UNIVERSIDAD DE LOS ANDES

FACULTAD DE INGENIERÍA
DOCTORADO EN CIENCIAS APLICADAS

MODELO ADAPTATIVO DE PERFILES ACADÉMICOS BASADO EN COMPETENCIAS, USANDO MINERÍA SEMÁNTICA.

Autor:

Ing. MSc. Alexandra González Eras

Director:

Dr. Jose Aguilar Castro

TESIS DOCTORAL PRESENTADA ANTE LA UNIVERSIDAD DE LOS ANDES COMO REQUISITO FINAL PARA OPTAR AL TÍTULO DE DOCTOR

2021

A mis padres por ser apoyo
constante en mis esfuerzos

A mi esposo e hijos por ser el
motor y motivo de todos mis pasos

RESUMEN

En este trabajo doctoral se presenta un “Modelo adaptativo de perfiles académicos basado en competencias, usando minería semántica”, para la identificación, alineamiento y recomendación de competencias. En este trabajo, se utilizan técnicas de Procesamiento de Lenguaje Natural, Minería Semántica y Aprendizaje Ontológico, con el objetivo de generar un modelo que permita la gestión automática de las competencias en diferentes contextos y desde fuentes de información diversas. Surge de la necesidad de sobrellevar la ambigüedad lingüística y semántica de los perfiles profesionales y académicos en cuanto a competencias y sus componentes conocimiento y habilidad. Para ello, el modelo está conformado por cuatro componentes: caracterización, extracción, comparación y actualización. A través de ellos se identifican y extraen los términos de competencia, conocimiento y habilidad, se comparan contra tesauros propios del dominio, y en función de la comparación, se alinean los perfiles con los términos de alto nivel de los tesauros. Con este resultado, se realiza la retroalimentación de los perfiles, tanto académicos como profesionales, para enriquecer las competencias en ambos contextos. Son determinantes en este trabajo las técnicas de Procesamiento de Lenguaje Natural que permiten la selección de los términos relevantes, las técnicas de Minería Semántica, Aprendizaje Ontológico y los modelos descriptivo y dialéctico que permiten el enriquecimiento de los términos extraídos, logrando de esta manera el manejo de las ambivalencias, tanto de los términos de competencia como de los perfiles profesionales y académicos. El presente trabajo representa una propuesta novedosa en el ámbito de la gestión de las competencias, porque permite el análisis de las competencias de perfiles profesionales y académicos en español desde una arquitectura adaptable al contexto de los perfiles. Además, no solo basándose en el enfoque tradicional ontológico de lógica descriptiva, sino desde el punto de vista de lógica dialéctica, lo cual le otorga una mayor flexibilidad en el reconocimiento de ambivalencias, y, por tanto, le da una mayor capacidad de enriquecimiento de los perfiles profesionales y académicos.

Palabras Claves: Minería Semántica, Procesamiento de Lenguaje Natural Aprendizaje Ontológico, Gestión de competencias.

AGRADECIMIENTO

Para el desarrollo de la tesis doctoral, conté con el soporte de varias personas, que de una u otra forma contribuyeron a su culminación, y a las cuales debo mi agradecimiento.

Primeramente, doy gracias a Dios por darme esta oportunidad, y por llenarme de sabiduría paciencia y fortaleza para seguir siempre adelante.

A mi tutor José Aguilar, gracias infinitas por todas sus orientaciones, su apoyo y sus valiosos conocimientos, que fueron base fundamental en el desarrollo del trabajo y en mi crecimiento académico, profesional y personal en esta etapa de mi vida.

A mis padres ejemplo de trabajo, humildad y honradez, gracias por todo lo que me han dado y me siguen dando cada día. También, a mis hermanos, familiares y amigos que siempre han estado pendientes del desarrollo de mi proyecto de doctorado y me han dado sus palabras de ánimo.

A mis compañeros de doctorado, con los que desarrollé proyectos y publicaciones, donde compartimos ideas y experiencias y que me brindaron su ayuda en los momentos en que la necesité, les deseo ¡muchos éxitos!

A la Universidad de Los Andes por permitirme participar en este importante programa doctoral y a su personal.

A todos los que de alguna manera me han brindado su ayuda y su apoyo, ¡Qué Dios los bendiga!

INDICE

INDICE DE FIGURAS

INDICE DE TABLAS

LISTA DE ABREVIATURAS

NLP: Natural Language Processing (Procesamiento de Lenguaje Natural).

IR: Information Retrieval (Recuperación de Información).

IE: Information Extraction (Extracción de Información).

CL: Computational Linguistic (Lingüística Computacional basada en corpus).

NER: Named Entity Recognition (Reconocimiento de entidades nombradas).

OCR: Optical Character Recognition (Reconocimiento óptico de caracteres).

NLG: Natural Language Generation (Generación de Lenguaje Natural).

LM: Language Model (Modelo de Lenguaje).

NLM: Neural Language Models (Modelos de Lenguaje Neuronal).

LRE: Lenguaje Recursivamente Enumerable.

LSC: Lenguaje Sensible al Contexto.

LLC: Lenguaje Libre de Contexto.

LR: Lenguaje Regular.

DL: Lógica Descriptiva.

SA: Sentiment Analysis. (Análisis de Sentimiento).

ML: Machine Learning (Aprendizaje Automático).

LOD: Linked Open Data.

KDD: Knowledge Discovery in Databases.

KDT: Knowledge Discovery in Text.

RCD: Reusable Competence Definitions (Definiciones Reusables de Competencias)

MD: Modelo Dialéctico de competencias

OC: Modelo ontológico de competencias

CAPITULO 1: CONTEXTUALIZACIÓN

Este capítulo realiza una introducción a la tesis, por lo cual realiza el planteamiento del problema que se abordará, presenta los trabajos más relevantes de la literatura para el desarrollo de la tesis, describe los objetivos de la misma para terminar, justificando la relevancia de la misma.

1.1. PLANTEAMIENTO DEL PROBLEMA

En general, el dinamismo del entorno laboral trae complicaciones no solamente a los profesionales que deben cumplir con los requisitos solicitados, sino a instituciones como las universidades, que deben ajustar las carreras profesionales ofertadas, para manejar la aparición de nuevas posiciones dentro de las empresas y la creciente necesidad de expertos en determinadas áreas [2,3]. Además, el masivo uso de Internet en actividades empresariales se ha extendido también al ámbito del reclutamiento de personal, con una creciente cantidad de información alojada en plataformas de empleo, en donde se ofertan plazas de trabajo cuyos requerimientos cambian rápidamente [1,6]. Como resultado, las universidades deben revisar permanentemente sus ofertas profesionales y los cursos de sus programas de estudios, a fin de mantener actualizadas sus ofertas con las competencias y los conocimientos requeridos por el mercado de trabajo [5,9].

Actualmente, la gestión de competencias se ha enfocado, en mayor medida, en el análisis entre las plazas de trabajo de las empresas y los perfiles de los profesionales candidatos a ocuparlas [32], con el fin de establecer, en primer lugar, cuáles son los conocimientos y habilidades (competencias) que requiere un rol y, en segundo lugar, si el candidato reúne las competencias necesarias para lograr un óptimo desempeño en el cargo a ocupar [10]. Esto ha dejado mucho por hacer en cuanto al tratamiento de las competencias entre una oferta de carrera y una oferta de trabajo [9], ya que no es claro que requerimientos del mercado laboral deben considerarse dentro de los perfiles de carrera [5]. En consecuencia, se hace necesario promover una gestión que garantice un adecuado tratamiento de esta información, con miras a solucionar las expectativas del contexto académico y laboral en cuando al manejo de competencias.

La gestión automática de competencias requiere de sistemas heterogéneos [31], alimentados por diversas fuentes de información (estructuradas, semiestructuradas y no estructuradas), que realicen procesos que permitan la identificación, alineamiento y recomendación de competencias; de cuyos resultados se obtienen los insumos para la formación de perfiles y catálogos de profesiones y carreras académicas, entre otros [10, 24, 105]. Al mismo tiempo, los sistemas deben manejar la ambigüedad que presentan los perfiles a diferentes niveles (léxico, sintáctico y semántico), de tal forma que sean capaces de adaptarse a las variaciones de las competencias en los dos contextos: académico y laboral [5]. Otro aspecto importante es que la mayoría de las investigaciones y el desarrollo de plataformas para la gestión de competencias se han realizado para el análisis de información en otros idiomas diferentes al español, lo cual deja una brecha en el estudio de las competencias en este idioma [12].

Particularmente, la gestión de competencias en los contextos académico y laboral se basa en procesos dinámicos diferentes, debido a que la interpretación que ambos contextos tienen

acerca de las competencias es diferente [1,5]. Lo anterior impide que instituciones de educación como las universidades, puedan reconocer con claridad las competencias requeridas por un perfil de trabajo, y así, establecer alineamientos con sus perfiles académicos.

Las consecuencias de la disociación de los perfiles académicos y laborales, debido a la ambigüedad que presentan las competencias, se reflejan en los siguientes aspectos: 1) la baja calidad de los perfiles profesionales [4], 2) la falta de mecanismos para determinar las competencias que no son explícitamente mencionadas en los perfiles [21,114] y, 3) la falta de plataformas tecnológicas que puedan realizar búsqueda y clasificación de ofertas de empleo o estudio basadas en competencias [21].

La integración de técnicas de Procesamiento de Lenguaje Natural, Minería Semántica y Aprendizaje Ontológico coadyuva a resolver los anteriores problemas a través de modelos semánticos que facilitan la representación y uso de las competencias y sus relaciones, según el contexto al que pertenezcan. La Tabla 1.1. presenta un resumen de la problemática del tema de tesis, y de las soluciones que se han dado en los contextos del Procesamiento de Lenguaje Natural, Minería Semántica y Aprendizaje Ontológico.

Tabla 1.1. Definición de la problemática

Problema	Causa	Efecto	Solución
No existe relación entre las competencias de los perfiles profesionales y académicos	Ambigüedad en los términos de competencias	Falta de mecanismos para determinar las competencias que no son explícitamente mencionadas en los perfiles	PLN: información estructurada, semiestructurada y no estructurada
	Diferentes definiciones sobre competencia entre los contextos académico y laboral	Baja calidad en la interpretación de los perfiles profesionales	Minería Semántica de competencias: frases, oraciones, términos ambiguos
	Las competencias se entienden como certificaciones títulos, áreas de conocimiento		
Heterogeneidad semántica de modelos ontológicos de competencias	No existen mecanismos para la búsqueda y clasificación de perfiles en función de competencias	Falta de plataformas que busquen perfiles en función de competencias	Aprendizaje ontológico: Ontologías, tesauros, diccionarios de competencias en idiomas, diferentes al español
Plataformas de gestión de perfiles académicos y laborales no clasifican su información en función de competencias	No existen mecanismos para lograr alineamientos entre competencias del entorno Académico y Laboral	Falta de gestión en las competencias que contienen los perfiles	Sistemas de gestión de competencias inteligentes: caracterización, extracción, alineamiento y retroalimentación de competencias

Entre los enfoques que se han utilizado para solventar estos problemas se encuentran las tecnologías semánticas (ontologías y tesauros) [9,10], con el propósito de modelar el contexto de la competencia, incrementando así el nivel de interoperabilidad semántica entre los sistemas

[1]. Sin embargo, la dinámica de las competencias, particularmente en el contexto laboral, ha provocado un problema de heterogeneidad semántica, debido a que las ontologías presentan diferencias en sus estructuras, en la denominación de sus componentes, y en los lenguajes en que están construidas [2].

Es entonces que el Aprendizaje Ontológico se presenta como una solución al problema, a través de la generación automática o semiautomática de ontologías, contribuyendo así a solucionar el problema de la construcción manual y la actualización de modelos ontológicos [20, 21]. Para ello se apoya en el Procesamiento de Lenguaje Natural (PLN) para el tratamiento de los textos, y en la Minería Semántica para analizar los contenidos textuales.

La Minería Semántica se apoya en fuentes externas como taxonomías, tesauros y cuerpos de conocimiento que describen y clasifican habilidades, competencias y tópicos de conocimiento en cada profesión; con el propósito del sobrellevar la falta de detalle en las descripciones de competencia [4], y así contribuir a la estandarización y alineamiento de los perfiles. Estos esquemas son vocabularios normalizados de uso general y, por tanto, limitados en la representación del contexto de competencia a la realidad académica o laboral [21, 106].

La base de la minería semántica son las medidas de similitud, que capturan la fortaleza de la relación semántica entre elementos lingüísticos (como palabras o conceptos) [7,17]. En la literatura, existen varios tipos de medidas para estimar la similitud o disimilitud semántica, según las estructuras de datos específicas usadas (como vectores, matrices o grafos), o de acuerdo al tipo de representación sobre el cual está basada la comparación (por ejemplo, las unidades de lenguaje como las palabras, oraciones, párrafos y documentos) [23].

El uso de estas técnicas en el contexto de las competencias, conlleva los siguientes retos:

- Considerando que la fuente de consulta es la Web, las técnicas de PLN soportan el tratamiento de documentos no estructurados, con el propósito de extraer términos desde frases que no describen claramente las competencias [24,19], ya sea por el uso de expresiones sinónimas que no corresponden con competencias, o por la diversidad que ofrece el lenguaje español para expresar las mismas.
- Las competencias son escritas en términos de certificaciones o títulos [1], o como resultado del aprendizaje dentro del contexto académico [25]. Es por esto que dentro de la estructura de los documentos se encuentran frases que presentan indicios de competencias con diferentes denominaciones. Por ejemplo, en los perfiles universitarios las competencias aparecen bajo títulos como: descripciones, campos ocupacionales y objetivos [1,71]; en cambio, en las ofertas laborales aparecen descritas como funciones, cargos, conocimiento y habilidad [4]. Estas frases candidatas son pobres en términos de competencia que permitan su comparación, en su lugar, se encuentran descripciones de habilidades o áreas de conocimiento [25].
- Otra de las razones por las que es difícil comparar perfiles son las diversas interpretaciones que en ambos contextos tienen acerca de las competencias, y la forma en que los actores representan las mismas en los textos [10, 19]. Para las universidades, las competencias son el resultado de un proceso de aprendizaje, en el cual un individuo ha alcanzado un nivel cognitivo en un tema de conocimiento [5], mientras que, para las empresas, las competencias representan la capacidad para realizar una tarea [6], o nivel de dominio de un ámbito de conocimiento [3,35]. Estas interpretaciones de competencia pueden presentarse en los perfiles de varias formas, de acuerdo al conocimiento y al estilo del redactor/(es) del perfil [25, 40].

De esta manera, la propuesta de doctorado pretende desarrollar un esquema para la retroalimentación de perfiles de carrera mediante la integración de varios dominios: Minería Semántica, Aprendizaje Ontológico y PLN; con el objetivo de generar modelos que permitan el reconocimiento y la extracción de los patrones de competencia desde los textos en Internet, la comparación de las competencias de cada contexto, y en base a los resultados obtenidos, definir mecanismos de aprendizaje/actualización de perfiles académicos y profesionales. Así, esta propuesta, con la combinación de esas técnicas, propone procesos automáticos y semiautomáticos de gestión de competencias, que pueden ser aplicados en diferentes contextos y para diferentes fuentes de información, de tal forma de proporcionar una retroalimentación a los perfiles académicos y profesionales.

1.2. ANTECEDENTES

A continuación, se presentan algunos estudios sobre el análisis semántico de competencias, que nos permiten conocer las tendencias actuales en esta área. La Tabla 2 presenta un resumen de las investigaciones más importantes vinculadas a este trabajo. Organizaremos la presentación del estado de arte en varios grupos:

En primer lugar, se aborda el tema del aprendizaje ontológico y como se ha aplicado en la gestión de competencias.

En [1] se presenta una propuesta que usa un modelo ontológico para determinar las competencias de acuerdo a 3 niveles: el mercado laboral, el individuo y sus conexiones sociales. Este modelo se actualiza a través de un proceso de mantenimiento semiautomático, que permite la detección de tendencias con respecto a la importancia de la competencia y los perfiles de trabajo emergentes. Para ello, emplean métodos de minería de datos, análisis de redes sociales y recuperación de información. De este modo, permite a los gestores de recursos humanos reaccionar oportunamente a las necesidades cambiantes del mercado laboral en cuanto a tareas de contratación, así como en el desarrollo de competencias, especialmente, para la planificación de carreras profesionales. El correcto funcionamiento del sistema demanda la participación de los usuarios, además, la decisión final si un término es una competencia próxima o simplemente una palabra nueva para una conocida, tiene que ser hecha por un experto humano.

El articulo [2] propone una plataforma de gestión de competencias. El núcleo lo constituye una ontología de competencia vocacional (VCO), que es usada para conciliar modelos de competencia contextualizados e intercambiar información de empleo anotada con competencias. El ámbito específico de este caso de estudio es la gestión de un ciclo de vida de una ontología VCO. Finalmente, ellos usan el estándar de competencia multilingüe RCD para la representación de competencias.

En [4] y [35], los autores presentan un modelo para la definición de los recursos humanos en el tiempo. En particular, se usa la lógica descriptiva para representar y razonar acerca de la experiencia, habilidades y competencias, y capturar información sobre las fuentes de información sobre habilidades y competencias. El framework establece una ontología para la gestión en base a axiomas lógicos, de las habilidades de los individuos que deben ser evaluadas,

además de inferir competencias usando diferentes fuentes de información, cuya fiabilidad es validada por medio de axiomas formales de confianza definidos en [28,29].

El trabajo presentado en [30] presenta un sistema de información para almacenar, evaluar y razonar sobre las competencias de los estudiantes en las universidades, para planificar sus cursos en base a las brechas existentes entre las competencias esperadas y las competencias obtenidas. El sistema se basa en una representación de habilidades y una ontología de competencias. Además, el sistema apoya a los estudiantes en la producción de perfiles de competencia para las solicitudes de empleo, la cual se basa en HR-XML para permitir un intercambio de datos. Las competencias referenciadas se definen en una ontología, y la presencia o fuerza de una competencia puede ser atestiguada por evidencias (notas, certificados). Si los estudiantes permiten el acceso a su perfil, los empleadores pueden utilizar estos perfiles para encontrar candidatos adecuados para las vacantes de empleo. El modelo ontológico se extiende en [27] con conceptos de descripciones de trabajo y conceptos, para evaluar competencias y sus evidencias en diferentes niveles, los que son alineados al estándar HR-XML para definir perfiles de competencia.

En segundo lugar, se reseñan algunos estudios relacionados con el uso de técnicas de PLN para la identificación de competencias.

En [1] se propone un esquema de PLN que usa técnicas de extracción de información y reconocimiento de entidades, para obtener desde perfiles de candidatos información sobre títulos de trabajo, empleadores, organizaciones y etiquetas de competencia. Estas últimas son evaluadas según listas predefinidas desde tesauros (Germa-Net) y diccionarios (Wortschatz), usando medidas de frecuencia basada en modelos de espacio vectorial para la selección y filtrado de las mismas. Luego, las etiquetas se asocian a conceptos que son mezclados entre sí, usando la medida de distancia de Levenshtein para integrar conceptos con raíces lingüísticas similares.

En el trabajo [19] se usa un framework de gestión de competencias que utiliza herramientas y tecnologías semánticas para la población de ontologías de perfiles de expertos desde fuentes estructuradas y no estructuradas, por medio de la aplicación de una anotación semántica y de algoritmos de extracción de información, como son: D2RQ server, TopBraid Composer, UIMA, entre otros. El objetivo es la construcción de modelos de competencias individuales y empresariales en forma de ontologías, así como la definición de diferentes formas de búsqueda y recuperación de datos desde la Web Semántica.

El trabajo [21] presenta un resumen de trabajos sobre el uso de tecnologías semánticas en procesos de eRecruitement para el alineamiento de candidatos y ofertas de empleo, el cual se fundamenta en el desarrollo de una ontología de recursos humanos, que tiene como conceptos principales el candidato, el empleador, la descripción del trabajo y del perfil profesional. La interoperabilidad entre descripciones de trabajo y perfiles de candidato se logra por medio de vocabularios estandarizados, que proveen conceptos colectivos para describir títulos ocupacionales, habilidades requeridas y cualificaciones educacionales. Otros enfoques usan ontologías para alinear candidatos y ofertas de empleo, combinadas con algoritmos de aprendizaje ontológico. Además, existen ontologías en el dominio de los recursos humanos, como ProPer Ontology, KOWIEN Ontology [31] y Knowledge Nets [32], centradas en casos específicos de eRecruitment.

A continuación, se reseñan algunos estudios relacionados con la Minería Semántica de competencias.

En el artículo [1] se presentan dos procesos de minería semántica, el primero tiene por objetivo la detección de relaciones de competencia jerárquicas usando heurísticas, basadas en el análisis de n-gramas y patrones léxico-sintácticos, donde se aplicaron los patrones léxico-semánticos para la detección de relaciones jerárquicas (hipónimos) definidos en [33]. El segundo proceso tiene que ver con la identificación de relaciones y reglas de competencia asociativas en base a la frecuencia de coocurrencia. Para ello se usó el algoritmo A-priori [34], para calcular las combinaciones de competencias en conjuntos de dos conceptos. Se obtiene una red semántica donde los clústeres de conceptos de competencias tienen una frecuencia mínima de 0.01, que permite identificar empresas vecinas cuyos perfiles de trabajo comparten conceptos de competencia.

En [35] se presenta un sistema para el alineamiento de ofertas de empleo y perfiles de candidatos bajo un enfoque que usa una ontología que representa en lógica descriptiva los perfiles y ofertas, con sus respectivas nociones de: habilidades presentadas y requeridas, títulos y experiencia. El alineamiento se realiza por medio de un método deductivo que determina el tipo de alineación entre los perfiles, en base a medidas de similitud de distancia y estructurales, cuyos resultados se usan para rankear candidatos.

El trabajo [26] usa 3 medidas de similitud: Coseno, Coeficiente Dice e Índice de Jaccard; y dos medidas de distancia: Euclidiana y Manhattan, con 3 tipos de modelos espacio vectorial (frecuencia absoluta, frecuencia relativa y valores TF-IDF), para determinar cómo el uso de diferentes medidas y espacios vectoriales afecta la detección de resúmenes de candidatos seleccionados previamente. Se analiza un corpus etiquetado manualmente, y se realiza un análisis de varianza para determinar cómo las 5 medidas consideran la predilección de los resúmenes, y si estos perfiles tienen más en común que otros.

Los autores de [25] presentan una implementación para la recomendación de competencias. El núcleo del sistema lo constituye el modelo ontológico TELOS, el cual fue extendido para la comparación y recomendación de competencias en base a los siguientes componentes: conocimiento, habilidad y performance. Para ello, se determina la cercanía semántica de conocimiento por medio de la proximidad de los conceptos, determinando si son más generales o específicos (definiciones en Lógica Descriptiva DL). Para la comparación de competencias, se usan heurísticas que relacionan los siguientes componentes de las competencias: habilidad, conocimiento y performance, que son usadas por agentes para la recomendación de competencias.

Finalmente, se reseñan algunos sistemas de soporte para la representación y gestión de perfiles profesionales en los ámbitos académico y laboral.

En [9], los autores presentan a CUSP (Course and Unit of Study Portal), un sistema que apoya el diseño de habilidades genéricas y requisitos de acreditación para universidades, para una carrera determinada. Para ello usa un conjunto de frameworks que permiten el mapeo de objetivos de aprendizaje representados en ontologías ligeras de definiciones de la carrera y estándares. CUSP es una herramienta para la toma de decisiones de los coordinadores de carrera, que ayuda a identificar vacíos de conocimiento y de requisitos de acreditación, además de inconsistencias progresivas de aprendizaje.

En el artículo [36] se resalta la necesidad de usar información de las colocaciones de empleo para actualizar las competencias en el contexto académico, y se presenta un modelo empírico para identificar y evaluar las competencias genéricas adquiridas por los estudiantes en su aprendizaje. El estudio emplea una muestra de 351 informes realizados por supervisores de prácticas en empresas. Antes de probar las hipótesis, el instrumento de medida fue evaluado mediante una regresión parcial de mínimos cuadrados. Como resultado, se proporciona información útil para profesores y profesionales, en una herramienta de aprendizaje y evaluación propuesta para las prácticas profesionales.

El trabajo [21] presenta la plataforma LO-MATCH, una herramienta semántica diseñada con dos propósitos: 1. reconocer, validar y certificar aprendizajes previos, y con ellos construir currículums de candidatos y 2. alinear los curriculums con ofertas de trabajo. Para ello, la herramienta usa WordNet (Versión en inglés) para expresar resúmenes, necesidades de compañías y cualidades de acuerdo a una ontología compartida, en donde los candidatos escogen las salidas de aprendizaje que mejor los describen, luego la plataforma presenta una lista de ofertas de trabajo que fueron rankeadas de acuerdo con las salidas de aprendizaje del candidato, que fueron alineadas con aquellas requeridas por el puesto de trabajo.

En el trabajo [37] se presenta un sistema para la gestión de competencias para parques tecnológicos donde se automatiza el proceso de búsqueda del residente que puede satisfacer las posibles tareas del cliente. Cada residente es descrito por un perfil, que es accesible para otros usuarios del sistema de gestión de competencias. El perfil consiste en varias competencias y niveles de habilidades.

En la Tabla 1.2. se observa una comparación de la propuesta con trabajos anteriores según las técnicas de NLP usadas, la forma de representación de conocimiento y habilidad, las medidas de similitud usada, la información extra considerada (tesauros, etc.), y las métricas de validación del modelo ontológico. Se puede observar que existen coincidencias entre los trabajos anteriores y nuestra propuesta en el uso de medidas de similitud léxicas, pero la nuestra es la única que usa similitud semántica. En cuanto al uso de información extra, muchos usan tesauros, o lexicones. Así también, en las técnicas de NLP que se emplean, la mayoría usa técnicas de Reconocimiento de Entidades Nombradas (NER), y nuestro trabajo es el único basado en patrones lingüísticos. En cuanto a la validación ontológica, nuestra propuesta es la única que usa los criterios de Completitud y Robustez para determinar la utilidad y calidad del proceso de extracción de términos, y su uso para poblar la ontología de competencias OC, en lugar de la métrica de Precisión. Otra diferencia radica en el idioma de las fuentes de los datos, donde nuestra propuesta es la única para el español. Finalmente, la mayoría usa ontologías, y nuestra propuesta es la única que combina documentos de la web con perfiles en un modelo semántico.

Según los trabajos previos, se ve que los enfoques de análisis de competencias se realizan en su mayoría entre ofertas de empleo y perfiles de candidatos, para detectar las competencias para cubrir plazas de empleo, Además, por lo general están en inglés. Por otro lado, los tesauros, los vocabularios y los estándares que facilitan el alineamiento se encuentran también en ese idioma. Adicionalmente, son pocas las plataformas que permiten el análisis de competencias en el contexto académico y que, además, estén orientadas hacia la planificación de cursos o planes curriculares. Por otro lado, no contemplan el uso de ofertas de empleo como fuente de consulta de requerimientos laborales. En consecuencia, son estructuras estáticas que no consideran el dinamismo del entorno para el cual están formando profesionales.

Tabla 1.2. Análisis de trabajos relacionados

Trabajo	NLP	Representa conocimiento y habilidad	Medida de similitud	Tesauros extra	Fuente de datos	Validación
[1]	Patrones	Competencias como entidades	Levenshtein	Germa-Net Wortshatz	Ofertas laborales	Experto
[19]	NER	Competencias como entidades	--------	------	Perfiles de expertos	-------
[21]	NER	Títulos, habilidades	--------	ProPer, KOWIEN Knowledge Nets	Ofertas de empleos CV	Precisión
[26]	TF-IDF	Vectores de términos	Coseno, Dice, Jaccard Manhattan Euclideana	-------	CV	Precisión
[24]	NER	Competencias como entidades	Levenshtein WordNet	LinkedData ontología competencia	Perfiles	Precisión y relevancia
[20]	NER	Entidades y acciones	Léxicas WordNet	Ontología competencia	Corpus	Precisión Recall
[38]	NER	Competencias como entidades	Semántica	Lexicón y Onomasticón	Texto de la Web	Precisión y relevancia
[37]	NER	Ontología de perfiles de competencia	Léxicas	WordNet	Texto de la Web	Precisión
[9]	----	Ontologías de carrera	------	Estándares	-------	--------
Nuestra propuesta	Patrones	Competencias conocimiento y habilidad	Léxicas Semántica	DISCO II BLOOM	Ofertas laborales y carrera	Completitud Robustez Entropía

1.3. OBJETIVOS

1.3.1. Objetivo General

Desarrollar un modelo adaptativo de perfiles académicos basado en competencias desde ofertas de trabajo, mediante el uso de técnicas de minería semántica.

1.3.2. Objetivos Específicos

- Realizar un análisis sobre los perfiles académicos basados en competencias, con el fin de realizar una estandarización de conceptos para los contextos académico y laboral.
- Estudiar aspectos teóricos y prácticos relacionados con el análisis de competencias, desde los ámbitos del Procesamiento de Lenguaje Natural, Minería Semántica y Aprendizaje Ontológico.

- Caracterizar los modelos ontológicos de los perfiles universitarios y laborales, en base a competencias.
- Establecer esquemas para la extracción de competencias en base a patrones semánticos, para los contextos académico y laboral.
- Definir procesos de aprendizaje de ontologías de perfiles académicos basados en competencias.

1.4. JUSTIFICACIÓN E IMPORTANCIA

La justificación del presente trabajo se realiza desde las siguientes perspectivas:

Desde el ***enfoque teórico,*** esta investigación se realiza con el propósito de aportar al conocimiento existente en el dominio de la gestión de competencias, especialmente entre los contextos académico y laboral, donde se necesita establecer mecanismos de alineamiento de los perfiles profesionales y académicos, que permitan el enriquecimiento de las ofertas académicas con las competencias solicitadas por el entorno laboral. Otro aspecto importante es que se realiza el tratamiento de perfiles en español, donde se plantean esquemas de análisis de competencias desde la perspectiva lingüística y semántica, lo cual constituye un aporte al conocimiento en esta área.

Desde el ***enfoque metodológico,*** esta investigación propone el uso de técnicas y métodos de PLN, Minería Semántica y Aprendizaje Ontológico para la caracterización, extracción, alineamiento y retroalimentación de competencias de perfiles profesionales, así como el establecimiento de métricas y medidas que permiten validar los resultados obtenidos en cada una de las fases del modelo; de esta manera, se plantean técnicas y esquemas para todo el proceso de análisis de los perfiles profesionales.

Desde el ***enfoque práctico***, esta investigación busca aportar una solución al problema de ambigüedad existente en la definición de competencias en los contextos académico y laboral mediante un modelo de análisis de perfiles profesionales que permita su alineamiento en función de las competencias, conocimientos y habilidades que contienen. Es así que, se identifican aquellas competencias comunes y no comunes, y se establecen procesos de validación y retroalimentación de perfiles académicos que contribuyen a la actualización de las ofertas académicas de las instituciones de educación superior.

1.5. ORGANIZACIÓN DE LA TESIS

En este capítulo se describió el planteamiento del problema, los antecedentes donde se muestran diferentes estrategias de análisis de competencias desde la perspectiva del PLN, Minería Semántica y Aprendizaje ontológico, los objetivos de la investigación y su importancia. En el capítulo 2 se establecen los aspectos teóricos relacionados con la tesis, a saber, PLN, Minería Semántica y Aprendizaje Ontológico. En el capítulo 3 se describe el modelo de análisis de competencias, sus propiedades, arquitectura y componentes. En el capítulo 4 se desarrollan algunos casos de estudio, a fin de mostrar la aplicación del modelo propuesto en diferentes escenarios. Por último, en el Capítulo 5 se presentan las conclusiones del trabajo, los trabajos futuros y una discusión de los resultados obtenidos.

De esta manera, a través de la tesis se presentan las principales contribuciones de este trabajo, las cuales son:

- El desarrollo de una arquitectura para el análisis de competencias de perfiles profesionales y académicos, que comprende las fases de caracterización, extracción, comparación y retroalimentación de competencias.
- La definición de un modelo ontológico para la identificación de competencias, en función de patrones lingüísticos de habilidad y conocimiento.
- La definición de un modelo dialéctico para el análisis de la ambigüedad de los perfiles profesionales en cuanto a competencias y sus componentes (habilidad y conocimiento)
- La definición de un esquema de alineamiento de perfiles profesionales y académicos contra tesauros, en función de medidas de similitud de términos de competencias.
- La definición de un esquema de retroalimentación de perfiles profesionales y académicos en función de medidas de completitud y robustez semántica

CAPITULO 2: MARCO TEÓRICO

En este capítulo se presenta el marco teórico de la tesis. Primero, se describe el concepto de competencias en los entornos académico y laboral, lo cual permite el entendimiento de sus características y componentes; segundo, se presenta el área de PLN, donde se revisan las técnicas, modelos y tipos de análisis que se utilizan para el tratamiento de texto no estructurado, que permiten el reconocimiento de competencias en los perfiles profesionales y académicos; y tercero, se explica el área de Minería Semántica, que provee los modelos y técnicas que apoyan los procesos de análisis de las ambigüedades semánticas de las competencias, desde la lógica descriptiva tradicional y la lógica dialéctica.

2.1 GESTION DE COMPETENCIAS

2.1.1 Competencias, currículo y perfiles profesionales

El enfoque de competencias ha surgido, específicamente en el ámbito laboral, a partir de la necesidad de preparar a un individuo para desenvolverse de acuerdo a un perfil, que define las funciones, conocimiento y habilidades para desempeñar un puesto de trabajo [41,25]. Una competencia es "un conjunto de conocimientos utilizados para realizar una tarea", la cual se desarrolla con un nivel de habilidad [5,40]. En el contexto académico, esta definición se extiende al considerar la competencia como “la movilización, la puesta en acción, de un conjunto de capacidades -cognitivas, procedimentales y actitudinales- que se deben integrar en contextos específicos” [39,44].

Un perfil profesional basado en competencias define la identidad del individuo (sea este un Neo-graduado o un candidato a una plaza laboral), y explica las características principales de la profesión específica [45]. En el contexto laboral, las organizaciones utilizan modelos y catálogos para describir competencias de núcleo (core competences), que contienen descripciones de habilidades y conocimientos esenciales, y niveles de rendimiento para cada rol, grupo de trabajo o departamento [46,47]. Estos modelos son usados en procesos de reclutamiento y capacitación de personal, a través de la creación de perfiles de competencias y ofertas de empleo [1,2].

En el contexto académico, las instituciones educativas realizan diseños curriculares y estructuran programas de formación profesional [42], según requerimientos del mercado laboral [9,48], que describen el desarrollo de las competencias a diferentes niveles: 1. Macrocompetencias (proporcionan la estructura general de una profesión); 2. Mesocompetencias (son la base para el desarrollo de perfiles profesionales); y, 3. Microcompetencias (establecen las unidades de competencia e indicadores que definen los planes de estudios del perfil) [5,49].

Los perfiles reúnen competencias en forma de enunciados, que declaran el conocimiento y/o la habilidad para aplicar ese conocimiento en un ámbito específico [19, 43]. Estos enunciados se pueden encontrar en los documentos de perfiles profesionales, identificados bajo descriptores que no corresponden a competencias [2,4]. En la Tabla 2.1. se presenta una muestra de los posibles elementos de los perfiles; es claro que las competencias pueden ser interpretadas como parte de descripciones, habilidades, conocimientos y campo ocupacional [5].

Tabla 2.1. Elementos de los documentos de perfiles profesionales

Perfil profesional (Oferta de empleo)	Perfil profesional (perfil académico)
Perfil: títulos académicos, certificaciones, experiencia en el cargo **Conocimientos:** tópicos y áreas de conocimiento, experiencia en tópicos de conocimiento, certificados **Habilidades:** genéricas (soft-skills) experiencia en dominios específicos	**Perfil de competencias:** habilidades, conocimientos, áreas de conocimiento **Objetivos:** ámbito donde se aplica el perfil. **Perfil del egresado:** área de conocimiento que dominará **Campo ocupacional:** rol, cargo, función que podrá desempeñar en el contexto laboral

2.1.2. Conceptualización de la noción de competencias para este trabajo

Considerando las múltiples definiciones de competencia que existen, tanto en el contexto académico como en el laboral, es necesario realizar algunas precisiones para delimitar el alcance de la investigación del trabajo doctoral:

- Se determinan dos tipos de perfiles: *académicos*, los cuales describen, habilidades, conocimiento y campos ocupacionales del futuro profesional [30,5]; y *ofertas de trabajo*, los cuales describen las funciones, actividades, habilidades y especializaciones requeridas para una plaza de trabajo [2,46].
- El estudio se enfoca en las competencias de dominio, que son entendidas en el contexto académico como *competencias específicas* de la profesión, mientras que en el contexto laboral se consideran *competencias técnicas o funcionales*. Las competencias específicas reflejan resultados de aprendizaje [45] y están orientadas a las características técnicas de cada ocupación o puesto de trabajo de acuerdo con cada área temática [40]. Las competencias funcionales reflejan capacidades, "las habilidades y conocimiento que los individuos pueden tener, con el fin de ser aptos para trabajos particulares" [2].
- El concepto de competencia que se asume es "la habilidad con que un profesional se desenvuelve en un área de conocimiento específico" [25]. Aplicado este concepto a cada ámbito, podemos afirmar que existe una coincidencia en lo que tiene que ver a los siguientes elementos de competencia: *el conocimiento*, ya que se enmarca en tópicos que forman parte de una profesión y que son necesarios para desenvolverse en ella [2,50]; y *la habilidad*, porque constituye la capacidad o acción de usar el conocimiento para actuar con éxito en el desarrollo de una actividad [45].

En consecuencia, para cumplir con el objetivo de esta investigación, se analizarán las competencias funcionales o específicas, tomando en cuenta solamente el conocimiento y habilidad como elementos de competencia, ya que son comunes a los dos ámbitos (académico y laboral) [19].

2.1.3. Ambigüedad en las competencias

En la práctica común, la representación de las competencias siempre se ha realizado por medio de declaraciones lingüísticas, que no describen formalmente los dominios de conocimiento, además de no ser aptas para procesos computacionales [25]. Además, la estructura de las frases que contienen las competencias es diferente en cada tipo de perfil. En la Tabla 2.2. se muestran ejemplos de las estructuras textuales que se encuentran en los perfiles. Como se puede apreciar,

las expresiones resaltadas de rojo representan habilidades, mientras que las expresiones resaltadas en azul representan conocimiento.

Tabla 1.2. Ejemplos de competencias encontrados en los perfiles laborales y académicos

Competencias del perfil profesional	Competencias de perfil académico
Diseñar arquitecturas basadas en computación en paralelo.	Conocimientos de software basado en computación distribuida, montaje y utilización de redes de interconexión entre equipos de cómputo.
Abordar proyectos de sistemas informáticos.	Conocimiento de sistemas operativos Linux y Windows, dispositivos periféricos y equipos electrónicos involucrados en el control de procesos industriales.
Interactuar con bases de datos y lenguaje SQL.	Diseñar y administrar sistemas de comunicación de datos, además de la toma de decisiones y en la difusión de las mejores opciones de desarrollo de software.
Diseño de bases de datos en cualquier plataforma.	Evaluar, formular, ejecutar y monitorear proyectos de sistemas computacionales y de procedimientos para la ejecución de los mismos.

Aunque estos enunciados evidencian la presencia de habilidades y conocimiento, algunas oraciones presentan más de un verbo para denotar niveles de habilidad, y utilizan palabras similares para expresar conocimiento. En consecuencia, establecer una comparación de perfiles en base a estos enunciados, implica el desarrollo de alineamientos de tópicos de conocimiento, para encontrar la similitud entre tópicos ambiguos; y un proceso de síntesis de los tópicos de habilidad, con el propósito de seleccionar aquella que represente a la competencia como tal.

Dado que los ejemplos presentados corresponden a casos de ambigüedad de nivel léxico y semántico, es necesario el uso de técnicas del Procesamiento de Lenguaje Natural, Minería semántica y Aprendizaje ontológico que permiten el reconocimiento de los tópicos de conocimiento-habilidad contenidos en los enunciados de competencias, y el tratamiento de los significados ambivalentes, mediante modelos semánticas. Todos estos temas serán abordados en los siguientes apartados de este capítulo.

2.2 PROCESAMIENTO DEL LENGUAJE NATURAL

El procesamiento de Lenguaje Natural (PLN) es el uso de las computadoras para entender los lenguajes humanos, en donde “el proceso de entender” significa que el computador pueda reconocer y usar información expresada en lenguaje humano [51]. Con este propósito, el PLN combina técnicas de la Inteligencia Artificial (IA), aprendizaje automático, inferencia estadística y lingüística aplicada [12], para hacer posible que el computador realice con la información expresada en lenguaje humano, tareas como[1]: análisis de sentimiento, traducción automática, reconocimiento y clasificación de entidades nombradas (por ejemplo, personas, lugares, instituciones, fechas) (NER, por sus siglas en inglés), sistemas de diálogo, entre otras cosas [17].

A continuación, se presenta el procesamiento textual, que trata sobre la problemática, el reconocimiento y representación de unidades de significado, y los procesamientos estadísticos y lingüísticos de texto. Luego, se explican los tres modelos para el análisis y clasificación de términos y entidades y, por último, se revisan las características de los cinco tipos de análisis en PLN que pueden realizarse sobre el texto.

[1] PLN, tomado de http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural

2.2.1. Procesamiento Textual

El procesamiento textual consiste en el traslado de los textos de un documento hacia una forma intermedia, que permita su tratamiento mediante procesos computacionales [11]. Esta forma intermedia debe garantizar la integridad en la representación del documento, preservando información relevante como, por ejemplo, entidades, relaciones entre conceptos, etc. [51].

Existen métodos para la estructuración de un documento, por ejemplo, sus unidades de significado (palabras, términos, oraciones, párrafos), las porciones de texto en función de un concepto, entre otros [11]. Una vez definidos los fragmentos de texto, se realiza un análisis sintáctico, que puede ser [53]: un análisis parcial de las oraciones para estudiar las características sintácticas de las oraciones; un análisis superficial (shallow parsing) para determinar la estructura morfológica de las frases; el troceado (chunking) para obtener segmentos de texto no superpuestos que contienen partículas nominales o verbales, entre otros. Finalmente, con los textos extraídos se procede a la creación de una forma intermedia de representación. La Tabla 2.3. presenta un resumen de las típicas formas intermedias, con una descripción de su uso, en función de sus características [13], y en el apartado 2.2.2.1. se describe la forma intermedia basada en los n-gramas.

Tabla 2.3. Formas de procesamiento textual

Forma Intermedia	Características
Palabra	Unidad mínima de significado, se obtiene por medio de Tokenización (Pre-procesamiento mínimo). No representa el conocimiento semántico del documento.
Bolsa de palabras	El documento es una colección de palabras claves, con una representación binaria (0 ausencia, 1 presencia) y real (frecuencia de aparición de la palabra). Pre-procesamiento mínimo, requiere el manejo de sinónimos (diccionarios).
Concepto (término)	Representación de un objeto en un contexto específico, y presenta dependencia del contexto
Taxonomía de términos	Clasificación jerárquica de términos, con asociaciones de diferente granularidad,
Grafo semántico	Representación de conceptos como nodos, y sus relaciones semánticas como arcos,
Red “IS-A”	Representación jerárquica de conceptos,
Grafo conceptual	Representación de conceptos por tipos y sus relaciones conceptuales,
Ontología	Marco conceptual de conceptos y sus relaciones,
Frase	Secuencia de palabras con cierto nexo sintáctico
N Frases	Secuencia de N elementos, donde cada uno es una frase,
Párrafo	Conjunto de frases que concluyen en un punto y aparte,
Documento	Conjunto de párrafos, frases, términos, con una misma idea fundamental
N-grama	Representación distribuida de las palabras de un documento

2.2.1.1. Problemáticas en el procesamiento de lenguaje natural

Los problemas que se presentan en el PLN tienen mucho ver con la ambigüedad que presentan los textos, y las representaciones que se pueda lograr del mismo durante los diferentes niveles de análisis que se realizan en el procesamiento (ver Tablas 2.5. y 2.7.). Según [13], estos problemas se pueden resumir en los siguientes grupos:

Ambigüedad: la cual se presenta a diversos niveles:

1. **Léxico,** ya que una palabra puede tener varios significados (por ejemplo, banco de mueble y banco de institución); se puede dar solución a todos los casos de ambigüedad mediante el uso de diccionarios, gramáticas, bases de conocimiento, y correlaciones estadísticas [12].
2. **Estructural,** ya que de una oración con más de un sintagma preposicional[2], se genera más de un árbol sintáctico [54]. Son dos los ámbitos de investigación en este caso: a. Adjunto de sintagmas preposicionales (ej: [(sistemas operativos)][(para móviles)], en donde la preposición "para" determina el inicio del sintagma preposicional); y, b. Preposiciones en expresiones multipalabra (ej: sistemas de comunicación de redes distribuidas, en donde las preposiciones "de" generan árboles como [sistemas]{[(de, comunicación)] [(de, redes, distribuidas)]} o [(sistemas, de, comunicación)] [(de, sistemas, distribuidos)] [12]. Para resolver este tipo de ambigüedad se usa aprendizaje automático basado en corpus, tales como técnicas como árboles de decisión, redes neuronales, entre otros. [55]
3. **Pragmático,** donde el significado semántico de la oración no es el descrito por la estructura lingüística de la misma [14]. Las soluciones a este problema se centran en mantener la coherencia del discurso, y para ello, se combinan técnicas de PLN (análisis morfo-léxico y sintáctico) y Aprendizaje Automático (LSA[3], entre otros), para analizar la intención de texto [56].

Detección de separación entre las palabras: El PLN necesita la separación de los textos en unidades de significado (Tabla 2.3.), como el caso de palabras y términos, de tal forma que se mantenga un sentido lógico, tanto gramatical como contextual [13,55].

Captura errónea de palabras: que provocan variaciones en el lenguaje debido a, por ejemplo: acentos extranjeros, dialectos, modismos, errores de digitación, errores en la lectura de textos, entre otros [55].

2.2.1.2. Reconocimiento y extracción de texto

Las técnicas de NLP son muy utilizadas, tanto para facilitar la descripción del contenido de los documentos, como para representar la consulta formulada por el usuario [57, 58]. En particular, para lo último, puede ser usado en sistemas de recuperación de información, los cuales, según [15], llevan a cabo las siguientes tareas (ver Fig. 2.1.):

1. Indexación de la colección de documentos, mediante la representación de cada documento como un vector que reúne un conjunto de términos relevantes;
2. Representación de la consulta, donde la pregunta del usuario se analiza con el fin de identificar la necesidad de información del usuario;
3. Alineamiento, en donde el sistema compara cada documento con la consulta, y presenta al usuario los documentos cuyas descripciones más se asemejan a la descripción de su consulta.
4. Ordenamiento, en donde los resultados suelen ser mostrados en función del grado de similitud entre las descripciones de los documentos y la consulta. La Figura 2.1. muestra en detalle este proceso.

[2] Sintagma preposicional: está formado por una preposición seguida de un sintagma nominal (sustantivos o preposiciones)

[3] LSA: Latent Semantic Analysis (Análisis Semántico Latente)

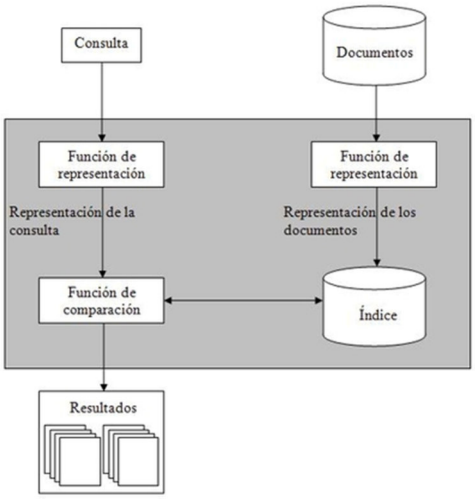

Figura 2.1. Arquitectura de un sistema de recuperación de información [15].

La aplicación de las técnicas de PLN en la extracción y análisis de información textual, se puede realizar desde dos enfoques: Procesamiento estadístico y Procesamiento lingüístico.

2.2.1.3 Procesamiento estadístico del lenguaje natural

Según [14], se caracteriza porque cada documento está descrito por un conjunto de palabras clave, denominadas términos índices o "bolsa de palabras" (o "bag of words"), en donde todas las palabras de un documento son términos índices para ese documento, las cuales son ponderadas, normalmente, por su frecuencia de aparición en el documento [15]. En este modelo, el procesamiento de los documentos consta de las siguientes etapas:

1. ***Preprocesamiento*** de los documentos: consiste fundamentalmente en preparar los documentos para su parametrización, eliminando aquellos elementos que se consideran superfluos (etiquetas, etc.). También, aplica modelos de lenguaje (normalización) sobre los documentos, como es el caso de n-gramas, los cuales son secuencias contiguas de n elementos de una unidad de significado (párrafos, oraciones). Los elementos pueden ser fonemas, sílabas, letras, palabras, entre otros [17]; este modelo de lenguaje se explica en el apartado 2.2.1.

2. ***Parametrización:*** una vez se han identificado los términos relevantes, se realiza el paso de la colección de documentos en una matriz documento-término, donde cada fila es un documento (vector de características), cada columna es una unidad de significado (término) y el valor asignado refleja la importancia que produce el término [22]. Para ello, se realiza una cuantificación estadística de las características (términos) de los documentos, es decir, la aparición de los mismos en los documentos de la colección, la cual puede ser binaria (1 presencia, 0 ausencia) o su frecuencia (número de apariciones del término en el documento) [14] (ver Figura 2.3.).

En esta fase, se pueden incluir ponderaciones de los términos según el propósito del análisis, como por ejemplo, establecer la relevancia de un término en función de la frecuencia de

aparición del mismo en la colección de documentos; para este propósito se usan las funciones TF-IDF[4][15] y Okapi BM25[5] [60], en las que se establece un valor de relevancia para el término en función de su frecuencia en el documento y de su importancia para todos los documentos de una colección; estas funciones se explican en el apartado 2.2.1.5.

2.2.1.4. Procesamiento lingüístico del lenguaje natural

Según [8], se fundamenta en la aplicación de técnicas y reglas que codifican de forma explícita las características lingüísticas de los textos, los cuales son analizados según los niveles lingüísticos[6], con herramientas que incorporan al texto las anotaciones propias de cada nivel [15]. La Figura 2.2. presenta el conjunto de elementos que comprende el procesamiento lingüístico de lenguaje natural dentro del esquema de trabajo de Stanford CoreNLP [16]. Como se observa, el texto sin procesar ingresa a un flujo de ejecución que comprende: Preprocesamiento, donde el documento se divide en oraciones (Sentence Splitting) y en palabras (Tokenización); Análisis Morfológico (Etiquetado gramatical), que determina la forma, clase o categoría gramatical de cada palabra; Análisis Sintáctico, donde se establecen las funciones de las palabras o grupos de palabras dentro de la oración; y, Análisis Semántico, que estudia a las palabras en función de su significado (Reconocimiento de entidades nombradas, Resolución de co-referencias). El resultado se refleja en el texto enriquecido con anotaciones. En el apartado 2.3. se explica con mayor detalle estos análisis.

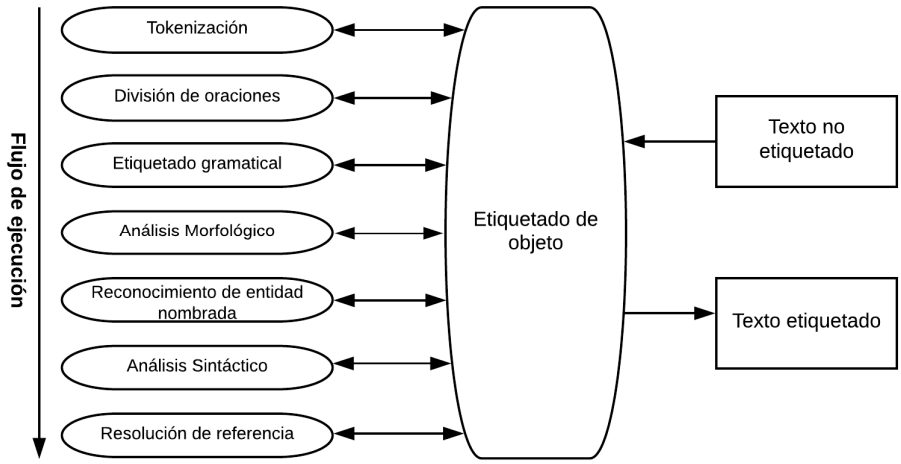

Figura 2.2. Procesamiento lingüístico del lenguaje natural [16]

2.2.1.5. Filtrado de términos

Otro aspecto relevante del procesamiento estadístico del lenguaje natural es el filtrado de términos, el cual se utiliza para la selección, obtención, indexado y cálculo de relevancia de información [15]. Generalmente, se utilizan vectores de características de documentos

[4] TF-IDF: Frecuencia del término – Frecuencia Inversa del Documento.
[5] Okapi BM25: Variación de TF-IDF que considera la longitud de los documentos y la frecuencia de términos.
[6] Nivel lingüístico: nivel de análisis del lenguaje natural: fonético, morfológico, sintáctico y semántico.

(palabras, términos, frases o conceptos), espacios lineales multidimensionales [14], y esquemas de ponderación de términos, para encontrar la posición que ocupa un término en relación a un documento o grupos de documentos [22]. Entre los esquemas utilizados se encuentran: TF-IDF[7][15], PMI [23] y Okapi BM25[8] [60], los cuales se explican a continuación:

TFD-IDF: El esquema de pesos más usado es la frecuencia de aparición de los términos en los documentos (Term Frequency / Inverse Document Frequency; TF-IDF) para expresar el peso relativo del rasgo o término w en el vector asociado a un documento d, y se calcula según la ecuación (2.1) , donde idf(w) se calcula según la ecuación (2.2).

$$tfidf(w,d) = tf(w,d) * idf(w) \quad (2.1)$$

$$idf(w) = log\frac{N}{df(w)} \quad (2.2)$$

En donde tf(w,d) es la frecuencia del término (cantidad de ocurrencias de la palabra w en un documento d), idf(w) es la frecuencia inversa de documentos (cantidad de documentos donde aparece la palabra w pero de forma inversa, debido a que se le otorga mayor peso a las palabras que ocurren en una menor cantidad de documentos), df(w) es la frecuencia de documentos (cantidad de documentos que contienen la palabra w), y N representa la cantidad total de documentos en el corpus [62,16].

PMI: La información mutua puntual (Pointwise mutual information) es una medida derivada de la teoría de la información, que usa estadísticas a nivel de corpus para reflejar el significado de las coocurrencias. Ha sido utilizada, por ejemplo, para medir la orientación semántica de frases [22]. Dada una palabra w y otra palabra v, PMI entre w y v se define en la ecuación (2.3) como:

$$PMI(w,v) = log\ \frac{(p(w,v))}{p(w)p(v)} \quad (2.3)$$

Donde p(w, v) es la probabilidad de que w y v co-ocurran, por ejemplo, en un mismo contexto, y p(w) y p(v) son las probabilidades de aparición de las palabras w y v, respectivamente [63,23].

Okapi BM25: es una función de ranking de documentos, se basa en el concepto de bolsa de palabras mediante al cual se ordenan los documentos de una colección según su relevancia con respecto a un término dado [60]. Al igual que TF-IDF, Okapi BM25 utiliza la frecuencia con que el término se presenta en los documentos [62], pero es más sensible a la variación en la longitud del documento, lo cual hace que los procesos de filtrado de términos sean más precisos en relación a TF-IDF [16]. De esta forma, la función Okapi BM25 se define mediante las ecuaciones (2.4) y (2.5):

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_iD) + k_1.(1-b+b.\frac{|D|}{avgdl})} \quad (2.4)$$

$$IDF(q_i) = \log\frac{N - n'(q_i) + 0.5}{n'(q_i) + 0.5} \quad (2.5)$$

[7] TF-IDF: Frecuencia del término – Frecuencia Inversa del Documento.
[8] Okapi BM25: Variación de TF-IDF que considera la longitud de los documentos y la frecuencia de términos.

Donde $f(q_i, D)$ es la frecuencia de aparición en el documento D de los términos que aparecen en la consulta Q, $|D|$ es la longitud del documento D (en número de palabras), y avgdl es la longitud media de los documentos en la colección sobre la cual estamos realizando la búsqueda. k_1 y b son parámetros que permiten ajustar la función a las características concretas de la colección, estandarizando los valores $k_1 = 2.0$ o 1.2 y $b = 0.75$ [64]. $IDF(q_i)$ es el peso IDF (inverse document frequency) de las palabras clave que aparecen en la consulta, donde N es el número total de documentos en la colección, y $n'(q_i)$ es el número de documentos que contienen la palabra clave q_i.

2.2.1.6. Generación de textos

La generación de textos es "el subcampo de inteligencia artificial y lingüística computacional que se ocupa de la construcción de sistemas informáticos que puedan producir textos comprensibles en inglés, u otros idiomas humanos, desde una representación no lingüística subyacente de información" [117]. Desde este punto de vista, se pueden dar dos tipos de generación: de **texto a texto,** en donde se toman textos como entrada y se genera un nuevo texto, y, de **datos a texto**, en la cual, a partir de datos se generan nuevos textos [18]. Las aplicaciones de estos dos tipos de generación son en el ámbito de la traducción automática, fusión y resumen de textos, corrección ortográfica, reportes de documentos (noticias, etc.), entre otros.

Según [117, 18], un sistema de generación comprende los siguientes componentes:

1. ***Planificador de textos:*** donde se decide qué información debe incluirse, y se determina en qué orden se presentará en el texto en construcción.
2. ***Planificador de oraciones:*** donde se decide qué presentar en frases individuales, encontrando las palabras y frases correctas para expresar la información.
3. ***Realizador de párrafos:*** donde se seleccionan las palabras y frases para identificar/describir/analizar el dominio objeto de estudio, para lo cual se combinan todas las palabras y frases.

2.2.2. Modelado del Lenguaje

El modelado se constituye en un mecanismo para definir la estructura del lenguaje, con el propósito de identificar las unidades de significado, y manejar la complejidad de la sintaxis o semántica del texto [18]. Entre los modelos clásicos del lenguaje se encuentran: los modelos probabilísticos y los modelos lógicos.

2.2.2.1. Modelos probabilísticos

Los modelos probabilísticos se pueden clasificar en los siguientes:

2.2.2.1.1 Basado en conteo

Definen una distribución de probabilidad sobre unidades de significado (palabras, términos), en función de la ocurrencia de las mismas en los textos [14]. El método más conocido es el de n-gramas. Un modelo n-grama usa las n-1 palabras anteriores para predecir la siguiente. Para ello, se requiere de la estimación a priori de una secuencia de palabras [18]. Los modelos n-gramas más utilizados son: unigramas, que se restringen a la probabilidad de una sola palabra; bigramas, en donde una palabra es estadísticamente dependiente de la palabra temporalmente anterior, y trigramas, en donde una palabra es estadísticamente dependiente de las dos palabras temporalmente anteriores [14]. La Figura 2.3. presenta un ejemplo de la obtención de N-gramas,

en el cual se presentan los 3 casos típicos: unigramas (N=1), bigramas (N=2) y trigramas (N=3); como se observa, los unigramas representan una palabra, mientras que los bigramas y trigramas grupos de 2 o 3 palabras que se encuentran en secuencia dentro de la unidad de significado (en este caso, una oración).

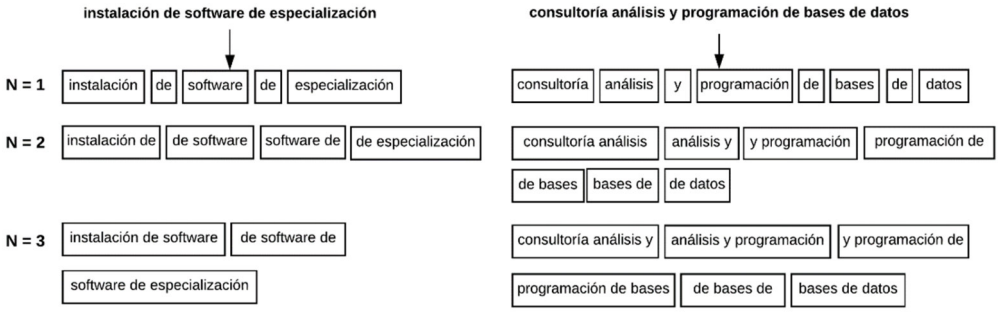

Figura 2.3. Generación de N-gramas

La Figura 2.4. presenta un ejemplo del procesamiento estadístico de dos documentos d1 y d2 [17], usando los términos obtenidos en la Figura 2.3, del proceso de formación de N-gramas [14]. Para cada término se obtiene el número de apariciones en los documentos, que se consolida en una matriz de frecuencias con los términos en d1 y d2. Como se observa, el espacio vectorial para d1 y d2 muestra la frecuencia de cada término trasladada a coordenadas (ej: programación (7,7)). Se puede establecer una ponderación sobre la relevancia de los términos, en función de la distancia que existe entre ellos; por ejemplo, usando la distancia euclidiana entre dos puntos (la raíz cuadrada de la sumatoria de la diferencia al cuadrado de las coordenadas en los ejes de dos puntos). Los resultados se reflejan en la matriz de distancias, que contiene en la intersección de filas y columnas la distancia por cada par de puntos, Como se observa, el par "***análisis-programación"*** tiene una distancia menor en relación a otros pares de términos (**1,41**); le sigue ***"análisis- bases de datos***" (**2,00)**, y el par con mayor distancia es ***"instalación-programación"*** (**6,40**).

Matriz de frecuencias

palabras	d1	d2
instalación	3	2
software	4	5
especialización	4	7
consultoría	2	7
análisis	6	6
programación	7	7
bases de datos	6	4

Matriz de distancias euclidianas para d1 y d2

	instalación	software	especialización	consultoría	análisis	programación	bases de datos
instalación	.....						
software	3,16	.....					
especialización	5,10	2,00	.....				
consultoría	5,10	2,83	2,00	.....			
análisis	5,00	2,24	2,24	4,12	.....		
programación	6,40	3,61	3,00	5,00	1,41	.....	
bases de datos	3,61	2,24	3,61	5,00	2,00	3,16	.....

Figura 2.4. Modelo de lenguaje basado en conteo

2.2.2.1.2 De espacio continuo

Los modelos de espacio continúo predicen que palabras son cercanas entre sí [18]. Estos modelos resuelven el problema de la escasez de datos del modelo N-gramas, representando palabras como vectores (Word embeddings: incrustaciones de palabras)[9] mediante una red neuronal [118]. Las incrustaciones de palabras obtenidas a través de estos modelos demuestran que las palabras con una similitud semántica son cercanas en el espacio vectorial inducido por el modelo [22]. Los modelos de espacio continúo son de dos tipos: 1. ***Basado en la red neuronal hacia adelante*** (Feed-Forward Neural Network Based Models), que trata con los problemas de escasez de datos; y 2. ***Basado en la red neuronal recurrente*** (Recurrent Neural Network Based Models), que aborda el problema del contexto limitado de los documentos.
La Figura 2.5. presenta un ejemplo de este modelo, donde el objetivo es determinar el vecino de la palabra ***x***. Para ello, ocurren los siguientes pasos: **(1)** cada palabra es representada como un ***one-hot vector*** (vector de 0 y 1 con tantas posiciones como tamaño tenga la unidad de significado), donde el valor 1 indica la posición de ***x*** en la unidad de significado (ej: en el conjunto de palabras, rey está en la posición 1). **(2)** La tabla contiene los vectores one hot que serán la entrada del modelo (palabra one hot vector), así como la salida esperada (vecino one hot vector); los vectores se generan según el tamaño de la ventana, en este caso 2, que definen la vecindad de las palabras (por ej. "rey" es vecino de "valiente" y de "hombre", "valiente" es vecino de "rey" y de "hombre"; y, "hombre" es vecino de "valiente" y de "rey"). **(3)** En el entrenamiento, cada vector ingresa en el modelo neuronal, y la capa oculta de la red (en este caso de dos neuronas) establece la probabilidad de vecindad de las palabras mediante una medida de entropía cruzada (Softmax cross entropy)[10]. **(4)** El resultado es un modelo de espacio vectorial de dos dimensiones que representa a las palabras vecinas (word embeddings); por ejemplo, la vecindad de las palabras rey y hombre se muestra en pares de coordenadas que determinan su posición dentro de la unidad de significado (oración) [97].

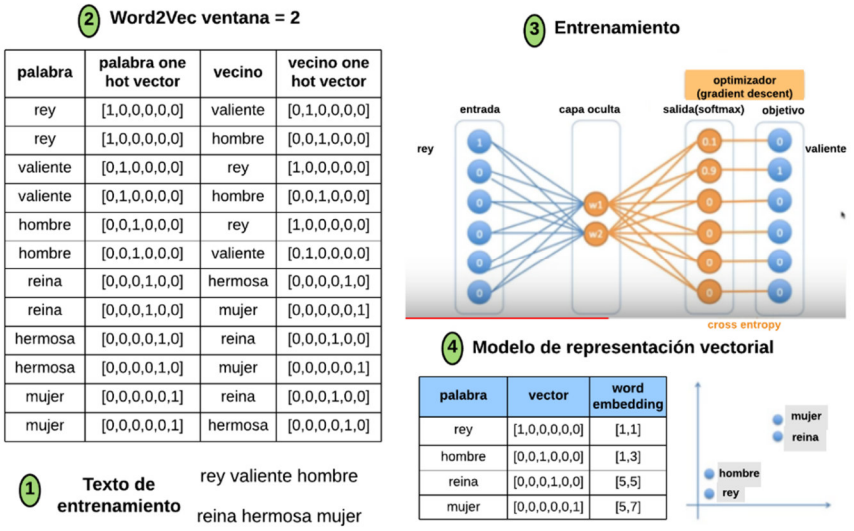

Figura 2.5. Modelo de lenguaje de espacio continuo [118]

[9] Word Embeddings: representación más popular del vocabulario de documentos. Es capaz de capturar el contexto de una palabra en un documento.
[10] Cross entropy: la entropía cruzada entre dos distribuciones de probabilidad p y q sobre el mismo conjunto de eventos, mide la capacidad de la distribución p para predecir un evento que ocurre en la distribución q.

2.2.2.2 Modelos condicionales

Un Campo Aleatorio Condicional (Conditional Random Field, CFR) es un modelo estocástico utilizado habitualmente para etiquetar y segmentar secuencias de datos, o extraer información de documentos [119]. Dada una secuencia de datos, el CFR asigna una etiqueta para cada elemento, y calcula la probabilidad de la secuencia correcta de etiquetas, condicionada por las observaciones. Se puede representar con un grafo no dirigido, en el que cada vértice represente una variable aleatoria (palabra), y cada arista indique una dependencia entre las variables en los vértices que conecta (pudiéndose establecer una probabilidad condicional entre ellas). La Figura 2.6. presenta un ejemplo donde se tiene una entrada de varias palabras "El Sr. Barack Obama habló en New York, Obama dirige la ONU", el objetivo es asignar etiquetas de entidades a las palabras (ej: persona, lugar, organismo, acción, conector). Entonces, se establece la probabilidad de que la etiqueta de la palabra sea ***"persona"***, y se observa las dependencias de la palabra (ej: palabra en mayúscula, palabra previa es Sr.). De esta forma, se asigna la etiqueta "persona" (BP) a la palabra ***x***.

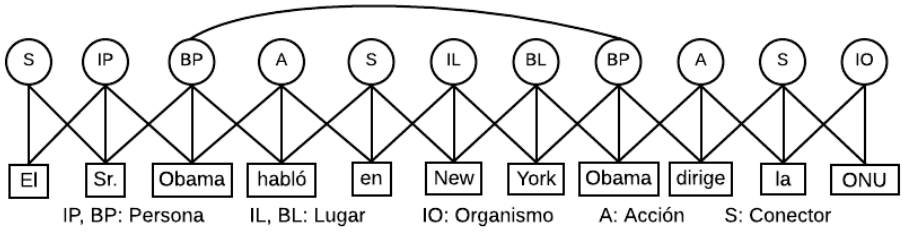

Figura 2.6. Modelo de lenguaje condicional (CRF)

2.2.2.3. Modelos lógicos

Los modelos lógicos representan las restricciones del lenguaje, de tal forma que permiten modelar dependencias lingüísticas[11] [121]. Existen dos tipos de modelos: 1. Gramáticas, que reúnen las reglas y elementos que definen un lenguaje y, 2. Ontologías, que permiten la representación de conceptos y relaciones de un dominio (las cuales se explican en el apartado 2.4.5 Aprendizaje Ontológico).

2.2.2.3.1. Gramáticas formales

Una gramática formal "es una estructura matemática con un conjunto de reglas de formación que definen las cadenas de caracteres admisibles en un determinado lenguaje formal o lengua natural" [128]. Sobre esta estructura, se determina un relativo orden de las palabras en las oraciones [18]; es por esto que, a una gramática formal se le considera como una generadora de lenguajes, o como base para un "reconocedor": una función que determina si una cadena cualquiera pertenece a un lenguaje, o es gramaticalmente incorrecta [54]. De esta forma, una gramática debe tener la capacidad de generar un conjunto de descripciones estructurales que representen todas las frases de una lengua [122].

En el contexto del PLN, las gramáticas se fundamentan en dos tipos de componentes:

[11] Dependencia lingüística: es la noción de que una unidad lingüística (ej: palabra) está conectada con otras dentro de la oración por enlaces directos.

1. ***Constituyente***: que es una palabra, o secuencia de palabras, que funciona en conjunto como una unidad dentro de la estructura jerárquica de una oración, la cual está definida por reglas recursivas que conectan categorías sintácticas (clasificación de las palabras de acuerdo a la función que cumplen dentro de una oración; ej: sustantivo, verbo, etc.), unidas por reglas recursivas que definen unidades de significado. Se representan mediante un árbol sintáctico[12] [128].
2. ***Dependencia:*** es la noción de que una unidad sintáctica (ej: palabra) está conectada con otras dentro de la oración por enlaces directos. Se toma al verbo como centro de la estructura, y las demás palabras están conectadas directa o indirectamente al verbo [14]; estás conexiones se determinan en función de un corpus anotado sintácticamente (ej: Treebank de dependencias[13]) [54].

Existen diferentes tipos de gramáticas formales, [128]: categoriales (C-gramáticas), de estructura sintagmática (ES-gramáticas) y gramáticas asociativas (A-gramáticas) [121]. Según la complejidad estructural del lenguaje, las ES-gramáticas se clasifican siguiendo la jerarquía de Chomsky en tipos 0 a 4, siendo las de menor tipo las menos complejas y las del tipo 1 las usadas por los lenguajes naturales [122]. La Tabla 2.4. presenta un resumen de las características de los tres tipos de gramáticas.

Tabla 2.4. Características de las gramáticas formales

Gramática	Característica	Reglas de producción
Categoriales ES-gramáticas	Formada por Constituyentes. Se representa por Árbol sintáctico	**Oración:** Sintagma Nominal y Sintagma Verbal O → SN + SV **Sintagma Nominal:** Determinante, Sustantivo y Complemento SN → Det + S + C **Sintagma Verbal:** Auxiliar y Grupo Verbal SV → Aux + GV **Grupo Verbal:** Verbo y Complementos GV → V[+ Complementos] **Complemento Directo:** Sintagma Nominal o Sintagma Preposicional CD → SN + SP **Complemento Indirecto:** Sintagma Preposicional CI → SP **Sintagma Preposicional:** Preposición y Sintagma Nominal SP → Prep + SN
A-gramáticas	Usan dependencias lingüísticas. Se representan por un grafo de dependencias	**Dependencia Morfológica:** Una palabra o parte de ella, influye en la forma de la otra palabra. **Dependencia Sintáctica:** El verbo determina la estructura de las palabras dependientes. **Dependencia Semántica:** Los argumentos[14] de un predicado[15] son semánticamente dependientes de ese predicado.

Para explicar el uso de las gramáticas, la Figura 2.7 presenta 2 ejemplos de análisis de la oración ***"gerenciar centros de cómputo";*** usando las consideraciones de la Tabla 2.4. En el árbol sintáctico se observa como los componentes de la oración, **habilidad** y **conocimiento,** se asocian con sintagmas verbales y nominales, respectivamente, y así, según las reglas de producción se identifican constituyentes (SP, V, S, Prep) y sus instancias en el texto. En el grafo de

[12] Arbol Sintáctico: es la representación gráfica de los diferentes constituyentes de una oración.
[13] Universal Dependencies Treebank, https://universaldependencies.org/
[14] Argumento: es la expresión que completa el significado de un preficado.
[15] Predicado: unidad semántica que toma uno o más argumentos y los relaciona entre sí.

dependencias se observa una dependencia sintáctica al verbo **gerenciar,** que es la cabeza de la estructura, y una dependencia semántica entre el predicado **gerenciar** y los argumentos **centros** y **centros de cómputo**, donde el significado puede ser ***gerenciar centros*** o ***gerenciar centros de cómputo***.

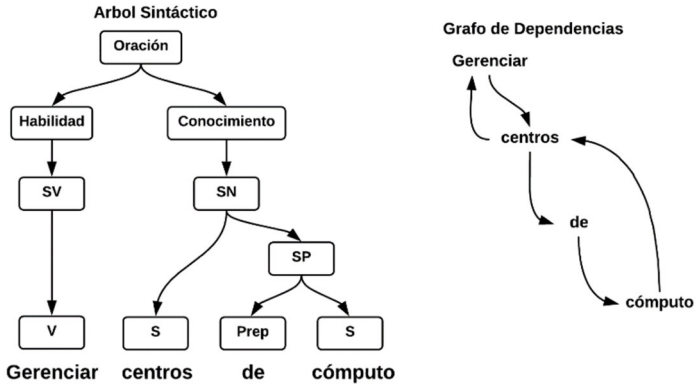

Figura 2.7. Aplicación de gramáticas en minería de texto

2.2.3. Tipos de análisis en el PLN

Para los autores de [54], el método más explicativo para presentar lo que sucede dentro de un sistema de PLN es mediante el enfoque de niveles de análisis del lenguaje. Según [14], se identifican cuatro niveles de análisis del lenguaje que derivan cuatros tipos de análisis: morfológico, sintáctico, semántico, y pragmático. Recientemente, se ha añadido el análisis sentimental.

2.2.3.1. Análisis morfológico

La morfología considera la estructura interna de las palabras (sufijos, prefijos, raíces, flexiones), y el sistema de categorías gramaticales de los idiomas (género, número, etc.) [18]. Consiste en determinar la forma, clase o categoría gramatical de cada palabra dentro de una oración [128]; el cual es ejecutado por etiquetadores (taggers) que asignan a cada palabra su categoría gramatical, a partir de los rasgos morfológicos identificados [16]. Por ejemplo, el análisis morfológico de la oración de la Figura 2.7, revela que cada palabra tiene información de su estructura interna, por ejemplo "gerenciar" (***raíz*** gerencia y ***flexión*** r), así como de su categoría gramatical (***tipo:*** sustantivo, ***género:*** femenino, y ***número:*** singular).

2.2.3.2. Análisis Sintáctico

Identifica la presencia de unidades lingüísticas superiores, como oraciones, constituyentes o sintagmas, para determinar la estructura sintáctica del texto. En este punto se aplican gramáticas (parsers), cuya estructura depende del objetivo del análisis sintáctico, que conlleva el estudio de todas las combinaciones de las categorías gramaticales. Por ejemplo, un *análisis superficial* (Shallow Parsing) identifica las estructuras más significativas: sintagmas nominales, sintagmas verbales y preposicionales, entidades; y un *análisis completo* (Full Parsing), además de obtener todos constituyentes, identifica "chunks" (fragmentos de información con significado, ej: entidades) [14]. También, el análisis puede darse considerando constituyentes y dependencias, como se muestra en el ejemplo de la Figura 2.7.

2.2.3.3. Análisis Semántico

Estudia el significativo que cada elemento tiene por sí mismo dentro de la oración; obtiene la representación semántica de las frases a partir de los elementos que la forman. Para ello, se realiza el enriquecimiento de los conceptos en base a diccionarios, tesauros y taxonomías; como es el caso WordNet Wikipedia y Dbpedia [23]. Por ejemplo, en la oración de la Figura 2.7 "**Gestionar centros de cómputo**", se usa un diccionario para encontrar los sinónimos de ***gestionar*** (**dirigir, administrar**) y ***centros*** (**salas, aulas**) y, de esta manera enriquecer la oración. Este proceso es común en el desarrollo de búsquedas semánticas en la Web [58].

2.2.3.4. Análisis pragmático.

Estudia la interpretación de la estructura de la oración, para determinar el verdadero significado dentro de un contexto específico [54]. El análisis pragmático considera el análisis de correferencias (anáforas), las cuales se ocupan de hacer coincidir los pronombres con los sustantivos o nombres a los que se refieren mediante un análisis de dependencias; ej: **"María" fue a clase, "ella" tomó el autobús**; un estudio de las dependencias sintácticas en la oración alinea al sustantivo ***María*** con el pronombre ***ella*** [18]. Otro caso es el análisis de discurso, en el cual se estudia a la metáfora, que consiste en el desvío del significado de un término; ej: **"nos venían metiendo la mano al bolsillo"**, significa ***"sacarnos lo que no les pertenece"*** [128].

2.2.3.5. Análisis de sentimiento

El análisis de sentimiento busca la identificación de la actitud de un interlocutor o un escritor con respecto a algún tema contextual general de un documento [17]. También conocido como *minería de opinión*, trata del estudio de las opiniones, actitudes y emociones, de las personas hacia individuos, eventos o temas [123]. Desde el punto de vista de la minería de texto, el análisis de sentimientos es una tarea de clasificación masiva de documentos de manera automática, en función de su **"Polaridad"** (es decir la connotación positiva, negativa o neutra del lenguaje usado en el documento) [124]. Otra clasificación del sentimiento más avanzada se hace según estados emocionales ("enfado", "tristeza", o "felicidad") [17].

Según [123], el análisis de sentimiento puede hacerse de tres maneras:

1. ***Nivel de documento,*** en donde se considera al documento como una unidad básica de información, que expresa una opinión o sentimiento positivo o negativo;
2. ***Nivel de oración,*** se clasifica el sentimiento expresado en cada oración, según la polaridad de las palabras que la conforman, y;
3. ***Nivel de aspecto,*** se clasifica el sentimiento de las oraciones según aspectos específicos de entidades (personas, instituciones), por ejemplo, la popularidad de un político según sus características.

Según [17], hay cuatro enfoques en el análisis de sentimiento:

1. Localización de palabras clave, se clasifica el texto en categorías de afecto según la presencia de palabras no ambiguas (ej: ira, alegría, tristeza) propias de cada categoría, tomadas desde por ejemplo un diccionario de sentimientos, o de la misma colección de textos;
2. Afinidad léxica, se detectan no solo palabras de afecto obvias, sino que se asigna a palabras arbitrarias una probable "afinidad" a emociones particulares;

3. Métodos estadísticos, ayudan a clasificar aquellas palabras ambiguas en una categoría de afecto, mediante técnicas de aprendizaje automático (LSA[16], BOW[17], entre otras);
4. Técnicas a nivel de concepto, tratan de obtener la característica sobre la cual se opinó mediante dependencias sintácticas, combinadas con modelos de conocimiento (ontologías y redes semánticas) para detectar información relevante pero no totalmente visible, ej: en la oración **"Me gusta el Iphone 9",** se establece la polaridad positiva hacia la entidad Iphone 9 y sus características implícitas (portable, seguro, etc.).

La Tabla 2.5 presenta un resumen de los niveles de análisis, en función a la ambigüedad detectada en los textos.

Tabla 2.5. Características de los componentes de PLN

Nivel	Objetivo	Caso de Ambigüedad
Morfológico (Léxico)	Formar palabras, derivar unidades de significado	Ambigüedad léxica: una palabra puede tener más de una categoría gramatical.
Sintáctico	Formar oraciones	Ambigüedad estructural: posibilidad de asociar a una oración más de una estructura sintáctica.
Semántico	Derivar significado de oraciones (independiente)	Polisemia: una palabra puede tener más de un significado. Sinonimia: posibilidad de utilizar términos distintos para representar un mismo significado.
Pragmático	Derivar significado de oraciones relativo al significado que se observa en el contexto	Metáfora: no puede realizarse una interpretación literal y automatizada de los términos utilizados. Anáfora: presencia en la oración de pronombres y adverbios que hacen referencia a algo mencionado con anterioridad
Sentimental	Derivar polaridad de palabras u oraciones	Palabras con categoría de afecto ambiguo

2.3. MINERÍA SEMÁNTICA

La minería semántica se encarga de extraer conocimiento semántico desde diferentes fuentes semánticas, como lo son páginas web, contenidos sin estructura en la web, contenidos estructurados en la web, grafos anotados, ontologías, entre otros [17]. La Minería Semántica se divide en tres grandes grupos:

1. ***Minería de datos semántica (MDS):*** nace con el propósito de incorporar contenido semántico a los datos [8]. La MDS comprende dos pasos: 1. *Enriquecimiento semántico*, el cual usa ontologías o fuentes semánticas para realizar el mapeo de los datos, y enriquecerlos por medio del alineamiento con equivalencias; y 2. *Identificación de patrones* que permiten aportar un nuevo conocimiento al contexto, aplicando técnicas de Minería de Datos.
2. **Minería de la web semántica (MWS):** nace de la integración de dos áreas de conocimiento, como lo son la web semántica y la minería web. Según [7], la web semántica es usada para darle significado a los datos que se encuentran en la Web, es expresada en lenguajes como

[16] LSA: Latent Semantic Analysis
[17] BOW: Bolsa de Palabras

OWL, RDF, e incluso XML. Por otro lado, la minería web comprende la aplicación de técnicas de Minería sobre los datos en la web para la extracción de patrones, y pueden ser:

- Minado de contenido, es una forma de Minería de Texto, que se aplica al contenido desplegado en la Web.
- Minado de la estructura, estudia el enlace de las páginas web usando técnicas de análisis de grafos, para extraer conocimiento de qué páginas son más centrales, puntos débiles de conectividad, entre otros.
- Minado del uso de la web, se enfoca en minar el historial de uso de los usuarios; se busca descubrir los patrones de comportamiento en las consultas que hacen a una página, los movimientos que se hacen entre páginas, entre otros.

2.3.1. Web Semántica

La Web Semántica es una extensión de la actual Web, en la cual la información se da con un significado en lenguajes universales, lo que facilita que la información sea procesada por las máquinas [61]. Esto implica que los datos descritos y enlazados para establecer una semántica, siguen estructuras gramaticales y lingüísticas definidas [7]. Por ello, la Web Semántica proporciona mayor significado a la información, lo que garantiza una mayor eficiencia en el enlace y búsqueda de recursos en la Web [17].

Uno de los componentes relevantes dentro de la Web Semántica son los datos enlazados, que reúnen un conjunto de prácticas para publicar e interconectar datos estructurados en la Web, las cuales, según [17], son:

1. Utilizar URIs para identificar las cosas;
2. Aprovechar el HTTP de la URIs para buscar información a través de esos nombres;
3. Incluir enlaces a otros URIs, para que puedan descubrir más cosas.

Entre las tecnologías de la web semántica que se usan para el enlazado de datos, se encuentran: 1. Lenguajes de modelado, como RDF, RDFSchema y OWL; y 2. Lenguajes de consultas como SPARQL [17]. Además, la Web Semántica considera el uso de vocabularios LOD (Linked Object Data, por sus siglas en inglés), con el propósito de estandarizar la representación de la información.

2.3.2 Fuentes Semánticas

Una forma de resolver casos de ambigüedad textual es alinear las unidades de un texto con estructuras semánticas conocidas, obteniendo así su similitud [23]. Las fuentes semánticas normalmente utilizadas son: tesauros, taxonomías y diccionarios de preferencia dentro del mismo contexto e idioma [17]. Tomando en consideración el dominio de las competencias, existen pocas fuentes semánticas que permitan la desambiguación de los términos de competencia en el área de Computación y en español, por lo que se estudia solo dos fuentes semánticas: el tesauro DISCO II[18] y un tesauro basado en la taxonomía BLOOM[19] [36].

El tesauro DISCO II es un estándar internacional utilizado en la creación de perfiles de competencia en los campos laboral y educativo, que tiene una versión en español. Además, el área de Ciencias de la Computación incluye declaraciones que parafrasean competencias, que contienen elementos de conocimiento que representan resultados de aprendizaje [71]. Es un

[18] DISCO II, available online in http://disco-tools.eu/disco2_portal/projectInformation.php
[19] BLOOM, relacionado a los sinónimos de las habilidades descritas en la taxonomía de BLOOM

vocabulario controlado, y las relaciones existentes entre los términos son de tres tipos: 1. equivalencias semánticas (sinónimos), 2. relaciones jerárquicas, que establecen relaciones de hiperonimia[20], hiponimia[21] y meronimia[22] y 3. relaciones por asociación, que especifican cualquier otra relación contextual, semántica o de uso [17, 18].

La alineación de un término de conocimiento con el tesauro DISCO II requiere una similitud entre el término y un nivel taxonómico del árbol de sinónimos. La Figura 2.8. presenta tres casos de similitud con el diccionario de sinónimos DISCO II, donde los términos pertenecen al mismo subárbol dentro del diccionario de sinónimos y, por lo tanto, tienen el mismo nivel jerárquico superior en el árbol. Así, por ejemplo, términos como "computación en red" y "computación paralela", además de tener una similitud léxica (por la palabra computación), tienen una relación de meronimia porque comparten el mismo subárbol dentro del tesauro (caso 1). Este es también el caso entre "Geoinformática" y "Procesamiento de datos geográficos", que tienen una relación de sinonimia (caso 2), y para "análisis de base de datos" y "modelado de datos", existe una relación de hiperonimia / hiponimia porque estos términos son parte del subárbol correspondiente a "conocimiento de bases de datos" (caso 3). En consecuencia, para lograr la alineación de dos términos de conocimiento, el primer paso es encontrar ese término en el árbol cuya similitud léxica para cada tema es alta, y luego determinar el grado de similitud entre los subárboles de cada uno de ellos.

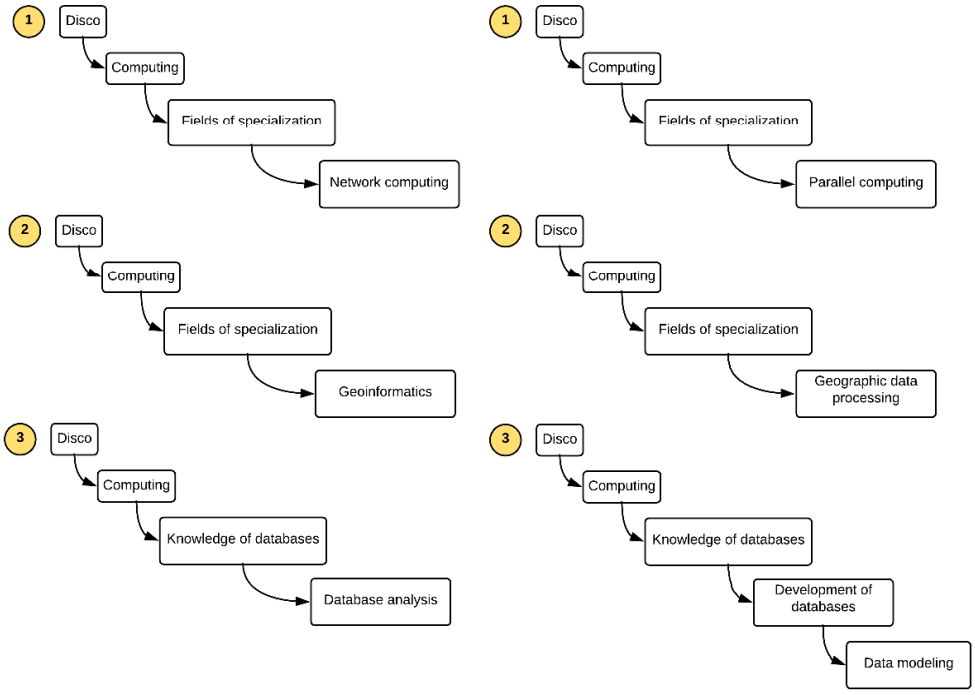

Figura 2.8. Diferentes tipos de similitud de términos según el tesauro DISCO II, en casos de meronimia (1), sinonimia (2) e hiperonimia / hiponimia (3).

Para realizar la alineación de los términos de habilidad, existe un diccionario de sinónimos construido sobre la base de la taxonomía de Bloom, propuesto en [17], que contiene 6 niveles cognitivos (conocimiento, comprensión, aplicación, análisis, síntesis, evaluación), 255 verbos

[20] Hiperonimia: relación entre una palabra de carácter más general y otra de carácter más específico
[21] Hiponimia: relación entre una palabra de carácter más específico y otra de carácter más general
[22] Meronimia: relación "parte de" entre los significados de dos palabras dentro del mismo campo semántico

asociados con cada nivel cognitivo, y aproximadamente 800 sinónimos relacionados con cada verbo. Las relaciones entre los verbos de este tesauro corresponden a la pertenencia a un nivel cognitivo, ya sea por su inclusión en el conjunto de verbos relacionados, o en el conjunto de sinónimos.

Del mismo modo, la alineación de un término de habilidad con el tesauro BLOOM requiere obtener una similitud del término con los niveles taxonómicos del tesauro. La Figura 2.9. presenta dos casos de similitud de términos de habilidad en las alineaciones con el tesauro BLOOM. Como se observa, existe una relación de sinonimia entre "articular" y "componer", ya que independientemente de si pertenecen a diferentes grupos de verbos relacionados (ensamblar y escribir), están bajo el nivel cognitivo "Síntesis", que determina que ambos términos de habilidades son similares (caso 1). Del mismo modo, existe una relación de similitud entre "programa" y "desarrollo", porque están bajo el nivel cognitivo "Aplicación", aunque no pertenecen al mismo conjunto de verbos o sinónimos relacionados (caso 2). En resumen, la alineación de dos términos de habilidad se obtiene, primero, encontrando el grupo en el tesauro en el que se encuentra cada tema, y luego, determinando si tienen el mismo nivel cognitivo, o están en un nivel cognitivo más alto, pero con componentes comunes.

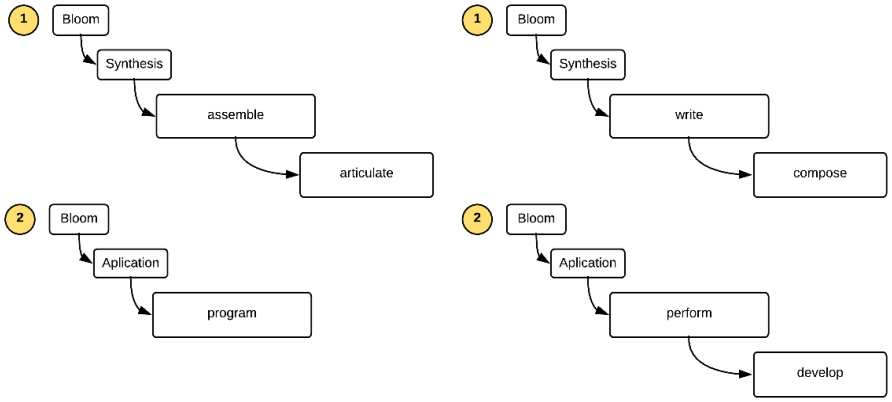

Figura 2.9. Casos de sinomia para diferentes términos de habilidad según el tesauro BLOOM.

2.3.3. Minería de Texto

La Minería de Texto se concentra en la obtención de información no explícita en un conjunto de textos, a través de la identificación de patrones y correlaciones [8]. En [11, 17], se entiende a la minería de texto como el proceso de extracción de información que no se encuentra presente de forma obvia en la colección de documentos objeto de estudio.

Según [13], los problemas que confronta la aplicación de la minería de textos son los siguientes: 1. Falta de una estructura homogénea del texto, que permita su procesamiento automático sin pérdida de información; 2. Fuentes heterogéneas y distribuidas de los documentos: BD documentales, redes sociales, páginas web; 3. Multilingüismo de diferentes conjuntos de documentos y dentro de una colección de textos; 4. Dependencia del contexto y del dominio, implica el uso de diccionarios, tesauros u ontologías específicas a dicho contexto.

Según [18, 11, 17], existen varias aplicaciones de la Minería de Texto, de las cuales destacan:

- ***Extracción de información:*** permite la obtención de información relevante de cantidades extensas de textos, permitiendo definir entidades y sus relaciones, revelando información semántica significativa.
- ***Análisis de sentimientos o minería de opiniones:*** revela información importante sobre un tema específico.
- ***Clasificación de documentos:*** agrupa los documentos, obteniendo información descriptiva de cada grupo para facilitar la comprensión de cada uno de ellos.
- ***Elaboración de resúmenes automáticos:*** obtiene la descripción general de un conjunto de documentos pertenecientes a un tema en específico. En ese sentido, puede hacer resúmenes extractivos (formados por unidades de información extraídas de los textos o chunks), o resúmenes abstractos (formados por unidades de información generadas desde el texto, combinada con otros textos).
- ***Extracción de conocimiento:*** plasma la información extraída en modelos de conocimiento.

2.3.4. Métodos de Similitud Semántica

Los métodos de similitud semántica "son herramientas matemáticas utilizadas para estimar cuantitativa o cualitativamente la fuerza de la relación semántica entre unidades de lenguaje, conceptos o instancias, a través de una descripción numérica o simbólica obtenida, de acuerdo a la comparación de información, formalmente o implícitamente" [23].

Los métodos semánticos pueden usarse para hacer comparaciones de diversas unidades de lenguaje, como son: palabras, conceptos/clases, instancias (abstractas), semánticamente caracterizadas en una relación de conocimiento [80]. Se ha utilizado una terminología extensa en la literatura para referirse a la noción de medida de similitud [22]. La Tabla 2.6. presenta el significado de los términos comúnmente utilizados.

Tabla 2.6. Definiciones de similitud.

Término	Definición
Relación semántica (semantic relatedness)	Es la fuerza de las interacciones semánticas entre dos elementos, sin restricción respecto a los tipos de enlaces semánticos considerados.
Similitud semántica (semantic similarity)	Especializa la noción de relación semántica, considerando únicamente las relaciones taxonómicas en la evaluación de la fuerza semántica entre dos elementos.
Distancia semántica (semantic distance)	Toma en cuenta todas las interacciones semánticas entre los elementos comparados. Estas medidas respetan las propiedades matemáticas de las distancias.
Disimilitud semántica (semantic dissimilarity)	Se entiende como la inversa de la similitud semántica
Distancia taxonómica (taxonomical distance)	Corresponde a fuerza semántica de dos elementos (conceptos) ordenados en una estructura taxonómica, según el camino más corto que existe entre ellos

En la Tabla 2.7 se observa algunas de las propiedades de estas definiciones, tanto para distancia como para similitud [23, 22]. Como se observa, la función de distancia entre dos elementos ***x*** y ***y*** de un dominio debe ser mayor a 0; así mismo, el valor de distancia es el mismo tanto desde ***x*** a ***y***, como desde ***y*** a ***x***, siendo el valor máximo 1 y mínimo 0, y la distancia de ***x*** a ***x*** = 0. De igual forma, la función de distancia debe respetar dos propiedades: 1. identidad propia, y, 2. desigualdad del triángulo. En cuanto a la similitud, se observa que cumple las mismas

propiedades de la distancia acerca de valor no negativo, simetría y normalización (0/1); pero, en el caso de la reflexividad e identidad propia, el valor de similitud es máximo (1). Aunque la función no toma en cuenta la propiedad de desigualdad del triángulo, si incluye la integridad.

Tabla 2.7. Propiedades matemáticas de la distancia y la similitud.

Propiedad	Distancia	Similitud
No negativa	dist(x,y)>= min and min = 0	sim(x,y)>= min and min = 0
Simétrica	dist(x,y) = dist(y,x)	sim(x,y) = sim(y,x)
Reflexiva	dist(x,x) = 0	sim(x,x) = max
Normalizada	0 <= dist(x,y) <= 1 min = 0, max = 1	0 <= sim(x,y) <= 1 min = 0, max = 1
Identidad propia	dist(x,y) = 0 y solo sí x = y	sim(x,y) = max si y solo sí x = y
Desigualdad del triángulo	dist(x,y) <= dist(x,z) + dist(z,y)	----------
Integridad	----------	sim(x,y)<= sim (x,x)

La clasificación de los métodos semánticos se puede realizar, considerando los siguientes aspectos:

- ***El tipo de elementos que las medidas quieren comparar***: unidades de lenguaje, conceptos, clases o instancias.
- ***Las configuraciones semánticas usadas para extraer la semántica requerida por la medida***: textos no estructurados o semi estructurados, tesauros y ontologías.
- ***Las evidencias semánticas y presunciones consideradas durante la comparación***: relacionadas con el tipo de medida semántica usada en la comparación.
- ***La forma canónica adoptada para manejar un elemento***: normalizaciones de elementos.

La Tabla 2.8 presenta un vistazo general de la clasificación de las medidas semánticas en base a las consideraciones descritas anteriormente, en donde se identifican las siguientes medidas: distribucionales, cuyo objetivo es la obtención de similitudes y relaciones semánticas; y las basadas en conocimiento, que buscan la obtención de distancias semánticas y taxonómicas.

Tabla 2.8. Resumen de medidas de similitud.

Enfoques	Elementos	Tipos	Medidas
Medidas Distribucionales	• Textos no estructurados o semi estructurados • Unidades de lenguaje: palabras, oraciones, párrafos, documentos. • Contexto: matrices de co-ocurrencia • Pesos basados en frecuencias • Técnicas de reducción de dimensiones • Vectores de contexto	**Geométricas:** que evalúan las posiciones relativas de dos palabras en un espacio semántico definido por vectores de contexto	• Producto escalar • Distancia de Manhattan • Distancia Euclidiana • Similitud Coseno • Medidas de correlación
		Basadas en conjuntos y Probabilísticas: que analizan los contextos en los que la palabra ocurre	• Índice Dice • Coeficiente Jaccard • Maximum likelihood estimate (MLE). • Pointwise Mutual Information (PMI)
		Co-ocurrencias profundas: resalta relaciones profundas entre palabras como segundas co-ocurrencias	• Latent semantic analysis (LSA). • Hyperspace Analogue to Language (HLA).

			• Syntax or dependency-based model. • Random indexing
Basadas en conocimiento	1. Representaciones de conocimiento: tesauros, taxonomías, ontologías 2. Unidades de lenguaje: palabras/términos, conceptos, grupos de conceptos, instancias semánticamente caracterizadas	**Basadas en análisis de grafos**: elementos son comparados por sus interconexiones, considerando la semántica contenida en las relaciones	• Basada en estructuras • Basada en características • Teoría de información
		Representaciones de conocimiento múltiple: determinar Relaciones semánticas, alineamiento de instancias	• Enfoque estructural • Basado en características • Basado en información teórica • Híbrido

A continuación, se presenta un caso de cálculo de similitud de competencias en base a medidas léxicas y basadas en tesauros [17]. La Tabla 2.9. presenta una propuesta de algoritmo, para el cálculo de la similitud entre dos entidades C y C', usando el tesauro DISCO II[i] . Para calcular los niveles de los árboles n (árbol de la entidad C) y m (árbol del tesauro) se usa la medida de **distancia de Levenshtein** (***distancia léxica***). Luego, se determina la ***similitud taxonómica*** basada en la similitud entre ancestros (SA), hermanos (SS) e hijos (SD), de los árboles de C y C' con mayor valor de ***similitud léxica***, usando el **coeficiente de Sorensen**. A continuación, se explican las medidas:

La distancia léxica entre dos entidades C_j y C'_i viene dada por el número de cambios de caracteres que deben realizarse para que la entidad C_j se convierta en la entidad C'_i [74]. El valor de esa medida es máximo cuando el número de cambios es cero (C_j y C'_i son iguales), y es min en el caso contrario.

$$Dis_{lex}(C_j, C'_i) = \begin{cases} \max(C_j, C'_i) & Si\ Dis_{lex}(C_j, C'_i) = 0 \\ \min(C_j, C'_i) & Si\ Dis_{lex}(C_j, C'_i) > 0 \end{cases} \quad (2.6)$$

La similitud léxica entre dos entidades C_j e C'_i es dos veces el número de pares de caracteres que son comunes a ambas entidades, divididos por la suma del número de pares de caracteres en las dos entidades [75].

$$Sim_{lex}(C_j, C'_i) = \frac{2x\ |pares(C_j)\ \cap\ pares(C'_i)|}{|pares(C_j) + pares(C'_i)|} \quad (2.7.)$$

La similitud taxonómica de dos entidades C_j y C'_i está dada por la sumatoria de las similitudes de ancestros, hermanos e hijos de la entidad C_j, divididas por 3. Entonces, para cada par de entidades se obtiene un valor de similitud que se encuentra en el rango de 0 a 1, donde 0 representa no similitud y 1 representa alta similitud. Las medidas de *Sim(Anc$_i$(C$_j$),Anc$_j$(C'))*, *Sim(Sin$_i$,Sin'$_j$)* y *Sim(Des$_i$,Des'$_j$)*, se calculan con la medida de similitud léxica (Sorensen).

$$Sim_{sem}(C_j, C'_i) = \frac{SA(C_j, C'_i) + SS(C_j, C'_i) + SD(C_j, C'_i)}{3} \quad (2.8.)$$

$$SA(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim\left(Anc_i(C_j), Anc_1(C')\right), \dots, Sim\left(Anc_i(C_j), Anc_n(C')\right) \quad (2.9.)$$

$$SS(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim(Sin_i, Sin'_1), \dots, Sim(Sin_i, Sin'_n) \quad (2.10)$$

$$SD(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim(Des_i, Des'_1), \dots, Sim(Des_i, Des'_n) \quad (2.11.)$$

Tabla 2.9. Algoritmo de alineamiento de entidades C y C' contra tesauros

```
Inicio
Variables
string C, C'
double SM, SA, SS, SD, Sim
integer n, m, med
struct tesauri
Procedimiento
  get (C, C')
  calcular n level of (C)
  calcular m level of (C')
  For i=1 to n
   SA= SA + calcular max ancestro (Ci, C'j)/n ∀j=1 to m
   For i=1 to n
      SS = SS + calcular max siblings (Ci, C'j)/n ∀j=1 to m
   For i=1 to n
     SD = SD + calcular max descendientes (Ci, C'j)/n
  SM = calcular similitud taxonómica(C, C')
get SM
```

La Tabla 2.10. presenta la aplicación del algoritmo para diferentes entidades [17]. El primer caso representa a dos entidades con los mismos padres y hermanos (se observa en la Figura 2.10 que no tienen descendencia, por lo que el valor de SD es 1), por lo cual tienen una alta similitud. Para los casos restantes, la medida varía de acuerdo a la similitud de las entidades en los diferentes niveles del tesauro. Por ejemplo, la Figura 2.11 presenta el caso de baja similitud entre las entidades software y sistemas informáticos, ya que estas entidades no comparten elementos de la taxonomía del tesauro. De acuerdo con los resultados, cuando SM tiene un valor superior a 0.3, entonces se consideran a las entidades similares, pero si el valor es cercano al 1, se consideran sinónimas.

Tabla 2.10. Resultados de la medida de similitud sobre entidades.

C	C'	SA	SS	SD	SM
Computación en paralelo	Computación distribuida	1	1	1	1
Computación en paralelo	Computación en red	0,82	0,12	0,05	0,33
Computación en paralelo	Software	0,02	0,03	0,02	0,02
Sistemas informáticos	Software	0,2	0,05	0,1	0,09

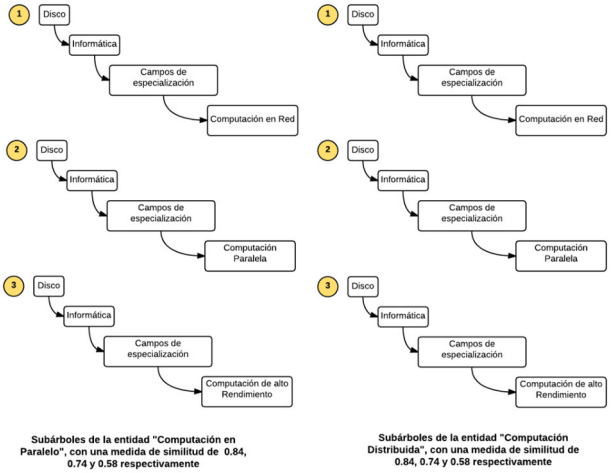

Figura 2.10. Caso de alta similitud entre las entidades computación en paralelo y computación distribuida

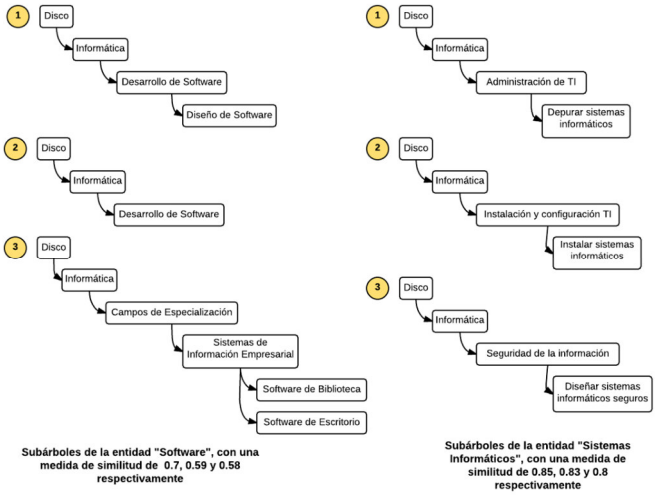

Figura 2.11. Caso de baja similitud entre las entidades software y sistemas informáticos

2.3.5. Aprendizaje Ontológico

El Aprendizaje Ontológico "se define como el conjunto de métodos y técnicas utilizadas para construir una ontología desde cero, enriqueciendo o adaptando una ontología existente de manera semiautomática, usando fuentes de conocimiento distribuidas y heterogéneas, permitiendo una reducción en el tiempo y esfuerzo necesario en el proceso de desarrollo ontológico" [61]. El aprendizaje ontológico es un proceso multidisciplinario ya que en él intervienen disciplinas como el Aprendizaje Automático, PLN, Recuperación de Información, entre otros [78].

Entre los retos que conlleva el desarrollo del aprendizaje ontológico se encuentran: 1. El desarrollo de un aprendizaje ontológico completamente automático, 2. La falta de plataformas de desarrollo comunes para ontologías, 3. Lo difícil del descubrimiento de relaciones de

granularidad fina entre conceptos, y 4. Un aprendizaje ontológico a escala Web, para abordar el cuello de botella de adquisición de conocimiento.

2.3.5.1. Fases del aprendizaje ontológico

En [125] se proponen las fases de un proceso del aprendizaje ontológico, considerándola como un proceso semi-automático que requiere la intervención humana. En la Figura 2.12. se muestran las fases del proceso de aprendizaje, las cuales se explican a continuación [126]:

Figura 2.12. Fases del Aprendizaje Ontológico

- ***Importación y reutilización:*** Tiene por objetivo desarrollar mecanismos y estrategias para importar y reutilizar conceptualizaciones de un dominio, mediante la fusión de estructuras o la definición de reglas de mapeo.
- ***Extracción:*** Consiste en la selección de conceptos del dominio, sus entradas léxicas, y la generación de la taxonomía de conceptos, mediante técnicas de agrupamiento.
- ***Poda*****:** el boceto de la ontología final, resultante de la importación, reutilización y extracción, es reducida para caracterizar la ontología final para el dominio deseado.
- ***Refinamiento*****:** Consiste en incorporar nuevas entradas léxicas o conceptos, de acuerdo a las necesidades específicas de los usuarios del dominio, o a las actualizaciones del dominio, para refinar la ontología.

En la Tabla 2.11. se mencionan los diferentes aspectos a considerar en un mecanismo de Aprendizaje Ontológico.

Tabla 2.11. Aspectos a considerar en un mecanismo de Aprendizaje Ontológico

Aspecto	Consideraciones
Elementos Aprendidos	Términos, Conceptos, relaciones, axiomas
Punto de Partida	Antecedentes o Conocimiento previo, Tipo de Entrada Estructuradas, Semi-estructuradas, No-estructurados
Pre-procesamiento	Convertir la entrada en una estructura adecuada para aprender, como el pre-procesamiento lingüístico que elimina siglas, abreviaturas
Método de Aprendizaje	Técnica de Aprendizaje: Estadístico, Lógico, Lingüistico, Basado en patrones, Heurístico, Híbrido
Tareas de Aprendizaje	Clasificación, Agrupamiento, Aprendizaje de reglas, Análisis de Conceptos, Población de ontologías
Dirección de Aprendizaje	Ascendente, Descendente
Grado de Automatización	Manual, Automático o Semi-automático.

Resultado	Lo que construye el proceso de aprendizaje: ontologías o estructuras intermedias que sirven de apoyo a otros procesos de aprendizaje
Evaluación	Evaluación del proceso de aprendizaje o la evaluación del resultado del proceso (ontología resultante)

Existen cuatro tipos de resultados en el aprendizaje de ontologías, cada uno de los cuales es pre-requisito para obtener el siguiente [58, 78]: términos, conceptos, relaciones (taxonómicas y no taxonómicas) y axiomas.

1. ***Términos:*** Los términos son realizaciones léxicas que pueden ser simples (es decir, una sola palabra) o complejos (es decir, varias palabras). Las principales tareas asociadas son **preprocesar textos** usando modelos de lenguaje (ver apartado 2.1.2), y **extraer términos** usando análisis de relevancia (ver apartado 2.3).
2. ***Conceptos:*** Puede ser cualquier cosa sobre lo que se dice algo. Los conceptos se forman por agrupación de términos similares. Las tareas principales son **formar conceptos** (determinando la estructura de las palabras mediante análisis sintáctico o *clustering*, usando medidas de similitud); y, **etiquetar conceptos** (implica el uso de diccionarios y tesauros (ej: WordNet), para encontrar la categoría a la que corresponde).
3. ***Relaciones:*** se usan para modelar las interacciones entre los conceptos en una ontología. Las tareas principales son: **construcción de jerarquías**, es decir, las relaciones taxonómicas is-a (hiperónimo/hipónimo), mediante relaciones predefinidas del conocimiento, modelos estadísticos basados en la similitud semántica entre conceptos y patrones lingüísticos y lógicos; **descubrimiento de relaciones no taxonómicas,** que son las interacciones entre conceptos como meronimia, roles temáticos, atributos, posesión y causalidad**;** y **etiquetado de relaciones no taxonómicas,** que depende principalmente del análisis de estructuras sintácticas y dependencias (ver apartado 2.3).
4. ***Axiomas:*** que son proposiciones u oraciones que siempre se toman como verdaderas, y permiten usar los elementos ontológicos existentes para describir/analizar contextos, y definir restricciones. La tarea principal implica **descubrir axiomas** mediante las relaciones conocidas que satisfacen ciertos criterios.

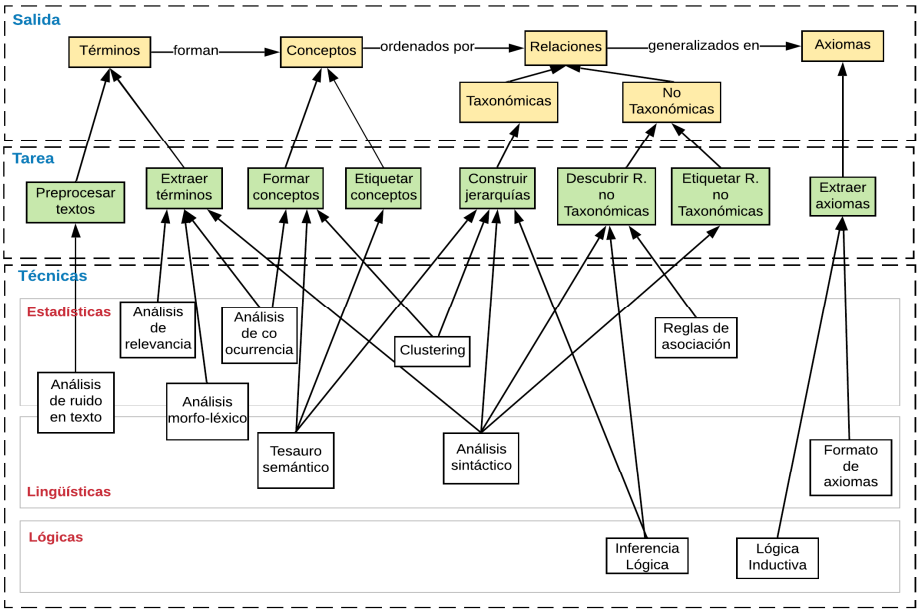

Figura 2.13. Resultados del Aprendizaje Ontológico

Como se observa en la Figura 2.13., en procesos de aprendizaje ontológico se establecen como salida los elementos de una ontología (términos, conceptos, relaciones y axiomas), a partir de tareas que se requieren para extraerlos (por ejemplo, para obtener los términos se debe pre-procesar los textos y extraer los términos), y para la ejecución de las tareas, se identifican las técnicas y recursos (por ejemplo, para extraer términos se pueden usar técnicas estadísticas y lingüísticas).

2.3.5.2. Ontologías

Las ontologías son parte vital de los procesos de aprendizaje ontológico, porque proporcionan la estructura lógica y semántica sobre la cual se realizarán los procesos de aprendizaje (términos, conceptos, relaciones y axiomas) [78]. También, las ontologías constituyen otro mecanismo de representación del conocimiento de un contexto, mediante lenguajes formales [17], como es el caso de la Lógica Descriptiva (Description Logic DL) [76], de tal forma que se crean dos conjuntos:

- TBox: contiene las definiciones de conceptos y axiomas.
- ABox: contiene las definiciones de los individuos o términos.

Según [77], se formaliza la representación de la ontología a través de las siguientes definiciones:

Definición 1: Una ontología es una estructura O = (C,R,A,Ax), donde C representa los conceptos del dominio, R las relaciones entre los conceptos, A los atributos, y Ax los axiomas.

- Los conceptos se pueden estructurar/organizar en forma no taxonómica o taxonómica (jerarquía de conceptos).
- Las relaciones representan los vínculos entre los conceptos. Por lo general, las relaciones son del tipo binarias, donde el primer argumento de la relación se conoce como el dominio, y el segundo argumento se conoce como el rango.
- Los atributos son las características que describen a un concepto.
- Los axiomas o reglas, son definidas en un lenguaje lógico, y describen las restricciones/condiciones que deben cumplir las relaciones/conceptos de una ontología dada.

La aplicación de esta definición se presenta en el siguiente ejemplo extraído de [17]. Sea una ontología O=(C,R,A,Ax), donde:

- El conjunto de conceptos es:
  C:= {Competencia, Conocimiento, Habilidad, Patrón_Conocimiento, Similitud_Conocimiento, Instancia_Conocimiento, Patrón_Habilidad, Cobertura_Habilidad, Instancia Habilidad}.
- El conjunto de relaciones es: R:={estaFormadoPor, tieneDespues}.
- El conjunto de atributos es: A:= {perteneceA, tieneNombre, tieneEtiqueta}.
- Las relaciones entre conceptos son: {Competencia estaFormadoPor Habilidad, Competencia estaFormadoPor Conocimiento, Habilidad tieneDespues Conocimiento}

Algunos ejemplos de axiomas definidos en DL para este modelo, tomados de [19], se presentan en la Tabla 2.12.

Tabla 2.12. Axiomas definidos en Lógica Descriptiva.

Concepto	Axioma
Competencia es la unión de una habilidad (H) y un conocimiento (Co).	$\forall C, H, Co \;\; C \equiv Co \; U \; H$
Conocimiento es un Patrón_Conocimiento (PC)	$\forall Co, PC \;\; Co \sqsubseteq PC$
Habilidad es un Patrón_Habilidad (PH)	$\forall H, PH \;\; H \sqsubseteq PH$

2.3.5.3. Evaluación de modelos ontológicos

Para evaluar la calidad de una ontología [66], se pueden usar las siguientes medidas:

Completitud. Una ontología OC es considerada completa con referencia al documento id_i, si contiene todos los términos relevantes extraídos del documento id_i. La Ecuación (2.12) presenta la definición de Completitud.

$$Completitud(OC, id_i) = \frac{\sum_{i=1}^{m} Trelevantes(id_i) \cap Términos(OC)}{\sum_{i=1}^{m} Trelevantes(id_i)} \quad (2.12)$$

En donde: *Trelevantes(id_i)* son los términos cuyo *Score(id_i,C_j)* se encuentra en el rango definido en el U_R (entre 0.3 y 1), y *Terminos(OC)* son los términos en la OC.

Robustez. Una ontología OC se considera robusta con referencia al conjunto de documentos, si sus términos C_j son relevantes para documento id_i de la colección (ver ecuación (2.13)).

$$Robustez(OC, id_i) = \frac{\sum_{i=1\&C_j \in Terminos(OC)}^{m} Score(id_i, C_j)}{|Terminos(OC)|} \quad (2.13)$$

En donde, si ese promedio es menor a 0.3, implica que se han usado términos no relevantes para poblar a OC.

Para establecer un análisis cualitativo de los resultados de calidad, se usa la escala de valoración propuesta en [73], que considera 3 criterios: Alto (valor mayor que 0.5), Medio (valor entre 0.3 y 0.5) y Bajo (valor inferior a 0.3).

Entropía: La entropía determina la cantidad de información que contiene una fuente de información X. [72] (ec (2.14)).

$$H(X) = -\sum_{i=1}^{k} P(x_i) log_2 P(x_i) \quad (2.14)$$

Donde: X es la fuente de información a evaluar, k es el total de los posibles estados o escenarios, x_i estado o escenario i, $P(x_i)$ es la probabilidad del estado i.

2.3.6. Minería Ontológica

La minería ontológica (OM) consiste en la extracción de patrones de comportamiento, de conocimiento, entre otras características, para construir o enriquecer ontologías [127]. Así, la minería ontológica permite explorar técnicas de extracción de conocimiento global sobre un conjunto de ontologías [61], y de esta manera, manejar la heterogeneidad semántica [4]. Según [17], las técnicas de la Minería Ontológica son:

- ***Alineación ontológica***, consiste en realizar la comparación (matching) entre los conceptos de las ontologías que se estén analizando. Para ello, la comparación se fundamenta en el cálculo de medidas de similitud, que pueden ser: lingüísticas (nombres de entidades), entre propiedades (clases) y grafos (estructura taxonómica), según la fortaleza estructural, de sus propiedades y de la cantidad de información [129], entre otras. Por ejemplo, en la Tabla 2.11. se presenta el proceso de alineamiento de entidades mediante medidas de similitud,

realizando en primer lugar una comparación lingüística de los nombres, y luego una comparación estructural del subárbol al que pertenecen. Es así que se identifican los mejores alineamientos entre ellas, cuando la medida de similitud sobrepasa un umbral establecido [17].

- ***Enlazado ontológico***, también llamado *mapping*, el cual establece relaciones de identidad, entre entidades de la ontología O1 y de la ontología O2, mediante sus características comunes (ej: propiedades superclase_de, subclase_de) [130]. El resultado del mapping se refleja en una ontología de enlace que contiene las entidades y propiedades equivalentes en las ontologías, a través de la que se conectan las dos ontologías. Por ejemplo, O3 tomaría a las entidades "computación en paralelo" y "computación distribuida" como equivalentes (computación en paralelo := computación distribuida) [61].
- ***Mezcla ontológica***, es el proceso donde varias ontologías dentro de un mismo dominio se unen para estandarizar el conocimiento, hacer crecer el conocimiento, o tener el conocimiento total de manera local. Se puede realizar una mezcla débil de ontologías, donde se puede dejar conceptos de las ontologías sin ser mezclados, o mezcla fuerte, que se hace en dos partes, una primera parte donde se realiza la mezcla débil, y una segunda parte donde se incorporan los conceptos y relaciones dejadas por fuera [17]. Ambas formas requieren de la validación de experto, para verificar los conceptos que son incorporados en la ontología resultante [130]. Por ejemplo, una mezcla débil de las ontologías O1 y O2 toma en cuenta solo los pares de entidades cuyo valor de similitud supere un umbral establecido. En cambio, una mezcla fuerte tomaría en cuenta todos los pares, sin considerar si el valor de similitud es bajo o alto.

2.3.7. Modelos Dialécticos

Los lenguajes formales de representación del conocimiento, como la Lógica de Primer Orden (First Order Logic)[23] o la Lógica Descriptiva (Descriptive Logic)[24] [85], no representan los casos de contradicción existentes en un contexto [70, 84,131]. Para su tratamiento, se deben usar otros formalismos como la Lógica Dialéctica (Dialetheic Logic), la cual es un lenguaje formal para describir ambigüedades lingüísticas. En [88], los autores la describen como una lógica que permite a sus axiomas ser verdaderos o falsos o, a diferencia de la lógica descriptiva, verdaderos y falsos al mismo tiempo [70]. La ley de la lógica dialéctica es la contradicción, y en ese sentido, la lógica trata las contradicciones, atenuándolas, ya que es tolerante a inconsistencias, y permite que las contradicciones sean válidas [113, 89].
Según [83], se consideran los siguientes casos de ambigüedad dialéctica:

- **Vaguedad:** falta de claridad, precisión o exactitud en los fenómenos del lenguaje natural [82]. ej. En la oración "Él es calvo", no se puede negar que una persona con cero cabellos sea calva, como tampoco se puede negar que una persona con 10000 cabellos sea calva.
- **Declaraciones contingentes sobre el futuro:** son declaraciones sobre eventos futuros, acciones, estados, etc. Para calificar como contingente, el evento, estado, acción o lo que esté en juego, no debe ser imposible ni inevitable [90]. ej. "mañana habrá una guerra" puede ser cierto o falso, ya que han ocurrido ambos casos en el pasado.
- **Falla de una presuposición:** supone algo que no es realmente cierto [92]. ej. "andar solo en el bosque en la noche es peligroso porque las brujas te hechizan ", presupone que las brujas existen y son capaces de hechizar a las personas.
- **Discurso ficticio:** es tomar decisiones según ciertos supuestos reales o imaginarios (lógicas imaginarias no aristotélicas) [91]. ej. "Las vacas están volando", se puede decir que es falso

[23] First-order logic: usa cuantificadores sobre objetos no lógicos y permite el uso de oraciones que contienen variables.
[24] Descriptive Logic: modela conceptos, roles e individuos y sus relaciones.

porque nuestras creencias nos indica que las vacas no vuelan, pero podría ocurrir que las vacas están siendo transportadas en un avión, y la respuesta sería verdadera.

- **Razonamiento contrafáctico:** La idea básica es que el significado de las afirmaciones causales se puede explicar en términos de condicionales contrafactuales de la forma "Si A no hubiera ocurrido, entonces C no habría ocurrido"[93]. ej. “Si no hubiese salido, entonces hubiera aprobado y ahora no tendría que estudiar para el recuperativo”, esto ayudaría a los sistemas a aprender de los errores para hacer los correctivos en un futuro.

La Tabla 2.13. presenta ejemplos de los cinco fenómenos dialécticos con sus axiomas y ejemplos, en el cual el caso 1 presenta vaguedad por la falta de exactitud en las propiedades que determinan cuan cerca puede estar una persona de un lugar en función del tiempo que le toma llegar a él. La respuesta es ambigua, ya que la concepción de ***cerca*** y ***pronto*** es distinta para cada usuario. El caso 2 presenta una declaración contingente sobre el futuro, ya que la persona puede estar en el lugar**,** pero el lugar puede no necesariamente estar abierto [112]. El caso 3 supone un fallo en la presuposición, porque el uso de las expresiones como “no hay nadie más” y pronombres como “él”, que aparecen como sujetos gramaticales, pueden provocar errores en la interpretación ya que no se puede presuponer que el Rey de Francia es sabio, solo porque no hay nadie más que sea Rey de Francia [113, 59].

Tabla 2.13. Ejemplos de Axiomas definidos en Lógica Dialéctica.

N°	Fenómeno	Axioma	Ejemplo
1	Vaguedad	Si localización A tiene actividad en tiempo B y usuario se encuentra en localización C en tiempo cercano a B, y A es cerca de B, el usuario puede ir a la actividad.	Si hay un concierto en un auditorio y es a las 5:00pm y Juan está en una plaza a las 4:45pm, entonces Juan puede ir al concierto en el auditorio.
2	Declaraciones contigentes sobre el futuro	Si un usuario X realiza una actividad en una ubicación Y en tiempo A, entonces la ubicación Y está abierta en tiempo A.	Si Oscar está en un concierto en el estadio, entonces el estadio está abierto.
3	Fallo en la presuposición	Una entidad X posee la propiedad F y no hay ninguna otra entidad Y que sea distinta de X y que posea la propiedad F entonces X posee la propiedad G	Existe un Rey en Francia y no hay nadie más que sea rey de Francia y él es sabio
4	Discurso ficticio	Una entidad X tiene una característica Y y ninguna entidad tiene la característica Y	Una jirafa vuela y ninguna otra jirafa vuela
5	Razonamiento contrafáctico	Si hubiera habido un objeto X entonces habría habido un objeto Y	si hubiera habido un lápiz, entonces habría habido una libreta

De igual forma, el caso 4 presenta discurso ficticio establece una situación ficticia con respecto a un hecho, en el ejemplo, se conoce que las jirafas no pueden volar, pero si en el contexto se asume que una jirafa vuela, entones se puede suponer que para ese contexto el hecho podría ser verdad. En cuanto al caso 5 presenta razonamiento contrafáctico ambos son condicionales con antecedente y consecuente, aunque el contrafáctico imagina mundos posibles sobre la base de la negación del antecedente (si hubiera habido un lápiz) [52].

2.3.7.1. Evaluación de modelos dialécticos

En general, en los modelos dialécticos se reconocen dos tipos de eventos: evento dialéctico (**sd_{ji}**) que corresponde a aquel cuyo valor de verdad, según los axiomas e instanciaciones del modelo dialéctico, es siempre verdadero; y, evento no dialéctico a aquel que su valor de verdad, según los axiomas e instanciaciones del modelo dialéctico es falso. En base a estos eventos, se define para cada documento **id_i,** que **sd_{ji} es igual a 1** cuando el evento dialéctico j (término ambivalente) es reconocido por el modelo dialéctico MD. Para evaluar los modelos dialécticos se propone la siguiente medida:

Robustez: El modelo MD se considera robusto dialecticamente con referencia a un documento id_i, si reconoce todos sus términos dialécticos. La ecuación (2.15) presenta la definición de robustez, donde n_i representa el número de términos en el documento id_i.

$$RobustezD(DM, id_i) = \sum_{j=1}^{n_i} \frac{sd_{ji}}{n_i} \quad (2.15)$$

CAPITULO 3: MODELO ADAPTATIVO DE ANÁLISIS DE COMPETENCIAS

En este apartado, se presenta el aporte fundamental de la tesis, el modelo adaptativo de análisis de competencias. Para ello, se describe en detalle la arquitectura propuesta y sus componentes, tanto para el caso de funcionar bajo el esquema de lógica descriptiva como dialéctica.

3.1. ARQUITECTURA GENERAL PARA EL ANÁLISIS DE COMPETENCIAS

La Figura 3.1. presenta la arquitectura general propuesta, en la cual se presentan cuatro componentes:

- ***Caracterización***, en este componente se desarrollan dos modelos de conocimiento: el primero basado en lógica descriptiva [68], que es la base para la creación del modelo ontológico, el cual está orientado al reconocimiento de los elementos de conocimiento y habilidad de las competencias en los documentos considerando patrones lingüísticos [66.67]; y, el segundo centrado en lógica dialéctica [70], en el cual se describen los fenómenos ambiguos que existen en los perfiles profesionales mediante axiomas [69]. Ambos modelos utilizan fuentes semánticas para completar los hechos del dominio, en este caso, dos tesauros en español (DISCO II y BLOOM) [71,67]. De esta manera, se cubre los problemas de ambigüedad lingüística y semántica de las competencias detectados en los perfiles profesionales [69,75].
- ***Extracción***, este componente realiza el preprocesamiento y filtrado de términos desde los perfiles profesionales y académicos. Particularmente, en este componente se usan dos enfoques de PLN: el enfoque lingüístico, según el cual se reconocen los términos de conocimiento y habilidad mediante patrones [8,16,15,24], y el enfoque probabilístico mediante N-gramas (N-grams) [15,14]. Complementariamente, se utilizan técnicas de Recuperación de Información [11,12] para filtrar los términos de competencia según su relevancia dentro de la colección de perfiles profesionales analizados [60]. La combinación de estos enfoques permite la optimización del proceso de reconocimiento de términos, en comparación con los resultados obtenidos con los modelos de espacio continuo y condicionales, que requieren grandes colecciones de documentos de gran longitud (cantidad de palabras) [17,118, 119].
- ***Comparación,*** en este componente se aborda la similitud de términos de los perfiles contra términos de los tesauros DISCO II y BLOOM desde la perspectiva distribucional [23], mediante el uso de métricas de distancia (Levenshtein) y similitudes léxicas (Coeficiente Dice) [74,75] y, desde la perspectiva basada en conocimiento [23], a través de una medida de similitud taxonómica usada en la teoría de colonias de hormigas (algoritmo ACO) [73]. Para obtener los alineamientos relevantes, se complementa el enfoque con el uso de técnicas de Recuperación de Información [57,58] para filtrar los términos de competencia [60]. De esta forma, se logra el alineamiento de los términos de competencia no solamente desde la estructura lingüística de los mismos, sino considerando también la similitud

semántica entre ellos, resolviendo aspectos de ambigüedad como son meronimia, sinonimia e hiperonimia/hiponimia [15, 128].

- ***Actualización***, en este componente se desarrolla la población ontológica del modelo de conocimiento basado en lógica descriptiva [76,77] y la generación de los hechos para el modelo basado en lógica dialéctica [83], tomando los términos relevantes obtenidos en el componente de comparación. A continuación, se determina la capacidad de los modelos para representar el dominio de las competencias en función de dos medidas de calidad: Completitud y Robustez [73]. De esta forma, se logra con los modelos el análisis de las competencias no solamente desde un enfoque descriptivo [78], sino desde la determinación de los casos de ambigüedad [83].

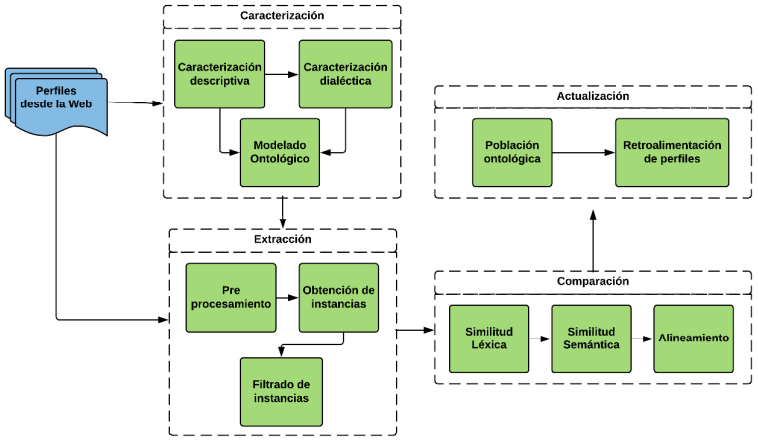

Figura 3.1. Arquitectura del esquema de actualización de ontologías de competencias

A continuación, se explica en detalle cada uno de los componentes de la arquitectura, empezando por su descripción para el caso en que funciona basado en lógica descriptiva (apartado 3.2), y seguidamente, el caso cuando funciona basado en lógica dialéctica (apartado 3.3).

3.2 ARQUITECTURA DE ANÁLISIS DE COMPETENCIAS BASADA EN LÓGICA DESCRIPTIVA

3.2.1. Caracterización

Esta fase comprende 3 actividades: la caracterización descriptiva, donde se realiza la revisión de los perfiles profesionales o académicos a nivel lingüístico; la caracterización dialéctica, donde se realiza el análisis de los perfiles desde la perspectiva de los cinco fenómenos dialécticos: vaguedad del lenguaje natural, fallo en la presuposición, razonamiento contrafáctico, discurso ficticio y declaraciones contingentes sobre el futuro (esto será presentado en la sección 3.3); y el modelado ontológico, donde se realiza la construcción de la ontología que representa al dominio de los perfiles profesionales y académicos analizados.

3.2.1.1. Caracterización descriptiva

El objetivo de la caracterización descriptiva es determinar la estructura lingüística de las competencias y sus componentes (conocimiento y habilidad). Con estas estructuras, se definen

patrones utilizados para extraer términos de los documentos recolectados desde la Web. La Tabla 3.1. presenta los patrones de competencia, conocimiento y habilidad en lógica descriptiva [65]. Por ejemplo, el primer patrón describe en lógica de primer orden una frase nominal que representa una competencia, habilidad o conocimiento. Con estos patrones, se extraen desde los documentos de la web, los posibles términos para poblar la ontología de competencias [20] (apartado 3.2.4.).

Tabla 3.1. Patrones para el reconocimiento de términos de competencia y sus componentes

Descripción	Lógica de Primer Orden	Ejemplo
La frase nominal (NP) puede ser una competencia (Co), un tópico de conocimiento (C) o una habilidad (H)	1. $\forall NP, NC \ \ NP \equiv NC$ 2. $\forall NP, SP, NC \ \ NP \equiv NC \sqcap SP \sqcap NC$ 3. $\forall NC, SP, AQ, NP \ \ NP \equiv NC \sqcap SP \sqcap NC \ \sqcap AQ$ 4. $\forall NC, SP, NP \ \ NP \equiv NC \sqcap SP \sqcap NC \ \sqcap SP \sqcap NC$ 5. $\forall NP, NC, AQ \ NP \equiv NC \ \sqcap AQ$	Administración. Experto en Java. Desarrollo de software. Conocimiento sistemas operativos. Conocimiento base de datos. Administración base de datos.
La frase verbal (VP) puede ser una competencia o una Habilidad	6. $\forall VMN, NC, SP, VP \ VP \equiv VMN \sqcap NC \sqcap SP \sqcap NC$	Administrar bases de datos.

3.2.1.2. Modelado ontológico

En esta etapa se desarrolla el modelo ontológico basado en los elementos de competencia encontrados en los perfiles, siguiendo la metodología de desarrollo ontológico definida en [61]. Particularmente, la ontología de competencias OC tiene la estructura presentada en la Figura 3.2 (clases, propiedades y relaciones). Algunas de las clases principales son:

- **Categoría gramatical:** contiene las subclases para el Análisis Morfo-léxico, como son: adjetivo, conjunción, preposición, sustantivo y verbo. Además, tiene asociadas propiedades como tieneCategoria, tieneGenero, tieneNumero;
- **Cobertura de conocimiento:** sus dieciseis subclases son los tópicos del nivel superior del tesauro DISCO del dominio de Informática, por ejemplo, Gestion_Proyectos_TI; tiene asociadas propiedades como tieneCoberturaConocimiento y perteneceA;
- **Dominio cognitivo:** sus siete subclases corresponden a los niveles cognitivos de la taxonomía de BLOOM, tiene asociadas propiedades como tieneCoberturaHabilidad y perteneceA
- **Patrón:** cada subclase representa uno de los axiomas de la Tabla 3.1. Asi, define patrones de competencias, habilidad y conocimiento.
- **Perfil:** contiene las subclases de Perfil_Academico y Perfil_Laboral, además de propiedades como perteneceA, y;
- **Competencia:** tiene por subclases CompetenciaAcademica, CompetenciaLaboral.

Con respecto a los componentes de las competencias, se define una clase para cada una. En el caso de **Habilidad** tiene las subclases de Cobertura_Habilidad, Instancia_Habilidad, Patron_Habilidad, y propiedades como tienePatronHabilidad y tieneDominioCognitivo. Esas subclases permiten definir una habilidad.

- **Cobertura de Habilidad:** corresponde a aquellos términos de habilidad de los perfiles académicos y laborales que pertenecen al mismo nivel del tesauro BLOOM.
- **Instancia de Habilidad:** corresponde a los términos de habilidad encontrados en los perfiles.
- **Patrón de Habilidad:** corresponde al patrón lingüistico relacionado con el término de habilidad.

En el caso de **Conocimiento,** contiene las subclases de Instancia_Conocimiento, Patron_Conocimiento y Similitud_Conocimiento, y propiedades como tienePatronConocimiento y tieneCoberturaConocimiento.

- **Instancia de Conocimiento:** corresponde a los términos de conocimiento encontrados en los perfiles.
- **Patrón de Conocimiento:** corresponde al patrón lingüístico relacionado con el término de conocimiento.
- **Similitud de Conocimiento:** corresponde a aquellos términos de conocimiento de los perfiles académicos y laborales que pertenecen al mismo nivel del tesauro DISCO.

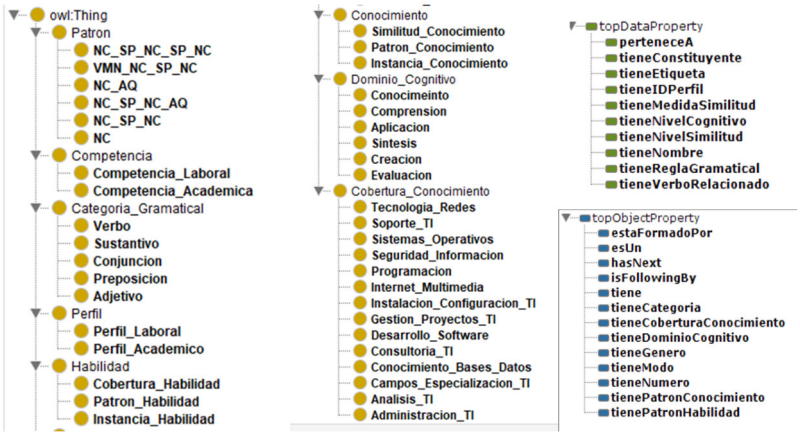

Figura 3.2. Clases y propiedades de la ontología de competencias OC

3.2.2. Extracción

En esta etapa se realiza la identificación y obtención de los términos de competencia (conocimiento y habilidad) mediante las siguientes etapas: preprocesamiento, análisis morfoléxico y filtrado de términos.

- **Preprocesamiento:** consiste en la preparación de los textos para el análisis. Para ello, se eliminan las etiquetas HTML (encabezados, números, fechas, metadatos), dejando solo aquellos textos que están dentro de las etiquetas de tipo <p> </ p>.
- **Análisis Morfoléxico:** se realiza la formación de oraciones y palabras (tokens), las cuales se normalizan (mayúsculas, lematización, etc.) [14] y se etiquetan con una categoría gramatical [20]. Para ello, se utilizan herramientas de PLN[25] para el desarrollo de cada paso, como son [16]: Stanford CoreNLP[26], Spacy[27] y Textacy[28].

[25] PLN: Procesamiento de Lenguaje Natural
[26] Stanford CoreNLP disponible en https://stanfordnlp.github.io/CoreNLP/pipelines.html
[27] Spacy disponible en https://spacy.io/
[28] Textacy disponible en https://spacy.io/universe/project/textacy

- **Filtrado de términos:** se aplica los patrones de la Tabla 3.1. para extraer los términos de conocimiento, habilidad o competencia de los documentos. Luego, se realiza el análisis de la relevancia de cada uno de los términos reconocidos según su frecuencia de aparición en el perfil y en el conjunto de documentos.

En concreto, el macroalgoritmo de esta fase es:

Tabla 3.2. Macroalgoritmo de fase de extracción de términos

Inicio
Entrada
Colección de perfiles profesionales P
Lista de Patrones descriptivos PD según la Tabla 3.1.
Procedimiento
1. Para toda la colección de perfiles P
1.1. Realizar el preprocesamiento del perfil(id_i)
1.2. Realizar el Análisis Morfo-léxico del perfil(id_i)
1.3. Para todo el perfil(id_i)
1.3.1. Para toda la lista de patrones PD
1.3.1.1. Obtener el término C_j según el patrón(Pd_k)
1.3.1.2. Agregar C_j al dataset de términos DI según el Enunciado 3.1.
1.3.2. Fin Para
1.4. Fin Para
2. Fin Para
3. Para todo DI
3.1. Obtener el término (C_j)
3.2. Calcular la relevancia del término C_j *Score(id_i,C_j)* según la Definición 3.1.
3.3. Si *Score(id_i,C_j)* es menor que el umbral de relevancia U_R según el Enunciado 3.2.
3.3.1. Eliminar C_j del Dataset DI
3.4. Fin Si
4. Fin Para
Salida
Dataset de términos relevantes DI
Fin

Como se indica en el macroalgoritmo, sobre cada uno de los perfiles de la colección P se realizan los procesos de preprocesamiento, análisis morfoléxico y filtrado de términos. El filtrado usa la frecuencia de términos iguales y semejantes, para establecer la relevancia de cada uno en el conjunto de perfiles, siendo seleccionados para la fase de alineamiento con tesauros, los términos cuya relevancia superan el umbral U_R. A continuación, se explican los enunciados y definiciones que soportan el macroalgoritmo de extracción.

Enunciado 3.1. Sea DI={(c,id_1,C_1,Pd_1,r,L_1),⋯, (c,id_n,C_n,Pd_n,r,L_n)} el dataset que contiene los términos obtenidos en la fase de extracción, donde c indica el tipo de perfil (1: académico, 2: laboral), id_i el identificador del perfil, C_n el término C_j que fue extraído, Pd_i el patrón descriptivo del término, r el valor de relevancia del término en la colección de perfiles, y L_n su nivel en el tesauro resultado del análisis de similitud léxica.

Definición 3.1. *El valor de relevancia de un término* consiste en la posición del mismo dentro de la colección de términos analizados. En concreto, el valor de relevancia de un término C_j del Perfil id_i, está dado por la ecuación (3.1). Esta es una medida utilizada en la recuperación de información, conocida como Okapi BM25, que ordena por relevancia los documentos en función

del tema que contienen [64]. La métrica clásica TF-IDF tiene en cuenta la frecuencia de aparición de un término dentro de la colección de documentos [81], siendo Okapi BM25 más sensible, porque también tiene en cuenta la longitud de los documentos [60]

$$Score(id_i, C_j) = \sum_{j=1}^{n} IDF(C_j).\frac{f(C_j, id_i).(k_1+1)}{f(C_j, id_i) + k_1.(1 - b + b.\frac{|D|}{avgdl})} \quad (3.1)$$

En donde, $f(C_j,id_i)$ es la frecuencia de aparición del término C_j en el Perfil id_i; $|D|$ es el número de términos en el Perfil id_i; avgdl es longitud media de los perfiles que conforman la colección; k_1 y b son parámetros para ajustar las diferencias de longitud de los perfiles; $IDF(C_j)$ es el peso dado al término C_j en la colección; n es el número de perfiles de la colección.

Definición 3.2. *La frecuencia de aparición de un término* consiste en la cantidad de veces que este término se presenta en el perfil, considerando palabras iguales o semejantes, según la medida de distancia de la Definición 3.7. y el umbral U_R del Enunciado 3.2 La ecuación (3.2) explica el cálculo de $f(C_j,id_i)$:

$$f(C_j, id_i) = \sum_{k=1}^{m} C_j = C_k \quad (3.2)$$

En donde, m representa el total de términos C en el Perfil id_i, y C_k cada uno de los términos en id_i [14].

Definición 3.3. *El peso de un término C_j $IDF(C_j)$* está dado por la frecuencia inversa del mismo en relación a la colección de perfiles (ecuación (3.3)):

$$IDF(C_j) = \log \frac{n - n'(C_j) + \delta}{n'(C_j) + \delta} \quad (3.3)$$

Donde, $n'(C_j)$ es el número de perfiles que contienen el término C_j, y δ es un parámetro de ajuste al peso dado a un término según su frecuencia en la colección de perfiles y la longitud de documentos [60].

Enunciado 3.2. Se establece el umbral U_R entre 0.3 y 1, tal que el término C_j es relevante si su valor de Score supera dicho umbral U_R. Cabe mencionar que la definición del valor de U_R se realizó según el método utilizado en [79], aplicando diferentes umbrales a la colección de términos para escoger el que reúne la mayor cantidad de términos válidos.

3.2.3. Comparación

La comparación se realiza alineando los términos relevantes contra los términos de tesauros. Para ello, se usan dos tipos de medidas de similitud, unas de similitud léxica, como es el caso de las medidas Levenshtein y Coeficiente de Dice (Sorensen); y, otras semánticas, como es el caso de las inspiradas en el algoritmo ACO. Finalmente, se realiza el alineamiento de los perfiles en función de los términos de conocimiento y habilidad del tesauro que se encuentran presentes en cada perfil profesional.

3.2.3.1. Similitud Léxica

Se establece una similitud léxica entre los términos relevantes de los perfiles y cada término perteneciente a los tesauros DISCO y BLOOM. Para ello, se consideran dos medidas, la primera llamada Dislex(C_j,C'), que usa la medida Levenshtein para determinar la distancia de edición entre el término del perfil C_j y el término del tesauro C' [74]; y la segunda llamada Simlex(C_j,C'), que usa el coeficiente de Sorensen para determinar cercanías entre C y C' según la similitud de sus pares de caracteres [75]. Para explicar el proceso de similitud léxica se presenta a continuación el siguiente macroalgoritmo:

Tabla 3.3. Macroalgoritmo de la Similitud Léxica

Inicio
Entrada
Dataset DI según el Enunciado 3.1.
Árbol del tesauro DISCO D según el Enunciado 3.3.
Procedimiento
1. Para todo el Dataset DI
1.1. Obtener el término de conocimiento (C_j)
1.2. Para todo el Árbol de términos D
1.2.1. Obtener el término del Árbol D (C'_i)
1.2.2. Calcular la distancia de edición entre C_j y C'_i *Dislex(C_j,C'_i)* según Definición 3.6.
1.2.3. Si *Dislex(C_j,C'_i)* es mayor al umbral de distancia U_D según Enunciado 3.5.
1.2.3.1. Calcular la similitud léxica entre los términos de conocimiento C_j y C'_i *Simlex(C_j,C'_i)* según Definición 3.7.
1.2.3.2. Si *Simlex(C_j,C'_i)* es menor al umbral léxico U_L según Enunciado 3.6.
1.2.3.2.1. Eliminar C_j del Dataset DI
1.2.3.3. Si no
1.2.3.3.1. Asignar el nivel L_k al término C_j en el Dataset DI
1.2.3.3. Fin Si
1.2.4. Si no
1.2.4.1. Eliminar C_j del Dataset DI
1.2.5. Fin Si
1.3. Fin Para
2. Fin Para
Salida
Dataset DI de términos alineados léxicamente
Fin

De esta forma, los términos relevantes ingresan en el proceso de similitud léxica donde son comparados con los términos del tesauro DISCO mediante la medida de distancia léxica Dislex(C_j,C'_i). Si el valor de similitud es mayor al umbral de distancia U_D, entonces se calcula la medida de similitud léxica *Simlex(C_j,C'_i).* Al final, solo los términos que superen el umbral léxico U_L permanecen en el dataset DI. A continuación, se presentan las definiciones que contiene este macroalgoritmo.

Enunciado 3.3. Sea D={(C'_1, L_1),⋯, (C'_n, L_n)} un conjunto de términos de conocimiento organizados en jerarquía, tal que un término C' pertenece a un nivel L_k del tesauro.

Definición 3.4. Un término C' tiene asociado un conjunto de frases F' descritas según ec. (3.4); una frase F'_i es una colección de términos descritos según ec. (3.5), un término C'_i puede tener

cero o más frases de competencia, y cada frase de competencia F'_i contiene uno o más términos de conocimiento Co'_i o habilidad H'_i.

$$C'_i = \{F'_i, \dots, F'_n\} \quad (3.4)$$

$$F'_i = \{(H'_i, Co'_i), (H'_n, Co'_n)\} \quad (3.5)$$

Por ejemplo, el término "administración del software" de la Tabla 3.4. contiene la frase "administrar aplicaciones de software", donde H' es "administrar" y C' es "aplicaciones de software". Se observa entonces que para cada término C' del tesauro DISCO contiene un conjunto de frases F' que a su vez están formadas por términos de habilidad y conocimiento.

Tabla 3.4. Ejemplo de la Enunciado 3.3. tomado del tesauro DISCO II

C'	F'	H'	Co'
administración del software	administrar aplicaciones de software	Administrar	aplicaciones de software

Enunciado 3.4. Sea $B=\{(TH_1, L_1),\cdots,(TH_n, L_n)\}$ un conjunto de términos de habilidad organizados jerárquicamente, en donde un término TH_k pertenece a un nivel L_k dado del tesauro, y tiene asociado un conjunto de verbos V relacionados, descritos según ec. (3.6).

$$TH_i = \{V_i, \dots, V_n\} \quad (3.6)$$

Definición 3.5. Un verbo V_i tiene una colección de sinónimos S (ver ec. (3.7)).

$$V'_i = \{S_i, \dots, S_n) \quad (3.7)$$

Por ejemplo, el término "aplicación" de la Tabla 3.5. contiene al verbo "comprobar" y a su sinónimo "verificar". Se observa entonces que para cada término TH del tesauro BLOOM contiene un conjunto de verbos V que a su vez están formadas por sinónimos S.

Tabla 3.5. Ejemplo del Enunciado 3.4. tomado del tesauro BLOOM

TH	V	S
Aplicación	Comprobar	verificar

Definición 3.6. *La distancia de edición* entre los términos C_j y C'_i viene dada por el número de cambios de caracteres que deben realizarse para que C_j se convierta en C'_i (ver ec. (3.8)) [74]. El valor de la medida es máximo cuando el número de cambios es cero (C_j y C'_i son iguales), y es mínimo en el caso contrario.

$$Dis_{lex}(C_j, C'_i) = \begin{cases} \max(C_j, C'_i) & Si\ Dis_{lex}(C_j, C'_i) = 0 \\ \min(C_j, C'_i) & Si\ Dis_{lex}(C_j, C'_i) > 0 \end{cases} \quad (3.8)$$

Definición 3.7. *La similitud léxica* entre dos términos C_j e C'_i es dos veces el número de pares de caracteres que son comunes a ambos términos, divididos por la suma del número de pares de caracteres en los dos términos [15, 75] (ver 3.9).

$$Sim_{lex}(C_j, C'_i) = \frac{2\ x\ |pares(C_j)\ \cap\ pares\ (C'_i)|}{|pares(C_j) +\ pares\ (C'_i)} \quad (3.9)$$

Enunciado 3.5. Se establece un umbral de distancia U_D igual a 4, como la mínima distancia de edición que puede existir entre C_j y C'_i para considerar que tienen una similitud. El valor de U_D de 4 se selecciona siguiendo el proceso desarrollado en [79], donde se usan diferentes umbrales para luego escoger el que reúne el mayor número de términos válidos.

Enunciado 3.6. Se establece un umbral U_L igual a 0.4, como la mínima similitud léxica que puede existir entre C_j y C'_i para considerar que tienen una similitud. El valor de U_L de 0.4 está basado en el trabajo de [79], donde se aplican varios umbrales para escoger el que reúne la mayor cantidad de términos.

3.2.3.2. Similitud semántica

La similitud semántica busca determinar si los términos relevantes de los perfiles profesionales comparten el mismo dominio semántico de los términos de los tesauros con los cuales fueron alineados léxicamente. Para ello, se realizan los siguientes procesos: comparación estructural de términos de conocimiento contra el tesauro DISCO, comparación estructural de términos de habilidad contra el tesauro BLOOM.

- **Comparación con el Tesauro DISCO:** Se establece una similitud semántica entre los términos relevantes mediante la alineación de los pares de términos seleccionados por la similitud léxica contra el Tesauro DISCO II, basado en el esquema propuesto en [65], (ver Tabla 3.7.). Así para cada par de términos C_j y C'_i se verifica la similitud de sus ancestros, hermanos e hijos en el tesauro [73]. Luego, para los términos con una alta similitud semántica, se obtiene la similitud de sus términos de habilidad contra el tesauro BLOOM.
- **Comparación con el Tesauro BLOOM:** Se calcula la similitud semántica entre los términos de habilidad seleccionados en el proceso de similitud léxica con los términos de habilidad del tesauro BLOOM [66], lo cual implica identificar el nivel cognitivo que es raíz del subárbol en donde se localizan los términos comparados
- **Alineamiento de perfiles:** se realiza el filtrado de los términos de conocimiento y habilidad, utilizando el umbral Us, que permite seleccionar aquellos términos con una mayor medida de similitud semántica. Luego, se obtiene la frecuencia de los términos en los perfiles, los cuales se alinean con los términos raíz de los tesauros según la medida de similitud que sirve para establecer alineamientos de los perfiles. Se establece de acuerdo con el término raíz del árbol de sinónimos al que pertenecen [20].

Para explicar el proceso de similitud semántica se presenta a continuación el siguiente macroalgoritmo:

Tabla 3.6. Cálculo de la Similitud Semántica

Inicio
Entrada
Dataset DI según Enunciado 3.1.
Árbol del tesauro DISCO D de términos de conocimiento según el Enunciado 3.3.
Árbol del tesauro BLOOM B términos de habilidad según el Enunciado 3.4.
Procedimiento
1. Para todo el Dataset DI
1.1. Obtener el término de conocimiento (C_j) según Pd_j
1.2. Para todo el Arbol D
1.2.1. Calcular la similitud semántica entre C_j y C'_i *Simsem(C_j,C'_i)* según Definición 3.8.

1.2.2. Obtener el término raíz del subárbol del tesauro DISCO md de máxima *Simsem(C_j,C'_i)* según Enunciado 3.7.
1.2.3. Agregar al dataset de términos de conocimiento TC C_j ,máx *Simsem(C_j,C'_i)*, md, PC_j, id_i y c según Enunciado 3.8.
1.3. Fin Para
2. Para todo el Dataset DI
2.1. Obtener el término de habilidad (H_j) según Pd
2.2. Para todo el Arbol B
2.2.1. Calcular la similitud semántica entre H_j y H'_i *Simsem(H_j,H'_i)* según la Definición 3.8.
2.2.2. Obtener el término raíz del subárbol del tesauro BLOOM mb donde se encuentran los términos de habilidad H_j,H'_i según Enunciado 3.9.
2.2.2. Agregar al dataset de términos de habilidad TH H_j, máx *Simsem(H_j,H'_i),*mb,PH_j,id_i y c según Enunciado 3.10.
2.3. Fin Para
3. Fin Para
4. Para todo el dataset de términos de conocimiento TC
4.1. Obtener el término de conocimiento (C_i)
4.2. Si la medida de similitud semántica de C_i Ms (eq. (3.10)) es menor que el umbral de similitud semántica U_s según el Enunciado 3.11.
4.2.1. Descartar el término de conocimiento C_i
4.2.2. Descartar el término de habilidad H_i asociado con el término C_i
4.3. Fin Si
5. Fin For
6. Para todo el dataset de términos de conocimiento TC
6.1. Para cada perfil id_i
6.1.1. Calcular la relevancia del perfil id_i Score (id_i, md_j) según la Definición 3.9.
6.1.2. Agregar al dataset TC, Score (id_i, md_j), md_j, id_i y c según Enunciado 3.8.
6.2. Fin For
7. Fin For
8. Para todo el dataset de términos de habilidades TH
8.1. Para cada perfil id_i
8.1.1. Calcular la relevancia del perfil id_i Score (id_i, mb_j) según la Definición 3.9.
8.1.2. Agregar al dataset TH, Score (id_i, mb_j), mb_j, id_i y c según Enunciado 3.10.
8.2. Fin For
Salida
Dataset TC
Dataset TH

Fin

Como se observa en el macroalgoritmo, el proceso inicia con la comparación de los términos de conocimiento y habilidad contra los tesauros DISCO y BLOOM, respectivamente. Para ello, se usa la medida de similitud semántica *Simsem(C_j,C'_i).* Finalmente, se realiza el alineamiento de perfiles calculando el Score (id_i, mb_i) de cada perfil con base en los alineamientos de sus términos con los tesauros. A continuación, se presentan las definiciones que contiene este macroalgoritmo.

Definición 3.8. *La similitud semántica* de dos términos C_j y C'_i está dada por la sumatoria de las similitudes de ancestros, hermanos e hijos del término C_j, divididas para tres [73]. Entonces, para cada par de términos se obtiene un valor de similitud que se encuentra en el rango de 0 a 1, donde 0 representa no similitud y 1 representa alta similitud (ver ec. (3.10, 3.11.,3.12, 3.13)).

$$Sim_{sem}(C_j, C'_i) = \frac{SA(C_j, C'_i) + SS(C_j, C'_i) + SD(C_j, C'_i)}{3} \quad (3.10)$$

$$SA(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim\left(Anc_i(C_j), Anc_1(C')\right), \ldots, Sim\left(Anc_i(C_j), Anc_n(C')\right) \quad (3.11)$$

$$SS(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim(Sin_i, Sin'_1), \ldots, Sim(Sin_i, Sin'_n) \quad (3.12)$$

$$SD(C_j, C'_i) = \frac{1}{n}\sum_{i=1}^{n} \max(Sim(Des_i, Des'_1), \ldots, Sim(Des_i, Des'_n) \quad (3.13)$$

En donde: $Anc_i(C_j)$: ancestro i del término C_j; $Anc_j(C')$: ancestro j del término C'; $Sim(Anc_i(C_j),Anc_j(C'))$: medida de similitud entre ancestros de los términos C_j y C'; Sin_i: hermano i del término C_j; Sin'_j: hermano j del término C'; $Sim(Sin_i,Sin'_j)$: medida de similitud entre los hermanos de los términos C_j y C'; Des_i: hijo i del término C_j; Des'_j: hijo j del término C'; $Sim(Des_i,Des'_j)$: medida de similitud entre los hijos de los términos C y C'; n: máximo nivel de similitud léxica entre C_j y C'.

Enunciado 3.7. Se considera que md representa al término raíz del subárbol de DISCO, en donde los términos C_j y C' alcanzan un valor de similitud máxima.

Enunciado 3.8. TC={(c,id_1,C_1,PC_1,Ms_1,md,SP_1),⋯,(c,id_n,C_n,PC_n,Ms_n,md, SP_n)} es el dataset de términos de conocimiento resultado de la fase de comparación del modelo, el cual registra para cada perfil y término de conocimiento, el valor de similitud semántica (Ms), término subárbol del tesauro DISCO y el Score (id_i, md_i) alcanzado por el perfil id_i para cada término raíz del subárbol del tesauro (SP_i).

Enunciado 3.9. mb representa el término raíz del subárbol de BLOOM, en donde los términos habilidad H_j y H'_i alcanzan un valor de similitud máxima.

Enunciado 3.10. TH={(c,id_1,H_1,PH_1,Ms_1,mb,SP_1),⋯, (c,id_n,H_n,PH_n,Ms_n,mb, SP_n)} es el dataset de términos de habilidad resultado de la fase de comparación, que guarda los valores de similitud máxima de cada término de habilidad (Ms), además de el término raíz del subárbol del tesauro BLOOM y el score del perfil para cada término raíz del subárbol del tesauro (SP_i).

Enunciado 3.11. Se establece el umbral U_S igual a 0,45, que corresponde a la medida mínima que puede existir entre los términos C y C 'para considerar que tienen una similitud semántica. El valor de Us se define observando el resultado del cálculo de similitud en 100 casos, basado en el trabajo realizado en [65].

Definición 3.9. El valor de relevancia de un perfil consiste en la posición del perfil dentro de la colección de perfiles analizados, en función de los términos de conocimiento y habilidad que contiene. En particular, el valor de relevancia de un perfil de id_i de acuerdo con el término del tesauro md_i, viene dado por la ecuación (3.14), que es una variación de la ecuación 3.1 aplicada, en este caso, a los términos del tesauro que se encuentran en el perfil id_i, tal como se indica a continuación:

$$Score(id_i, md_j) = \sum_{i=1}^{n} IDF(md_j).\ \frac{f(md_j, id_i).(k_1+1)}{f(md_j, id_i) + k_1.(1 - b + b.\frac{|D|}{avgdl})} \quad (3.14)$$

Dónde *f(md$_j$,id$_i$)* es la frecuencia de md$_j$ término en el perfil id$_i$ según la definición 3.10; |D| es el número de términos en el perfil id$_i$ (longitud del perfil id$_i$); avgdl es la longitud promedio de los perfiles que componen la colección; k_1 y b son parámetros ajustables de la función Score (id$_i$, md$_j$) al conjunto de perfiles según características específicas (frecuencia de términos y extensión del documento, respectivamente) [64]; IDF(md$_j$) es el peso dado al término md$_j$ en la colección de perfiles, de acuerdo con la Definición 3.11.; y, n es número de perfiles en la colección.

Definición 3.10. La frecuencia de aparición de un término md$_j$ consiste en el número de términos de conocimiento raíces de los subárboles del tesauro que contiene el perfil. La frecuencia de aparición del término *md$_j$* en el perfil *id$_i$* está dada por la ecuación (3.15):

$$f(md_j, id_i) = \sum_{k=1}^{m} md_j = md_k \tag{3.15}$$

Donde m representa el total de términos *md$_j$* encontrados en el perfil *id$_i$*, y md_k cada uno de los términos en id_i [14].

Definición 3.11. El peso de un término md$_j$ IDF(md$_j$) viene dado por la frecuencia inversa del mismo en relación con la colección de perfiles, que se presenta en la siguiente ecuación (3.16):

$$IDF(md_j) = log \frac{n - n'(md_j) + \delta}{n'(md_j) + \delta} \tag{3.16}$$

Donde n es el número de perfiles en la colección, *n'(md$_j$)* es el número de perfiles que contienen el término *md$_j$*, y δ es un parámetro de ajuste al peso dado a un término, de acuerdo con las características de su frecuencia en la colección de perfiles y la longitud de los documentos [60].

La Tabla 3.7 presenta ejemplos del cálculo de la similitud semántica y del alineamiento de perfiles contra términos del tesauro (md y mb), en función de la similitud de los términos que contiene el perfil. Por ejemplo, "proyecto informático", "lenguaje sql" y "lenguaje java" tienen una similitud con el término del tesauro "Programación", perteneciendo al mismo contexto, porque los valores de M_s superan el umbral $U_s > 0{,}45$. En cuanto a habilidad, se aprecia, que los términos "dirigir" y "planificar" tienen una relación semántica con el término del subárbol del tesauro BLOOM "Síntesis" (relación de sinonimia).

Tabla 3.7 – Cálculo de la similitud semántica en términos de conocimiento (C) y habilidad (H) con DISCO II y BLOOM, respectivamente y el alineamiento de los perfiles a los términos de los tesauros

id	C	Ms	Md	H	Ms	mb
id_1	actividad de especificación	0.76	Desarrollo de software	dirigir	0.8	Síntesis
id_2	proyecto informático	0.53	Programación	planificar	1	Síntesis
id_{21}	lenguaje sql	0.61	Programación	programar	0.9	Aplicación
id_{21}	lenguaje java	0.58	Programación	conocer	1	Evaluación

3.2.4. Actualización

La fase de actualización tiene por objetivo realizar la representación de los términos de competencia, conocimiento y habilidad, mediante los modelos semánticos y axiomas definidos

en la fase de caracterización. Para ello, se realizan los siguientes procesos: población ontológica y retroalimentación de perfiles.

- **Población ontológica:** consiste en relacionar los términos que fueron obtenidos en las fases de extracción y comparación con cada una de las clases correspondientes de la ontología OC [94]. Luego se evalúa la ontología resultado mediante métricas de Completitud y Robustez ontologica, según se indica en [66], definidas en función de la relevancia de los términos tanto para los perfiles como para los tesauros DISCO y BLOOM. Además, se calcula la Entropía del modelo OC [120]. La entropía permite estimar la cantidad de información que aportan algunos conceptos a un concepto objetivo específico [116]. En este contexto se usa esta definición para decir que una ontología, cuando es muy ambigua no contiene información útil. Entonces, intuitivamente se supone que cuando la entropía de la ontología disminuye significativamente hay menos incertidumbre en su contenido.
- **Retroalimentación de perfiles:** consiste en aplicar los axiomas del modelo dialéctico (ver sección 3.3) sobre los términos de competencia, conocimiento y habilidad para el reconocimiento de eventos dialécticos. Luego se evalua la Robustez dialéctica del modelo dialéctico, definida en función de la capacidad del modelo para reconocer eventos dialécticos en los perfiles profesionales.

Para explicar el proceso de actualización, se presenta a continuación el siguiente macroalgoritmo:

Tabla 3.8. Macroalgoritmo de la fase de Actualización

Inicio
Entrada
Dataset TC según el Enunciado 3.8.
Dataset TH según el Enunciado 3.10.
Modelo Ontológico OC según la Figura 3.2.
Modelo dialéctico MD según los Axiomas del apartado 3.3.2.
Procedimiento
1. Para todo el dataset TC
1.1. Obtener el término de conocimiento (C_j)
1.2. Aplicar los axiomas de población ontológica de la ontología OC según la Tabla 3.1.
1.3. Aplicar los axiomas de población ontológica de la ontología OC según el Enunciado 3.12.
2. Fin Para
3. Para todo el dataset TH
3.1. Obtener el término de habilidad (H_j)
3.2. Aplicar los axiomas de población ontológica de la ontología OC según la Tabla 3.1.
3.3. Aplicar los axiomas de población ontológica de la ontología OC según el Enunciado 3.12.
4. Fin Para
5. Para todos los términos de la ontología OC
5.1. Calcular la Completitud (OC,id_i) según la Definición 3.14.
5.2. Calcular la Robustez ontológica (OC,id_i) según la Definición 3.15.
5.3. Calcular la Entropía (OC) según la Definición 3.16.
6. Fin Para
7. Para todos los axiomas dialécticos de MD
7.1. Aplicar los axiomas 3.1., 3.2., 3.3., 3.4. y 3.5. del modelo MD
7.2. Calcular la Robustez dialéctica (MD,id_i) según la Definición 3.17.
7.3. Obtener la Proporción (Pr_MD) según el Enunciado 3.23.
8. Fin Para
Salida
Modelo ontológico validado
Modelo dialéctico validado
Fin

En el macroalgoritmo se muestra que la población ontológica del modelo OC se realiza desde los datasets TC y TH (que contienen toda la información sobre los términos de conocimiento y habilidad resultado de la fase de comparación) y mediante los axiomas definidos en la Tabla 3.1 y el enunciado 3.12. Además, se realiza el proceso de retroalimentación con la aplicación de los axiomas dialécticos del modelo MD sobre los términos de TC y TH. Al final, se evalúa los dos modelos semánticos OC y MD con las métricas de Completitud, Robustez y Entropía definidas para cada uno de ellos. A continuación, se presentan las definiciones que contiene este macroalgoritmo.

Enunciado 3.12. La población de la ontología OC para las clases cobertura de conocimiento y dominio cognitivo se realiza según las siguientes definiciones:

Definición 3.12. La cobertura de conocimiento se presenta si el término C tiene una medida de similitud Ms contra un término C' mayor que el umbral Us, entonces C pertenece al término raíz del tesauro md_j).

Definición 3.13. El dominio cognitivo se presenta si el término H tiene una medida de similitud Ms contra un término H' mayor que el umbral Us, entonces H pertenece al término raíz del tesauro mb_j.

Definición 3.14. Una ontología de competencia OC es considerada completa con referencia al perfil profesional o académico id_i, si contiene todos los términos relevantes extraídos del perfil id_i. La Ecuación (3.17) presenta la definición de Completitud.

$$Completitud(OC, id_i) = \frac{\sum_{i=1}^{m} \boldsymbol{Trelevantes(id_i) \cap Términos(OC)}}{\sum_{i=1}^{m} \boldsymbol{Trelevantes\ (id_i)}} \quad (3.17)$$

Donde *Trelevantes(id$_i$)* son los términos cuyo *Score(id$_i$,C$_j$)* (ver ec. 3.1) se encuentra en el rango definido en el U_R (entre 0.3 y 1) y *Términos(OC)* son los términos en la ontología OC.

Definición 3.15. Una ontología de competencia OC se considera robusta ontológicamente con referencia al conjunto de perfiles profesionales y académico, si sus términos C_j son relevantes para el perfil id_{ii} (ec 3.18).

$$RobustezO(OC, id_i) = \frac{\sum_{i=1\ \&C_j \in Terminos(OC)}^{m} Score(id_i, C_j)}{|Terminos(OC)|} \quad (3.18)$$

Enunciado 3.13. Se establece un umbral de retroalimentación U_{RE} de 0.3 que corresponde a la medida mínima de Completitud y Robustez que debe alcanzar un perfil id_i para demostrar que los términos son relevantes para poblar la ontología OC.

Enunciado 3.14. Se establece una escala de valoración cualitativa del resultado de las medidas de Completitud y Robustez, que considera 3 criterios: Alto (valor mayor que 0.5), Medio (valor entre 0.3 y 0.5) y Bajo (valor inferior a 0.3) [73].

Definición 3.16. La entropía determina la cantidad de información en la ontología OC. Considerando que los términos de los perfiles pueden ser ambiguos, la entropía define la incertidumbre que cada perfil id_i aporta al modelo OC [120] (ec (3.19)).

$$H_{OC}(id_i) = \sum_{j=1}^{m} P(C_j) log_2 P(C_j) \quad (3.19)$$

Donde: $P(C_j)$ corresponde a la probabilidad de que el término $C_j \in id_i$ incluido en el OC (es un término relevante) no sea tanto verdadero como falso, y m es el número de términos en el perfil.

Enunciado 3.15. Se establece que, si el valor de la entropía $H_{OC}(id_i)$ es cero, entonces el perfil id_i no introduce incertidumbre en el modelo OC.

3.3. MODELO DIALÉCTICO DE ANÁLISIS DE COMPETENCIAS

En este apartado se presenta el modelo dialéctico para el análisis de competencias, en el cual se realiza como primer paso la definición de los fenómenos dialécticos de las ambigüedades encontradas en los perfiles profesionales. A continuación, se procede a su caracterización según axiomas, generando así el modelo dialéctico (MD) y, finalmente, se propone la validación de los modelos mediante la medida de Robustez (ver definición 3.15).

3.3.1. Definición de los fenómenos dialécticos en el contexto de las competencias

Los lenguajes formales para representar el conocimiento, como la lógica de primer orden o la lógica descriptiva, no representan casos de contradicción en un contexto [84]. Por el contrario, los modelos de lógica dialéctica describen ambigüedades lingüísticas a través de axiomas que pueden ser verdaderos o falsos al mismo tiempo [88]. De esta manera, los axiomas dialécticos permiten que las contradicciones y ambivalencias sean válidas dentro de un modelo formal [89]. Según [83], existen los siguientes fenómenos dialécticos: vaguedad, fallo de una presuposición, razonamiento contrafáctico, discurso ficticio y declaraciones contingentes sobre el futuro. En particular, estos cinco casos se presentan en los perfiles de competencia:

- **Caso 1: Vaguedad o falta de claridad, precisión o exactitud en el lenguaje natural**: En el caso de las competencias, los patrones lingüísticos de las frases nominales y verbales que identifican las competencias de las habilidades y el conocimiento pueden ser los mismos [82,83].
- **Caso 2: Declaraciones contingentes sobre el futuro:** Las declaraciones se desarrollan sobre eventos futuros, acciones, declaraciones, etc. [90]. Este fenómeno ocurre en frases verbales que generalmente describen competencias y habilidades. En este caso, la frase está formada por varios verbos que, considerando sus sinónimos, se encuentran en diferentes niveles de habilidad y proceso cognitivo, lo que no permite establecer qué habilidad desarrollará la competencia en el futuro cercano.
- **Caso 3 Discurso ficticio:** Implica tomar decisiones de acuerdo con las creencias de las personas [91]. En el caso de las competencias, y sus componentes de conocimiento y habilidades, es común que el editor de perfiles coloque estos 3 componentes en secciones de un documento, como descripción, campo ocupacional, y no precisamente como competencias, conocimientos o habilidades.
- **Caso 4. Fallo de la presunción**: Implica una suposición de algo que no es realmente cierto [92]. En el caso de las competencias, si el término se usa incorrectamente en una sección del perfil, de tal manera que la presuposición que se hace sobre el término es incorrecta; es decir, según la interpretación del editor es de un tipo, pero que al final es de otro tipo.
- **Caso 5 Razonamiento contrafáctico:** Es el caso en el que el significado de los enunciados causales puede explicarse en términos de condicionales contrafácticos de la forma "Si A no hubiera ocurrido, entonces C no habría ocurrido" [93]. En el contexto de las competencias, el razonamiento contrafactual se aplica en los supuestos hechos al alinear los términos de las competencias con los términos del tesauro de acuerdo con las medidas de similitud léxica, y establecer umbrales para determinar altas similitudes.

3.3.2. Caracterización dialéctica

Este módulo corresponde al componente de la arquitectura propuesta en el apartado 3.1, en cual se desarrolla un modelo de conocimiento para cada uno de los fenómenos dialécticos. Para ello, se utiliza la herramienta RM3 de [83], en donde los casos ambiguos se describen mediante tres components: ***axiomas***, que corresponden a las reglas dialécticas que definen cada caso; ***hechos***, que son las entradas al modelo formadas por las instancias extraídas de los perfiles; y ***conjeturas***, que se activan durante el razonamiento para realizar la interpretación del caso. Adicionalmente, se require de tesauros o taxonomías para complementar los hechos del modelo, en este caso los tesauros DISCO II (para el componente de conocimiento) y BLOOM (para el componente de habilidad) [66,67].

3.3.2.1 Caso 1: Vaguedad del lenguaje natural

En el primer caso, la ambigüedad se obtiene de los patrones de lenguaje de las frases nominales y verbales, que pueden interpretarse como conocimiento, habilidad o competencia. La base de conocimiento con la que se realiza este análisis fue anotada por expertos en [67], y donde el axioma relacionado con los problemas de vaguedad se presenta en el siguiente enunciado:

Enunciado 3.16. Si el término T tiene un patrón P como conocimiento y P se interpreta como una Habilidad (C1) o Competencia (C2), entonces T tiene un patrón ambivalente***.***

Axioma 3.1. El axioma ***"tienePatronAmbivalente"*** establece la relación entre los patrones lingüísticos de los términos, dependiendo de si el término T tiene un patrón P que representa el conocimiento, pero eso, cuando se interpreta es diferente (como habilidad (C1) o competencia (C2)), por lo que hay una ambivalencia. Por lo tanto, independientemente de si el patrón lingüístico del término indica que es una frase nominal que corresponde al conocimiento, el término se interpreta como una habilidad o competencia (Ver Tabla 3.9.).

Tabla 3.9. Axioma 3.1. en formato RM3

Axiomas	fof(tienePatronAmbivalente,axiom,(! [T,P,C1,C2] : ((tienePatron(T, P) & esInterpretadoComo(T, C1) & esInterpretadoComo(P, C2) & esDiferente(C1,P) & esDiferente(C2,P)) => tienePatronAmbivalente(T, P)))).
Hechos	fof(tienePatron1, axiom, tienePatron(conocimiento_de_hardware, nc_sp_nc)). fof(esInterpretadoComo1, axiom, esInterpretadoComo(conocimiento_de_hardware, competencia)). fof(esInterpretadoComo2, axiom, esInterpretadoComo(conocimiento_de_hardware, habilidad)). fof(tienePatron1, axiom, tienePatron(nc_sp_nc, conocimiento)). fof(esDiferente1, axiom, esDiferente(conocimiento, competencia)). fof(esDiferente2, axiom, esDiferente(conocimiento, habilidad)).
Conjeturas	Sif "SZS status Theorem for FOF" termino tiene patron ambivalente fof(conjetura,conjecture, (tienePatronAmbivalente(conocimiento_hardware, nc_sp_nc))).

La Tabla 3.10. muestra tres ejemplos de la ambivalencia de estos patrones, que se consideran frases nominales de la forma NC-SP-NC y NC-SP-NC-AQ, y que representan al término de conocimiento; pero, según la concepción del editor, cada uno de los términos se interpreta como una habilidad (experto en Java y conocimiento de hardware) o una competencia (desarrollo de software) [67]. De esta manera, la estructura lingüística de las frases nominales es ambivalente, de acuerdo con la interpretación que el editor tiene sobre el conocimiento y la habilidad.

Tabla 3.10. Casos de Vaguedad según el Enunciado 3.16.

Término	Patrón lingüístico	Interpretación según en patrón	Interpretación según el editor
Conocimiento de hardware	NC-SP-NC	Conocimiento	Habilidad
Experto en Java	NC-SP-NC	Conocimiento	Habilidad
Desarrollo de software	NC-SP-NC	Conocimiento	Competencia

De igual forma, la Figura 3.3. presenta la aplicación del Axioma 3.1 en los ejemplos de la Tabla 3.10, para el caso del término "conocimiento de hardware". El proceso comienza con la definición de hechos como fof (tienePatron1, axiom, tienePatron(conocimiento_de_hardware, nc_sp_nc)), sobre los cuales los axiomas realizan las interpretaciones, desde axiomas básicos como fof (esInterpretadoComo2, axiom, esInterpretadoComo (conocimiento_de_hardware, habilidad)), hasta llegar a la conjetura, que es un axioma que interpreta los hechos basándose en el axioma fof(conjetura, conjecture, (tienePatronAmbivalente(conocimiento_de_hardware, nc_sp_nc))). En este caso, el resultado de la conjetura es cierto, porque según los hechos, el término "conocimiento hardware" tiene un patrón de conocimiento, pero se interpreta como una habilidad.

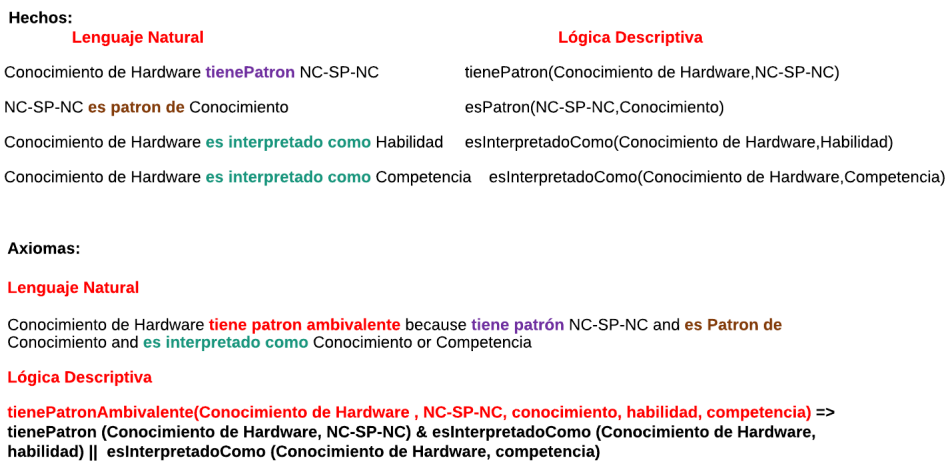

Figura 3.3. Aplicación del Axioma 3.1. sobre los ejemplos de la Tabla 3.10.

3.3.2.2. Caso 2: Declaraciones contingentes sobre el futuro

El caso 2 considera la ambigüedad que existe entre los términos de habilidad cuando pertenecen a dos niveles y procesos cognitivos diferentes. El tesauro de referencia para este análisis es el tesauro BLOOM [67], en el que cada nivel cognitivo tiene asociado un conjunto de términos relacionados. Entonces, un término puede ser sinónimo de un término relacionado que pertenece a un nivel cognitivo dado y que, a su vez, pertenece a un nivel cognitivo diferente. Los enunciados que definen este caso de ambigüedad son:

Enunciado 3.17. Si el término Th es sinónimo del término tesauro Tb y Th y Tb tienen diferentes niveles cognitivos Nc1 y Nc2, entonces Th pertenece a varios niveles cognitivos.

Enunciado 3.18. Si el término Th1 es sinónimo del término Th2 y Th1 y Th2 pertenecen a diferentes niveles cognitivos Nc1 y Nc2, entonces Th1 y Th2 tienen varios niveles cognitivos.

Enunciado 3.19. Si el término T es sinónimo de los términos Th1 y Th2 y Th1 y Th2 pertenecen a diferentes niveles cognitivos Nc1 y Nc2, entonces T tiene varios niveles cognitivos.

Axioma 3.2. El axioma ***"perteneceVariosNivelesCognitivos"*** establece que si el término Th es sinónimo de Tb, y los niveles cognitivos de Th (Nc1) y Tb (Nc2) son diferentes, entonces pueden pertenecer a varios niveles cognitivos (Nc1 y Nc2) debido a verbos relacionados que pertenecen a varios niveles cognitivos. De esta manera, se identifica la contradicción del término Th sobre a qué nivel cognitivo pertenece; en consecuencia, dado que Th pertenece a varios niveles cognitivos, entonces se dice que también pertenece a varios procesos cognitivos (Tabla 3.11).

Tabla 3.11. Axioma 3.2. en formato RM3

Axiomas	fof(terminoPerteneceVariosNivelesCognitivos, axiom,(! [Th,Tb,Nc1,Nc2] : ((esSinonimo(Th,Tb) & perteneceNivelCognitivo(Th,Nc1)& perteneceNivelCognitivo(Tb,Nc2) & esDiferente(Nc1,Nc2)) => terminoPerteneceVariosNivelesCognitivos(Th)))). fof(terminosVerbosRelacionadosPertenecenVariosNivelesCognitivos,axiom,(! [Th1,Th2,Nc1,Nc2] : ((perteneceNivelCognitivo(Th1,Nc1)& perteneceNivelCognitivo(Th2,Nc2) & esDiferente(Nc1,Nc2)) => terminosVerbosRelacionadosPertenecenVariosNivelesCognitivos(Th1, Th2)))). fof(terminosPertenecenVariosNivelesCognitivos, axiom,(! [Th1,Th2] : ((terminosVerbosRelacionadosPertenecenVariosNivelesCognitivos(Th1, Th2) | terminoPerteneceVariosNivelesCognitivos(Th1) | terminoPerteneceVariosNivelesCognitivos(Th2)) => terminosPertenecenVariosNivelesCognitivos(Th1,Th2)))).
Hechos	fof(esDiferente1, axiom, esDiferente(síntesis, conocimiento)). fof(esDiferente2, axiom, esDiferente(síntesis, aplicación)). fof(esDiferente3, axiom, esDiferente(aplicación, síntesis)).

	fof(esSinonimo1, axiom, esSinonimo(diseñar, bosquejar)). fof(esSinonimo2, axiom, esSinonimo(diseñar, planificar)). fof(esSinonimo3, axiom, esSinonimo(planificar, bosquejar)). fof(perteneceNivelCognitivo1, axiom, perteneceNivelCognitivo(diseñar, síntesis)). fof(perteneceNivelCognitivo2, axiom, perteneceNivelCognitivo(bosquejar,síntesis)). fof(perteneceNivelCognitivo22, axiom, perteneceNivelCognitivo(bosquejar,conocimiento)). fof(perteneceNivelCognitivo3, axiom, perteneceNivelCognitivo(planificar, aplicación)).
Conjeturas	Si "SZS status Theorem for FOF" termino Pertenece Varios Niveles Cognitivos fof(conjetura,conjecture, (terminoPerteneceVariosNivelesCognitivos(diseñar, planificar))). fof(conjetura,conjecture, (terminoPerteneceVariosNivelesCognitivos(diseñar))).

En la Tabla 3.12. se observa que, para la competencia "Diseñar y administrar sistemas", "diseñar" pertenece al nivel cognitivo 3 y "administrar" al nivel cognitivo 5, ambos dentro de diferentes niveles y procesos cognitivos (inferior y superior, respectivamente). Por lo tanto, la competencia es ambigua y no es posible establecer los mecanismos de enseñanza, o incluso, el proceso de evaluación del aprendizaje, porque no está claro en qué nivel y proceso cognitivo debe considerarse la competencia a corto o mediano plazo.

Tabla 3.12. Casos de Declaraciones contingentes sobre el futuro en los perfiles profesionales

Frase Verbal	Nivel Cognitivo 1	Nivel Cognitivo 2	Proceso cognitivo 1	Proceso cognitivo 2
Diseñar y administrar sistemas	Diseñar	Administrar	Inferior	Superior
Operación y mantenimiento de centros de cómputo	Operar	Mantener	Inferior	Superior

En la Figura 3.4. se presenta la aplicación del Axioma 3.2 en los ejemplos de la Tabla 3.12, en la que la base de conocimiento está formada por los términos del tesauro Bloom descrito en [67]. Como se observa en el caso de los términos “diseñar” y “planificar”, se parte del hecho de que sus niveles cognitivos son diferentes mediante el axioma fof (esDifferente2, axioma, esDiferente (síntesis, aplicación)). Luego, se establece la relación de sinonimia entre los términos (con el axioma fof (esSinonimo2, axiom, esSinonimo (diseñar, planificar)), y de la pertenencia de cada término a un nivel cognitivo (mediante el axioma fof (perteneceNivelCognitivo1, axiom, perteneceNivelCognitivo (diseñar, síntesis))). De esta manera, como se muestra en la Figura 3.4., la base de conocimiento para la interpretación se construye para la conjetura fof (conjetura, conjecture, (terminoPerteneceVariosNivelesCognitivos (diseñar, planificar))), que tiene un valor de verdadero porque "diseñar" y "planificar" son sinónimos, aunque pertenecen a diferentes niveles cognitivos ("síntesis" y "aplicación", respectivamente).

Hechos:

Lenguaje Natural	Lógica Descriptiva
Diseñar **es sinónimo de** bosquejar	esSinonimo(diseñar,bosquejar)
Diseñar **es sinónimo de** planificar	esSinonimo(diseñar,planificar)
Diseñar **pertenece al nivel cognitivo** síntesis	perteneceNivelCognitivo(diseñar,síntesis)
Bosquejar **pertenece al nivel cognitivo** conocimiento	perteneceNivelCognitivo(bosquejar,conocimiento)
Planificar **pertenece al nivel cognitivo** aplicación	perteneceNivelCognitivo(planificar,aplicación)
Planificar **pertenece al nivel cognitivo** síntesis	perteneceNivelCognitivo(planificar,síntesis)

Axiomas:

Lenguaje Natural

Diseñar **pertenece al mismo nivel cognitivo de** planificar porque **es sinonimo de** planificar y planificar **pertenece al nivel cognitivo** síntesis y diseñar **pertenece al nivel cognitivo** síntesis

Diseñar **pertenece a varios niveles cognitivos** porque **es sinonimo de** bosquejar y diseñar **pertenece al nivel cognitivo** síntesis y bosquejar **pertenece al nivel cognitivo** conocimiento

Lógica Descriptiva

perteneceMismoNivelCognitivo(diseñar,planificar, síntesis)=> esSinonimo(diseñar,planificar) & perteneceNivelCognitivo(diseñar,síntesis) & perteneceNivelCognitivo(planificar,síntesis)

perteneceVariosNivelesCognitivos(diseñar,bosquejar,síntesis,conocimiento)=> esSinonimo(diseñar, bosquejar) & perteneceNivelCognitivo(diseñar,síntesis) & perteneceNivelCognitivo (bosquejar,conocimiento)

Figura 3.4. Aplicación del Axioma 3.2. sobre los ejemplos de la Tabla 3.12.

3.3.2.3. Caso 3: Discurso ficticio

En el tercer caso, se analizan los discursos ficticios que existen en los perfiles, debido a que el editor coloca los términos de competencia y conocimiento en secciones que no tienen relación con ellos. La base de conocimiento para este análisis fue etiquetada por expertos en [67], donde se presentan las contradicciones. Es así que, el enunciado para este caso es:

Enunciado 3.20. Si el término T se encuentra en la sección del documento C1 y T es un componente C2 y C1 es diferente de C2, entonces T es una frase ficticia.

Axioma 3.3. El axioma ***"esFraseFiciticia"*** establece la contradicción en la narrativa de un perfil cuando el término T se encuentra en la sección del documento C1, que es un componente C2, y C1 es diferente de C2, lo que provoca una declaración no real sobre el término T, lo cual confirma que T es una frase ficticia. De esta manera, se altera el significado de la frase que contiene T, generando así un discurso ficticio en el perfil (Tabla 3.13.).

Tabla 3.13. Axioma 3.3. en formato RM3

Axioma	fof(esFraseFicticia, axiom,(! [T,P,C1,C2] : ((estaUbicado(T, C1) & esComponente(T, C2) & esDiferente(C1,C2)) => esFraseFicticia(T)))).
Hechos	fof(esComponente1, axiom, esComponente(control_de_procesos_industriales, conocimiento)).

	fof(esComponente2, axiom, esComponente(control_de_procesos_industriales, competencia)). fof(estaUbicado1, axiom, estaUbicado(control_de_procesos_industriales, competencia)). fof(esDiferente1, axiom, esDiferente(competencia, conocimiento)). fof(esDiferente2, axiom, esDiferente(conocimiento, competencia)).
Conjeturas	Si "SZS status Theorem for FOF" es Frase Ficticia fof(conjetura,conjecture, (esFraseFicticia(control_de_procesos_industriales))).

La Tabla 3.14. muestra casos ambiguos obtenidos, donde la competencia “Planificar y manejar proyectos de software” fue colocada por el editor de perfil como un antecedente. Un caso similar es el término de conocimiento "Control de procesos industriales", que se colocó en la sección de competencia. En consecuencia, depende mucho de la interpretación y el conocimiento del editor, para reconocer una competencia, o el conocimiento y la habilidad de sus componentes, lo que genera una ficción en la redacción del perfil profesional.

Tabla 3.14. Casos de Discurso ficticio en los perfiles profesionales

Término	Componente	Sección del documento
Control de procesos industrials	Conocimiento	Competencia
Desarrollo de aplicaciones computacionales	Conocimiento	Perfil de carrera
Planificación y manejo de proyectos de software	Conocimiento	Antecedente

En la Figura 3.5. se observa la aplicación del Axioma 3.3 en los ejemplos de la Tabla 3.14, en donde a través del axioma se reconocen las ambigüedades del término "control de procesos industriales". Para ello, se parte de que el término es un componente de "conocimiento" (mediante el axioma fof (esComponente1, axiom, esComponente (control_de_procesos_industriales, conocimiento))), que se encuentra en la sección "competencia" del documento (mediante el axioma fof (estaUbicado1, axiom, estaUbicado (control_de_procesos_industriales, competencia))), siendo diferentes "conocimiento" y "competencia" (con el axioma fof (esDiferente2, axiom, esDiferente (conocimiento, competencia))). Con base en los anteriores hechos, la conjetura fof (conjectura, conjeture, (esFraseFicticia (control_de_procesos_industriales, competencia)) tiene un valor de verdadero, porque al mismo tiempo " control_de_procesos_industriales " es un componente de "conocimiento" y se identifica como una "competencia".

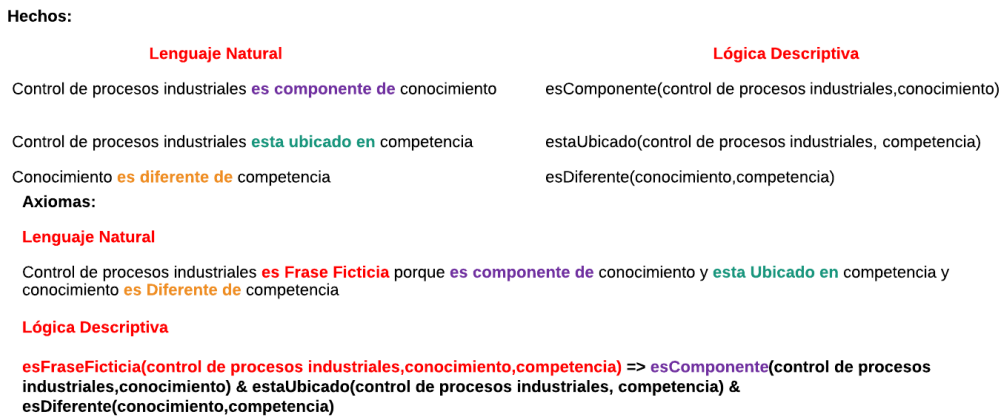

Figura 3.5. Aplicación del Axioma 3.3. en los ejemplos de la Tabla 3.14.

3.3.2.4 Caso 4: Fallo en la presunción

En el cuarto caso, la contradicción que existe en los perfiles se analiza a partir de la suposición hecha por el editor de perfiles acerca de un término en una sección del documento, cuando el término es de otro tipo y no corresponde al lugar donde se encuentra. Por lo tanto, el término es usado incorrectamente por el editor. La base de conocimiento con la que se realiza este análisis fue etiquetada por expertos, y el enunciado para este caso se plantea a continuación:

Enunciado 3.21. Si el término T se encuentra en la sección del documento C1 y T tiene un patrón C2 y C1 es diferente de C2, entonces T es una falla de Presuposición.

Axioma 3.4. El axioma ***"isPresuposicionFallida"*** establece la contradicción en el uso del término T, que se encuentra ubicado en la sección del documento C1, que tiene un patrón C2, y que tanto las secciones C1 y C2 son diferentes. De esta manera, la suposición del editor sobre el término T es fallida, porque hace un mal uso del término en el documento (Tabla 3.15.), creando un presupuesto fallido que causa constradicción.

Tabla 3.15. Axioma 3.4 en formato RM3

Axioma	fof(esPresuposicionFallida, axiom,(! [T,P,C1,C2] : ((tienePatron(T, P) & estaUbicado(T, C1) & esPatron(P, C2) & esDiferente(C1,C2)) => esPresuposicionFallida(T))))..
Hechos	fof(tienePatron1, axiom, tienePatron(conocimiento_ en_java, nc_sp_nc)). fof(estaUbicado1, axiom, estaUbicado(conocimiento_en_java, competencia)). fof(esPatron1, axiom, esPatron(nc_sp_nc, conocimiento)). fof(esDiferente1, axiom, esDiferente(competencia, conocimiento)).
Conjeturas	Si "SZS status Theorem for FOF" termino es Presuposicion Fallida fof(conjetura,conjecture, (esPresuposicionFallida(conocimiento_en_java))).

En la Tabla 3.16. el término "Conocimiento en Java" se asume como "Experiencia", cuando de hecho se interpreta como una "Habilidad". Del mismo modo, se supone que "control de hardware" que se supone es un "antecedente", se interpreta como un "conocimiento", y así para los otros casos. Entonces para cada término, la suposición hecha por el editor de perfil es incorrecta, en relación a la interpretación del experto.

Tabla 3.16. Casos de Fallo de la presunción en los perfiles profesionales

Término	Presuposición	Interpretación por experto (patrón)
Desarrollar aplicaciones informáticas	Perfil profesional	Habilidad
Desarrollar programas de computadora	Competencia	Habilidad
Planificación y manejo de proyectos de software	Antecedente	Habilidad
Control de Hardware	Antecedente	Conocimiento
Conocimiento en Java	Experiencia	Habilidad

En la Figura 3.6. se observa la aplicación del Axioma 3.4 a los casos de la Tabla 3.16, donde los axiomas detectan la ambigüedad del término "conocimiento en Java", que tiene un patrón de conocimiento y se localiza o utiliza como una habilidad (experiencia). Asi, se comienza con el hecho de que el término tiene un patrón de conocimiento "nc_aq" (fof (tienePatron1, axiom, tienePatron (conocimiento_en_java, nc_aq))), que se encuentra en la sección "experiencia" del documento (fof (estaUbicadoEn1, axiom, estaUbicadoEn (conocimiento_en_java, experiencia))), siendo diferentes "conocimiento" y "experiencia" (fof (esDiferente1, axiom, esDiferente (experiencia, conocimiento))). Basado en los anteriores hechos, la conjetura fof (conjetura, conjecture, (esPresuposiciónFallida (java_knowledge))) tiene un valor de verdadero porque al mismo tiempo " conocimiento_en_java " tiene un patrón de "conocimiento" y se identifica como una "experiencia" porsu ubicación, lo cual determina una ambigüedad.

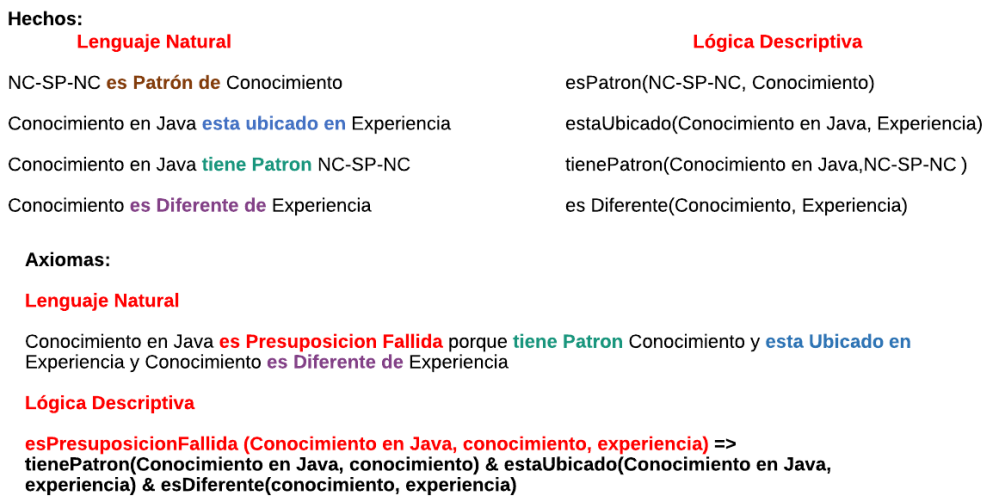

Figura 3.6. Aplicación del Axioma 3.4. en los ejemplos de la Tabla 3.16.

3.3.2.5. Caso 5: Razonamiento contrafáctico

En este caso trata la contradicción que existe con respecto a la pertenencia de un término de conocimiento a un tema de un tesauro, debido a los umbrales utilizados en las medidas de similitud. En este caso, se utiliza DISCO II como tesauro de referencia [66,67], para ello se propone el siguiente enunciado:

Enunciado 3.22. Si el término T tiene una medida de similitud Ms con un tema Tr mayor que el umbral Us, entonces pertenece al tema raíz del tesauro TD.

Axioma 3.5. El axioma ***"terminoPerteneceVariosTopicos"*** describe que un término T, que pertenece a un tópico (término raíz del subárbol del tesauro), también pertenece a varios tópicos si la medida de similitud es mayor que el umbral establecido (Tabla 3.17). Es aquí donde cambiar el umbral genera contradicciones porque un término T que para el umbral 1 era ambiguo, para el umbral 2 puede ya no serlo.

Tabla 3.17. Axioma 3.5. en formato RM3

Axioma	fof(terminoPerteneceTopico,axiom,(! [T,Tr,Ms,Us,TD] : ((relacionMedida(T, Tr, Ms, TD) & esMayorQue(Ms, Us)) => terminoPerteneceTopico(T, TD)))). fof(terminoPerteneceVariosTopicos,axiom,(! [T,TD1,TD2] : ((terminoPerteneceTopico(T, TD1) & terminoPerteneceTopico(T, TD2) & esDiferente(TD1, TD2)) => terminoPerteneceVariosTopicos(T)))).
Hechos	fof(relacionMedida1, axiom, relacionMedida(software, programación, ms0_48, td1)). fof(relacionMedida2, axiom, relacionMedida(software, instalacion_de_software, ms0_30, td1)). fof(relacionMedida3, axiom, relacionMedida(software, depuracion_de_software, ms0_52, td12)). fof(umbral, axiom, umbral = ms0_45). fof(esMayorQue1, axiom, esMayorQue(ms0_48 , umbral)). fof(esMayorQue2, axiom, esMayorQue(ms0_52 , umbral)). fof(esDiferente1, axiom, esDiferente(td1 , td12)).
Conjeturas	Si "SZS status Theorem for FOF termino Pertenece a Topico fof(conjetura,conjecture, (terminoPerteneceTopico(software, td1))). fof(conjetura,conjecture, (terminoPerteneceTopico(software, td12))). Si "SZS status Theorem for FOF" termino Pertenece Varios Topicos fof(conjetura,conjecture, (terminoPerteneceVariosTopicos(software))).

En la Tabla 3.18. se plantea la siguiente hipótesis: “Un término de perfil y un término de un tesauro de competencia pertenecen al mismo tópico de conocimiento cuando la medida de similitud entre ellos excede el límite de 0,45”. Es así que, dos de los tres casos pertenecen al mismo tópico porque la medida de similitud excede el umbral de 0.45. Pero si cambia el valor límite a 0.51, solo el caso "software" versus "Programación" cumple con la hipótesis.

Tabla 3.18. Casos de Razonamiento contrafáctico en los perfiles profesionales

Término	Tópico	Dominio	Similitud
Software	Depuración de software	Programación	0.53
	Instalación de software	Configuración e instalación de TI	0.50
	Desarrollo de aplicaciones de software	Desarrollo de software	0.41

En la Figura 3.7. se observa la aplicación del Axioma 3.5. a los casos de la Tabla 3.18., donde los axiomas detectan la ambigüedad del término "software" que puede pertenecer a más de un dominio, según la medida de similitud que alcanza contra los términos del tesauro (Programación o Depuración de software). Para ello, se comienza con las definición de los hechos fof (relacionMedida1, axiom, relacionMedida (software, programacion, ms0_48, td1)) y fof (relacionMedida3, axiom, relacionMedida (software, depuracion_de_software, ms0_52, td12)), que definen que el término “software” tiene una medida de similitud de 0,48 con “programación” y de 0,52 con “depuracion_de_software”. Otro hecho es que las medidas de similitud de 0,48 y 0,52 son mayores que el umbral (0,45), y también, que los hechos "td1" y "td12" son "diferentes". De esta manera, como se muestra en la Figura 3.7., la base de conocimiento se construye para la interpretación del axioma fof (conjectura, conjecture, (terminoPerteneceTopico (software, td12))), que es el axioma base para la conjetura fof (conjectura, conjecture, (terminoPerteneceVariosTopicos (software))), que tiene un valor de verdadero porque "software" pertenece a los tópicos "programación" y "depuración de software". En síntesis, el valor del umbral es subjetivo, causando errores y ambivalencias en la interpretación de la pertenencia de un término a un dominio del conocimiento.

Figura 3.7. Aplicación del Axioma 3.5. en los ejemplos de la Tabla 3.18.

3.3.3. Validación del modelo dialéctico

El proceso de validación comienza con la aplicación de los axiomas de los cinco modelos dialécticos para cada término de la colección de perfiles académicos y profesionales, después se realiza el cálculo de las medidas de Robustez y Entropía.

Como preparación para el proceso de análisis dialéctico, los perfiles de la colección se etiquetan según la ubicación de los términos en el documento, y la opinion de experto sobre el uso del término y su significado en el context de perfil profesional [67].

Además, en el caso 5, se considera una medida de similitud léxica para establecer la relación entre el término de conocimiento del perfil académico o profesional (T) y los términos del tesauro DISCO II (Tr), que es una variación de la ecuación 3.9, donde que determina la cercanía de acuerdo con la similitud de los pares de caracteres de los términos [75] según la Definición 3.7 (3.20):

$$Sim_{lex}(T, Tr) = \frac{2\ x\ |pares(T) \cap pares(Tr)|}{|pares(T)| + |pares(Tr)|} \quad (3.20)$$

De este modo, para cada par de T y Tr se comparan sus caracteres, y se obtiene un valor de similitud entre cero y uno, donde cero no representa similitud y uno representa alta similitud.

Por otro lado, en el caso 2, se determina la pertenencia del término de habilidad a un nivel cognitivo (tema raíz del subárbol del tesauro BLOOM), si se incluye dentro de su grupo de verbos relacionados (nivel 1 del tesauro BLOOM), o sus sinónimos (nivel 2 del tesauro BLOOM).

En general, se reconocen dos tipos de eventos: términos dialécticos, que son aquellos en los que los axiomas del modelo MD (apartado 3.3.2) devuelven un valor de verdad positivo (verdadero); y, términos no dialécticos, cuando los axiomas dan un valor de verdad negativo (falso). Con base en estos términos, se define para cada perfil id_i, sd_{ji} = 1 cuando el evento dialéctico j (término ambivalente) es reconocido por el modelo MD, y sd_{ji} = 0 en caso contrario.

Para la evaluación de los axiomas del modelo MD, se utilizan las siguientes medidas:

Definición 3.17. El modelo MD se considera robusto dialecticamente con referencia a un perfil profesional o académico id_i, si reconoce todos sus términos dialécticos. La ecuación (3.21) presenta la definición de robustez, donde n_i representa el número de términos en el perfil id_i.

$$RobustezD(DM, id_i) = \sum_{j=1}^{n_i} \frac{sd_{ji}}{n_i} \quad (3.21)$$

Enunciado 3.23. La proporción de términos dialécticos (Pr_MD(id_i)) corresponde a los términos reconocidos por el modelo MD sobre el total de términos relevantes en el modelo OC.

Enunciado 3.24. Si la proporción de casos ambiguos detectados por MD se acerca al valor de entropía de la ontología, entonces se dice que MD identifica las ambigüedades presentes en la ontología OC.

CAPITULO 4: CASOS DE ESTUDIO

En este capítulo se presenta el funcionamiento de la arquitectura definida en el capítulo anterior a través de dos casos de estudio: el primero centrado en el análisis de perfiles profesionales y académicos según la lógica descriptiva, y el segundo realizando el análisis de los perfiles según la lógica dialéctica. El capítulo explica en detalle el comportamiento de la arquitectura para cada caso de estudio, y al final, presenta una discusión de los resultados obtenidos.

4.1. CASO 1: ANÁLISIS DE PERFILES PROFESIONALES SEGÚN LÓGICA DESCRIPTIVA

4.1.1 Procesamiento de los datos del experimento

Para el desarrollo del caso, se toman como entrada 35 documentos en español: 20 perfiles académicos, obtenidos de portales universitarios (id_1, ..., id_{20}), y 15 ofertas de trabajo, obtenidas de portales de empleo en internet (id_{21}, .. id_{35}). De cada perfil, se seleccionan extractos de texto que están bajo secciones como, descripción, objetivos, competencias, habilidades, y conocimiento. En estas oraciones hay elementos de competencia, tales como habilidades y conocimientos, con los cuales se crea una colección (corpus) para cada perfil, registrado con un identificador (id_i). Además, el conjunto de frases que componen el corpus se identifica como (P_j) y el tipo de perfil puede ser de dos tipos (C = 1 si es perfil académico y c=2 si es perfil laboral). Cada frase P_j es un conjunto de términos que pueden ser de conocimiento (C_j) o de habilidad (H_j). La Tabla 4.1 presenta un ejemplo de una colección de perfiles profesionales.

Tabla 4.1. Extracto de la colección de perfiles profesionales

id_i	c	P
id_1	1	Diseñador y administrador de sistemas de comunicación de datos. Abordar proyectos de automatización computacional. Abordar proyectos de automatización computacional, mandos de máquinas eléctricas. Integrar equipos, en operación y mantenimiento de sistemas electrónicos.
id_2	1	Desarrollo de aplicaciones computacionales. Diseño y manejo de base de datos estadísticas.
id_{21}	2	Interacción con bases de datos y lenguaje SQL. Programar y desarrollar en lenguaje Java POO. Conocimientos de JRE 5, 6 y 7 y de programación concurrente en Java. Conocimiento Avanzado en Java. Conocimiento en Lenguaje SQL. Análisis y diseño de base de datos

El primer paso es el preprocesamiento de los textos, para obtener los términos de conocimiento y habilidad, de acuerdo con el procedimiento indicado en la arquitectura general (ver apartado 3.1). Comienza con el desarrollo de un análisis, para reconocer los términos de conocimiento y habilidad, basados en patrones lingüísticos (ver apartado 3.2.1.1) [65]. La Figura 4.1. presenta un ejemplo del análisis para la primera oración de los perfiles de la Tabla 4.1., donde los términos

de conocimiento se reconocen de acuerdo con patrones, que están formados por sustantivo, preposición o adjetivo ([NC], [NC- SP-NC], [NC-NC], [NC-AQ])[4]; mientras que los términos de habilidad por patrones con verbo, sustantivo, preposición o secuencias conjuntas ([VMN], [NC-SP], [NC-CC-NC]) [29] [14,94].

P_j	c	H			C		
Diseñador y administrador de sistemas de comunicación de datos.	1	Diseñador	Y	Administrador	Comunicación	de	datos
		NC	**CC**	**NC**	**NC**	**SP**	**NC**
Desarrollo de aplicaciones computacionales	1	Desarrollo		de	Aplicaciones		computacionales
		NC		**NC**	**NC**		**AQ**
Interacción con bases de datos	2	Interacción		con	Bases	de	datos
		NC		**CC**	**NC**	**SP**	**NC**
Conocimiento avanzado en Java	2	Conocimiento		Avanzado	De		Java
		NC		**AQ**	**SP**		**NC**

Figura 4.1. Ejemplo del preprocesamiento de datos

Como resultado, se detectaron 93 instancias de conocimiento y 70 instancias de habilidad en los perfiles académicos. Por otro lado, en los perfiles profesionales se detectaron 204 instancias de conocimiento y 96 de habilidades. En la Figura 4.2. se presenta un extracto de la información generada con los patrones. En particular, de las instancias de conocimiento (columna C) y habilidades (columna H) extraídas.

c	id	P_j	H	C
1	id_1	Diseñador y administrador de sistemas de comunicación de datos	Diseñador y administrador	Comunicación de datos
1	id_2	Desarrollo de aplicaciones computacionales	Desarrollo	Aplicaciones computacionales
2	id_{21}	Interacción con bases de datos	Interaccion	Bases de datos
2	id_{21}	Conocimiento avanzado de Java	Conocimiento	Java

Figura 4.2. Extracción de términos de conocimiento y habilidad del dataset experimental

4.1.2 Fuentes semánticas

Para la parte experimental, se usarán los tesauros DISCO II [71] y BLOOM [67]. La Tabla 4.2. presenta los términos raíz de los subárboles del tesauro DISCO II contra los cuales se alinean los perfiles, por ser los que corresponden a la subárea de Informática del tesauro. Del mismo modo, la Tabla 4.3. presenta los términos raíz de los subárboles del tesauro BLOOM, que corresponden a los niveles cognitivos definidos en la taxonomía de Bloom, que son comúnmente utilizados en el contexto de los perfiles analizados para describir las acciones o habilidades que deben realizarse o tenerse sobre el conocimiento para alcanzar la competencia [50].

Tabla 4.2. Definición de los subárboles del tesauro DISCO II

	md
Tc_1	Instalación y configuración TI
Tc_2	Desarrollo de software
Tc_3	Campos de especialización en TI
Tc_4	Consultoría de TI
Tc_5	Análisis TI

[29] Formato CONLL, VMN: verbo, CC: conjunción, SP: preposición, AQ: Adjetivo, NC: Sustantivo

Tc_6	Programación
Tc_7	Conocimientos de bases de datos
Tc_8	Sistemas operativos
Tc_9	Gestión de proyectos TI
Tc_{10}	Administración de TI
Tc_{11}	Internet y multimedia
Tc_{12}	Informática
Tc_{13}	Seguridad de la información
Tc_{14}	Soporte TI
Tc_{15}	Tecnología de red

Tabla 4.3. Definición de los subárboles del tesauro BLOOM

	mb
Th_1	Conocimiento
Th_2	Comprensión
Th_3	Aplicación
Th_4	Síntesis
Th_5	Creación
Th_6	Evaluación

4.1.3 Extracción de términos a analizar

Ya identificado los términos de conocimiento, habilidad y competencia en la fase de preprocesamiento, en esta fase se realiza el filtrado de términos considerando los macro algoritmos que se detallan en el apartado 3.2.2 (Tabla 3.2.). La Tabla 4.4. presenta algunos de los términos seleccionados, considerando los parámetros de filtrado de términos de k_1=1.2 b=0.75 y δ=0.5 (ver ec. 3.1). Recordemos que k_1 y b son parámetros para ajustar las diferencias de longitud de los perfiles (según las características del corpus de documentos), y δ es un parámetro de ajuste al peso dado a un término según su frecuencia en la colección de perfiles y la longitud de los documentos [60]. Los valores para k1, b y δ se tomaron de los resultados de la conferencia TREC[30], donde determinaron los mejores valores de esos parámetros para la recuperación de texto.

La Tabla 4.4. presenta el cálculo del filtrado para los términos de la Tabla 4.2 utilizando el Score(id_i,C_j) (ec. 3.1), para lo cual se establece el número de perfiles que contienen al término C_j ($n'(C_j)$), el peso del término IDF(C_j), y el score de los términos en los perfiles id_1, id_2 e id_3 (ejemplos de perfiles para el cálculo de score). Como se observa, los valores del score de algunos términos son mayores al umbral U_R (0.3), y esta tendencia se mantiene en los perfiles donde el término se encuentra con mayor frecuencia, por ejemplo, “aplicación computacional”, en id_1 con 0.32 y id_2 con 0.72.

Por otro lado, algunos términos en los perfiles laborales no están presentes en algunos perfiles académicos (por ejemplo, el termino del perfil laboral id_{21} “base datos” no está en id_1 y id_3) y viceversa. También, el peso que se asigna al término tiene una relación con el valor de $n'(C_j)$, tal que si disminuye el número de perfiles que contienen el término C_j se incrementa el peso IDF dado al término. Cabe mencionar que para seleccionar los términos que se utilizan en la fase de

[30] TREC: Text Retrieval Conference

comparación con fuentes semánticas, se considera que el término pase el umbral al menos en uno de los perfiles.

Tabla 4.4. Cálculo de filtrados de términos para la colección de perfiles

c	id	Término C_j	$n'(C_j)$	$IDF(C_j)$	id_1	id_2	Id_3
1	id_1	comunicación de datos	9	0.54	0.44	0	0.66
1	id_2	aplicación computacional	12	0.39	0.32	0.72	0
2	id_{21}	base de datos	14	0.31	0	0.57	0
2	id_{21}	Java	11	0.44	0.3	0	0.2

4.1.4. Comparación con las fuentes semánticas

En este apartado se presenta la comparación de perfiles contra los tesauros presentados en el apartado 4.1.2. Para ello, se realiza el cálculo de la similitud léxica y semántica de los términos, su alineamiento contra los términos raíz de los tesauros, y se obtiene la relevancia de un perfil en función de los términos raíz de los tesauros que contiene. Para ello, se siguen los macroalgoritmos detallados en el apartado 3.2.3., tomando como referencia a id_1, id_2 e id_{21} de la colección de perfiles.

4.1.4.1 Similitud Léxica

La similitud léxica se obtiene entre los términos de conocimiento y el tesauro DISCO, a través de las medidas de similitud presentadas en las ecuaciones (3.8) y (3.9) del macroalgoritmo de la Tabla 3.3. (apartado 3.2.3.1.). Particularmente, la distancia calculada entre el término C de conocimiento del perfil analizado y el término C' del tesauro (de acuerdo con las definiciones 3.6. y 3.7. y el umbral U_L) permite identificar el término del tesauro DISCO con el cual el término C tiene una mayor similitud léxica. Esto es importante para el siguiente componente de la arquitectura porque determina el nivel del tesauro donde se encuentra el término C' y, por lo tanto, el subárbol en el que se realizará el cálculo de la medida taxonómica.

La Tabla 4.5. presenta un ejemplo del cálculo de la similitud léxica para los términos del conjunto de datos de la Figura 4.2., donde para cada término C se identifica el subárbol del tesauro que tiene una mayor posibilidad de relacionarse con el dominio de dicho término. Por ejemplo, "bases de datos estadísticas" con "bases de datos" y "Java" con "Lenguaje Java" comparten el mismo contexto, por lo que puede suponerse que existe una similitud semántica entre estos términos. Particularmente, el cálculo de la similitud léxica entre esos términos se hace según la Definición 3.7. (ec. 3.9). En la Tabla 4.5., L es el nivel del tesauro donde está el termino C' con el que se obtuvo la mayor similitud léxica con C.

Tabla 4.5. Ejemplo de cálculo de la similitud léxica para los términos de los perfiles

c	id	C	C'	$Sim_{lex}(C,C')$	L
1	id_1	Comunicación de datos	Sistemas de comunicación de datos	0.6	3
1	id_2	Aplicaciones computacionales	Aplicaciones informáticas	0.7	3
2	id_{21}	Bases de datos	Base de datos relacional	0.8	2
2	id_{21}	Java	Lenguaje Java	0.5	2

4.1.4.2 Similitud Semántica

La medida calculada entre el término C de conocimiento del perfil analizado y el término C' del tesauro DISCO, según la Definición 3.8. (apartado 3.2.3.2.), permite identificar el término raíz (general) del tesauro DISCO con el cual los términos de los perfiles tienen una mayor similitud. La Tabla 4.6. muestra un ejemplo del cálculo de la similitud semántica para los términos del conjunto de datos de la Figura 4.2. Por ejemplo, "aplicaciones computacionales" y "Java" tienen una relación de similitud con el término general del tesauro "Programación", que confirma que pertenecen al mismo contexto, en este caso al subárbol de Programación. Esto se obtiene mediante el promedio de las sumatorias de la medida de similitud léxica Sim_{lex} del término "java" para los términos ancestros (SA, ver ec. 3.11), hermanos (SS, ver ec. 3.12) e hijos (SD, ver ec. 3.13) de los subárboles donde se ubique el término "lenguaje java". Luego se identifica el término raíz del subárbol del tesauro (md) según la similitud semántica máxima obtenida (Ms).

El valor de similitud semántica del término "java" muestra que el término existe tal como está escrito en el tesauro DISCO, por tanto, el término "lenguaje Java" obtiene un valor de similitud semántica de 1, que es muy lejano al obtenido solo con la medida de similitud léxica (0.5). De igual forma, existen casos de términos como "bases de datos", en donde hay una pequeña variación entre sus similitudes léxica y semántica (0.8 y 0.7, respectivamente), pero igualmente se considera que pertenece al término raíz "Conocimiento de bases de datos" (md) porque la medida de similitud Ms (0.7) supera el umbral U_S, el cual se establece en 0.45 (definido según lo detallado en [65,79]). Según ese umbral y los valores de Ms obtenidos, existe una similitud bastante clara entre los términos de los perfiles y tesauros (ver Tabla 4.6).

Tabla 4.6. Ejemplo de cálculo de la similitud semántica para los términos de conocimiento con el tesauro DISCO

c	id	C	Ms	md
1	id_1	Comunicación de datos	0.67	Instalación y configuración TI
1	id_2	Aplicaciones computacionales	0.59	Programación
2	id_{21}	Bases de datos	0.7	Conocimiento de bases de datos
2	id_{21}	Java	1	Programación

Del mismo modo, la Tabla 4.7. presenta el cálculo de la medida de similitud para el término de habilidad H del perfil bajo estudio y el término H' del tesauro, utilizando la Definición 3.8. Para calcular las similitudes de H con los términos ancestros, hermanos e hijos del subárbol del tesauro donde se encuentra H', se identifica el término raíz del subárbol del tesauro (mb) en donde se obtuvo el valor máximo de la similitud (Ms). Según la medida calculada, los términos "administrador", "diseñador" e "interacción" tienen una relación semántica con el término raíz "Síntesis", según el tesauro BLOOM. En el caso del término de habilidad "desarrollar", es evidente que existe una relación de similitud dentro del contexto del subárbol del término "Aplicación", igual que conocimiento con el término raíz "conocimiento".

Tabla 4.7. Ejemplo de cálculo de la similitud semántica para los términos de habilidad con el tesauro BLOOM

c	id	H	Ms	mb
1	id_1	Administrador	0.7	Síntesis
1	id_1	Diseñador	0.8	Síntesis
1	id_2	Desarrollo	0.8	Aplicación
2	id_{21}	Interacción	0.7	Síntesis
2	id_{21}	Conocimiento	1	Conocimiento

4.1.4.3. Alineamiento

En esta fase se utilizan las medidas de similitud semántica obtenidas en el apartado 4.1.4.2 y el macroalgoritmo del apartado 3.2.3.2., para determinar la alineación de los perfiles de la colección (según sus términos de conocimiento y habilidad) con los tesauros DISCO y BLOOM. Con los términos alineados, se establecen los términos raíz de los tesauros alrededor de los cuales los perfiles se relacionan y aquellos con los que no tienen relación. A continuación, se muestra un ejemplo de este proceso en 3 documentos: id_1, id_2, y id_{21}.

La Tabla 4.8. y la Figura 4.3. muestran el resultado de la alineación de los perfiles en función de los términos de conocimiento, de acuerdo con la definición 3.14, considerando que k1 = 1.2 b = 0.75 y δ = 1 (esos valores permiten manejar las diferencias de longitud de los documentos y valores de frecuencias bajas en los términos en el conjunto de perfiles). Por ejemplo, el score de Tc_6 (Programación) en la Tabla 4.8. se calcula obteniendo el número de perfiles que contienen a Tc_6 ($n'(md_j)$), que en este caso son 12 perfiles, para después calcular el peso $IDF(md_j)$ (en este caso 0.391), el cual se multiplica por la frecuencia de aparición del término en el perfil id_2 (que es igual a 1), ajustada por los parámetros k1 y b, según la Definición 3.14., para obtener el valor de Score de Tc_6 en el perfil idi_2, en este caso de 0.59.

Se observa que los perfiles id_2 e id_{21} están alineados con los términos Tc_6 y Tc_7 (Programación y Conocimiento de bases de datos, respectivamente). El perfil id_{21} presenta una mayor alineación con el término "Programación" (0.62 versus 0.59), mientras que id_2 tiene una mayor alineación con "Conocimiento de bases de datos" (0.41 versus 0.18). También, existe una alineación entre id_1 e id_2 con el término Tc_3 (Campos de especialización en TI), donde id_1 tiene el valor más alto (0.29 contra 0.16). Los resultados anteriores indican que el perfil académico id_2 cubre parcialmente los requisitos de la oferta de trabajo id_{21} (recordar que desde id_{21} son perfiles de trabajo); no es el caso de id_1 que posee poca alineación con id_{21}. Además, el valor alto que alcanza id_{21} con el término Tc_6 (Programación) da una primera retroalimentación del contexto laboral al contexto académico, enfatizando la importancia que las compañías le dan a este tema dentro de sus ofertas de trabajo.

Tabla 4.8. Ejemplo de cálculo del alineamiento de perfiles y términos de conocimiento

Término raíz md	$n'(md_j)$			$IDF(md_j)$			$Score(id_i,md_j)$		
	c		Total	c		Total			
	1	2		1	2		id_1	id_2	id_{21}
Tc_1	3	1	4	0.802	1.176	0.923	1.33		
Tc_2	6	5	11	0.465	0.415	0.436	0.54		
Tc_3	12	4	16	0.082	0.528	0.235	0.29		0.16
Tc_5	3	5	8	0.802	0.415	0.595			0.4
Tc_6	3	9	12	0.802	0.087	0.391		0.59	0.62
Tc_7	5	10	15	0.556	0.021	0.271		0.41	0.18

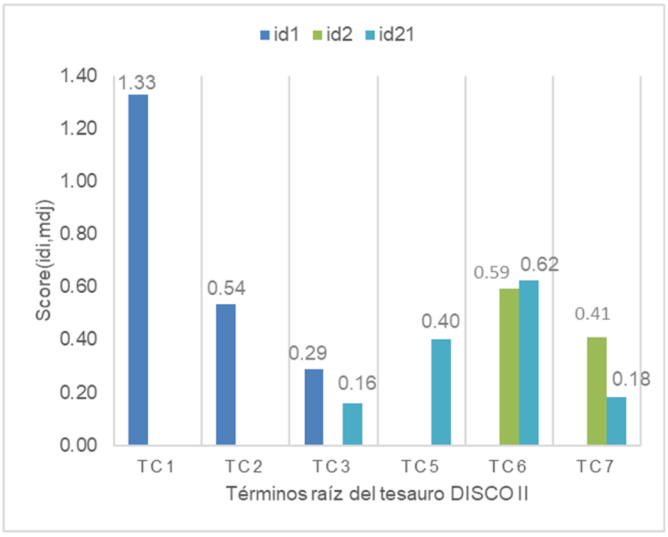

Figura 4.3. Alineamiento de los perfiles id_1, id_2 e id_{21} según términos de conocimiento

Por otro lado, el valor de score del término Tc_1 (Instalación y configuración de TI) en id_1 excede el valor de 1 (1.33) porque el número de perfiles que contienen el término Tc_1 dentro de la colección ($n'(md_j)$) es bajo, con respecto a los otros términos (Tc_1 se presenta en 3 perfiles académicos y 1 perfil laboral). En consecuencia, tiene un mayor peso ($IDF(md_j)$ es 0.923). Eso significa que según el Score de Tc_1 para id_1 no resulta relevante al momento de determinar la importancia de este término del tesauro en la colección de perfiles, y se puede considerar como un dominio de conocimiento aislado en relación a los otros dominios que se presentan en los perfiles.

Del mismo modo, la Tabla 4.9. y la Figura 4.4. presentan la alineación de los perfiles en función de los términos de habilidad, de acuerdo con la Definición 3.9. (apartado 3.2.3.3), considerando $k_1 = 1.2$ $b = 0.75$ y $\delta = 1$.

Tabla 4.9. Ejemplo de cálculo del alineamiento de perfiles y términos de habilidad

Término raíz mb	$n'(md_j)$			$IDF(md_j)$			$Score(id_i,md_j)$		
	c		Total	c		Total			
	1	2		1	2		id_1	id_2	id_{21}
Th_3	12	6	18	0.08	0.28	0.30	0.46	0.41	0.12
Th_4	1	4	4	1.32	0.57	0.86			0.38
Th_5	18	13	31	0.07	0.06	0.06	0.10	0.11	0.10
Th_6	8	13	21	0.31	0.06	0.23	0.23		0.30

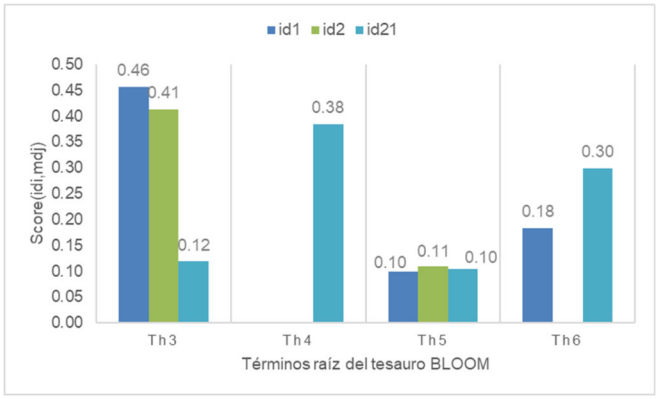

Figura 4.4. Alineamiento de los perfiles id_1, id_2 and id_{21} según términos de habilidad

Se observa que los perfiles id_1, id_2 e id_{21} están alineados con el término Th_3 (Aplicación), siendo id_1 el que presenta un mayor valor (0.46, contra 0.41 y 0.12), lo que indica que los perfiles académicos le dan gran importancia a la aplicación del conocimiento. También hay una alineación entre id_1, id_2 e id_{21} con el término Th_5 (Creación), donde los tres tienen valores muy cercanos (0.10, 0.11 y 0.10, respectivamente), lo que indica que los perfiles académicos cubren a id_{21} en términos de la capacidad de crear conocimiento. Del mismo modo, id_{21} e id_1 se alinean con el tema Th_6 (Evaluación), destacando esta habilidad como un requisito del contexto laboral, que también está presente en el perfil académico id_1, pero en un nivel inferior (0.30 frente a 0.18, respectivamente). Finalmente, el término Th_4 (Síntesis) es una habilidad solicitada por las empresas, que no se considera en los perfiles académicos id_1 e id_2.

Las Figuras 4.5. y 4.6. presentan los resultados de la alineación de cada uno de los perfiles (id_1 ..., id_{35}) con los términos raíz del tesauro DISCO II (Tc_1, ..., Tc_{15}). Como se ve, el promedio de los documentos de la colección se enfoca en los términos: "Desarrollo de software" (Tc_2), "Campos de especialización de TI" (Tc_3), "Análisis de TI" (Tc_5), "Programación" (Tc_6), " Conocimiento de bases de datos "(Tc_7) y" Sistemas operativos "(Tc_8). Algunos términos, como "Gestión de proyectos de TI" (Tc_9), "Administración de TI" (Tc_{10}) o "Tecnología de red" (Tc_{15}), tienen una alta alineación con uno o varios de los perfiles, pero en general, sus promedios en la colección son bajos. En general, los términos con el promedio más alto de alineación en los perfiles son los comprendidos en el intervalo Tc_1 a Tc_8. Con los otros términos, el promedio es menor o no existe alineación, como en el caso de Tc_4 (Consultoría de TI). También, hay algunos perfiles que tienen un gran alineamiento con varios de los términos del tesauro DISCO II, como los id_{24} y id_{25}, o poca alineación, como el id_{10}.

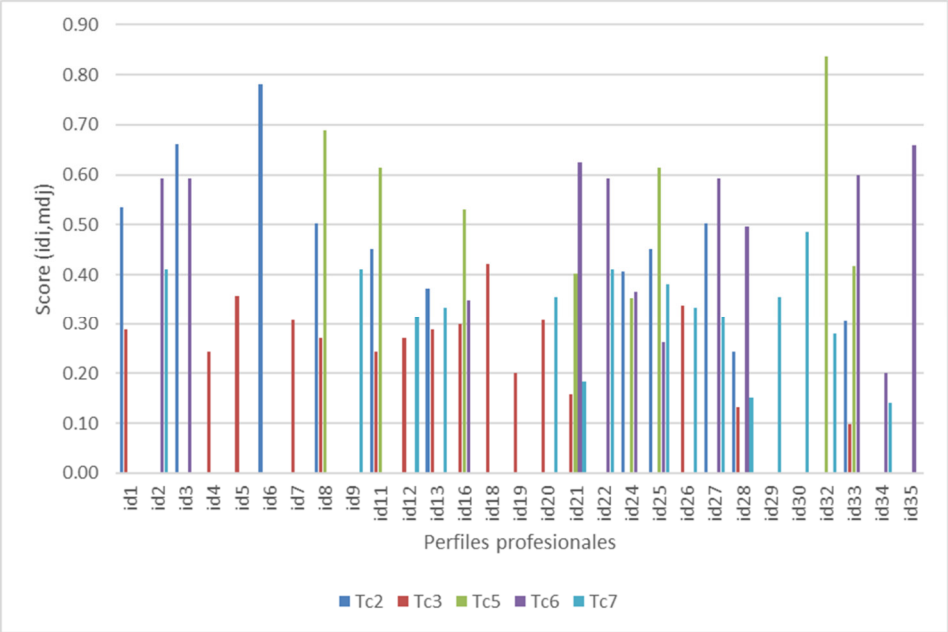

Figura 4.5. Alineamiento de perfiles con los términos de conocimiento (Tc_1 a Tc_7) del tesauro DISCO II

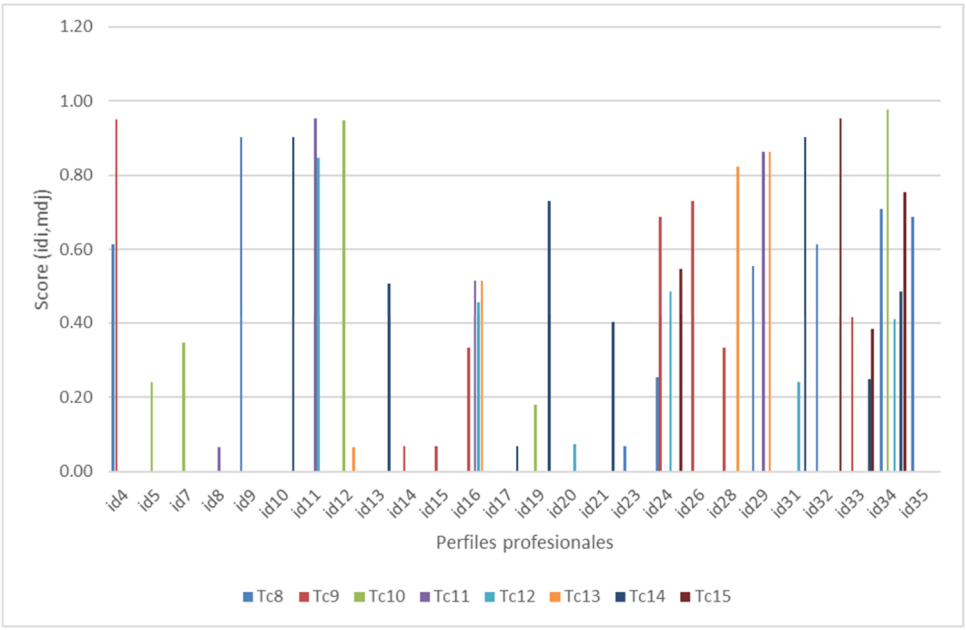

Figura 4.6. Alineamiento de perfiles con los términos de conocimiento (Tc_8 a Tc_{15}) del tesauro DISCO II

La Tabla 4.10. presenta los valores de alineación de los perfiles (documentos) para los temas con mayor promedio de alineación. En resumen, la colección de perfiles presenta una tendencia hacia los primeros 8 términos raíz del tesauro DISCO II. Por ejemplo, para el término Tc_2 se puede ver que los documentos tienen valores de alineación de mayor a menor de la siguiente manera: id_6 (0.78), id_3 (0.66), id_1 (0.54), id_8 e id_{27} (0.5), id_{11} e id_{25} (0,45), id_{24} (0,41), etc. Esto significa que el dominio de conocimiento Tc_2 para este grupo de perfiles es el mismo, por lo que se puede afirmar que existen semejanzas entre los componentes de conocimiento que contiene cada uno de los documentos, y que tanto los perfiles académicos como los laborales de este grupo están semejantemente rankeados según el valor del score alcanzado para este dominio.

Tabla 4.10. Resultados del alineamiento de perfiles con el tesauro DISCO II

PERFIL	Tc_1	Tc_2	Tc_3	Tc_5	Tc_6	Tc_7	Tc_8
id_1	0.33	0.54	0.29				
id_2					0.59	0.41	
id_3		0.66			0.59		
id_4			0.24				0.62
id_5			0.36				
id_6		0.78					
id_7			0.31				
id_8		0.50	0.27	0.69			
id_9						0.41	0.90
id_{11}		0.45	0.24	0.62			
id_{12}			0.27			0.31	
id_{13}	0.79	0.37	0.29			0.33	
id_{16}	0.29		0.30	0.53	0.35		
id_{18}			0.42				
id_{19}			0.20				
id_{20}			0.31			0.36	
id_{21}			0.16	0.40	0.62	0.18	
id_{22}					0.59	0.41	
id_{24}		0.41		0.35	0.36		0.25
id_{25}	0.62	0.45		0.62	0.26	0.38	
id_{26}			0.34			0.33	
id_{27}		0.50			0.59	0.31	
id_{28}		0.24	0.13		0.50	0.15	
id_{29}						0.36	0.56
id_{30}						0.49	
id_{32}				0.84		0.28	0.62
id_{33}		0.30	0.10	0.42	0.60		
id_{34}					0.20	0.14	0.71
id_{35}					0.66		0.69

La Figura 4.7. presenta los resultados de la alineación de perfiles de acuerdo con los términos de habilidades del tesauro BLOOM. Se observa que los documentos de la colección se enfocan en los términos raíz "Aplicación" (Th_3), "Síntesis" (Th_4), "Creación" (Th_5) y "Evaluación" (Th_6), y los términos con mayores promedios son Th_3. y Th_6. En cuanto a los otros términos, el promedio es muy bajo o no existe alineación, como en el caso de Th_1 (Conocimiento) y Th_2 (Comprensión). También se ve que muchos perfiles tienen ninguna alineación con los términos del tesauro BLOOM, como id_2, id_4 e id_6, entre otros.

La Tabla 4.11. presenta el valor para los términos con mayor promedio de alineación en los documentos de la colección de perfiles (como se dijo en el párrafo anterior, los términos Th_3, Th_4, Th_5 y Th_6 del tesauro BLOOM). Por ejemplo, para el término Th_3, los perfiles con mayores valores son: id_{18} (0.55), id_{17} e id_{10} (0.47), id_1 (0.46), id_{14} (0.41), id_{20} (0.37), id_{19} (0.33), id_{12} (0.30), id_{27} (0.15), id_{16} e id_{25} (0.14), id_{21} (0.12), id_{28} (0.10), id_{34} (0.08) y id_{33} (0,06).

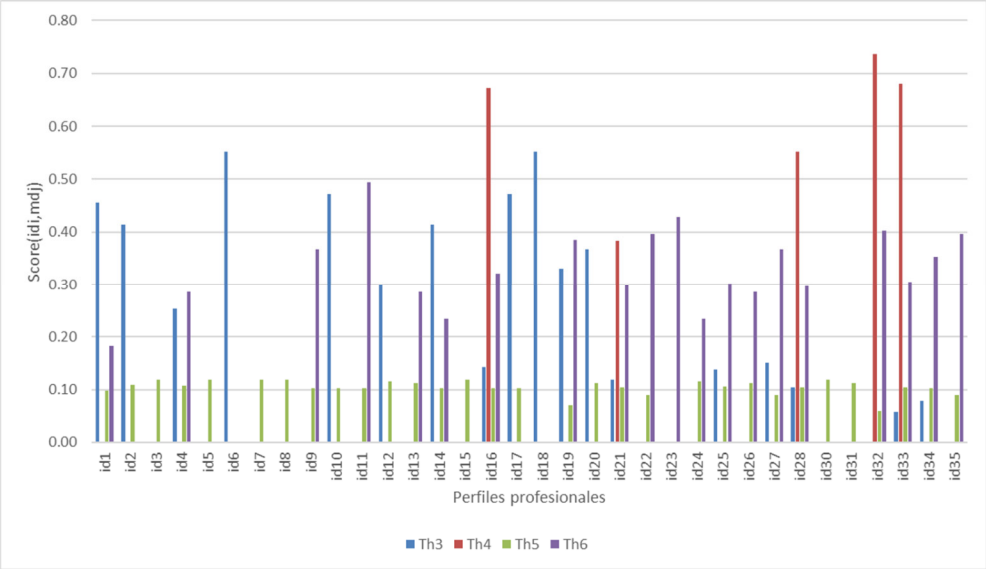

Figura 4.7. Alineamiento de la colección de perfiles según los términos de habilidad

Tabla 4.11. Resultados del alineamiento de perfiles con el tesauro BLOOM

PERFIL	TH_3	TH_4	TH_5	Th_6
id_1	0.46		0.10	0.18
id_2				
id_3			0.12	
id_4				
id_5			0.12	
id_6				
id_7			0.12	
id_8			0.12	
id_9			0.10	0.37
id_{10}	0.47		0.10	
id_{11}			0.10	0.49
id_{12}	0.30		0.12	
id_{13}			0.11	0.29
id_{14}	0.41		0.10	0.23
id_{15}			0.12	
id_{16}	0.14	0.67	0.10	0.32
id_{17}	0.47		0.10	
id_{18}	0.55			
id_{19}	0.33		0.07	0.39
id_{20}	0.37		0.11	
id_{21}	0.12	0.38	0.10	0.30
id_{22}			0.09	0.40
id_{23}				0.43
id_{24}			0.12	0.23
id_{25}	0.14		0.11	0.30
id_{26}			0.11	0.29
id_{27}	0.15		0.09	0.37
id_{28}	0.10	0.55	0.10	0.30
id_{30}			0.12	
id_{31}			0.11	
id_{32}		0.74	0.06	0.40

id_{33}	0.06	0.68	0.10	0.30
id_{34}	0.08		0.10	0.35
id_{35}			0.09	0.40

Con los resultados obtenidos en esta fase, se puede establecer qué perfiles profesionales están alineados con un mismo término del tesauro, y cuál es la fuerza de estos alineamientos. Por ejemplo, para el término raíz Th_6, los perfiles organizados según su fortaleza de alineamiento son: id_{11} (0.49), id_{23} (0.43), id_{22} (0.40), id_{32} (0.40), id_{35} (0.40), id_{19} (0.39), id_9 (0.37), id_{27} (0.37), id_{34} (0.35), id_{16} (0.32), id_{33} (0.30), id_{21} (0.30), id_{13} (0.29), id_{25} (0.30), id_{28} (0.30), id_{26} (0.29), id_{14} (0.23), id_{24} (0.23) e id_1 (0.18). Por otro lado, existe similitud en algunos de los términos de habilidad de los perfiles (por ejemplo, id_{22} y id_{32}), por tanto, están en el mismo dominio cognitivo.

Además, se puede hacer retroalimentación entre ellos para determinar qué perfiles académicos cubren los dominios de habilidad. Por ejemplo, en el id_{11}, la habilidad "generar soluciones técnicas" conlleva un nivel cognitivo de complejidad alto (creación, nivel 6 de Bloom); en cambio en el id_{27}, el término "desarrollar aplicaciones Web" implica un nivel de complejidad medio (aplicación, nivel 3 de Bloom). Así, id_{11} puede cubrir las necesidades de habilidad de id_{27} (porque están alineados a Tc_6).

De igual forma, se puede establecer qué habilidades requieren los perfiles laborales; por ejemplo, en el id_{35} el término "programar con Forms" implica un nivel de complejidad medio (aplicación, nivel 3 de Bloom) que el id_9 no puede satisfacer (por ejemplo, con el término "conocimiento de ingeniería de sistemas", que está relacionado a un nivel de complejidad bajo); lo cual permite deducir qué competencias requieren las ofertas de empleo y qué perfiles académicos pueden cubrirlas o no. Además, es posible identificar qué universidades tienen sus perfiles académicos alineados con ofertas de trabajo, como en el caso de los perfiles id_{11} e id_{19}, ambos tienen un nivel de complejidad alto (Evaluación, nivel 6 de Bloom por los términos "generar soluciones técnicas" y "mantenimiento de sistemas informáticos", respectivamente). Estos resultados se pueden utilizar en diferentes contextos, tales como la planificación de carreras profesionales, reclutamiento de personal, entre otros dominios.

4.1.5. Actualización

En este componente se realiza la actualización del modelo ontológico OC detallado en el apartado 3.2.1.2., utilizando la información extraída en la fase de comparación (ver Tablas 4.6. y 4.7.). Seguidamente, se valida el modelo OC mediante las métricas de Completitud, Robustez y Entropía (definidas en el apartado 3.2.4.).

La Figura 4.8. presenta un extracto de la población ontológica de OC con los términos extraídos de los perfiles id_1, id_2 e id_{21}, considerando los axiomas de la Tabla 3.1. Como se observa, las clases "Instancia_Habilidad" e "Instancia de conocimiento" contienen a los términos de conocimiento y habilidad que fueron seleccionados en la fase de extracción, para la comparación con los tesauros. También, la clase "Patron_Conocimiento" contiene los patrones relacionados con los términos, en este caso, para identificar términos de conocimiento. También, la clase "Categoría Gramatical" contiene las palabras que conforman los términos según su categoría gramatical, particularmente, las categorías "Sustantivo" y "Adjetivo".

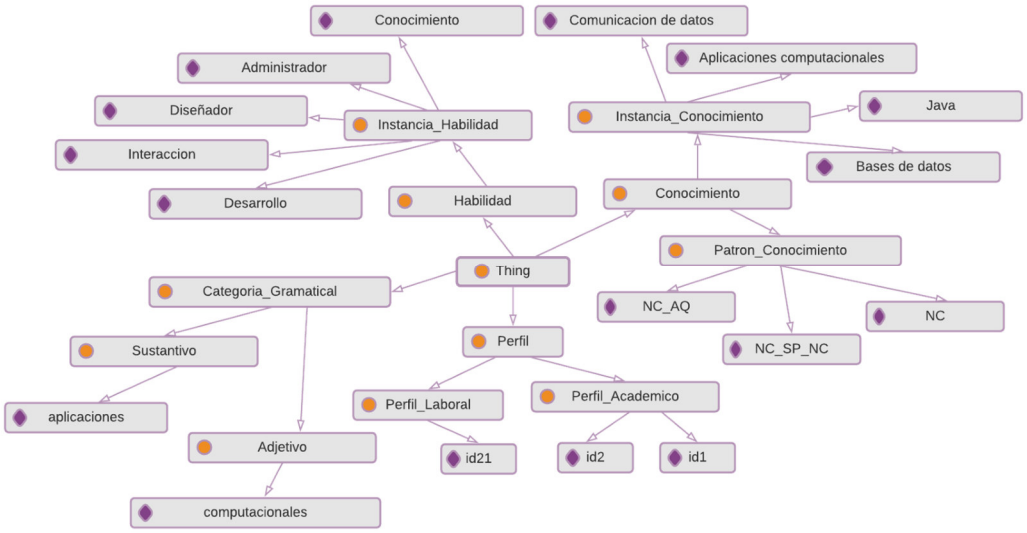

Figura 4.8. Extracto de la Ontología OC, para los perfiles id_1, id_2 e id_{21} según los axiomas de la Tabla 3.1.

La Figura 4.9. presenta un extracto de la población del modelo OC, según las definiciones 3.12. y 3.13., para los términos de conocimiento y habilidad de los perfiles id_1, id_2 e id_{21}, cuya medida de similitud semántica supera el umbral $U_S > 0.45$ (ver Tablas 4.6. y 4.7.). Como se observa, la clase "Dominio_Cognitivo", contiene a los términos de habilidad que pertenecen a las clases "Síntesis", "Aplicación" y "Evaluación", relacionadas con el tesauro BLOOM. La pertenencia se establece según el valor de la medida de similitud máxima entre el término del perfil y el término del tesauro. Por ejemplo, a la clase "Síntesis" pertenecen los términos "Administrador", "Diseñador" e "Interacción" (ver Tabla 4.7.). En cuando a la clase "Cobertura_Conocimiento", contiene las clases "Instalación_Configuracion_TI", "Programación" y "Conocimiento_Bases_Datos"; y a la clase "Programación" le pertenecen los términos "Aplicaciones computacionales y "Java" (ver Tabla 4.6.).

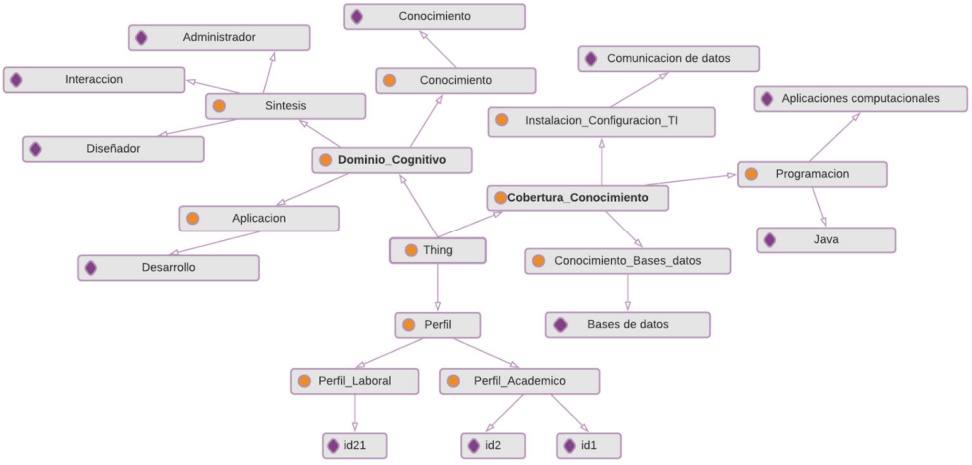

Figura 4.9. Extracto de la Ontología OC, para los perfiles id_1, id_2 e id_{21} según las definiciones 3.12. y 3.13.

El siguiente paso después de la población ontológica consiste en la validación del modelo ontológico. La Tabla 4.12. presenta los resultados generales de la calidad de la ontología en cuanto a su completitud y robustez ontológica. Para interpretar los resultados, se considera la

escala de valoración cualitativa definida en el Enunciado 3.14. En general, para el 71.44% de perfiles, la ontología tiene una completitud Alta, y, para el 22.85% de los perfiles una completitud Media; la completitud de nivel Bajo corresponde a dos perfiles (5.71%). Eso implica que, en promedio, todos los términos relevantes son usados para poblar el modelo ontológico OC, un resultado que indica la utilidad del proceso. En cuanto a la robustez ontológica, el 28.57% de los perfiles tienen una relevancia Alta, mientras que el 68.57% de los perfiles tienen una relevancia media. En este caso, prácticamente todos los componentes con los que se pobló el modelo ontológico son relevantes para los perfiles usados. Con respecto al umbral $U_{RE} > 0.3$ (Enunciado 3.13.), se observa que, para la medida de Completitud, solamente 5 de los 35 perfiles no superan el umbral (id_{30}, id_{33}, id_{24}, id_{16} e id_{23}). En cambio, los otros 30 perfiles sí aportan términos relevantes para la población del modelo OC. De la misma forma, solamente 4 perfiles (id_{18}, id_{22}, id_6 e id_{23}) tienen valores de Robustez ontológica menores a 0.3. Sin embargo, 31 perfiles contribuyen con términos relevantes para poblar la ontología OC. En consecuencia, desde la perspectiva de estas métricas, la ontología OC ha sido poblada con términos relevantes.

Tabla 4.12. Cálculo de la Completitud y Robustez ontológica del modelo ontológico OC

Perfil	Completitud	Valor
id_2	1.00	Alto
id_{14}	0.93	
id_{15}	0.87	
id_{19}	0.85	
id_{18}	0.81	
id_4	0.77	
id_8	0.76	
id_6	0.71	
id_{11}	0.71	
id_{29}	0.71	
id_5	0.7	
id_{12}	0.7	
id_7	0.68	
id_{20}	0.67	
id_9	0.63	
id_{10}	0.63	
id_{17}	0.63	
id_{27}	0.63	
id_{13}	0.61	
id_{21}	0.61	
id_{26}	0.59	
id_{32}	0.58	
id_{25}	0.54	
id_{31}	0.51	
id_3	0.5	
id_{35}	0.48	Medio
id_{22}	0.47	
id_{28}	0.46	
id_1	0.42	
id_{34}	0.32	
id_{30}	0.29	
id_{33}	0.27	
id_{24}	0.26	
id_{16}	0.19	Bajo
id_{23}	0.14	

Perfil	Robustez	Valor
id_{24}	0.94	Alto
id_{21}	0.66	
id_{30}	0.66	
id_{28}	0.65	
id_{25}	0,6	
id_1	0.53	
id_{16}	0.57	
id_{32}	0.53	
id_{27}	0.5	
id_{33}	0.5	
id_{13}	0.49	Medio
id_{34}	0.48	
id_{26}	0.47	
id_{11}	0.46	
id_{35}	0.46	
id_{10}	0.45	
id_{17}	0.45	
id_8	0.43	
id_4	0.42	
id_{29}	0.42	
id_3	0.42	
id_{14}	0.4	
id_9	0.38	
id_{19}	0.37	
id_{12}	0.36	
id_{15}	0.36	
id_2	0.39	
id_7	0.33	
id_{31}	0.33	
id_5	0.32	
id_{20}	0.3	
id_{18}	0.28	
id_{22}	0.26	
id_6	0.24	
id_{23}	0	Bajo

4.2. CASO 2: ANÁLISIS DE PERFILES PROFESIONALES SEGÚN LA LÓGICA DIALÉCTICA

4.2.1 Procesamiento de los datos del experimento

Para el desarrollo del experimento se toma como entrada la colección de 35 documentos en español: 20 perfiles académicos, obtenidos de portales universitarios (id_1, ..., id_{20}), y 15 ofertas de trabajo, obtenidas de portales de empleo en internet (id_{21}, .. id_{35}), que se usaron en Caso 1 (apartado 4.1.). La Tabla 4.13. presenta un extracto de los perfiles de la colección (id_1, id_2,id_{21}), en donde a cada término de habilidad y conocimiento del perfil se asocia información de contexto, como es el caso de la sección del documento donde fue encontrado (Ubicación), el patrón lingüístico del término, y la validación del experto sobre qué tipo de componente es (habilidad, conocimiento o competencia). Esta información constituye los hechos de los axiomas de los casos dialécticos 1, 3 y 4.

De igual forma, se considera como hechos del caso 5 del MD la información sobre el alineamiento de los términos de conocimiento con el tesauro DISCO II, cada uno con su respectiva medida de similitud semántica y término raíz del subárbol del tesauro (md). Además, se considera como hechos del caso 2, la información sobre el alineamiento de los términos de habilidad, con su correspondiente nivel cognitivo (mb). Cabe señalar que está información es utilizada por el modelo dialéctico, en cada uno de sus axiomas, para determinar la ambigüedad de los términos de los perfiles.

Tabla 4.13. Extracto del dataset para el MD

c	id	H	C	Ubicación	Patrón	Experto
1	id_1	Diseñar Administrar	Sistemas de comunicación	Antecedentes	NC-SP-AQ	Conocimiento
1	id_2	Desarrollar	aplicaciones computacionales	Perfil de la carrera	NC-SP-NC-AQ	Conocimiento
2	id_{21}	Interacción	bases de datos	Requisitos	NC-SP-NC	Conocimiento
2	id_{21}	Conocer	Java	Experiencia	NC-SP-NC	Conocimiento

Una vez definido el dataset para el experimento, se realiza la ejecución de los axiomas de los cinco casos dialécticos del modelo MD, detallados en el apartado 3.3.2., utilizando la herramienta RM3 [83], para identificar los eventos dialécticos sd_{ji} existentes en los perfiles.

4.2.2 Validación del modelo dialéctico

En este apartado se realiza la validación de los resultados obtenidos por los axiomas del modelo dialéctico MD, en función de la medida de Robustez dialéctica de la Definición 3.17. (apartado 3.3.3), la cual determina la capacidad del modelo para identificar términos dialécticos en los perfiles.

La Tabla 4.14. presenta los resultados de la medida de Robustez dialéctica del modelo MD, para los diferentes fenómenos del lenguaje natural definidos por los axiomas. Así, en el caso 1 dialéctico de la sección anterior (Vaguedad o falta de claridad, precisión o exactitud en el lenguaje natural), el 93% de los perfiles tienen valores de robustez dialéctica que oscilan entre 0.8 y 1, lo que indica que MD reconoce todos los términos etiquetados como dialécticos. Esta tendencia se mantiene para los casos 2, 3 y 4, donde alrededor del 90% de los perfiles tienen

valores de robustez entre 0.8 y 1, lo que indica que los axiomas y la base de hechos del modelo MD permiten identificar los términos dialécticos relevantes en los perfiles. Cabe mencionar que los valores de Robustez dialéctica se mantienen en el rango de 0.8 a 1, tanto para el grupo de perfiles académicos (id_1 - id_{20}), como de los perfiles profesionales (id_{21} – id_{35}).

De igual forma, en el caso 5 (Razonamiento contrafáctico), el umbral no afecta el valor de robustez dialéctica de cada perfil, ya que se pueden reconocer los términos dialécticos, independientemente del umbral utilizado para determinar los términos relevantes (el cual establece si la ontología OC tendrá un mayor o menor número de términos). Finalmente, se observan perfiles en los que se mantiene la medida de robustez dialéctica para los cinco casos (id_{20}), y en algunos perfiles no se aplica el cálculo de la Robustez por no tener términos dialécticos (n/a). Por ejemplo, id_{12} e id_{15} para el caso 3; e id_{29} para los casos 1, 2, 3 y 4.

Tabla 4.14. Cálculo de la Robustez dialéctica del MD

id	Caso 1	Caso 2	Caso 3	Caso 4	Caso 5				
					U1 0.2	U2 0.3	U3 0.4	U4 0.5	U5 0.6
id_1	0.86	0.83	0.82	0.81	0.83	0.81	0.77	0.78	0.78
id_2	0.8	0.87	0.8	0.8	0.8	0.8	0.81	n/a	n/a
id_3	0.8	1.00	0.87	0.67	0.8	0.8	0.8	0.8	0.8
id_4	0.88	0.67	0.88	0.88	0.75	0.8	0.8	0.8	0.8
id_5	0.87	1.00	0.93	0.83	0.83	0.8	0.8	0.8	0.8
id_6	0.8	1.00	0.8	0.8	0.8	0.8	0.8	0.8	n/a
id_7	0.87	0.83	0.93	0.93	0.83	0.8	0.8	0.8	0.8
id_8	0.85	0.89	0.85	0.85	0.75	0.75	0.73	0.75	0.73
id_9	0.9	1.00	0.8	0.7	0.83	0.8	0.8	0.8	n/a
id_{10}	0.8	1.00	0.8	0.8	0.8	0.8	0.8	n/a	n/a
id_{11}	0.9	1.00	0.8	0.9	0.8	0.82	0.8	0.8	1
id_{12}	0.76	0.83	n/a	0.86	0.76	0.8	0.8	0.8	0.9
id_{13}	0.84	0.95	0.84	0.84	0.84	0.84	0.84	0.8	0.8
id_{14}	0.8	0.67	0.96	0.8	0.8	0.8	0.8	0.77	n/a
id_{15}	0.8	1.00	n/a	0.8	0.8	0.8	1	n/a	n/a
id_{16}	0.67	1.00	0.89	0.87	0.88	0.87	0.87	0.87	0.8
id_{17}	0.86	0.67	0.8	0.9	0.8	0.8	0.8	n/a	n/a
id_{18}	0.91	1.00	0.86	0.8	0.8	0.8	0.8	n/a	n/a
id_{19}	0.68	0.87	0.86	0.86	0.85	0.87	0.85	0.8	0.8
id_{20}	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
id_{21}	0.81	0.81	0.59	0.79	0.81	0.83	0.85	0.8	0.8

id_{22}	0.87	0.87	0.87	0.8	0.8	0.79	0.8	n/a	n/a
id_{23}	0.81	n/a	0.86	0.8	0.8	0.8	n/a	n/a	n/a
id_{24}	0.87	1.00	0.81	0.75	0.84	0.87	0.87	0.87	0.8
id_{25}	0.82	1.00	0.8	0.84	0.87	0.87	0.87	0.8	0.8
id_{26}	0.7	0.89	0.75	0.7	0.75	0.73	0.73	0.78	1
id_{27}	0.87	0.83	0.87	0.87	0.79	0.72	0.77	0.77	0.77
id_{28}	0.89	0.95	0.81	0.8	0.86	0.84	0.86	0.88	0.8
id_{29}	n/a	n/a	n/a	n/a	0.87	0.87	0.87	0.87	0.8
id_{30}	0.8	1.00	0.8	0.8	0.8	0.8	n/a	n/a	n/a
id_{31}	0.83	0.83	0.83	0.83	0.83	0.8	0.8	0.8	n/a
id_{32}	0.87	0.67	0.87	0.87	0.8	0.8	0.8	0.8	0.8
id_{33}	0.79	0.89	0.69	0.8	0.88	0.87	0.82	0.83	0.84
id_{34}	0.85	1.00	0.82	0.83	0.85	0.86	0.88	0.82	0.85
id_{35}	0.81	0.83	0.68	0.8	0.79	0.77	0.8	0.8	0.8

4.3 COMPARACIÓN DE LA ONTOLOGIA OC CON LOS RESULTADOS DEL MD

En este apartado se quiere evaluar la relación entre lo detectado como ambigüedad por el MD y la calidad de la información almacenada por OC. Para ello, se compara la entropía del OC (que indica el nivel de incertidumbre presente en OC) con la proporción de eventos dialecticos detectados por el MD en OC.

La Tabla 4.15. presenta el cálculo de la entropía de los perfiles para el OC, donde la entropía es cero cuando se considera que todos los términos relevantes del perfil id_i no introducen incertidumbre. En los resultados se observa que una gran mayoría tiende a cero (entre 0,5 y cero).

Además, la proporción de términos reconocidos por el modelo MD sobre el total de términos relevantes en la ontología OC ($Pr\_MD(id_i)$) es cero cuando no reconoce los términos relevantes como dialécticos (por ejemplo, id_2), y esta medida tiende a uno (1) cuando reconoce un gran número de términos como términos dialécticos (por ejemplo, id_{13}). Como se ve en los resultados, Pr_MD sigue la métrica de entropía; por ejemplo, para los perfiles id_2 e id_8, ambos valores son cero, lo que indica que no hay incertidumbre en el modelo OC, detectado por Pr_MD cuando su valor es cero, que significa que los términos no son dialécticos.

Para el resto de los perfiles, se observa que ambos valores son cercanos, lo que significa que la determinación de los términos dialécticos se correlaciona con la entropía de la OC. Este resultado es muy importante, porque indica que MD puede ser utilizado para determinar la incertidumbre presente en una ontología; además, permite analizar cuál de los tipos de ambigüedad está presente en la ontología (algo que no se puede hacer con la métrica de entropía).

Tabla 4.15. Análisis de MD y OC desde la Entropía

id	$H_{OC}(id_i)$	$Pr\_MD(id_i)$
id_1	0.5	0.39
id_2	0	0
id_3	0.39	0.32
id_4	0	0.16
id_5	0.39	0.28
id_6	0	0
id_7	0.19	0.09
id_8	0	0
id_9	0	0.14
id_{10}	0	0.14
id_{11}	0	0.23
id_{12}	0.26	0.20
id_{13}	0.06	0.08
id_{14}	0	0.09
id_{15}	0	0
id_{16}	0.40	0.38
id_{17}	0.5	0.34
id_{18}	0	0
id_{19}	0	0.08
id_{20}	0	0.09
id_{21}	0.53	0.49
id_{22}	0	0.14
id_{23}	0.53	0.53
id_{24}	0.31	0.47
id_{25}	0	0.14
id_{26}	0	0.19
id_{27}	0.53	0.47
id_{28}	0	0.16
id_{29}	0.19	0
id_{30}	0.35	0.52
id_{31}	0.19	0.09
id_{32}	0.5	0.46
id_{33}	0.46	0.39
id_{34}	0.44	0.41
id_{35}	0.53	0.47

CAPITULO 5: ANÁLISIS DE RESULTADOS

En el presente capítulo se realiza una comparación de nuestro trabajo con otros trabajos en el área de la gestión de competencias. Además, se realiza una contextualización del contexto del problema de estudio con enfoques tradicionales, y su relación con los componentes del modelo propuesto.

5.1. COMPARACIÓN CON OTROS TRABAJOS

En la Tabla 5.1. se presenta la comparación de nuestra propuesta con otros trabajos, considerando los siguientes criterios:

- **Modelo de conocimiento utilizado**. El modelo de conocimiento establece la estructura semántica utilizada para representar la información
- **Componentes de la competencia considerados**, determinan los elementos involucrados en el análisis (por ejemplo: habilidades, conocimiento, actitudes).
- **Método de reconocimiento de componentes:** referente a la técnica de PLN que se usa para la extracción de los componentes de competencia.
- **Estrategia de desambiguación**, define los métodos utilizados para lidiar con la ambigüedad en las unidades de información
- **Fuentes semánticas (tesauros)**, indican las bases de conocimiento utilizadas para apoyar el proceso de desambiguación
- **Fuentes de datos,** indican el origen de la información analizada en cada trabajo
- **Métricas de validación**, establecen las medidas utilizadas para verificar los resultados obtenidos en cada investigación

Tabla 5.1. Comparación de la propuesta con otros trabajos

Trabajo	Modelo de conocimiento	Componente de competencia	Método de reconocimiento de componentes	Estrategia de desambiguación	Fuentes semánticas (tesauros)	Fuentes de datos	Métricas de validación
[30]	Ontología	Conocimiento	NER	Promedio	WordNet	Ofertas de empleo	Experto
[25]	Ontología	Habilidad	NER	Axiomas LPO	Linked Data competence ontology	Perfiles académicos	Precisión y relevancia
[35,4]	Ontología	Conocimiento	Patrones	Medidas de similitud	Competence ontology	Corpus	Precisión Recall
[38]	Ontología	Conocimiento	Patrones	Medidas de similitud	Lexicon y Onomasticon	Texto Web	Precisión relevancia
[73]	Ontología	Conocimiento	---	Algoritmo ACO	-----	Texto Web	Precisión
[101]	Ontología	Verbos	Parones	Coseno Dice measures	Adesse Wordnet Sensem	Casos de verbos	Pearson Ratio
[102]	Ontología	Conocimiento y habilidad	NER	----	Europass, IEEE	Currículos	Fuzzy-based aproach
[114]	Ontología	Habilidad, objetivos de aprendizaje	---	Por expertos	ACM, IEEE, TEQSA	----	Precisión
[102]	Ontología	Conocimiento, habilidad y actitud	NER	Por razonadores	RCD, IEEE	Currículos	Precisión
[100]	---	Títulos de trabajo, niveles de educación, habilidad	NER	Por recomendación	---	Currículos	Precisión
[99]	---	Conocimiento habilidad	NER	Medidas de distancia	---	Ofertas de trabajo	Precisión
[98]	Ontología	Competencia	Patrones	---	---	Currículos	---
[96]	Ontología	Competencia	NER	Medidas de distancia	---	Perfiles	Precisión
Nuestra propuesta	Ontología y modelo dialéctico	Conocimiento y habilidad	Patrones lingüísticos	Axiomas dialécticos y medidas de simiitud	DISCO II BLOOM	Ofertas de trabajo y perfiles académicos	Completitud Robustez ontológica y dialéctica Entropía

En el modelo de conocimiento, la mayoría de los trabajos utilizan ontologías para la representación de los componentes de conocimiento y habilidad, lo que limita estos trabajos al campo de la Lógica Descriptiva. Nuestra propuesta utiliza, adicional a la ontología de perfiles profesionales y académicos, un modelo dialéctico para el reconocimiento de casos de contradicción. Comparando los resultados de completitud, robustez y entropía obtenidos, es claro que el modelo dialéctico resuelve el problema de detectar ambiguedades porque tiene la capacidad de reconocimiento de las mismas y analizarlas (determiner el tipo de ambiguedad presente). En relación a los componentes analizados, muchos trabajos coinciden en el análisis de los componentes, siendo el conocimiento uno de los más analizados. Nuestro trabajo considera conocimiento, habilidad y competencia, que se analizan a partir de los cinco fenómenos dialécticos.

Con respecto al método de reconocimiento de los componentes, es frecuente encontrar que varios de los trabajos utilizan técnicas de PLN, como es el caso del reconocimiento de entidades nombradas (NER). Esto sucede porque las competencias se asocian con descripciones de títulos y áreas de conocimiento o experiencia, para los cuales existen diccionarios o tesauros que se usan como soporte en el proceso de reconocimiento. En otros casos se observa el uso de patrones para identificar competencias, que luego son comparadas con diccionarios, taxonomías o tesauros de dominio en el idioma en que se encuentra el perfil. Nuestro modelo utiliza patrones lingüísticos que han sido adaptados al idioma español, y a las características lingüísticas de los componentes de competencia encontrados en los perfiles. Esto, con el fin de incrementar la precision en el reconocimiento de términos.

Con respecto a las estrategias de desambiguación, varios métodos están orientados al uso de modelos de espacio vectorial con medidas de similitud y algoritmos, que alinean los componentes con tesauros, con el objetivo de eliminar la ambigüedad de los componentes analizados. En nuestra propuesta, también se usan medidas de similitud contra tesauros, pero adicionalmente, se realiza un estudio de la ambigüedad de los componentes basado en lógica dialéctica, lo que permite obtener otra perspectiva de los perfiles profesionales: detectar las contradicciones que presentan los elementos de competencia. En cuanto a los tesauros, la mayoría de ellos se encuentran en idiomas como el inglés y el alemán, lo que supone una limitación para el análisis y la comparación de los componentes de competencia en lengua española. En nuestra propuesta se considera el tesauro DISCO II, que tiene la ventaja de ser multilenguaje, y el tesauro BLOOM que ofrece varios sinónimos de los verbos de la taxonomía de BLOOM; lo que da flexibilidad para la replicabilidad de nuestro modelo para otros idiomas.

Normalmente, otros trabajos utilizan datos etiquetados (para la medida de precisión), o expertos que generan un “gold standard” para la comparación de los resultados de la población del modelo ontológico. Nuestra propuesta usa las medidas de Completitud y Robustez ontológica para establecer la calidad de modelo ontológico (OC), según la relevancia de los términos seleccionados para llenar la ontología. Además, mediante la Entropía, se determina la incertidumbre que cada perfil introduce a la OC. Por otra parte, la Robustez dialéctica del modelo dialéctico establece su capacidad para reconocer los términos ambiguos en los documentos.

5.2. CONTEXTUALIZACIÓN DE RESULTADOS

Uno de los procesos de gestión de competencias es la construcción de perfiles profesionales, de los cuales surge el conjunto de competencias del individuo ideal para ocupar con éxito un puesto

de trabajo [104]. Con esta información, las empresas planifican la formación que necesitan sus empleados, y las universidades desarrollan sus perfiles profesionales [1]. Debido a la dinámica del entorno empresarial y a la no estandarización de las competencias centrales de la mayoría de las empresas [105], las competencias tienen diferentes interpretaciones recogidas en las ofertas laborales, que son el resultado del conocimiento, la experiencia y las creencias del editor. De esta forma, las universidades reciben competencias ambiguas del entorno laboral, que dificultan las alineaciones entre perfiles académicos y profesionales [106].

Del mismo modo, los programas de grado universitario construyen las competencias de los alumnos, intentando gestionar la aparición de nuevos puestos y la creciente necesidad de expertos en determinadas áreas [2]. Al mismo tiempo, los programas deben estar alineados con las metas de aprendizaje definidas en estándares y marcos (por ejemplo, ACM[31], EQF[32], etc.) para cumplir con las regulaciones de las entidades gubernamentales. En general, las competencias (objetivos de aprendizaje) de estos cuerpos profesionales varían en sus descriptores, granularidad, especificidad y estructura [103, 114], generando ambigüedades en los procesos de desambiguación de habilidades y conocimientos dentro de los modelos ontológicos de competencias [102], y consecuentemente, en las descripciones de las titulaciones [107] y el temario de las asignaturas [108].

En la actualidad, la gestión de competencias encuentra complicaciones para comprender el significado real de la competencia en perfiles profesionales digitales no estructurados, donde la trascendencia de las competencias depende directamente del conocimiento y percepción del editor. Así, una de las principales limitaciones es la interpretación de competencias, que puede dar lugar a más de un significado. Por ejemplo, encontramos habilidades que describen diferentes niveles y procesos cognitivos al mismo tiempo y, las creencias del editor determinan suposiciones erróneas sobre habilidades o conocimientos. De esta forma, nos encontramos con ambigüedades en la alineación de términos de competencia, que constituyen un problema del lenguaje natural, afectando la construcción de modelos de competencias.

En general, la falta de claridad de competencias genera ambivalencias y malentendidos, que no permiten su uso en la creación de formación o currículos específicos [110]. El plan de estudios del proceso de aprendizaje requiere definiciones de dominio y alcance de conocimientos y habilidades para ofrecer una adecuada adquisición de competencias [109]. Por otro lado, las universidades actualizan los perfiles académicos con las nuevas necesidades del mercado laboral [105], pero si los requisitos laborales son ambiguos, cómo los perfiles pueden cumplir con las competencias deseadas del candidato [107] y, garantizar la relevancia de los procesos de validación de la adquisición de competencias [103, 114]. Entonces es necesario proponer modelos para el reconocimiento de ambigüedad competencial de los perfiles profesionales para afrontar los crecientes retos en el entorno empresarial y académico actual [106].

Los modelos semánticos, como las ontologías, modelan las relaciones entre las habilidades y los conocimientos, proporcionando el marco teórico y contextual para la creación de perfiles profesionales [102]. A pesar de esto, estos modelos enfrentan problemas de ambigüedad del lenguaje entre términos y conceptos que, durante la elicitación ontológica, provocan ambivalencias al unificar información de múltiples fuentes [2]. Además, las ontologías utilizan lenguajes formales, como la Lógica Descriptiva, para describir conceptos y sus relaciones asignándoles valores verdaderos o falsos [85], lo que limita la capacidad de la ontología para

[31] ACM: Association for Computing Machinery (Asociación de Maquinaria Computacional)
[32] EQF: European Qualification Framework

representar el significado del término ambiguo, dependiendo del contexto al que pertenece y los contextos relacionados. [86]. En consecuencia, las ontologías no son eficientes para la representación de la ambigüedad en los perfiles académicos y profesionales.

Por otro lado, la lógica dialéctica considera que las fórmulas lógicas pueden ser verdaderas, falsas o ambas, lo que permite modelar términos de competencia ambiguos. Según [70, 83], existen cinco fenómenos dialécticos, que se relacionan con la ambigüedad léxica del lenguaje natural (como es el caso de sinónimos, homónimos, hipónimos e hiperónimos), y cómo el editor usa estos términos ambiguos de la misma manera que sus creencias y conocimiento. Desde el punto de vista de las ontologías existen métodos de desambiguación ontológica para tratar casos de incertidumbre [52], vaguedad [69] e imprecisión [111], sin embargo, estos modelos no analizan los fenómenos dialécticos de competencias en los perfiles.

La lógica descriptiva juega un papel crucial en los modelos ontológicos que actualmente sustentan la gestión de competencias, como los lenguajes OWL Lite y OWL DL, que se basan fundamentalmente en esta lógica [102]. En general, la lógica descriptiva proporciona un formalismo de primer orden decidible, con una semántica declarativa simple y bien establecida, para capturar el significado de las características más populares de la representación estructurada del conocimiento [69].

Sin embargo, la lógica descriptiva no es confiable para la representación de la información que contiene incertidumbre y/o vaguedad y/o imprecisión, pues puede ocasionar inconsistencias en los procesos de razonamiento ontológico, debido a su incapacidad para manejar estas ambivalencias [85]. Por ejemplo, bajo evidencia incompleta y conocimiento parcialmente inconsistente, donde es imposible describir exactamente el estado existente, un resultado futuro o más de un resultado posible [86].

Los modelos ontológicos para la gestión de competencias presentan vaguedad e incertidumbre en su estructura a diferentes niveles: 1. Terminológica, porque no se establece exactamente a qué clase pertenece un individuo de la ontología debido a su ambigüedad léxica [1]. Este caso es muy frecuente en aquellos modelos de competencia donde una población ontológica se compone de fuentes no estructuradas, como es el caso de los perfiles profesionales [19]; y, 2. Estructural, donde la ambigüedad semántica de los conceptos de ontologías provoca alineamientos y mezclas de ontologías que no representan el dominio de las competencias analizadas [87]. En consecuencia, la gestión de competencias requiere de modelos ontológicos eficientes, y a la vez flexibles, para la detección y tratamiento de fenómenos ambiguos.

De esta manera, a través de la presente tesis se presentan las principales contribuciones de este trabajo que buscan responder a las diferentes limitaciones narradas en los párrafos anteriores, las cuales son:

- El desarrollo de una arquitectura para el análisis de competencias de perfiles profesionales y académicos, que comprende las fases de caracterización, extracción, comparación y retroalimentación de competencias.
- La definición de un modelo ontológico para la identificación de competencias, en función de patrones lingüísticos de habilidad y conocimiento.

- La definición de un modelo dialéctico para el análisis de la ambigüedad de los perfiles profesionales en cuanto a competencias y sus componentes (habilidad y conocimiento).
- La definición de un esquema de alineamiento de perfiles profesionales y académicos contra tesauros, en función de medidas de similitud de términos de competencias y algoritmos de ranking.
- La definición de un esquema de retroalimentación de perfiles profesionales y académicos en función de medidas de completitud, y robustez ontológica y dialéctica.

CAPITULO 6: CONCLUSIONES Y TRABAJOS FUTUROS

En el capítulo 6 se presenta las conclusiones del trabajo doctoral. Además, se realiza una contextualización de los trabajos futuros que pueden desarrollarse, tomando como punto de partida la investigación realizada.

6.1. CONCLUSIONES

En este trabajo se ha desarrollado un Modelo adaptativo de perfiles académicos basado en competencias usando minería semántica, cumpliendo los objetivos planteados al inicio de la investigación. En primer lugar, se han estudiado las metodologías, métodos, técnicas y herramientas relacionadas con la minería semántica, PLN y aprendizaje ontológico disponibles actualmente. Igualmente, se estudiaron los aspectos teóricos relacionados con la gestión de competencias, lógica descriptiva y lógica dialéctica, para caracterizar competencias y sus elementos en los contextos académico y laboral. Posteriormente, se definió un modelo para el análisis de perfiles profesionales y académicos provenientes de la Web. Finalmente, se verificó el esquema propuesto a través de su aplicación en casos de estudio.

El modelo propuesto está compuesto por cuatro componentes: caracterización, extracción, comparación y actualización, y cada una de ellos está definido por macroalgoritmos que describen los pasos para el tratamiento de perfiles profesionales y académicos en formato textual. El componente de caracterización genera dos modelos: el modelo ontológico de perfiles (OC) y el modelo dialectico (MD), los cuales representan los elementos de competencia encontrados en los documentos según la perspectiva de la lógica descriptiva y dialéctica, representando tanto la parte lingüística léxica y semántica de los componentes de competencia, conocimiento y habilidad. A continuación, el componente de extracción usa técnicas de PLN para el reconocimiento y extracción de los términos de competencia de cada uno de los perfiles. En el componente de comparación se alinean los términos contra dos tesauros (DISCO II y BLOOM), estableciendo correspondencias de los términos en los perfiles con tópicos de conocimiento y habilidad. Finalmente, en el componente de actualización se realiza la retroalimentación de los perfiles en cuanto a las áreas de conocimiento y niveles cognitivos detectados en ellos (tanto en los laborales como en los académicos); adicionalmente, se reconocen las ambigüedades que ofrecen diferentes significados a los términos de competencia, habilidad y conocimiento. Esto último, particularmente se hace con los axiomas del modelo dialéctico MD.

Entre los aspectos más importantes del esquema propuesto destacan el uso combinado de medidas de similitud léxicas y semánticas para el análisis de los términos de competencia, lo cual incrementa la capacidad del alineamiento de los términos contra tesauros. Además, la definición de umbrales en las diferentes fases permite filtrar aquellos términos que no cumplen con el umbral, haciendo que en cada fase se obtenga un conjunto de términos de competencia relevantes. De igual forma, la definición de las métricas de Completitud, Robustez y Entropía para validar los modelos es otro aporte. En el caso del modelo ontológico, la Completitud establece el aporte de los perfiles a OC tomando en cuenta la frecuencia de los términos

seleccionados para la población ontológica; y, la Robustez ontológica determina la relevancia de los términos escogidos según su alineamiento con los tópicos de los tesauros, estableciendo un ranking entre los perfiles. En cuando a la Entropía, esta medida se utiliza para definir la incertidumbre que cada perfil introduce a OC. Finalmente, la medida de Robustez dialéctica del modelo dialéctico, refleja la aparición de términos dialécticos en los perfiles, según los axiomas del modelo MD.

Cabe mencionar que en la literatura revisada existen varias propuestas que analizan perfiles profesionales en función de competencias, pero no consideran la ambigüedad que existe en los perfiles, lo que dificulta la comprensión del significado de las competencias dentro del contexto; así también, las propuestas no consideran su aplicación en documentos en idioma español. El presente modelo cubre estas deficiencias con todo un esquema de análisis que es adaptable a los contextos académico y laboral, basado en una representación ontológica tradicional de las competencias y complementada con el análisis desde una perspectiva dialéctica formal de las ambigüedades existentes.

El modelo propuesto en este trabajo puede ser parte de sistemas automáticos para el desarrollo de competencias a lo largo de un programa de carrera, apoyando los procesos de validación y seguimiento del cumplimiento de competencias en las asignaturas del plan de estudios. También puede contribuir a la detección de ambigüedades léxicas en los estándares y marcos utilizados en el desarrollo de programas de grado. Así también, en entornos de aprendizaje inteligentes, para desarrollar rutas de aprendizaje flexibles para los estudiantes dentro y entre materias. De esta forma, nuestra propuesta da solución a la falta de capacidad de las universidades para el control automático de la adquisición de requerimientos laborales, y su implementación en las asignaturas a lo largo de toda la carrera.

También puede ser usado en el reclutamiento de personal, entre otros dominios. Particularmente, para comparar currículos con las ofertas del mercado laboral, y determinar aquellos candidatos con los requisitos adecuados para ocupar las plazas de trabajo. Por otro lado, las empresas pueden utilizar esta propuesta para evaluar la calidad y pertinencia de sus ofertas laborales, comparándolas con las de otras empresas, estableciendo requerimientos comunes y diferentes, determinando así la actualidad de las competencias de una plaza laboral. Además, con los resultados de la comparación de requisitos, se puede generar catálogos y perfiles de competencias actualizados, que pueden estandarizarse mediante el alineamiento con tesauros y cuerpos de conocimiento. De esta manera, las empresas pueden mantener actualizadas sus estructuras funcionales, y obtener buenos resultados en sus procesos de reclutamiento.

Los resultados obtenidos permiten la correcta interpretación de los perfiles académicos y profesionales digitales. En particular, la propuesta hibrida tiene una representación formal de competencias basada en lógica descriptiva, extendida con lógica dialéctica para encontrar ambigüedades y contradicciones en los componentes de habilidad y conocimiento. La validación del modelo dialéctico a través de la medida de Robustez dialéctica, permite determinar la capacidad del modelo para encontrar la presencia de eventos dialécticos en los perfiles. Además, la medida de Entropía permite establecer la cantidad de incertidumbre en un modelo ontológico en cuanto a los términos dialécticos que contiene, que luego pueden ser analizados usando el modelo dialéctico para determinar sus tipos de ambigüedades.

El modelo se ha desarrollado en un contexto en español para el dominio de la Informática. Puede extenderse a otros idiomas considerando la adaptación de los patrones lingüísticos, que

identifican los componentes de conocimiento, habilidad y competencia, de acuerdo con las características lingüísticas de la lengua donde se aplica nuestro modelo. Además, puede extenderse a otros dominios de conocimiento, utilizando otras áreas que contiene el tesauro DISCO II o complementarlo con otros tesauros.

6.2. TRABAJOS FUTUROS

Los trabajos futuros están orientados a la integración del modelo de conocimiento propuesto con modelos semánticos, como las ontologías basadas en datos enlazados, de manera que permitan un análisis más profundo de las competencias utilizando información obtenida de la Web. También, la multilingualidad que ofrece el tesauro DISCO permite el alineamiento de los perfiles profesionalidades en español con otros propuestos en otros idiomas, lo cual facilita el análisis de las competencias en diversos contextos laborales y académicos de países con idiomas oficiales diferentes. De igual forma, dado que la gran mayoría de tesauros y cuerpos de conocimiento se encuentran en inglés, se puede explorar alineamientos con esos tesauros, para enriquecer el análisis de los perfiles.

Por otro lado, nuestra propuesta puede integrarse con plataformas de e-learning para el análisis de la evolución de competencias a lo largo de un programa de carrera, validando el cumplimiento de competencias en las asignaturas del currículo (sistemas de tutoría inteligente), y detectando casos de ambigüedad en el planteamiento de habilidades y conocimientos. Además, se puede detectar ambigüedades en los estándares y marcos utilizados en el desarrollo de programas de grado (sistemas de diseño instruccional) y en entornos de aprendizaje inteligentes, para desarrollar syllabus que ofrezcan rutas de aprendizaje claras para los estudiantes dentro y entre materias.

De igual manera, en el contexto laboral, el modelo puede integrarse con plataformas de reclutamiento para crear nuevos modelos de análisis de currículos, detectando requisitos y ambigüedades entre las definiciones de competencias, conocimientos y habilidades. También, en la planificación de la capacitación continua, puede validar las competencias en los planes de entrenamiento de personal y en los manuales de funciones, buscando alineamientos y contradicciones.

Por último, se abre un campo de estudio sobre las ambigüedades en los diferentes documentos y cuerpos de conocimiento utilizados por las universidades y las empresas para describir las competencias, para lo cual se puede usar el modelo dialéctico de nuestra propuesta. Particularmente, se puede usar para verificar la presencia de los casos definidos en los axiomas, y también encontrar nuevos casos de contradicción que puedan incluirse en el modelo.

6.3. REFERENCIAS

[1] N. Malzahn, S. Ziebarth y H.U. Hoppe, «Semi-automatic creation and exploitation of competence ontologies for trend aware profiling, matching and planning,» *Knowledge Management & E-Learning*, vol. 5, nº 1, pp. 84-103, 2013.

[2] P. De Leenheer, S. Christiaens y R. Meersman, «Business semantics management: a case study for competency-centric HRM,» *Computers in Industry*, vol. 61, nº 8, pp. 760-775, 2010.

[3] E.Z. Zinder y I.G. Yunatova, «Conceptual framework, models, and methods of knowledge acquisition and management for competency management in various areas,» *International Conference on Knowledge Engineering and the Semantic Web,* pp. 228-241, 2013.

[4] M. Fazel-Zarandi y M. S. Fox, «Inferring and validating skills and competencies over time,» *Applied Ontology*, vol. 8, nº 3, pp. 131-177, 2013.

[5] C.K.S. Irvine y J.M. Kevan, «Competency-based education in higher education,» *Handbook of research on competency-based education in university settings,* pp. 1-27, 2017.

[6] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, F.J. García-Peñalvo y E. Tovar-Caro, «Competence gaps in software personnel: A multi-organizational study,» *Computers in Human Behavior*, vol. 29, nº 2, pp. 456-461, 2013.

[7] Q. Quboa y M. Saraee, «A State-of-the-Art Survey on Semantic Web Mining,» *Intelligent Information Management*, vol. 5, nº 1, pp. 10-17, 2013.

[8] P. Banerjee, «Semantic Data Mining,» *Encyclopedia of Data Warehousing and Mining*, pp. 1765-1770, 2009.

[9] R. Gluga, J. Kay y T. Lever, «Foundations for modeling university curricula in terms of multiple learning goal sets,» *IEEE Transactions on Learning Technologies*, vol. 6 nº 1, pp. 25-37, 2013.

[10] M.I. Enăchescu, «Screening the Candidates in IT Field Based on Semantic Web Technologies: Automatic Extraction of Technical Competencies from Unstructured Resumes,» *Informatica Económica*, vol. 23 nº 4, pp. 51-65, 2019.

[11] M.C. Justicia de la Torre, «Nuevas técnicas de Minería de Textos: aplicaciones,» Tesis Doctoral Universidad de Granada, 2017.

[12] P. Qi, Y. Zhang, Y. Zhang, J. Bolton y C.D. Manning, «Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,» *58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101-108, 2020.

[13] M. Justicia de la Torre, D. Sánchez Fernández, I.J. Blanco Medina y M.J. Martín-Bautista, «Text knowledge mining: An approach to text mining,» *ESTYLF*, 2008.

[14] C. D. Manning y S. Hinrich, «Foundations of statistical Natural Language Processing,», *Cambridge: MIT press*, vol. 999, 1999.

[15] M. Vallez y R. Pedraza-Jimenez, «El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines,» *Hipertext.net*, nº 5, 2007.

[16] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard y D. McClosky, «The Stanford CoreNLP natural language processing toolkit,» *52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55-60, 2014.

[17] J. Aguilar «Introducción a la Minería Semántica,» 2018.

[18] A. Gatt, Albert y E. Krahmer, «Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation,» *Journal of Artificial Intelligence Research,* vol. 61, pp. 65-170, 2018.

[19] V. Janev y S. Vranes, «Ontology-based Competency Management: The Case Study of the Mihajlo Pupin Institute,» *Universal Computer Sciences*, vol. 17, nº 7, pp. 1089-1108, 2011.

[20] C. Faria, I. Serra y R. Girardi, «A domain-independent process for automatic ontology population from text,» *Science of Computer Programming*, vol. 95, pp. 26-43, 2014.

[21] F.M. Hassan, I. Ghani, M. Faheem y A.A. Hajji, «Ontology matching approaches for eRecruitment,» *International Journal of Computer Applications*, vol. 51, nº 2, 2012.

[22] P.D. Turney & P. Pantel, «From frequency to meaning: Vector space models of semantics,» *Journal of Artificial Intelligence research*, vol. 37, pp. 141-188, 2010.

[23] S. Harispe, S. Ranwez, S. Janaqi y J. Montmain, «Semantic Measures for the Comparison of Units of Language, Concepts or Instances,» *LGI2P/EMA Research Center, Parc scientifique*, France, 2013.

[24] B. Sateli, F. Löffler, B. König-Ries y R. Witte, «ScholarLens: extracting competences from research publications for the automatic generation of semantic user profiles,» *PeerJ Computer Science*, vol. 3, 2017.

[25] G. Paquette, D. Rogozan y O. Marino. «Competency comparison relations for recommendation in technology enhanced learning scenarios,» de Proc *the 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2012)*, vol. 896, pp. 23–34. 2012.

[26] L.A. Cabrera, B. Durette, M. Lafon, J.M. Torres-Moreno, y M. El-Bèze, «How Can We Measure the Similarity Between Résumés of Selected Candidates for a Job?,» *the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, pp. 99, 2015.

[27] J. Dorn, N. Tabbasum y M. Pichlmair, «Ontology development for human resource management,» *the 4th International Conference on Knowledge Managements*, pp. 109 -120, 2007.

[28] J. Huang, «Knowledge provenance: An approach to modeling and maintaining the evolution and validity of knowledge,» *Doctoral dissertation, University of Toronto*, 2009.

[29] J. Huang y D. Nicol, «A calculus of trust and its application to PKI and identity management,» *8th Symposium on Identity and Trust on the Internet*. ACM, 23-37, 2009.

[30] J. Dorn, J. y M. Pichlmair, «A Competence Management System for Universities,» *the European Conference on Information Systems*, 2007.

[31] L. Dittmann, S. Zelewski, «Ontology-based Skills Management,» *the 8th World Multi-conference on Systemics, Cybernetics and Informatics*, vol. 4, pp. 190-195, 2004.

[32] C. Bizer, R. Heese, M. Mochol, R. Oldakowski, R. Tolksdorf y R. Eckstein, «The impact of semantic web technologies on job recruitment processes,» *the Conf Wirtschaftsinformatik*, pp. 1367-1381, 2005.

[33] S. Poria, B. Agarwal, A. Gelbukh, A. Hussain y N.M. Howard, «Dependency-based semantic parsing for concept-level text analysis,» *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 113-127, 2014

[34] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen y A. Verkamo, «Fast discovery of association rules,» *Advances in Knowledge Discovery and Data Mining*, pp. 307– 328, 1996.

[35] M. Fazel-Zarandi y M.S. Fox, «Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach,» *the 8th International Semantic Web Conference*, 2009.

[36] M. Gómez, E. Aranda y J. Santos, «A competency model for higher education: an assessment based on placements,» *Studies in Higher Education*, pp. 1-21, 2016.

[37] P. Montuschi, F. Lamberti, V. Gatteschi y C. Demartini, «A semantic recommender system for adaptive learning,» *IT Professional*, vol 17, nº 5, pp. 50-58, 2015.

[38] T. Rodríguez y J. Aguilar, «Aprendizaje ontológico para el marco ontológico dinámico semántico,». *DYNA*, vol. 81, nº 187, pp. 56-63, 2014.

[39] M. El Asame y M. Wakrim, «Towards a competency model: A review of the literature and the competency standards,» *Education and Information Technologies*, vol. 23, nº 1, pp. 225-236, 2018.

[40] A. Savanevičienė, D. Stukaitė y V. Šilingienė, «Development of strategic individual competences,» *Engineering Economics*, vol 58, nº 3, 2015.

[41] M. Poblete y A. Villa, «SEBSCO, una experiencia alternativa para evaluar competencias,» *Aula Abierta*, vol. 39, nº 3, pp. 15-30, 2011.

[42] M. Torres, F. Benavidez, «La importancia de la gestión curricular universitaria en programas a distancia, estudio institución de educación superior suramericana,» *Crescendo*, vol. 10, nº 1, pp. 13-34, 2019.

[43] Tuning Academy, «Modelo tradicional versus modelo por competencias,» [En línea] Disponible: https://historia1imagen.files.wordpress.com/2016/07/modelo-tradicional-versus-por-competencia.pdf. [Ultimo acceso: 15 de agosto 2017].

[44] J. Tardif, «Desarrollo de un programa por competencias: de la intención a la puesta en marcha,» *PROFESORADO*, vol. 12, nº 3, pp. 1-16, 2008.

[45] B. Ramírez, E. Grass, C. Ordóñez, C. González, «Propuesta de incorporación de competencias de formación en ingeniería,» *Guillermo de Ockham: Revista Científica*, vol. 15, nº 1, pp. 13, 2017.

[46] S. Sanghi, «The handbook of competency mapping: understanding, designing and implementing competency models in organizations,» *Publicaciones SAGE*, 2016.

[47] F. Draganidis y G. Mentzas, «Competency based management: a review of systems and approaches,» *Information Management & Computer Security*, vol. 14, nº 1, pp. 51-64, 2006.

[48] J. Blanco-González, Y. Ortega-González, M. Delgado-Fernández, L. Domínguez-Peña y M. González-Vengas, «Ontological models for professional competences management,» *Ingeniería Industrial*, vol. 32, nº 3, pp. 224-230, 2011.

[49] S. Tobón, «Formación integral y competencias,» vol. 227. *Editorial Macro*, 2015.

[50] G. Paquette, «An ontology and a software framework for competency modeling and management,» *Educational Technology & Society*, vol 10, nº 3, pp. 1-21, 2007.

[51] M.A. Covington, «Natural Language Processing for Programmers,» *Artificial Intelligence Programs*, 2012.

[52] C. Santamaría, O. Espino y R. Byrne, «Counterfactual and semifactual conditionals prime alternative possibilities,» *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 31, nº 5, pp. 1149-1154, 2005.

[53] V. Gupta y L. Gurpreet, «A survey of text mining techniques and applications,» *Journal of emerging technologies in web intelligence,* vol. 1, nº 1, pp. 60-76, 2009.

[54] E.D. Liddy, «Natural Language Processing,» *Encyclopedia of Library and Information Science*, 2001.

[55] S. Jusoh, «A study on nlp applications and ambiguity problems,» *Journal of Theoretical & Applied Information Technology*, vol. 96, nº 6, 2018.

[56] S. A. Crossley, L.K. Allen, K. Kyle y D. S. McNamara, «Analyzing Discourse Processing Using a Simple Natural Language Processing Tool,» *Discourse Processes*, vol. 51, nº 5-6, pp. 511-534, 2014.

[57] M. Sanderson, «Retrieving with good sense,» *Information Retrieval*, vol. 2, nº 1, pp. 49-69, 2012.

[58] R. Baeza-Yates, «Challenges in the Interaction of Information Retrieval and Natural Language Processing,» *the 5th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2014)*, vol. 2945, pp. 445-456, 2014.

[59] F. Sentis, «La presuposición como categoría pragmática: un caso de confrontación epistemológica,» *Onomázein: Revista de lingüística, filología y traducción*, vol. 6, pp. 105-148, 2011.

[60] L. Yuanhua and C. Zhai, «Lower-bounding term frequency normalization,» *20th ACM international conference on Information and knowledge management*, pp. 7-16, 2011.

[61] A. Gómez-Perez, Asuncion, M. Fernández-López y O. Corcho, «Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web", *Springer Science & Business Media*, 2006.

[62] C.C. Aggarwal y C. Zhai, «Mining Text Data,» *Springer*, 2012

[63] E. Grefenstette, «New Directions in Vector Space Models of Meaning,» *the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*, 2014.

[64] S. Robertson y H. Zaragoza, «The probabilistic relevance framework: BM25 and beyond,» *Now Publishers Inc*, 2009.

[65] A. González-Eras y J. Aguilar, «Semantic Architecture for the Analysis of the Academic and Occupational Profiles Based on Competencies,» *Contemporary Engineering Sciences*, vol. 8, pp. 1551- 1563, 2015.

[66] A. González-Eras y J. Aguilar, «Esquema para la actualización de Ontologías de Competencias en base al Procesamiento del Lenguaje Natural y la Minería Semántica,» *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. 17, pp. 433-447, 2019.

[67] A. González-Eras y J. Aguilar, «Determination of professional competencies using an alignment algorithm of academic and professional profiles, based on competence thesauri and similarity measures,» *International Journal of Artificial Intelligence in Education*, vol. 29, nº 4, pp. 536–567, 2019.

[68] I. Horrocks, P.F. Patel-Schneider y F. van Harmelen, «From SHIQ and RDF to OWL: the making of a web ontology language,» *J. Web Sem*, Vol. 1, nº 1, pp. 7-26, 2003.

[69] T. Lukasiewicz y U. Straccia, «Managing uncertainty and vagueness in description logics for the semantic web,» *Journal of Web Semantics*, vol. 6, nº 4, pp. 291-308, 2008.

[70] G. Sutcliffe y F.J. Pelletier, «JGXYZ: An ATP System for Gap and Glut Logics,» *International Conference on Automated Deduction*, pp. 526-537, 2019.

[71] H. Müller-Riedlhuber, «The European Dictionary of Skills and Competencies (DISCO): an Example of Usage Scenarios for Ontologies,» *I-SEMANTICS*, pp. 467-479, 2009.

[72] C. Shannon, «A mathematical theory of communications,» *Bell System Technical Journal*, pp. 379–423, 1948.

[73] M. Mendonça, N. Perozo y J. Aguilar, «An approach for Multiple Combination of Ontologies based on the Ants Colony Optimization Algorithm,» *Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*, pp. 140-145, 2015.

[74] V. I. Levenshtein, «Binary codes capable of correcting deletions, insertions, and reversals,» *Soviet physics doklady*, vol. 10, nº 8, pp. 707-710, 1966.

[75] F. Alqadah y R. Bhatnagar, «Similarity measures in formal concept analysis,» *Annals of Mathematics and Artificial Intelligence*, vol. 61, nº 3, pp. 245-256, 2011.

[76] S. Staab, R. Studer. «Handbook on ontologies,» *Springer Science & Business Media*, 2013.

[77] P. Horrocks, H. Patel-Schneider, S. Boley, B. Tabet, M.D. Grosof, «SWRL: A semantic web rule language combining OWL and RuleML,» *W3C Member submission*, vol. 21, . nº 79, 2004.

[78] P. Buitelaar, P. Cimiano y B. Magnini, «Ontology learning from text: An overview,» *Ontology learning from text: Methods, evaluation and applications*, vol. 123, pp. 3-12, 2005

[79] R. Dijkman, M. Dumas, B. Van Dongen, R. Käärik y J. Mendling, «Similarity of business process models: Metrics and evaluation,» *Information Systems*, vol. 36, nº 2, pp. 498-516, 2011.

[80] B. Van Dongen, R. Dijkman y J. Mendling, «Measuring similarity between business process models,» *Seminal Contributions to Information Systems Engineering*, pp. 405-419, 2013.

[81] K. Jones, S. Walker y S. Robertson, «A probabilistic model of information retrieval: development and comparative experiments,» *Information processing & management*, pp. 809-840, 2000.

[82] R. Sorensen, «Vagueness,» *Stanford Encyclopedia of Philosophy*, 2018.

[83] F.J. Pelletier, G. Sutcliffe, A.P. Hazen, «Automated Reasoning for the Dialetheic Logic RM3. *30th International Florida Artificial Intelligence Research Society Conference*, pp. 110-115, 2017.

[84] J. Aguilar, «Temporal Logic from the Chronicles Paradigm: learning and reasoning problems, and its applications in Distributed Systems,» *LAP Lambert Academic Publishing*, 2011.

[85] L.F. Sikos, «Description Logics in Multimedia Reasoning,» *Cham: Springer International Publishing,* 2017.

[86] M. Bourahla, «Reasoning over Vague Concepts,» *Lecture Notes in Computer Science*, pp. 591-602, 2015.

[87] H. Gasmi y A. Bouras, «Ontology-based education/industry collaboration system,» *IEEE Access*, vol. 6, pp. 1362-1371, 2017.

[88] J. E. Maybee, «Hegel's dialectics,» *Stanford Encyclopedia of Philosophy*, 2016.

[89] G. Pulcini y A.C. Varzi, «Paraconsistency in classical logic,» *Synthese*, vol. 195, nº 12, pp. 5485-5496, 2018.

[90] O. Peter y P. Hasle «Future Contingents,» *Stanford Encyclopedia of Philosophy*, 2015.

[91] M. Eklund, «Fictionalism,» *Stanford Encyclopedia of Philosophy*, 2017.

[92] D. I. Beaver, «Presupposition,» *Handbook of logic and language*, pp. 939-1008, 1997.

[93] P. Menzies, «Counterfactual theories of causation,» *Stanford Encyclopedia of Philosophy*, 2001.

[94] A. González-Eras, «Caracterización de las competencias en los contextos laboral y académico en base a tecnologías semánticas,» *tesis de Maestría Universidad Politécnica de Madrid*, 2017.

[95] I. Bosque y J. Gutiérrez-Rexach, «*Fundamentos de sintaxis formal*,» *Akal*, 2009.

[96] J. Rosa, M. Kich, L. Brito, «A Multi-Temporal Context-aware System for Competences Management,» *International Journal of Artificial Intelligence in Education*, vol. 25, pp. 455-492, 2015.

[97] M. Rau, «Do Knowledge-Component Models Need to Incorporate Representational Competencies?,» *International Journal of Artificial Intelligence in Education*, vol. 27, pp. 298-319, 2017.

[98] A. Smirnov, A. Kashevnik, S. Balandin, O. Baraniuc, V. Parfenov, «Competency Management System for Technopark Residents: Smart Space-Based Approach,» *Internet of Things, Smart Spaces, and Next Generation Networks and Systems,* pp. 15-24, 2016.

[99] M. Dane, «System and method for automatically processing candidate resumes and job specifications expressed in natural language into a normalized form using frequency analysis,» *U.S. Patent 8,117,024*, 2012.

[100] J. Hu, «Jobscan,» [En línea]. Disponible: https://www.jobscan.co/ [Último acceso: 21 de abril de 2018].

[101] L. Gil-Vallejo, I. Castellón y M. Coll-Florit, «Similitud verbal: Análisis comparativo entre lingüística teórica y datos extraídos de corpus,» *Revista Signos*, vol. 51, nº 98, pp. 310-332, 2018.

[102] S. Miranda, F. Orciuoli, V. Loia y D. Sampson, «An ontology-based model for competence management,» *Data & Knowledge Engineering*, vol. 107, pp. 51-66, 2017.

[103] A. González-Eras, R. Dos Santos y J. Aguilar, «Análisis de las contradicciones en las competencias profesionales en los textos digitales usando lógica dialéctica,» *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. E27, pp. 150-163, 2020.

[104] S. Guo, F. Alamudun y T. Hammond, «RésuMatcher: A personalized résumé-job matching system,» *Expert Systems with Applications*, vol. 60, pp. 169-182, 2016.

[105] R. Elchamaa, A. Mbaya, N. Moalla, Y. Ouzrout y A. Bouras, «Ontology for Continuous Learning and Support,» *Enterprise Interoperability*, vol. 8, pp. 191-202, 2019.

[106] C. Ramsauer, «Competencies of Production in SMEs in Assembly Industries in a Digital, Volatile Business Environment,» *Tehnički glasnik*, vol. 14, nº 3, pp. 388-395, 2020.

[107] I. Kondratova, H. Molyneaux y H. Fournier, «Design considerations for competency functionality within a learning ecosystem,» *International Conference on Learning and Collaboration Technologies,* pp. 124-136, 2017.

[108] A. González-Eras, O. Buendia, J. Aguilar, J. Cordero y T. Rodriguez, «Competences as services in the autonomic cycles of learning analytic tasks for a smart classroom,» *International Conference on Technologies and Innovation,* pp. 211-226, 2017.

[109] C. Guevara, J. Aguilar y A. González-Eras, «The Model of Adaptive Learning Objects for virtual environments instanced by the competencies,» *Adv. Sci. Technol. Eng. Syst. J*, vol. 2, nº 3, pp. 345-355, 2017.

[110] M. Kravcik, X. Wang, C. Ullrich y C. Igel, «Towards competence development for industry 4.0,» *International Conference on Artificial Intelligence in Education*, pp. 442-446, 2018.

[111] M. Faes y D. Moens, «Recent trends in the modeling and quantification of non-probabilistic uncertainty,» *Computational Methods in Engineering*, pp. 1-39, 2019.

[112] C. Jiménez, M. Jerez, J. Aguilar, R. García, «Linked Data and Dialetheic Logic for localization-aware applications,» *Contemporary Engineering Sciences*, vol. 12, nº 3, pp. 103 –116, 2019.

[113] S. Levinson, «Pragmática,» *Teide*, 1989.

[114] H. Chung y J. Kim, «An ontological approach for semantic modeling of curriculum and syllabus in higher education,» *International Journal of Information and Education Technology*, vol. 6, nº 5, pp. 365, 2016.

[115] D. Dubois y H. Prade, «Possibility theory: an approach to computerized processing of uncertainty ,» *Springer Science & Business Media*, 2012.

[116] E. Haque y F. Chiang, «Restoring Consistency in Ontological Multidimensional Data Models via Weighted Repairs,» *Procedia Computer Science*, vol. 159, pp. 1085-1094, 2019.

[117] E. Reiter y R. Dale, «Building natural-language generation systems,» Natural Language Engineering, vol. 3, pp. 57–87, 1997.

[118] Y. Bengio, R. Ducharme, V. Pascal y C. Jauvin, «A neural probabilistic language model,» *Journal of machine learning research*, vol.3, pp. 1137-1155, 2003.

[119] J. Zea, J. Luna, C. Thorne y G. Glavaš, «Spanish NER with word representations and conditional random fields,» *the sixth Named Entity Workshop*, pp. 34-40, 2016.

[120] M. Mendonça, N. Perozo, y J. Aguilar, «Ontological emergence scheme in self-organized and emerging systems,» *Advanced Engineering Informatics*, vol. 44, pp. 101045.

[121] A. Figueroa y J. Atkinson, «Contextual language models for ranking answers to natural language definition questions,» *Computational Intelligence,* vol. 28, nº 4, pp. 528 - 548, 2012.

[122] A. J. Gallego, «Perspectivas de sintaxis formal,» *Ediciones AKAL*, 2020.

[123] W. Medhat, A. Hassan y H. Korashy, «Sentiment analysis algorithms and applications: A survey,» *Ain Shams Engineering Journal* 5, nº 4, pp. 1093-1113, 2014.

[124] B. Liu, «Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data,» *Alemania: Springer*, pp. 412, 2007.

[125] P. Shvaiko y J. Euzenat, «Ontology matching: state of the art and future challenges,» *IEEE Transactions on knowledge and data engineering,* vol. 25, nº 1, pp. 158-176, 2011.

[126] A. Maedche, S. Staab, «Ontology learning for the semantic web,» *IEEE Intelligent systems,»* vol. 16, nº 2, pp. 72-79, 2011.

[127] R. Iqbal, M.A.A. Murad, A. Mustapha y N.M. Sharef, «An analysis of ontology engineering methodologies: A literature review,» *Research journal of applied sciences, engineering and technology*, vol. 6, nº 16, pp. 2993-3000, 2013.

[128] I. Bolshakov & A. Gelbukh, «Computational linguistics models, resources, applications,» *Series Ciencia de la computación*, 2004.

[129] J. Euzenat, P. Shvaiko, «Ontology Matching,», *Springer*, 2013.

[130] B. Niang, B. Bouchou, M. Lo, «Towards Tailored Domain Ontologies. Ontology Matching,» *the ISWC Workshop*, 2014.

[131] V. Gutiérrez-Basulto, J. Jung, C. Lutz y L. chröder, «Probabilistic description logics for subjective uncertainty,» Journal of Artificial Intelligence Research, vol. 58, pp. 1-66, 2017.

APENDICES

APENDICE A: ARTÍCULOS PUBLICADOS EN EL MARCO DE LA TESIS

1. A. González-Eras y J. Aguilar, «Semantic Architecture for the Analysis of the Academic and Occupational Profiles Based on Competencies,» Contemporary Engineering Sciences, vol. 8, pp. 1551- 1563, 2015.
2. C. Guevara, J. Aguilar y A. González-Eras, «The model of adaptive learning objects for virtual environments instanced by the competences,» Adv. Sci. Technol. Eng. Syst. J, vol 2, nº 3, pp. 345- 355, 2017.
3. A. González-Eras, O. Buendia, J. Aguilar, J. Cordero y T. Rodriguez, «Competences as services in the autonomic cycles of learning analytic tasks for a smart classroom,» I*nternational Conference on Technologies and Innovation,* pp. 211-226, 2017.
4. T. Rodriguez, J. Aguilar y A. González-Eras, «Opinion Mining using a Knowledge Extraction System from the Web,» *Contemporary Engineering Sciences*, vol 10, nº 17, pp. 829-840, 2017.
5. A. González-Eras y J. Aguilar, «Determination of professional competencies using an alignment algorithm of academic and professional profiles, based on competence thesauri and similarity measures,» *International Journal of Artificial Intelligence in Education*, vol. 29, nº 4, pp. 536–567, 2019.
6. A. González-Eras y J. Aguilar, «Esquema para la actualización de Ontologías de Competencias en base al Procesamiento del Lenguaje Natural y la Minería Semántica,» *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. 17, pp. 433-447, 2019.
7. A. González-Eras, R. Dos Santos y J. Aguilar, «Análisis de las contradicciones en las competencias profesionales en los textos digitales usando lógica dialéctica,» *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. E27, pp. 150-163, 2020.
8. A. Gonzalez-Eras, R. Dos Santos y J. Aguilar, «Evaluation of digital competence profiles using Dialetheic Logic,» *International Journal of Artificial Intelligence in Education,* En Revisión, 2021.