos para su entrenamiento, la dificultad para diferenciar entre relaciones sintagmáticas y paradigmáticas, y la opacidad de sus resultados. Por otro lado están las medidas que usan recursos estructurados de conocimiento tales como ontologías, taxonomías, tesauros, etcétera. Estas medidas tienen la ventaja de utilizar el conocimiento descrito por los expertos lo que permite aprovechar al máximo la información específica del dominio. El problema es la escasez de estos recursos debido a que generalmente no pueden usarse en otros dominios y construirlos es muy costoso. En respuesta han surgido estrategias híbridas que aplican varias medidas de ambos tipos y al final las combinan a través de un modelo de aprendizaje máquina entrenado sobre un corpus etiquetado. En general se observa una gran variedad de enfoques para medir la STS pero las basadas en modelos neuronales son las que han dominado en los últimos años; esto ha propiciado que el modelado del dominio y el conocimiento sintáctico se hayan relegado en favor de más datos y redes más complejas.

En el capítulo 3 se planteó el marco teórico de la tesis. Se abordaron dos aspectos fundamentales: la representación semántica y la abstracción como un mecanismo de simplificación. Primero, con fundamento en la teoría de Metalenguaje-Semántico-Natural (NSM) y en la Teoría-Texto-Significado (MTT) se propuso que no era necesario utilizar un marco simbólico externo al lenguaje natural para representar su propia semántica; es decir, se planteó representar el significado a través de las palabras y la sintaxis del lenguaje natural. Una de las principales ventajas es que los resultados de las manipulaciones son más fáciles de interpretar. En la segunda parte de este capítulo se analizó el proceso de abstracción exclusivamente como un mecanismo para ayudar en la medición de la similitud semántica entre textos. Se revisaron 14 teorías de abstracción[1] en las cuales se observaron 4 tipos de cambio de representación (ontológico, sin-

[1] Dentro del capítulo se describieron sólo aquellas teorías que se consideraron relevantes pero las demás pueden

táctico, axiomático y semántico) y 3 mecanismos para abstraer (abstracción, reformulación y aproximación). También se analizaron desde el punto de vista de la conservación o perdida de información. Se observó un falta de consenso en definir qué es abstracto y qué no los es, pero sí hay consenso en cuanto a que un proceso de abstracción es un mecanismo para simplificar las representaciones. Por esta razón se adoptó la siguiente definición de proceso de abstracción: "un mecanismo para el manejo de la complejidad de una representación permitiendo que se deriven diferentes representaciones abstractas dependiendo del uso que se les quiera dar".

En el capítulo 4 se abordó el objetivo de construir un corpus de similitud semántica específico al dominio de regulación transcripcional microbiana. Primero se diseñó el corpus considerando características como la política de muestreo, la representatividad, el tema, el tamaño, etc. También se diseñó una escala de anotación similar a la utilizada en SEMEVAL; es decir, cinco niveles (0-4) en donde 0 indica que las oraciones no comparten similitudes semánticas y 4, que son equivalentes. Para definir la escala se partió de una descripción genérica de cada nivel y después de un proceso de consenso y refinamiento se llegó a una descripción particular al dominio. A la par se expusieron ejemplos que evidenciaron la complejidad y especialización del lenguaje usado en este campo de estudio. Posteriormente se describieron las estadísticas disponibles para evaluar el nivel de interacuerdo y se expuso cuales eran las más adecuadas de acuerdo al tipo de corpus. El resultado fue un corpus de 171 pares de oraciones calificadas en una escala de similitud de 0 a 4 por un grupo de 7 expertos en el dominio. El diseño de la anotación fue no-totalmente-cruzado; es decir, no todos los anotadores calificaron a todos los pares de oraciones. A pesar de no ser muy grande y no estar balanceado, el corpus generado tiene un excelente grado de interacuerdo y, además, es el más grande en un dominio de biomédicas.

En el capítulo 5 se describen los experimentos realizados con medidas estado-del-arte en-

ser consultadas en el apéndice A.

trenadas y evaluadas en el corpus generado en el capítulo 4. Uno de los principales factores para la elección de las medidas fue la escasez de recursos para entrenamiento; se seleccionaron medidas basadas en heurísticas, o que requirieran pocos datos para entrenamiento, o que fueran no-supervisadas. Se implementó una estrategia que consistió en 7 medidas de tres tipos (textuales, basadas en recursos estructurados y basadas en recursos no-estructurados). Como medidas textuales se usaron Jaccard, Levenshtein y 3-Gramas. Como recursos estructurados se usaron WordNet y una ontología específica del dominio de RTM; y para medir la similitud entre sus términos se utilizó la métrica de Lin-98. Como recursos no-estructurados se utilizaron 2 colecciones de *embedings* de tipo GloVe, los pre-entrenados en lenguaje de dominio general y los entrenados en aproximadamente seis mil artículos del tema de regulación transcripcional, y se usó el coseno para medir la similitud entre *embeddings*. Las 7 medidas se combinaron utilizando modelos de aprendizaje máquina entrenados en el corpus de RTM. Los modelos de regresión que se probaron fueron *Random Forest*, *perceptrón* y *regresión lineal*. También se experimentó con los algoritmos de ensamble *Bagging* y *Voting*. El mejor desempeño lo tuvo *Voting* con una correlación de Pearson de 0.693. Se observó que el desempeño obtenido fue inferior al que se ha reportado para estrategias similares pero aplicadas a dominios generales.

Finalmente en el capítulo 6 se describe la propuesta de una métrica de STS con uso de abstracción (STS_{abs}). La métrica consiste en buscar abstracciones comunes a los consituyentes sintácticos de las oraciones comparadas y, posteriormente, usar la entropía generada por los procesos de abstracción para calcular la similitud semántica. El método consta de los siguientes pasos: 1) Generar el árbol de constituyentes de ambas oraciones; 2) Extraer las sub-frases de tipo NP, VP y PP; 3) Comparar las sub-frases del mismo tipo de ambas oraciones; 4) Aplicar una abstracción léxica y una sintáctica a cada par de sub-frases comparadas; 5) Para cada constituyente conservar la abstracción más específica (la menos general de las generalizaciones); y 6) Calcular la similitud entre las oraciones usando el valor de las abstracciones sintácticas

y restando la entropía inducida por las abstracciones léxicas. El método se describió a detalle incluyendo el pesudocódigo de cada paso y un ejemplo que se elaboró a la par. Esta métrica se evaluó en el corpus construido en el capítulo 4 y se obtuvieron resultados competitivos con las métricas estado-del-arte implementadas en el capítulo 5. STS_{abs} obtuvo resultados superiores a todas las medidas evaluadas individualmente e incluso a la estrategia combinada que usa un modelo de regresión lineal entrenado en el corpus. Otras estrategias de combinación como *Random Forest*, *Bagging* y *Voting* fueron ligeramente mejores siendo la última la que obtuvo el mejor desempeño, el cual fue sólo 2.5 % superior a STS_{abs}.

Con esta investigación se desarrollaron los siguientes publicaciones:

Publicaciones

1. Rinaldi, F., Lithgow-Serrano, O., López-Fuentes, A., Gama-Castro, S., Balderas-Martínez, Y. I., Solano-Lira, H., & Collado-Vides, J. (2015). An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database. Polibits, 52, 25–31. https://doi.org/10.17562/PB-52-3

2. Lithgow Serrano, O. W., Meza Ruiz, I. V., Orozco Camacho, A. M., Garcia Flores, J., & Buscaldi, D. (2016). LIPN-IIMAS at SemEval-2016 Task 1: Random Forest Regression Experiments on Align-and-Differentiate and Word Embeddings penalizing strategies. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 726–731). San Diego, California: Association for Computational Linguistics.

3. **(JCR)** Rinaldi, F., Lithgow, O., Gama-Castro, S., Solano, H., Lopez, A., Rascado, L. J. M., ... Collado-Vides, J. (2017). Strategies towards digital and semi-automated curation in RegulonDB. Database, 2017(1), 1–11. https://doi.org/10.1093/database/bax012. ** Incluye nota especificando misma contribución de los 2 primeros autores*

4. **(JCR)** Lithgow-Serrano, O. W., Gama-Castro, S., Ishida-Gutiérrez, C., Mejía-Almonte, C., Tierrafría, V., Martínez-Luna, S., ... Collado-Vides, J. (2017). Similarity corpus on microbial transcriptional regulation. Doi.org, 219014. https://doi.org/10.1101/219014

5. **(JCR)** Lithgow-serrano, O., & Colleado-vides, J. (n.d.). In the pursuit of semantic similarity for literature on microbial transcriptional regulation. Journal of Intelligent & Fuzzy Systems (JIFS), (Intelligent and Fuzzy Systems applied to Language & Knowledge Engineering).

7.2. Conclusiones

La hipótesis que se planteó en esta tesis fue que si generalizamos dos expresiones en lenguaje natural hasta una representación común, existe una correlación entre la indeterminación producida por la abstracción y la similitud semántica de ambas expresiones; se propuso que a mayor entropía menor cantidad de información compartida y por lo tanto una menor similitud semántica.

Para comprobarla se diseñó e implementó una estrategia que además de los procesos de abstracción hiciera uso de la menor cantidad de componentes relacionados a la semántica. Al ser evaluada obtuvo resultados superiores a todas las métricas estado-del-arte probadas de manera individual. Obtuvo una correlación 4 % superior a la mejor métrica basada en embeddings entrenados en seis mil artículos de regulación transcripcional (31M tokens), y 140 % superior a la otra métrica que también estaba basada en Wordnet. Debido a la magnitud de la diferencia de desempeño entre STS_{abs} y la otra métrica basada en Wordnet, y gracias a que sólo hace uso de otros dos componentes relacionados a la semántica, se propone que el éxito de la métrica se debe al uso de el análisis de constituyentes y, principalmente, al empleo de procesos de abstracción.

Por estas razones se puede concluir que la hipótesis de esta tesis es cierta: *existe una relación entre la indeterminación introducida al abstraer dos expresiones a una representación común y su similitud semántica.*

Los resultados muestran que la información sintáctica proporcionada por el análisis de constituyentes es un elemento clave en la semántica de oraciones tan complejas como las que se encontraron en la literatura de regulación transcripcional. El uso de constituyentes permite analizar el significado de la oración no a través de palabras individuales sino a través de frases que representan unidades semánticas. Esto va acorde a las hipótesis de que el significado está más asociado a las frases que se forman dentro de una oración que a sus palabras. Otra ventaja del uso de constituyentes es que permiten abarcar variaciones sintácticas tales como voz pasiva, preposición de constituyentes negativos, substitución de frase preposicional, preposición del complemento, dislocaciones izquierdas y derechas, etcétera. Por ejemplo, si dos oraciones tienen una frase verbal con el mismo significado pero en una está expresada en forma activa y en la otra en forma pasiva, los constituyentes superiores no coincidirán pero sí los sub-constituyentes nominales y el verbal.

El otro elemento clave es el uso de procesos de abstracción y el uso de la entropía para calcular una calificación de similitud. Los procesos de abstracción permitieron descartar información diferenciadora y llegar a representaciones comunes, y la entropía causada por dichas abstracciones se utilizó como indicio cuantitativo de la diferencia entre las oraciones. A mi mejor entender este es la primera propuesta de STS que usa la abstracción de manera explícita como herramienta en la medición de similitud. Además, este enfoque permite analizar el fenómeno con una perspectiva diferente abriendo la posibilidad de aprovechar las teorías existentes de abstracción para inspirar nuevas métricas.

Por ejemplo, el factor de especificidad que se introdujo en la métrica fue motivado por el

enfoque de una abstracción común y la observación de que sub-frases equivalentes de manera aislada podían diferir en cuanto a su relevancia dentro de sus respectivas oraciones. Por sí mismo esto representa un tema de estudio muy interesante y que a mi mejor entender tampoco ha sido abordado.

En resumen, la estrategia propuesta tiene las siguientes ventajas:

1. Aprovechan la información sintáctica

   a) No se basa sólo en el significado individual de las palabras, sino que considera a los constituyentes como unidades semánticas mayores.

   b) Al utilizar los constituyentes y sus anidaciones (sub-constituyentes) abarca múltiples variantes sintácticas.

   c) Hace uso del nivel de anidamiento como característica asociada a la relevancia de la sub-frase para el significado de la oración.

2. Debido a que sólo utilizar Wordnet como recurso externo se puede aplicar a dominios en donde hay escasez de recursos de entrenamiento.

3. A través de la generalización a abstracciones comunes expresadas en LN se favorece la explicabilidad de las calificaciones de STS.

7.3. Trabajo futuro

Existen muchas opciones para continuar esta línea de investigación. Los pasos siguientes más cercanos a la propuesta actual serían probar otros mecanismos de abstracción léxica y profundizar en la investigación del factor de especificidad de las sub-frases.

En relación a los mecanismos de abstracción léxica, un primer objetivo sería usar una taxonomía específica al dominio de regulación transcripcional. Actualmente se está utilizando únicamente una taxonomía de dominio general (Wordnet) por lo que términos especializados propios del dominio no están representados. Se podrían utilizar conjuntamente los dos recursos léxicos; primero se buscaría en el recurso más específico y en caso de que no se encontraran los términos, se utilizaría el recurso de dominio general (Wordnet). El segundo objetivo, sería implementar un mecanismo basado en el aprendizaje y uso de representaciones distribucionales de atributos conceptuales (no embeddings). En [197, 198], se propone, que a partir de corpora no etiquetado se aprendan representaciones distribucionales cuyas dimensiones correspondan a las características más relevantes del concepto (ej., para perro los atributos son: animal, mascota, pelo, etc.). Estas representaciones podrían ser utilizadas en conjunto con un método de abstracción aplicable a vectores tal como las memorias asociativas propuestas en [199].

Otra opción sería integrar *word-embbedings* entrenados sobre corpora propia del dominio; al tratarse de transformaciones lineales se podría usar como abstracción léxica al embedding más cercano a la diferencia entre los embeddings originales. Posteriormente calcular su entropía usando el IC de Resnik con base en la probabilidad de co-ocurrencia entre la abstracción y cada uno de los términos originales. Esto permitiría, aprovechar conocimiento específico al dominio aun cuando no se cuente con una ontología ad hoc.

Apéndice A

Otras teorías de abstracción

A.1. Teoría de Wright y Hale

Este modelo de abstracción [36, 37] se basa en el *principio de abstracción*. El cual presupone la existencia y entendimiento de una relación de equivalencia R sobre el dominio de una función f que representa la abstracción:

$$f(a) = f(b) \iff R(a, b) \tag{A-1}$$

Si dicho principio de abstracción existe, entonces existe un concepto K_f que agrupa los valores del rango de f:

$$x = instancia(K_f) \iff \exists y | x = f(y) \tag{A-2}$$

Por lo tanto, si x es una instancia de un concepto K_f que está expresado por f, y f sigue el principio de abstracción, entonces x es considerada una *entidad abstracta*. Es decir, en esta propuesta, una *entidad abstracta* se puede interpretar como una propiedad (*apreciación parcial*) común a un conjunto de objetos similares.

Un ejemplo de K_f es el concepto "dirección". La dirección de una línea a es idéntica a la dirección de una línea b, si y sólo si a es paralela a b. $\vec{a} = \vec{b} \leftrightarrow a||b$. [200]

A.2. Teoría de Hobbs

Hobbs [38] se basa en que el mundo puede ser descrito bajo una teoría de *lógica de primer orden* (τ_0) que represente los modelos de tipo *parte-de*, y el objetivo es generar a partir de τ_0 teorías más pequeñas y tratables. Si S_0 es el dominio de interpretación de τ_0, P_0 el conjunto de predicados y $R \subseteq P_0$, en donde R representa los predicados relevantes para el contexto de análisis; entonces:

$$\forall x, y \in S_0 : (x \sim y) \equiv (\forall p \in R : p(x) \equiv p(y)) \tag{A-3}$$

Define una *relación indistinguible*, por medio de la cual Hobbs busca particionar los elementos de S_0 en clases equivalentes. Es decir x y y son consideradas indistinguibles si no hay predicado relevante ($p \in R$) que las diferencie. Esto permite definir un mapeo f que reduzca τ_0 a un τ_1 más simple y menos detallado. Si para una relación dada, $f : S_0 \rightarrow S_1$ mapea los elementos de S_0 en sus clases equivalentes pertenecientes al conjunto S_1, entonces:

$$\forall x, y \in S_0 : (x \sim y) \Rightarrow (f(x) \equiv f(y)) \tag{A-4}$$

Así mismo, expresa la equivalencia de predicados por:

$$p(x) \Leftrightarrow \kappa_p(f(x)) \tag{A-5}$$

En donde $p \in P_0$, $x \in S_0$ y κ_p es el mapeo del predicado p.

Finalmente Hobbs también maneja otra forma de simplificación a la que llama *idealización*. Y que se basa en el mismo principio de *indistinguibilidad* pero con la variante de manejar un umbral de aceptación, entonces (A-4) cambia a:

$$\forall x, y \in S_0 : (x \sim y) \equiv \mid f(x) - f(y) \mid < \epsilon \tag{A-6}$$

La teoría de *granularidad* de Hobbs cae en la categoría de abstracción de dominio. En ella, teorías complejas son mapeadas hacia teorías menos detalladas, más simples y con un dominio más restringido. Por ejemplo, una realidad compleja de posiciones y tiempo continuo puede ser mapeada a un universo mucho más pequeño y simple de tiempo y posiciones discretas. De acuerdo con [35], con el propósito de simplificar algunas regularidades en el dominio objetivo, una abstracción puede introducir inferencias débiles sirviendo como proceso de idealización.

La teoría de granularidad se puede ver bajo el esquema de Giunchiglia y Walsh como un teoría de abstracción en donde Λ/Ω son invariantes. Se mapean objetos del dominio de Σ_1 en sus clases equivalentes en Σ_2.

A.3. Teoría de Tenenberg

[39] propone una abstracción de clasificación basada en el mapeo de predicados entre dos lenguajes dados $\mathfrak{L}_1$ y $\mathfrak{L}_2$ y sus conjuntos de predicados correspondientes $\mathbb{P}_1$ y $\mathbb{P}_2$. De tal manera que una función f mapea varios predicados en $\mathbb{P}_1$ al mismo predicado en $\mathbb{P}_2$.

$$f : \mathbb{P}_1 \rightarrow \mathbb{P}_2 \tag{A-7}$$

Una de las propiedades deseables de esta abstracción es a la que Tenenberg llama *solución ascendente*. Consiste en que las soluciones existentes en el espacio original de $\mathfrak{L}_1$ tienen una solución correspondiente en el espacio abstracto de $\mathfrak{L}_2$ pero no viceversa. Y lo que se busca es que el espacio abstracto de hecho tenga soluciones adicionales que permitan simplificar la búsqueda de soluciones.

Debido a que se pueden presentar inconsistencias en la teoría abstracta al colapsar dominios con propiedades contradictorias, Tenenberg introdujo un *mapeo de predicados restringido* en donde sólo aquellos predicados que conservan la consistencia respecto el espacio original son llevados al espacio abstracto. Esta restricción está basada en la intuición de que se desea que la interpretación de un predicado p_2 en el espacio abstracto sea equivalente a la unión de las interpretaciones de los predicados en $\mathbb{P}_1$ a los cuales representa. Para lograr esto, de los predicados originales se excluyen aquellos que hacen distinguibles aquellas relaciones que son colapsadas en el espacio abstracto.

A.4. Teoría de Lowry

En [40] parten de que "una buena representación *incorpora las restricciones relevantes* del problema al mismo tiempo que *oculta los detalles* superfluos", es decir, se busca que la representación haga explícitas aquellas características implícitas del problema que sean relevantes para su análisis.

El marco teórico propuesto en [40] y [149], considera tanto a las *especificaciones* como a los *algoritmos* como teorías. Por ende, la reformulación se plantea como el mapeo entre teorías. La aplicación consiste en encontrar, dada la especificación de un problema en una teoría de dominio, una especificación equivalente en un dominio más abstracto.

Se toma entidades con *equivalencia de comportamiento* como aquellas en donde a pesar de que no son necesariamente idénticas, poseen el mismo comportamiento respecto a su relación de entrada-salida (IO), y por lo tanto pueden ser substituidas unas por otras en el contexto de sus IO —no necesariamente en algún otro contexto—.

Al colapsar modelos con comportamiento equivalente en una teoría concreta, se produce la *abstracción de comportamiento*. En [40] se usan un esquema para identificar comportamiento de "entrada equivalente" (A-8) y otro para el comportamiento de "salida equivalente" (A-9).

$$In_1 \cong_{beh} In_2 \iff \forall Out : R(In_1, Out) \leftrightarrow R(In_2, Out) \tag{A-8}$$

$$Out_1 \cong_{beh} Out_2 \iff \forall In : R(In, Out_1) \leftrightarrow R(In, Out_2) \tag{A-9}$$

En [149] se proponen *kernel* y *homomorfísmo* como meta-métodos para generar teoremas de

comportamiento equivalente.

A.5. Teoría de Subramanian

Subramanian [41] expone que en la conceptualización de un dominio de análisis se requieren objetos, funciones y relaciones que representen las distinciones necesarias para describir el dominio respecto a un propósito (G), una descripción extensional del fenómeno.

Formalmente, una *conceptualización*, es una tripleta $\langle O, F, R\rangle$, en donde O representa el universo discursivo (conjunto de objetos, F), F es el conjunto de funciones $O^n \rightarrow O$, R es el conjunto de relaciones sobre O^m.

Por lo tanto, hacer un cambio en la conceptualización (reconceptualización) significa un cambio de estos elementos. Es decir, se obtiene una *reformulación* al cambiar la representación o codificación de las conceptualizaciones, lo cual se puede ver como un intercambio ontológico.

Una *codificación* ϵ de un una conceptualización C, es un conjunto de oraciones en el lenguaje $\mathfrak{L}_C$ tal que C es uno de los modelos de la interpretación de ϵ.

Debido a que una conceptualización original C_1 y una reconceptualización C_2 parten del mismo "mundo" (dominio), [41] plantean su propuesta de *reformulación* en torno a la definibilidad de C_2 a partir de C_1 y una conceptualización previa Δ (conocimiento previo).

Para que una reformulación sea "correcta", respecto a un conjunto de metas (G), se requiere que G se preserve a través del intercambio conceptual. Esto es, se considera a C_2 como una *reconceptualización correcta* de C_1 respecto a Δ y a un conjunto de relaciones objetivo G (metas), si se puede definir a G tanto en C_1 como en C_2. Se trata de una reformulación deductiva.

Por otro lado, consideran a C_2 como una *reformulación buena* de C_1 respecto a un "Problem Solver" PS, y restricciones de tiempo y espacio S, en el cálculo de las metas G en $\mathfrak{L}_{\mathfrak{C}_2}$; si existe una codificación ϵ de C_2 que permita el cálculo de g dentro de las restricciones S. La interpretación de g en C_2 es la meta de la relación G.

Desde la perspectiva ontológica, se busca economizar, transformando en tan pocos objetos y tan comprehensivos como sea posible mientras se preserve la consistencia de las restricciones.

En una conceptualización se adquiere un compromiso epistemológico al particionar el universo de una manera determinada por medio de distinciones específicas. Por ende, una conceptualización está justificada a través del análisis de los elementos conceptuales dentro de la resolución de una tarea específica.

Subramanian plantea que desde la perspectiva de la aplicación automatizada de la reformulación, el problema principal es que las interpretaciones deseadas de los términos y símbolos (distinciones) raramente están presentes en el sistema previamente modelado, y por lo tanto no existe una base lógica para realizar la reformulación.

Al enfocarse en la eficiencia computacional de la reformulación automatizada respecto a un propósito G, Subramanian identifica que se requiere una exploración guiada dentro del espacio de posibles conceptualizaciones de manera que esté justificada la introducción o eliminación de ciertos elementos conceptuales. Para esto propone que se asocie de manera directa el cambio en las conceptualizaciones con el cambio en las propiedades computacionales, bajo restricciones específicas S, en el contexto de un propósito G. De manera general, se basa en que ciertas distinciones en la formulación original, no son lógicamente necesarias para dar solución dentro de un propósito específico G, a esto le llama *explicación de irrelevancia* (A-10). Esta explicación de irrelevancia justifica porque algunas distinciones pueden ser colapsadas para obtener términos

más abstractos.

El método propuesto para generar abstracciones, a partir de minimizar distinciones irrelevantes para la tarea, se trata de una optimización local del colapsado de distinciones en el espacio de conceptualizaciones. Un proceso generativo que buscan reformular de tal manera que los estados de irrelevancia dejen de ser verdaderos en la nueva formulación y de esta manera mejoren el desempeño del sistema respecto a S mientras preservan a G. Dicha búsqueda está dirigida por el predicado de irrelevancia (A-10)

$$Irrelevant(f, g, T) \equiv (\frac{\Delta g}{\Delta f}|_T <= \epsilon) \tag{A-10}$$

Es decir, una distinción f es irrelevante para a un esquema objetivo g en el contexto de una teoría T, si al perturbar f en T, G no se ve afectado. Si en A-10 $\epsilon = 0$ se trata de una irrelevancia exacta, pero puede transformarse a una irrelevancia aproximada al asignar un valor de umbral de relevancia a ϵ.[1]

Concluyendo, los pasos propuestos para la generación de reformulaciones abstractas son:

1. Identificar estados de irrelevancia en la meta-teoría de una formulación.

2. Reducir la formulación por medio inferencias que minimicen las distinciones irrelevantes.

[1] Una irrelevancia aproximada se puede interpretar como la aplicación del concepto de irrelevancia en el marco de la teoría de *conjuntos difusos*

A.6. Teoría de Giunchiglia y Walsh

[42] define *abstracción* como el proceso de separar, extraer una representación abstracta o bosquejo de la representación real. Mapear la representación real de un problema en una representación considerada abstracta, la cual preserve ciertas características importantes para el problema a resolver. Este mapeo entre sistemas formales (eq. A-11) tiene como objetivo que la representación abstracta sea más simple de tratar que la real, en el contexto de un problema particular.

$$f : \Sigma_1 \Rightarrow \Sigma_2 \tag{A-11}$$

En donde Σ_1 representa el espacio base y Σ_2 el espacio "abstracto". f es la función de mapeo entre estos espacios la cual debe ser total porque se busca que se pueda traducir o mapear cualquier elemento del espacio base al espacio abstracto. Y además se desea que pueda ser calculada de computacionalmente.

Su propuesta de abstracción se centra en sistemas formales axiomáticos definidos como $\langle \Lambda, \Delta, \Omega \rangle$ en donde Λ es el lenguaje, Ω el conjunto de axiomas y Δ la maquinaria deductiva o reglas de inferencia.

Hay que notar que $\Omega \subseteq \Lambda$ y que en algunos casos, como en deducción natural, se puede carecer de axiomas por lo tanto Ω estará vacío. Así mismo, en sistemas no deductivos Δ estará vacío.

Si $\Sigma_1 = \langle \Lambda_1, \Delta_1, \Omega_1 \rangle$ y $\Sigma_2 = \langle \Lambda_2, \Delta_2, \Omega_2 \rangle$ son sistemas formales, decimos que $\Sigma_1 \subseteq \Sigma_2$ para indicar que $\Lambda_1 \subseteq \Lambda_2$, $\Omega_1 \subseteq \Omega_2$ y $\Delta_1 \subseteq \Delta_2$.

En cuanto a la conservación de ciertas propiedades

Si consideramos a $TH(\Sigma)$ como el conjunto de teoremas de Σ, podemos clasificar las abstracciones por su efecto en la demostrabilidad[2] de la siguiente manera:

1. Abstracción decreciente de teoremas (**TD**) si, para cualquier fbf α, $f_\Lambda(\alpha) \in TH(\Sigma_2)$ entonces, $\alpha \in TH(\Sigma_1)$. Todos los miembros de $TH(\Sigma_2)$ son producto del mapeo de sólo un subconjunto de los miembros de $TH(\Sigma_1)$.

2. Abstracción incremental de teoremas (**TI**) si sólo si, para cualquier fbf α, si $\alpha \in TH(\Sigma_1)$ entonces $f_\Lambda(\alpha) \in TH(\Sigma_2)$. Es decir, todos los miembros de $TH(\Sigma_1)$ están mapeados a un subconjunto de $TH(\Sigma_2)$. ej., Inducción.

3. Abstracción constante de teoremas (**TC**) si, para cualquier *fórmula bien formada (fbf)* α, $\alpha \in TH(\Sigma_1)$ si sólo si $f_\Lambda(\alpha) \in TH(\Sigma_2)$. Es decir, se mapean todos los miembros de $TH(\Sigma_1)$ a $TH(\Sigma_2)$ y $TH(\Sigma_2)$ no tiene otros miembros. Una TC es al mismo tiempo una TD y una TI.

En su propuesta, Giunchiglia y Walsh descartan el uso de abstracciones de tipo TC debido a que son muy fuertes y por lo tanto no desembocan en pruebas más simples en el sistema abstracto. De la misma manera, no usan abstracciones de tipo TD debido a que en ellas se pierde una propiedad importante e la completitud, hay teoremas en el espacio base que no están representados en el espacio abstracto.

De acuerdo al dominio sobre el cual se aplica la abstracción se puede clasificar en: *de proposiciones* si la teoría de abstracción es proposicional; *de predicados* en donde símbolos de predicados

[2]Es interesante observar que la probabilidad de la demostrabilidad puede ser vista como una función de creencia, i.e. $bel(A)$ es la suma de las masas que apoyan cualquier proposición B que implique A siempre y cuando no implique $\neg A$ [150]

en Σ_1 son mapeados a (posiblemente iguales) predicados en Σ_2; y *de dominio* en las que se mapean las constantes en la teoría base (dominio) a un probablemente conjunto más pequeño de constantes (dominio) en la teoría abstracta. Este última clasificación concuerda con la teoría de granularidad de Hobbs que propone también una abstracción de dominio (ver A.2).

En [42], se aclara que esta teoría busca preservar la lógica del sistema, es decir, realizar la asignación atómica hacia el espacio abstracto de sólo las FBFs y no de su estructura lógica intrínseca. Así mismo, sugiere que se debe escoger y/o construir un espacio de abstracción antes de realizar el proceso; i.e. construir un bosquejo de la abstracción (ó de la solución) y después ir refinando. Propone la construcción de jerarquías de abstracción, las cuales, de manera iterativa —1) seleccionar una abstracción, 2) generar el espacio abstracto, 3) usar o repetir — permiten ir refinando (o en cuanto a la cantidad de detalles, haciendo más granular) la representación.

A.7. Teoría de Levy y Nayak

[43] enfocan a la *abstracción* como mapeos en el modelo, es decir, al nivel de interpretación del dominio. Plantean dos etapas en el proceso, primero se abstrae un modelo del dominio, y posteriormente, un conjunto de formulas abstractas es construido para capturar el modelo.

Si $\mathfrak{L}_{base}$ es el lenguaje base y $\mathfrak{L}_{abs}$ el lenguaje abstracto entonces, las abstracciones son las funciones (π) que mapean del conjunto de interpretaciones en $\mathfrak{L}_{base}$ al conjunto de interpretaciones en $\mathfrak{L}_{abs}$:

$$\pi : Interpretaciones(\mathfrak{L}_{base}) \rightarrow Interpretaciones(\mathfrak{L}_{abs}) \tag{A-12}$$

Si existe una función auxiliar f_π que relaciona $\phi' \in \mathfrak{L}_{abs}$ a $\phi \in \mathfrak{L}_{base}$ tal que se pueda definir

$\phi' = \pi(Interpretacin(f_\pi(\phi')))$, entonces la abstracción está dada por fórmulas bien formadas, un π_P para cada $P \in \mathfrak{L}_{abs}$ y una $\pi_\forall$. En donde $\pi_p(I)$ —$I \in Interpretaciones(T_{base})$— expresa la relación de I en $Interpretaciones(T_{base})$.

Por otro lado, $\pi_\forall$ tiene una variable libre y su extensión sirve para definir el dominio abstracto. En este caso las Ps denotan la relación restringida al dominio de $\pi_\forall$. Es decir, dada una teoría en el lenguaje origen (T_{base}), T_{abs} es la imagen abstracta si y sólo si, para cada modelo I de T_{base} —$I \in Interpretaciones(\mathfrak{L}_{base})$— $\pi(I)$ es modelo de T_{abs} —$\pi(I) \in Interpretaciones(\mathfrak{L}_{abs})$—.

Levy y Nayak apuntan que este tipo de abstracción aplica a cualquier lenguaje con *semántica declarativa*. Esta semántica declarativa está dada por la *satisfacción* de las *interpretaciones* del lenguaje. En donde una interpretación I es un modelo de un conjunto de oraciones Σ si sólo si I satisface a cada oración del conjunto, $I \models \Sigma$. Si todo modelo de un conjunto de oraciones Σ_1 es modelo de un conjunto Σ_2, entonces se dice que Σ_1 implica Σ_2. Ejemplos de estos lenguajes son: lógica de primer orden, lógica propositiva, lenguajes con restricciones, etc.

A.8. Teoría de Fine

En una teoría un poco más filosófica, [44] re-analiza las teorías de Frege, principalmente relativas al *principio de contexto* y a la noción de *abstracción*, y propone un marco de principios lógicos de abstracción que soportan una teoría general. Dicha teoría está orientada sobretodo a probar la existencia y comportamiento de los abstractos[3] en general.

Su análisis comienza a partir de las propuestas de Frege que refieren que el propósito de una definición es asignar una interpretación a una expresión no interpretada, y puede consistir en la asignación de un referente, de un sentido o de ambos. En dicha asignación, la *importación*

[3] Productos de una abstracción

semántica asocia un sentido al término por definirse, mientras que la *importación referencial* tiene el objetivo de asignar un referente al mencionado término. Si la definición fue exitosa, se debe haber asignado un referente o un sentido o ambos.

Fine expresa que cualquier definición de un término implica la existencia de un dominio discursivo dentro del cual se aplicará el término. Por lo tanto el objetivo es la interpretación de un término indefinido, dentro de un dominio dado. Por ejemplo, en la función recursiva dada por $m + 0 = 0$ y $m + n' = (m + n')$, el dominio es el conjunto de los números naturales y no se requiere que la definición provea una interpretación del símbolo '$+$' sobre los reales, los racionales o cualquier otro tipo de objeto. Aplicando el mismo enfoque a la abstracción, el dominio sería el universo de todos los objetos y el operador tendría aplicación sobre los conceptos del dominio y sus objetos. Se señala que para que un principio tenga éxito determinando la interpretación del operador, no debe depender de la naturaleza de los individuos y por lo tanto debería poderse aplicar sobre otros universos en donde los individuos tengan una disposición diferente de lo que son.

En su análisis, busca que el principio no sólo determine la interpretación del operador bajo un dominio dado, sino que también pueda ayudar a determinar la composición de dichos universos. En lugar de que el principio sólo ayude a determinar una operación F para una selección apropiada de dominio M. Se plantea que el principio ayude a determinar M con base a un subdominio I de individuos (no-abstractos[4]). Es decir determinar F como función de M y también M como función de I. Así mismo, ayudar a determinar una función g de mapeo de los individuos de I hacia un dominio $M = g(I)$ de individuos y abstracciones. En general, se tendrían varios principios de abstracción, cada uno asociado a una función g. El dominio total (si existe) se obtendrá por la aplicación sucesiva de las funciones generadoras al dominio inicial

[4]Que no son producto de una abstracción

de no-abstractos hasta que no se puedan realizar más aplicaciones.

En la lógica de abstracción conceptual de Frege, resaltan principalmente dos principios. Un principio de abstracción extensional que asocia extenciones con conceptos, considerando dos extensiones iguales cuando los objetos que caen en los conceptos son los mismos, esto es, *Ley básica V* (A-13). Como se explica en [44], esta ley resulta en una inconsistencia al aplicarse la paradoja de Russell. Es por esto que Fine opta por usar en su lugar el segundo principio, *principio de Hume* (A-14). El cual al ser inmune a esta prueba le sirve como fundamento para los principios de abstracción de su propuesta.

$$\hat{x}F = \hat{x}G \leftrightarrow \forall x[F(x) \leftrightarrow G(x)] \tag{A-13}$$

De manera general presenta a un principio de abstracción como (A-14):

$$(\S F = \S G) \leftrightarrow \Phi(F, G) \tag{A-14}$$

En donde § es un operador que tiene como dominio conceptos (propiedades) y como rango objetos. Entonces, en el principio de Hume, $\S F$ es el número cardinal de F y $\Phi(F, G)$ se conserva si y sólo si hay una correspondencia uno a uno (biyectiva) entre F y G.

Fine formula que para que un principio de abstracción sea verdadero, debe existir una relación de equivalencia ($\Phi(F, G)$) entre la familia de conceptos, y además debe satisfacer la restricción de que Φ no sea *inflacionario*, es decir, que no haya mas entidades abstractas que objetos.

Al imponer la condición de que Φ sea no-inflacionaria, entonces le es posible llevar el principio de abstracción a una lógica de segundo orden. Sin embargo, al agregar un número ilimitado de

principios no-inflacionarios a una lógica de segundo orden, es posible que resulte en tener más clases equivalentes que objetos y que además distintas clases equivalentes apunten a distintas abstracciones (*hiper-inflación*). Este problema lo afronta, basándose en que una relación lógica es aquella que es invariante para todas las permutaciones del universo discursivo.

$$t \; Abstr_R \; C \tag{A-15}$$

Finalmente la operación de abstracción está simbolizada por un predicado *Abstr* que recibe un objeto t, un concepto C y una relación R (ej., Φ). Entonces, A-15 se lee como t es un abstracto de C respecto a R. Se dice que un concepto es abstraíble cuando produce un abstracto; una relación de tercer orden es invariante cuando su extensión permanece inalterada bajo cualquier permutación; y una relación de tercer orden es no-inflacionaria[5] cuando los conceptos no-identificados se pueden mapear uno-a-uno en los objetos.

Su teoría básica se basa en 3 axiomas:

1. El primero da condiciones suficientes y necesarias para que dos abstracciones sena la misma.

   Identidad:

$$(xAbstr_R C \;\&\; yAbstr_S D) \rightarrow [x = y \leftrightarrow (R(C,D) \;\&\; \forall E(R(C,E) \leftrightarrow S(C,E)))] \tag{A-16}$$

   En donde, dos abstractos asociados con, posiblemente diferentes, métodos de abstracción R y S; y diferentes conceptos C y D, serán el mismo si ambos métodos identifican a los dos conceptos y cada método identifica el mismo concepto con uno de los conceptos dados.

[5] Fine lo usa como método para identificar conceptos

2. El segundo da las condiciones suficientes para que una abstracción exista.

   Existencia:

$$Eq(R) \ \& \ Inv(R) \ \& \ NonInfl(R) \rightarrow Ap(R) \tag{A-17}$$

   Toda equivalencia no-invariante y no-inflacionaria es una forma aplicable de abstracción, resultando en abstracciones para todos los conceptos de primer orden.

3. El tercero, declara que respecto a un método específico de abstracción, ó ningún concepto es abstraíble o todos lo son.

   Aplicación:

$$Abstr_R(C) \rightarrow Abstr_R(D) \tag{A-18}$$

En la propuesta de Fine , dado un dominio I de individuos, la construcción resulta en una secuencia acumulativa de dominios; y se trata de una abstracción no-comprehensiva, en el sentido que no abstrae al mismo tiempo varias entidades (objetos, conceptos, relaciones o sus combinaciones).

A.9. Teoría de Ghidini y Giunchiglia

[47] proponen que la abstracción opera al nivel de representaciones y es aplicada antes que el sistema deductivo de una teoría formal. Por lo tanto, consideran que el modelo propuesto previamente por Giunchiglia y Walsh (ver A.6) es insuficiente al no poder incluir dicha premisa.

En esta teoría, la abstracción consiste en modelar a diferentes niveles de detalle —i.e. la representación base y la representación abstracta — un mismo fenómeno. Por lo tanto se define como una función de mapeo (f) entre dos lenguajes (conjuntos de Fórmulas Bien Formadas -

fbf), uno base $\mathfrak{L}_b$ y uno abstracto $\mathfrak{L}_a$ (A-19). En donde f sólo mapea *abstracciones atómicas* (i.e. fórmulas atómicas), por lo tanto, mantiene sin alterar la estructura lógica.

$$f : \mathfrak{L}_b \to \mathfrak{L}_a \tag{A-19}$$

Se busca que f sea *total* para que sea capaz de asignar cualquier fbf en $\mathfrak{L}_b$ a una fbf en $\mathfrak{L}_2$ —i.e. $\forall s \in \mathfrak{L}_b \exists f(s)$ —. De la misma manera se busca que sea *sobreyectiva* para que la representación abstracta sea completamente generada de la representación base —$\forall s' \in \mathfrak{L}_a \exists s \in \mathfrak{L}_b : f(s) = s'$ —. Por lo tanto $f(s) = s_0 \wedge f(s) = s_1 \to s_0 = s_1$.

Las abstracciones atómicas pueden clasificarse como *abstracciones de símbolos*, que operan colapsando símbolos (i.e constantes, funciones y predicados) (A-20) —corresponden a las abstracciones de granularidad propuesta por Hobbs (ver A.2) —; *abstracciones de espacio de parámetros*, las cuales actúan disminuyendo la dimensionalidad (A-21); y *abstracciones de verdad* que asignan símbolos de verdad en la representación abstracta a los predicados de la representación base (A-22).

$$s_1, ..., s_n \in \mathfrak{L}_b \to f(s_i) = s, \ \forall i \in [1, n] \tag{A-20}$$

$$p_1(x_1, ..., x_n) \in \mathfrak{L}_b \to p(x_1, ..., x_m) \in \mathfrak{L}_b \mid n \geq m \wedge f(p_1) = p \tag{A-21}$$

$$p(x_1, ..., x_n) \in \mathfrak{L}_b \to f(p(x_1, ..., x_n)) = T \tag{A-22}$$

La propuesta de [47] está basada en semántica de *contextos locales* o *modelos locales*, llamados así porque los razonamientos sobre el sistema (evaluaciones de verdad) son evaluados independientemente de otras perspectivas, es decir, se evalúan bajo perspectivas parciales o locales. Aunque no se descarta la posibilidad de que existan conflictos entre los estados recolectados desde diferentes perspectivas para un mismo sistema, se espera que dichas observaciones puedan coincidir en ciertos aspectos haciendo posible la reconstrucción de perspectivas "globales" del sistema analizado. De manera general se propone asociar a los lenguajes $\mathfrak{L}_b$ y $\mathfrak{L}_a$ un conjunto de interpretaciones (modelos) y usar como función de abstracción una asignación entre dichas interpretaciones, la cual formalmente la definen como *relación de compatibilidad*.

Tomando la definición común de *modelo* dentro de un lenguaje de primer orden ($\mathfrak{L}$), un *modelo local* es un par $\langle dom, \mathfrak{I} \rangle$ que corresponden al dominio y la función interpretación correspondientemente. Una *relación de dominio*, es entonces, la relación entre dominios de interpretación del modelo base y del abstracto (A-23). Las relaciones (ej., $r(a) = b$), al igual que las funciones de abstracción, se asumen como totales y sobreyectivas y representan una relación de lo que es verdad en ambos conjuntos de modelos.

$$r \subseteq dom_b \times dom_a \tag{A-23}$$

La *relación de compatibilidad*, relación existente entre los modelos base y abstracto, consiste en un conjunto no vacío de *pares compatibles*, con la restricción de no poder contener el par de conjuntos vacíos.

Dado M_b y M_a y una relación de dominio $r \subseteq dom_b \times dom_a$, un *par compatible* es definido como $c = \langle c_0, c_1 \rangle$ en donde c_i puede ser un modelo local de $m \in M_i \mid i \in [a, b]$ o el conjunto

vacío. En donde M_i contiene sólo modelos locales ($m = \langle dom_i, I \rangle$) que estén de acuerdo en la
320 interpretación de los términos, es decir que estos modelos sólo pueden diferir en la interpretación de predicados (A-24). Se impone el requisito de que si un término en el espacio base está relacionado a otro término en el espacio abstracto, entonces sus interpretaciones también deben estar relacionadas (A-25, Fig. A-1).

$$\forall\ (m_a = \langle dom_0, I_a \rangle \wedge m_b = \langle dom_0, I_b \rangle) \in M_0 \wedge \forall t \in \mathfrak{L}_0 \rightarrow t^{\mathfrak{I}_a} = t^{\mathfrak{I}_b} \tag{A-24}$$

$$f(c_1) = c \wedge f(c_2) = c \rightarrow f(c_1^{\mathfrak{I}}) = c^{\mathfrak{I}} \wedge f(c_2^{\mathfrak{I}}) = c^{\mathfrak{I}} \tag{A-25}$$

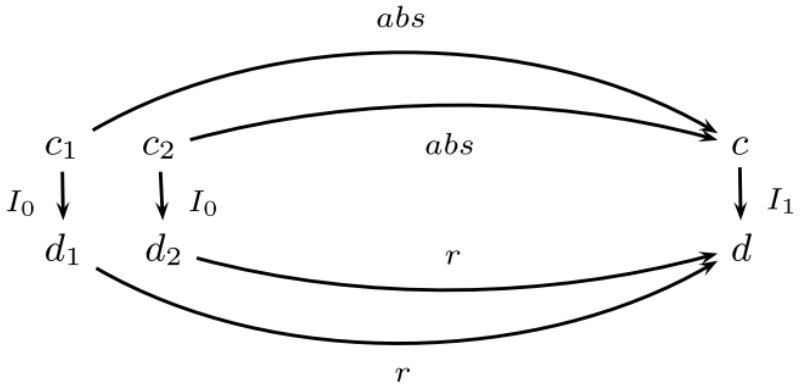

Figura A-1: Abstracción de términos y relación de dominio, tomada de [47]

A.10. Teoría de Saeger - teoría de canal

325 [35] propone un marco general para la abstracción en representación de conocimiento. Como base toman la teoría de *canal de información* [201], en la cual a su vez se propone marco matemático para analizar el flujo de información entre componentes de manera cualitativa (tipo

de información). En dicha teoría los componentes son representados por *clasificaciones*.

[35] define:

- *Clasificación* C como una tripleta $\langle typ(C), tok(C), \models\rangle$ en donde $typ(C)$ es un conjunto de tipos y $tok(C)$ es un conjunto de tokens, y $\models$ ($\models \subseteq tok(C) \times typ(C)$) es una relación de clasificación binaria.

- *Lógica local* L como una tripleta $\langle C, \vdash, N\rangle$ en donde C es una clasificación, $\vdash$ ($\vdash \subseteq typ(C) \times tok(C)$) una relación de consecuencia y N ($N \subseteq tok(C)$) el conjunto de *modelos normales* de L (N representa el conjunto de situaciones que la teoría L tiene el propósito de capturar).

- *Infomorfismo* $f : C \rightleftharpoons C'$ como un par de funciones $\langle f^\wedge, f^\vee\rangle$ tal que $f^\wedge : typ(C) \to typ(C')$, $f^\vee : tok(C') \to tok(C)$ y que satisfacen que $\forall \sigma \in typ(C), s \in tok(C') : f^\vee(s) \models_C \sigma \iff s \models_{C'} f^\wedge(\sigma)$. En donde $C = \langle typ(C), tok(C), \models_C\rangle$ y $C' = \langle typ(C'), tok(C'), \models_{C'}\rangle$ son un par de clasificaciones (Ver Fig. A-2). Los infomorfismos son la formalización de una correspondencia en la estructura de información entre dos clasificaciones diferentes y por lo tanto, permiten trasladar lógicas locales entre clasificaciones.

$$\begin{array}{ccc} typ(C) & \xrightarrow{\quad f^\wedge \quad} & typ(C') \\ \Big\downarrow \models_C & & \Big\downarrow \models_{C'} \\ tok(C) & \xleftarrow{\quad f^\vee \quad} & tok(C') \end{array}$$

Figura A-2: Infomorfismo $f : C \rightleftharpoons C'$

- *Canal* como una tupla $\langle C, \{f_i : C_i \rightleftharpoons C\}_{i\in I}\rangle$ que consiste en una familia indexada de infomorfismos f_i con un co-dominio (clasificación) C común. Es decir, un *canal* es una *clasificación* que sirve para conectar a otras clasificaciones.

Al utilizar *clasificaciones* para modelar la abstracción, en un *infomorfismo* $f^{\wedge}$ es una abstracción a nivel del lenguaje (mapeo sintáctico) y $f^{\vee}$ representa un mapeo al nivel de los modelos (nivel semántico). Pero en esta teoría, en lugar de concebir a la abstracción como un mapeo entre representaciones del sistema (clasificaciones) i.e. $f : C \to C'$, se propone trasladar, vía informorfismos ($f : C_B \rightleftharpoons C_C$ y $g : C_A \rightleftharpoons C_C$), a la teoría base C y a la teoría abstracta C' a un canal central C_C que representa a la abstracción, y que es en si misma una *lógica local* (Ver Fig. A-3). El proceso de abstracción puede ser visto como una teoría lógica de la alineación de C_A y C_B.

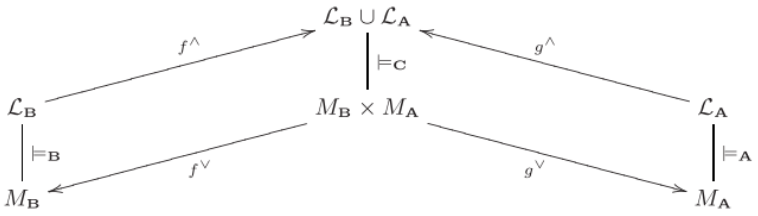

Figura A-3: Canal central binario $C_C = C_B + C_A$, en donde $\mathfrak{L}_x$ representa el conjunto de predicados (tipos —$typ(A+B)$ es la unión disjunta de $typ(A)$ y $typ(B)$—) y M_x el conjunto de modelos (tokens —$tok(A+B) = tok(A) \times tok(B)$—) en la *clasificación* x.

Al considerar a las abstracciones mismas como teorías, esta teoría permite que el mapeo pueda ser definido a nivel axiomático, sintáctico o semántico.

A.11. Teoría de Floridi

Floridi [48] retoma el enfoque de análisis de la realidad desde diferentes niveles, "nivelismo". Sobretodo apoya el uso de niveles desde el punto de vista epistemológico, en el que se observan diferentes niveles de observación o interpretación de un sistema.

En la teoría de Floridi, un *observable* refleja un enfoque particular respecto la entidad estudia-

da. Esta vista enfocada puede ser entendida como una simplificación y el método de abstracción es exitoso si se puede entender la entidad al conjuntar las simplificaciones generadas.

Es posible crear *observables* "complejos" combinando varias variables tipificas en un sólo vector. Por ejemplo, un observable "belleza" en el contexto musical puede ser la combinación de tipos como melodía, armonía, ritmo, etc. Sin embargo [48] explica que en la práctica asignar restricciones a este tipo de observables requiere una notación de proyecciones y por lo tanto opta por un enfoque de implementación más fácil, basado en colecciones o conjuntos de observables. A un conjunto finito, no vacío y conmutativo de observables es a lo que llama un *nivel de abstracción* o *LoA* por sus siglas en inglés. Un LoA puede ser discreto si y sólo si, todos sus observables son discretos, análogo si y sólo si todos sus observables son continuos o, de otro modo son considerados LoAs híbridos. Ligando el ejemplo anterior respecto a "evaluar" una pieza musical, en lugar del observable complejo de belleza, se tendría un LoA de belleza que incluiría los mismos observables, pero si además se quisiera evaluar la misma pieza musical desde un enfoque de popularidad se tendría otro LoA que incluyera observables como: compositor|cantante|dj, idioma, género musical, etc. Así se ejemplifica la capacidad de esta teoría de tener diferentes enfoques respecto al mismo sistema estudiado.

En el estudio de un sistema, además de la identificación de observables, se plantea la necesidad de especificar la relación que existe entre los observables de un LoA. Para esto utiliza predicados cuyas variables libres son los observables del LoA y aquellos valores de los observables que hacen verdadero al predicado son a los que llama *comportamientos del sistema*. Esto es, se describen las combinaciones de valores observables que son aceptables en el sistema. A un LoA con comportamientos definidos le llama *LoA moderado*.

En el análisis de sistemas discretos, el uso de predicados es indispensable debido a que no existe una relación continua entre el cambio de valor en un observable y el cambio de com-

portamiento del sistema. Sin embargo debido a la complejidad y tamaño de estos sistemas, frecuentemente es necesario hacer aproximaciones de dichos comportamientos. Uno de los aportes mencionados [48] del *método de abstracción*, es dar un formalismo a los sistemas discretos al permitir que los comportamientos aproximados puedan ser descritos de manera exacta en un LoA dado y sean los diferentes LoAs los que cambien en su grado de comprehensión de los comportamientos detallados del sistema, resultando en un LoA que incluya los comportamientos deseados.

Como los LoAs se pueden entender como diferentes representaciones o vistas de un sistema, Floridi introdujo el concepto de *Gradiente de Abstracciones* (GoA) como indicador que permitiera variar los LoA para hacer observaciones a diferentes niveles de abstracción. Formalmente consiste en un conjunto finito de LoAs moderados $\{L_i | 0 \geq i < n\}$ y una familia de relaciones $R_{i,j} \subseteq L_i \times L_j$ en donde $0 \leq i \neq j < n$, que asocia pares de observables de diferentes LoAs de manera que:

- la relación $R_{i,j}$ es el inverso de la relación $R_{j,i}$ para $i \neq j$
- el comportamiento p_j en L_j es al menos tan fuerte como el comportamiento traducido.

Dos tipos de GoAs son principalmente útiles para estudiar un sistema, los *GoAs disjuntos*, que son aquellos que describen un sistema como la combinación de varios componentes que no se sobreponen, i.e. si y sólo si tomando L_i en pares no tienen observables en común y las relaciones están vacías. Por otro lado los *GoAs anidados* son considerados así, si y sólo si las únicas relaciones no vacías son entre L_i y L_{i+1} para $0 \leq i < n-1$ y el inverso de $R_{i,i+1}$ es una función sobreyectiva de L_{i+1} a L_i. Intuitivamente quiere decir que en un GoA anidado los LoAs exhiben diferentes grados de granularidad sobre los mismos observables. Y cada observación abstracta tiene al menos una contraparte concreta. Se trata de un proceso tipo "top-down" en

el que se incrementa el detallado del modelo.

Un ejemplo concreto puede ser el estudio del sistema "semáforo" en donde un observable es el color de la luz del semáforo. LoA L_0 puede ser el que comprenda el color dentro del conjunto $\{rojo, amarillo, verde\}$, mientras que un L_1 del mismo GoA puede ser uno con una variable libre wl de tipo número real positivo correspondiente a la longitud de onda del color. Este segundo LoA sería moderado por un comportamiento dado por el predicado $((\lambda_{rojo} \leq wl \leq \lambda_{rojo'}) \vee (\lambda_{amarillo} \leq wl \leq \lambda_{amarillo'}) \vee (\lambda_{verde} \leq wl \leq \lambda_{verde'})$. Es decir, el tipo más abstracto L_0 sería una proyección de las bandas de color consideradas como ese color en el LoA más concreto L_1.

Al uso de los LoAs / GoAs para el análisis de un sistema se le denomina *método de abstracción*, y [48] plantea que su utilización acarrea ventajas como:

- permitir entender el significado de "conocimiento indirecto" en términos de conocimiento mediado por un LoA;
- al especificar un LoA se clarifica el entendimiento del sistema, se identifica las preguntas que sean relevantes;
- la entrada de un LoA es el sistema analizado (un conjunto de datos) mientras que la salida es un *modelo* del sistema (información);
- saber desde que LoA se está analizando el sistema, significa saber el alcance y los límites del modelo generado;
- al asociarse a un LoA, una teoría requiere tener los siguientes componentes i) un LoA que determina el rango de observables que se usarán, ii) el modelo como salida del sistema estudiado, iii) una estructura del sistema en una LoA dado. En consecuencia, la teoría es forzada a hacer explícito y clarificar su compromiso ontológico.

A.12. Teoría de Nivel de Interpretación

En el ámbito de la psicología, [49] se basa en la idea de que la misma entidad se puede representar mentalmente en diferentes niveles de *abstracción* y organizarse en un eje continuo de menor a mayor. Por un lado, las representaciones más abstractas comprehenden la idea central de la entidad pero no los detalles más específicos o periféricos. Por otro lado, las representaciones más concretas incluyen detalles específicos del contexto. A partir de esto, define *abstracción* como el proceso de representación mental que se enfoca en características "esenciales" de un objeto al mismo tiempo que omite detalles menos relevantes. Ésta provee un mecanismo por medio del cual los humanos pueden manejarse más allá de las experiencias.

La teoría se apoya el hecho de que la distancia psicológica —temporal, espacial, social, hipotética — influencia de manera sistemática la forma en que los humanos representan el mundo a su alrededor, es decir, el nivel de abstracción en la representación. Una observación importante es que al cambiar la distancia psicológica, y por lo tanto el nivel de representación del evento u objeto, se acarrea un impacto en la evaluación, análisis e inferencias derivadas.

También se afirma que representaciones más abstractas tienden a tener mayor estabilidad del significado que representan a través de diferentes contextos; incluyendo tiempo, área de conocimiento, cultura, etc.

Bibliografía

[1] J. Hutchins, "Retrospect and prospect in computer-based translation," no. September, 1999.

[2] A. M. TURING, "I.—COMPUTING MACHINERY AND INTELLIGENCE," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[3] L. A. Pineda, H. Calvo, L. Villaseñor, N. A. Castro, A. Gelbukh, Y. Hernández, H. Jiménez, M. Montes, D. Pinto, F. Sánchez, and G. Sidorov, *Lingüística Computacional*, ch. Lingüístic, pp. 91–125. Academia Mexicana de Computación (AMEXCOMP), 2017.

[4] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Similarity from Natural Language and Ontology Analysis," 2017.

[5] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," *Computacion y Sistemas*, vol. 20, no. 4, pp. 647–665, 2016.

[6] P. Turney, "Similarity of semantic relations," *Computational Linguistics*, no. February, 2006.

[7] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "*SEM 2013 shared task : Semantic Textual Similarity," *The Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, vol. 1, pp. 32–43, 2013.

[8] A. Molina, J. M. Torres-Moreno, E. SanJuan, G. Sierra, and J. Rojas-Mora, "Analysis and transformation of textual energy distribution," *Proceedings - 2013 12th Mexican International Conference on Artificial Intelligence, MICAI 2013*, no. November, pp. 203–208, 2013.

[9] S. Fernandez, E. SanJuan, and J. M. Torres-Moreno, "Textual energy of associative memories: Performant applications of Enertex algorithm in text summarization and topic segmentation," *Micai 2007 Advances in Artificial Intelligence*, vol. 4827, pp. 861–871, 2007.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," no. January 2013, 2013.

[11] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[12] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," in *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, jan 2003.

[13] Y. Goldberg, "Primer NN NLP," pp. 1–76, 2015.

[14] A. Neuraz, L. C. Llanos, A. Burgun, and S. Rosset, "Natural language understanding for task oriented dialog in the biomedical domain in a low resources context," 2018.

[15] O. Lithgow-Serrano, S. Gama-Castro, C. Ishida-Gutiérrez, C. Mejía-Almonte, V. H. Tierrafría, S. Martínez-Luna, A. Santos-Zavaleta, D. Velázquez-Ramírez, and J. Collado-

Vides, "Similarity corpus on microbial transcriptional regulation," *Journal of Biomedical Semantics*, vol. 10, p. 8, dec 2019.

[16] F. Nooralahzadeh, L. Øvrelid, and J. T. Lønning, "Evaluation of domain-specific word embeddings using knowledge resources," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 1438–1445, 2019.

[17] X. Zhu, *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005.

[18] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, pp. 1–114, jun 2012.

[19] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, oct 2010.

[20] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, *A survey of transfer learning*, vol. 3. Springer International Publishing, 2016.

[21] K.-g. Zhang, Y.-d. Zhang, and M. Wang, "A Unified Approach to Interpreting Model Predictions Scott," vol. 16, no. 3, pp. 426–430, 2012.

[22] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," no. Ml, pp. 1–13, 2017.

[23] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," pp. 1–11, 2019.

[24] H. Liu, Q. Yin, and W. Y. Wang, "Towards Explainable NLP: A Generative Explanation Framework for Text Classification," 2018.

[25] K. Allan, "Linguistic meaning," *Linguistic Meaning*, pp. 1–348, 2014.

[26] Allan Keith, *Linguistic meaning.* Routledge & Kegan Paul London ; New York, 1986.

[27] N. Riemer, "A history of semantics," in *Routledge Handbook of Semantics*, pp. 1–27, 2013.

[28] C. Goddard, "Natural Semantic Metalanguage : The state of the art," No. December, 2008.

[29] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing, Second Edition*, vol. 2. 2010.

[30] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[31] J. Hampton, "Concepts as prototypes," *Psychology of Learning and Motivation*, no. 2000, pp. 1–19, 2006.

[32] E. Rosch, "Natural categories," *Cognitive psychology*, vol. 350, 1973.

[33] R. Seising, "Views on Fuzzy Sets and Systems from Different Perspectives: Philosophy and Logic, Criticisms and Applications," apr 2009.

[34] J. Y. Son, L. B. Smith, and R. L. Goldstone, "Simplicity and generalization: Short-cutting abstraction in children's object categorizations," *Cognition*, vol. 108, no. 3, pp. 626–638, 2008.

[35] S. D. Saeger and A. Shimojima, "Channeling Abstraction," in *Abstraction, Reformulation, and Approximation*, pp. 124–138, 2007.

[36] C. Wright, *Frege's conception of numbers as objects.* Aberdeen University Press, 1983.

[37] B. Hale, *Abstract Objects.* Oxford, UK: Basil Blackwell, 1987.

[38] J. R. Hobbs, "Granularity," in *In Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 432–435, 1985.

[39] J. D. Tenenberg, "Preserving {Consistency} across {Abstraction} {Mappings}," *Proceedings of the 10th IJCAI*, vol. 14627, no. 3, pp. 1011–1014, 1987.

[40] M. Lowry, "The abstraction/implementation model of problem reformulation," *Proc. IJCAI*, pp. 1004–1010, 1987.

[41] D. Subramanian, "A theory of justified reformulations," *Change of representation and inductive bias*, pp. 147–167, 1990.

[42] F. Giunchiglia and T. Walsh, "A theory of abstraction," *Artificial Intelligence*, vol. 57, no. 2-3, pp. 323–389, 1992.

[43] P. P. Nayak and A. Y. Levy, "A semantic theory of abstractions," *14th International Joint Conference on Artificial Intelligence IJCAI-95*, pp. 196–203, 1995.

[44] K. Fine, *The limits of abstraction*. Clarendon Press, 2002.

[45] V. V. Cross, "Defining fuzzy relationships in object models: Abstraction and interpretation," *Fuzzy Sets and Systems*, vol. 140, no. 1, pp. 5–27, 2003.

[46] J. McCarthy, "Approximate Objects and Approximate Theories," *{KR}2000: Principles of Knowledge Representation and Reasoning*, pp. 519–526, 2000.

[47] C. Ghidini and F. Giunchiglia, "A Semantics for Abstraction," *Proc. 16th European Conference on Artificial Intelligence (ECAI)*, pp. 343–347, 2004.

[48] L. Floridi, "The method of levels of abstraction," *Minds and Machines*, vol. 18, no. 3, pp. 303–329, 2008.

[49] C. K. Soderberg, S. P. Callahan, A. O. Kochersberger, E. Amit, and A. Ledgerwood, "The effects of psychological distance on abstraction: Two meta-analyses.," *Psychological Bulletin*, vol. 141, no. 3, pp. 525–548, 2015.

[50] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. Del Moral-Chavez, F. Rinaldi, and J. Collado-Vides, "RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Research*, vol. 44, no. D1, pp. D133–D143, 2016.

[51] F. Rinaldi, O. Lithgow-Serrano, A. López-Fuentes, S. Gama-Castro, Y. I. Balderas-Martínez, H. Solano-Lira, and J. Collado-Vides, "An Approach towards Semi-automated Biomedical Literature Curation and Enrichment for a Major Biological Database," *Polibits*, vol. 52, pp. 25–31, jul 2015.

[52] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983.

[53] U. Hahn and T. M. Bailey, "What makes words sound similar?," *Cognition*, vol. 97, no. 3, pp. 227–267, 2005.

[54] R. Goldstone and J. Y. Son, "Similarity," in *The Cambridge handbook of thinking and reasoning*, no. 812, pp. 13–36, 2005.

[55] R. N. Shepard, "Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space," *Psychometrika*, vol. 22, no. 4, pp. 325–345, 1957.

[56] J. Carroll and M. Wish, "Models and methods for three-way multidimensional scaling," *Contemporary developments in mathematical psychology*, vol. 2, no. Measurement, psychophysics, and neural information processing, pp. 283–319, 1974.

[57] E. Rosch, "Principles of Categorization," *Cognition and Categorization*, pp. 27–48, 1978.

[58] R. M. Nosofsky, "Attention, Similarity, and the Identification-Categorization Relationship," *Journal of Experimental Psychology: General*, vol. 115, no. 1, pp. 39–57, 1986.

[59] D. L. Medin and M. M. Schaffer, "Context theory of classification learning," *Psychological Review*, vol. 85, no. 3, pp. 207–238, 1978.

[60] R. M. Nosofsky, "The generalized context model: an exemplar model of classification," *Formal Approaches in Categorization*, pp. 18–39, 2011.

[61] J. R. Firth, "A synopsis of linguistic theory 1930-55.," *Studies in Linguistic Analysis (special volume of the Philological Society)*, vol. 1952-59, pp. 1–32, 1957.

[62] M. Zarechnak and E. Coyne, "Semantic analysis of natural language statements," *Linguistics*, vol. 14, no. 182, pp. 73–81, 1976.

[63] M. Batet and D. Sánchez, "A Review on Semantic Similarity," *Encyclopedia of Information Science and Technology, Third Edition*, no. May 2016, pp. 7575–7583, 2014.

[64] J. O'Shea, Z. Bandar, and K. Crockett, "A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithms," *ACM Transactions on Speech and Language Processing*, vol. 10, no. 4, p. Article No. 19, 2013.

[65] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice Hall Press, 3rd ed., 2009.

[66] S. Janaqi, H. Sebastien, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Measures for the Comparison of Units of Language , Concepts or Instances from Text and Knowledge Representation Analysis A Comprehensive Survey and a Technical Introduction to Knowledge-based Measures Using Semantic Graph Analysis book for a more," vol. 1, no. 1, 2016.

[67] S. Ordoñez-Salinas and A. Gelbukh, "Information retrieval with a simplified conceptual graph-like representation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6437 LNAI, no. PART 1, pp. 92–104, 2010.

[68] Y. Kiyoki, *Information modelling and knowledge bases XVII*. No. v. 136, 2006.

[69] C. St-Jacques and C. Barrière, "Similarity judments: philosophical, psychological and mathematical investigations," *Coling-ACL Workshop on Linguistic Distances*, no. July, pp. 8–15, 2006.

[70] P. Resnik, "Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," vol. 11, pp. 95–130, 1999.

[71] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," vol. 1, p. 6, 1995.

[72] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity," *Demonstration Papers at HLT-NAACL 2004 on XX - HLT-NAACL '04*, no. July, pp. 38–41, 2004.

[73] T. Pedersen, "Information content measures of semantic similarity perform better without sense-tagged text," *Human Language Technologies: The 2010 Annual . . .*, no. June, pp. 329–332, 2010.

[74] G. A. Miller, D. Gross, and K. J. Miller, "Introduction to WordNet : An On-line Lexical Database *," vol. 3, no. 4, 1990.

[75] C. Fellbaum, "WordNet: an electronic lexical database.," *WordNet is available from http://www. cogsci. princeton. . . .*, 1998.

[76] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *American Journal of Physiology - Lung Cellular and Molecular Physiology*, vol. 274, pp. 133–138, 1994.

[77] P. Resnik, "WordNet and class-based probabilities," in *WordNet: An electronic lexical database*, pp. 239 – 263, 1998.

[78] D. Lin, "Automatic retrieval and clustering of similar words," *ACL '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 768–774, 1998.

[79] S. Patwardhan and T. Pedersen, "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts," *In: 11th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 1501, no. June 2014, pp. 1–8, 2006.

[80] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986*, pp. 24–26, 1986.

[81] G. Pirró and J. Euzenat, "A feature and information theoretic framework for semantic similarity and relatedness," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6496 LNCS, no. PART 1, pp. 615–630, 2010.

[82] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *Proceedings of the 21st national conference on Artificial intelligence*, vol. 1, pp. 775–780, 2006.

[83] F. Spitzer, "The Classification of Random Walk," in *Principles of Random Walk*, pp. 1–53, 1964.

[84] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[85] M. Gan, X. Dou, and R. Jiang, "From ontology to semantic similarity: calculation of ontology-based semantic similarity.," *TheScientificWorldJournal*, vol. 2013, p. 793091, 2013.

[86] S. A. Elavarasi, J. Akilandeswari, and K. Menaga, "A Survey on Semantic Similarity Measure," *International Journal of Research in Advent Technology*, vol. 2, no. 3, pp. 389–398, 2014.

[87] H. RUBENSTEIN and J. B. GOODENOUG, "Contextual Correlates of Synonymy," *Communications of the ACM*, vol. 8, no. 10, 1965.

[88] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[89] M. Sahlgren, *The Word-Space Model*. PhD thesis, Stockholm University, 2006.

[90] S. Clark, "Vector Space Models of Lexical Meaning," no. September, pp. 1–42, 2012.

[91] G. Salton, "Automatic text processing: the transformation, analysis, and retrieval of information by computer," jan 1989.

[92] S. Deerwester, S. Dumais, G. Furnas, and T. Ladauer, "Indexing by latent semantic analysis," *Journal of the American ...*, 1990.

[93] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods, Instruments, & Computers*, vol. 28, pp. 203–208, jun 1996.

[94] T. Hofmann, "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, pp. 50–57, 1999.

[95] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.

[96] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, "Distributed representations," *Parallel Distributed Processing*, pp. 77–109, 1986.

[97] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning." 2016.

[98] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Nips*, pp. 1–9, 2013.

[99] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2177–2185, 2014.

[100] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *Proceedings of the 21st national conference on Artificial intelligence*, vol. 1, pp. 775–780, 2006.

[101] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Behavior research methods*, vol. 41, no. 3, pp. 944–950, 1997.

[102] J. Mitchell and M. Lapata, "Composition in distributional models of semantics.," *Cognitive science*, vol. 34, no. 8, pp. 1388–1429, 2010.

[103] M. Baroni, R. Bernardi, and R. Zamparelli, "Frege in Space: A Program of Compositional Distributional Semantics," *Linguistic Issues in Language Technology*, vol. 9, no. 6, pp. 242–346, 2014.

[104] E. Grefenstette and M. Sadrzadeh, "Experimental support for a categorical compositional distributional model of meaning," *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1394–1404, 2011.

[105] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic Compositionality through Recursive Matrix-Vector Spaces," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, no. Mv, pp. 1201–1211, 2012.

[106] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *Proceedings of ACL*, pp. 1556–1566, 2015.

[107] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *International Conference on Machine Learning - ICML 2014*, vol. 32, pp. 1188–1196, 2014.

[108] T. M. V. Janssen, "Montague Semantics,"

[109] H. Schütze, "Automatic word sense discrimination," *Computational Linguistics*, vol. 24, pp. 97–123, mar 1998.

[110] S. Clark, B. Coecke, and M. Sadrzadeh, "A compositional distributional model of meaning," *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, no. Schuetze 1998, pp. 133–140, 2008.

[111] R. Muskens and M. Sadrzadeh, "Lambdas and Vectors," no. 1974, pp. 1–3, 2010.

[112] P. D. Turney, "The latent relation mapping engine: Algorithm and experiments," *Journal of Artificial Intelligence Research*, vol. 33, pp. 615–655, 2008.

[113] M. Sadrzadeh and R. Muskens, "Static and Dynamic Vector Semantics for Lambda Calculus Models of Natural Language," 2018.

[114] F. J. Pelletier, "Did Frege Believe Frege ' s Principle ?," *Journal of Logic, Language and Information*, vol. 10, pp. 87–114, 2001.

[115] H. Rott, "Words in contexts: fregean elucidations 1. i," *Linguistics and Philosophy*, pp. 621–641, 2000.

[116] F. Giunchiglia and P. Bouquet, "Introduction to contextual reasoning: An Artificial Intelligence perspective," *Perspectives on Cognitive Science*, vol. 3, no. section 4, pp. 138–159, 1997.

[117] G. Sundholm, "Tarski and Lesniewski on Languages with Meaning versus Languages without Use," *Philosophy and Logic. In Search of the Polish Tradition - Essays in Honour of Jan Wolenski on the Occasion of his 60th Birthday*, no. 1930, pp. 109–128, 2004.

[118] A. Copestake and A. Herbelot, "Lexicalised compositionality," pp. 1–34.

[119] L. W. Barsalou, "Perceptual Symbol Systems," vol. 30322, no. X, 1998.

[120] S. Laurence and E. Margolis, "Concepts and cognitive science," *Concepts: core readings*, 1999.

[121] G. Frege, *The Foundations of Arithmetic.* New York, NY, USA: Harper & Brothers, 2nd ed., 1953.

[122] M. Dummett, *Frege: Philosophy of Language*, vol. 49. Duckworth, 1973.

[123] M. Poesio, "Domain modelling and NLP: Formal ontologies? Lexica? Or a bit of both?," *Applied Ontology*, vol. 1, pp. 27–33, jan 2005.

[124] A. Wierzbicka, *Semantic Primitives.* (Frankfurt/M.)Athenäum-Verl, 1972.

[125] C. Goddard, "The universal syntax of semantic primitives," *Language Sciences*, vol. 19, no. 3, pp. 197–207, 1997.

[126] C. Goddard, "THE SEMANTICS OF COMING AND GOING 1 Cliff Goddard Introduction It is often assumed that the English motion verbs," *Time*, pp. 147–162, 1983.

[127] C. Goddard, "Semantic molecules," *Annual Meeting of the Australian Linguistic . . .*, 2007.

[128] I. Mel'čuk, "Anna Wierzbicka , Semantic Decomposition , and the Meaning-Text Approach," *Russian Journal of Linguistics*, vol. 22, no. 3, pp. 521–538, 2018.

[129] A. Žolkovskij and I. Mel'čuk, "On a Possible Methodology and Tools for Semantic Synthesis (of texts)," *Naučno-texničeskaja informacija*, vol. 5, pp. 23–28, 1965.

[130] N. Chomsky, *Syntactic structures.* MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 1957.

[131] S. Kahane, "THE MEANING-TEXT THEORY," 2001.

[132] G. Boleda and K. Erk, "Distributional semantic features as semantic primitives - Or not," *AAAI Spring Symposium - Technical Report*, vol. SS-15-03, pp. 2–5, 2015.

[133] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[134] P. Gärdenfors, "LANGUAGE AND THE EVOLUTION OF COGNITION," 1995.

[135] N. Chomsky, *Aspects of the theory of syntax.* Oxford, England: M.I.T. Press, 1965.

[136] J. F. Sowa, *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Pacific Grove, CA, USA: Brooks/Cole Publishing Co., 2000.

[137] R. Stern, *Hegel and the Phenomenology of Spirit.* 2002.

[138] J. Locke, *The Works of John Locke in Nine Volumes*, vol. 1. 1824.

[139] E. Husserl, "Philosophy of Arithmetic," p. 580, 1891.

[140] B. Hale and C. Wright, "The Metaontology of Abstraction," *Metametaphysics: new essays on the foundations of ontology*, no. 1983, pp. 178–212, 2009.

[141] L. Saitta and J. D. Zucker, *Abstraction in artificial intelligence and complex systems*, vol. 9781461470. 2013.

[142] K. Barsalou, L.W. and Wiemer-Hastings, "Situating Abstract Concepts," in *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, pp. 129–163, Cambridge: Cambridge University Press, 2005.

[143] D. Pecher, "The Perceptual Representation of Mental Categories," in *The Oxford Handbook of Cognitive Psychology*, pp. 1–19, 2013.

[144] L. Floridi, "The method of levels of abstraction," *Minds and Machines*, vol. 18, no. 3, 2008.

[145] P. Godfrey-smith, "Abstractions , Idealizations , and Evolutionary Biology," no. October 2006, pp. 1–15.

[146] M. J. de Vries, "Engineering Science as a "Discipline of the Particular"? Types of Generalization in Engineering Sciences," in *Philosophy and Engineering:: An Emerging Agenda* (I. Poel and D. Goldberg, eds.), pp. 83–93, Dordrecht: Springer Netherlands, 2010.

[147] H. Kamp and B. Partee, "Prototype theory and compositionality.," *Cognition*, vol. 57, pp. 129–91, nov 1995.

[148] Rosch, "Principles of Categorization," *University of California, Berkeley*, pp. 1–25, 1978.

[149] M. R. Lowry, "Algorithm synthesis through problem reformulation," pp. 432–436, 1987.

[150] P. Smets, "Probability of provability and belief functions," *Logique et Analyse*, vol. 6156, no. Drums Ii, pp. 1–16, 1991.

[151] L. Saitta and J.-D. Zucker, *Abstraction in Artificial Intelligence and Complex Systems*. 2013.

[152] Rosanna Keefe, *Theories of Vagueness*. 2003.

[153] M. Benerecetti, P. Bouquet, and C. Ghidini, "Contextual reasoning distilled," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, pp. 279–305, 2000.

[154] L. A. Pineda, "Abstraction, Entropy and Computing Formats," pp. 1–71, 2016.

[155] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, jun 2010.

[156] P. M. McCarthy and D. S. McNamara, "The User-Language Paraphrase Corpus," in *Cross-Disciplinary Advances in Applied Natural Language Processing*, pp. 73–89, IGI Global, 2011.

[157] W. B. Dolan and C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 9–16, Asia Federation of Natural Language Processing, 2005.

[158] V. Rus, M. Lintean, C. Moldovan, and W. Baggett, "The SIMILAR Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts," in *Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May*, pp. 23–25, 2012.

[159] D. Bernhard and I. Gurevych, "Answering learners' questions by retrieving question paraphrases from social Q&A sites," *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications - EANL '08*, no. June, pp. 44–52, 2008.

[160] G. Soğancloğlu, H. Öztürk, and A. Özgür, "BIOSSES: A semantic sentence similarity estimation system for the biomedical domain," in *Bioinformatics*, vol. 33, pp. i49–i58, 2017.

[161] J. Sinclair, "Developing Linguistic Corpora: a Guide to Good Practice," 2004.

[162] B. Karaoglan, T. Kisla, S. K. Metin, U. Hürriyetoglu, and K. Soleymanzadeh, "Using Multiple Metrics in Automatically Building Turkish Paraphrase Corpus," *Research in computing Science*, vol. 117, pp. 75–83, 2016.

[163] J. Cohen, "A power primer.," *Psychological bulletin*, vol. 112, pp. 155–9, jul 1992.

[164] M. Moinester and R. Gottfried, "Sample size estimation for correlations with pre-specified confidence interval," *The Quantitative Methods for Psychology*, vol. 10, pp. 124–130, sep 2014.

[165] C. L. Chuan and J. Penyelidikan, "Sample size estimation using Krejcie and Morgan and Cohen statistical power analysis: A comparison," *Jurnal Penyelidikan IPBL*, vol. 7, no. 1, pp. 78–86, 2006.

[166] D. Jurgens, M. T. Pilehvar, and R. Navigli, "Cross level semantic similarity: an evaluation framework for universal measures of similarity," *Language Resources and Evaluation*, vol. 50, no. 1, pp. 5–33, 2016.

[167] O. Lithgow-Serrano and J. Collado-Vides, "In the pursuit of semantic similarity for literature on microbial transcriptional regulation," *Journal of Intelligent & Fuzzy Systems*, vol. 36, pp. 4777–4786, may 2019.

[168] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, and I. Solti, "Building gold standard corpora for medical natural language processing tasks.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2012, pp. 144–53, 2012.

[169] J.-M. Torres-Moreno, G. Sierra, and P. Peinl, "A German Corpus for Text Similarity Detection Tasks," vol. 5, no. 2, 2017.

[170] K. A. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23–34, 2012.

[171] K. Gwet, "Inter-Rater Reliability : Dependency on Trait Prevalence and Marginal Homogeneity," *Statistical Methods for Inter-Reliability Assessment*, vol. 2, pp. 1–9, 2002.

[172] M. Vila, M. Bertran, M. A. Martí, and H. Rodríguez, "Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures," *Language Resources and Evaluation*, vol. 49, no. 1, pp. 77–105, 2014.

[173] P. K. Bhowmick, P. Mitra, and A. Basu, "An agreement measure for determining inter-annotator reliability of human judgements on affective text," *Proceedings of the Workshop on Human Judgements in Computational Linguistics - HumanJudge '08*, no. August, pp. 58–65, 2008.

[174] M. L. Mchugh, "Interrater reliability : the kappa statistic Importance of measuring interrater reliability Measurement of interrater reliability," *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276–282, 2012.

[175] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[176] J. J. Bartko, "The Intraclass Correlation Coefficient as a Measure of Reliability," *Psychological Reports*, vol. 19, no. 1, pp. 3–11, 1966.

[177] M. G. Kendall, *Rank correlation methods.* Oxford, England: Griffin, 1948.

[178] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[179] K. Gwet, "Kappa Statistic is not satisfactory for assessing the extent of agreement between raters," *Statistical Methods For Inter-Rater Reliability Assessmen*, no. 1, pp. 1–5, 2002.

[180] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, 1977.

[181] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples," *BMC Medical Research Methodology*, vol. 13, p. 61, dec 2013.

[182] D. Kahneman and A. Tversky, "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, vol. 3, pp. 430–454, jul 1972.

[183] C. E. Osgood, "The nature and measurement of meaning," *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, 1952.

[184] A. M. C. Isaac, "Objective Similarity and Mental Representation," *Australasian Journal of Philosophy*, vol. 91, pp. 683–704, dec 2013.

[185] G. Kondrak, "N -Gram Similarity and Distance," *Lecture Notes in Computer Science*, vol. 3772, pp. 115–126, 2005.

[186] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, pp. 39–41, nov 1995.

[187] M. A. Sultan, S. Bethard, and T. Sumner, "DLS @ CU : Sentence Similarity from Word Alignment," *SemEval2015*, vol. 2012, no. SemEval, pp. 241–246, 2015.

[188] G. D. Plotkin, "A note on Inductive Generalization," *Machine Intelligence*, pp. 153–163, 1970.

[189] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, "The Penn Treebank," p. 114, 1994.

[190] D. Gildea, "Dependencies vs. Constituents for Tree-Based Alignment," *Proceedings of EMNLP*, pp. 214–221, 2004.

[191] P. Resnik, "Semantic classes and syntactic ambiguity," in *HLT*, pp. 278–283, 1993.

[192] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," *Ecai*, vol. 16, no. Ic, p. 1089, 2004.

[193] B. Galitsky, J. Lluis, D. Rosa, and G. Dobrocsi, "Mapping Syntactic to Semantic Generalizations of Linguistic Parse Trees," *Artificial Intelligence*, pp. 168–173, 2010.

[194] B. Galitsky, "Machine learning of syntactic parse trees for search and classification of text," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 3, pp. 1072–1091, 2013.

[195] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[196] B. A. Galitsky, J. L. De La Rosa, and G. Dobrocsi, "Inferring the semantic properties of sentences by mining syntactic parse trees," *Data and Knowledge Engineering*, vol. 81-82, pp. 21–45, 2012.

[197] L. Bulat, S. Clark, and E. Shutova, "Modelling metaphor with attribute-based semantics," vol. 2, pp. 523–528, 2017.

[198] L. Fagarasan, E. M. Vecchi, and S. Clark, "From distributional semantics to feature norms : grounding semantic models in human perceptual data as," *Proceedings of the 11th International Conference on Computational Semantics, London, UK, April 15-17 2015*, pp. 52–57, 2015.

[199] L. A. Pineda, "A Distributed Extension of the Turing Machine," pp. 1–51.

[200] D. J. Anderson and E. N. Zalta, "Frege, boolos, and logical objects," *Journal of Philosophical Logic*, vol. 33, no. 1, pp. 1–26, 2004.

[201] J. Barwise and J. Seligman, *Information Flow: The Logic of Distributed Systems*. New York, NY, USA: Cambridge University Press, 1 ed., 2008.