PI1830-2-NLPSocial

PROCESAMIENTO DE LENGUAJE NATURAL PARA LA EVALUACIÓN DE PROBLEMAS SOCIALES

Juan Pablo Pajaro Hernandez

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2018

PI1830-2-NLPSocial

PROCESAMIENTO DE LENGUAJE NATURAL PARA LA EVALUACIÓN DE PROBLEMAS SOCIALES

Autor:

Juan Pablo Pajaro Hernandez

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Director

Rafael Andrés González Rivera

Comité de Evaluación del Trabajo de Grado

Alexandra Pomares Quimbaya

Jorge Andrés Alvarado Valencia

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~PI1830-2-NLPSocial

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
11,2018

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Lope Hugo Barrero Solano

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Angela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortíz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

A tres personas en particular. A la mujer de mis sueños y al mejor amigo que me ha dado la vida: Juana y Luis.

Y por supuesto, al gran Rafa, quien es admirado por todos en la Universidad.

CONTENIDO

ABSTRACT

The Logical Framework Approach (LFA) is often used to formulate and evaluate projects in the public sector in Latin America. The LFA proposes that in order to evaluate a social problem, the relationship between its causes, consequence and central problem must be demonstrated through a mental map called the problem tree. However, by definition a social problem is transdisciplinary, systemic and has multiple interests, make it difficult to evaluate its semantic analysis. The research proposes a web service, which based on the Natural Language Processing NLP, calculates a metric of semantic similarity between the sentences of the problem tree. The semantic similarity is useful for designing projects. The research was carried out following the Desing Science Research Model and ended with the validation in a use case with the application of "Technology Acceptance Model (TAM)".

RESUMEN

La Metodología Marco Lógico (MML) es con frecuencia la metodología utilizada para formular y evaluar proyectos de inversión en el sector público, en América Latina. La MML propone que para evaluar un problema social se debe demostrar la relación entre sus causas, consecuencia y problema central, a través de un mapa mental denominado árbol de problemas. Sin embargo, por definición un problema social contiene elementos transdiciplinarios, sistémicos y múltiples intereses que dificultad evaluar su análisis semántico. La presente investigación ilustra el desarrollo de un servicio web, que basado en el Procesamiento de Lenguaje Natural NLP, calcula una métrica de similitud semántica entre las oraciones del árbol de problemas. La cual es útil durante el diseño proyectos de inversión. La investigación se desarrolló siguiendo la metodología "Desing Science Research" y finalizó con la validación del servicio web en un escenario de uso con la aplicación de "Technology Acceptance Model (TAM)" a expertos del Departamento Nacional de Planeación.

RESUMEN EJECUTIVO

La Metodología Marco Lógico (MML) es con frecuencia la metodología utilizada para formular y evaluar proyectos de inversión en el sector público, en América Latina [1].

La MML propone que para describir un problema social se debe demostrar la relación entre sus causas, problema central y consecuencias, a través de un mapa mental denominado árbol de problema [1].

Sin embargo, construir un árbol de problemas durante el diseño de un proyecto de inversión es difícil. Por definición, un problema social contiene elementos transdiciplinarios, sistémicos y múltiples intereses que dificultad su análisis [2]. Transdiciplinario significa que no existe ninguna disciplina, o combinación de estas, que permita explicar completamente el problema o su solución. Sistémico hace referencia a que exhibe propiedades emergentes y requiere de soluciones de corto y largo plazo. Y múltiples intereses hace referencia a como estos dificultan las interacciones entre los involucrados en el problema social.

Ahora bien, el desarrollo de las tecnologías de información, ligada a la disponibilidad de altos volúmenes de datos, ha impulsado el desarrollo de técnicas computacionales que permiten procesar y acceder a la información de forma eficiente, incluyendo el análisis del lenguaje humano [3].

El procesamiento de lenguaje natural (NLP) es la combinación de las ciencias computacional y lingüística que busca apoyar las tareas de análisis del lenguaje humano, como su nombre lo expresa.

En este sentido, un árbol de problemas puede ser visto como la composición de oraciones que buscan lograr una representación del problema social, y para ello deben guardar relación entre sus conceptos. Esta relación entre conceptos puede ser evaluada a través del NLP.

De esta forma, la presente investigación diseño un servicio web que por medio de métodos de NLP obtiene una métrica de la relación entre conceptos del árbol de problemas. La cual es útil para la evaluación de árboles en el marco del diseño de proyectos de inversión.

Como objetivos específicos, primero se analizaron las características de los árboles de problemas de los proyectos de inversión formulados y aprobados por la Gobernación de Cundinamarca en 2017. Segundo, se determinaron los pipelines requeridos en el proceso de minería de texto, siguiendo las indicaciones de la literatura: entrenamiento, verificación y prueba. Tercero, se desarrolló el servicio web, donde se obtuviera como resultado la métrica similitud semántica de los árboles de problemas. Cuarto, se evaluó la utilidad del servicio web aplicando TAM con un grupo de expertos de la Subdirección de proyectos e información para la inversión pública SIIP del Departamento Nacional de Planeación.

INTRODUCCIÓN

El uso de la Metodología Marco Lógico (MML) impone, en el diseño de proyectos de inversión, la realización de árboles de problemas [1]. Que de acuerdo con los expertos en la MML realizar dicha tarea es difícil, en cualquier contexto. Debido a las características intrínsecas de los problemas sociales y la dinámica de las organizaciones públicas en Colombia.

Dicha dificultad para construir árboles de problemas está enmarcada en la dinámica de gestión de proyectos de inversión. Donde se reciben miles y se aprueba la financiación de estos con recursos públicos. Si dentro de estos proyectos existen algunos mal estructurados se generará como consecuencia la pérdida de los recursos.

De esta forma, resulta relevante proponer herramientas que apoyen el diseño de árboles de problemas y por ende la formulación de proyectos de inversión.

El desarrollo de las tecnologías de información, ligada a la disposición de volúmenes de datos ha impulsado el desarrollo de técnicas que permiten apoyar la toma de decisiones, cada vez más ligada al análisis del lenguaje humano.

La presente investigación buscó validar la utilidad de un servicio web que apoyara la validación de árboles de problemas para el diseño de proyectos de inversión.

El principal aporte de la investigación está dado en el escenario de uso realizado. El servicio web permitiría mejorar el desempeño de las organizaciones y profesionales sobre la validación de árboles de problemas durante el diseño de proyectos de inversión.

El documento se encuentra distribuido en cinco capítulos. El primero, expone de la relevancia de la investigación. El segundo, describe los pasos del estudio, según la metodología de investigación basada en el diseño, buscando una breve descripción de la aplicación de cada uno de sus ciclos. En el tercero, se expresa una breve definición de NLP, los principales métodos y tecnologías. Esta sección concluye con el pipeline NLP para la validación semántica de árboles problemas. El cuarto capítulo, expone el diseño del servicio web, el escenario de uso en el cual fue validado y los resultados. El quinto, describe las conclusiones de la investigación, así como una discusión sobre las potencialidades de aplicar NLP en el sector público.

1. DESCRIPCIÓN GENERAL

En esta sección se describe la relevancia de la investigación. La cual está en el marco en la Metodología Marco Lógico y por ende de interés para la gestión pública de proyectos en toda América Latina.

1.1. Relevancia

El sector público existe para promover o alcanzar el bienestar social de los ciudadanos, promover condiciones de igualdad. Es una estructura que emerge entre los agentes de un territorio conocida normalmente como Estado o Gobierno [4].

Se han logrado algunas soluciones no óptimas a problemas de desigualdad específicos. Se han establecido algunas herramientas; leyes, impuestos, subsidios, instituciones, canales de comunicación, entre otras. Pero sigue siendo un gran reto el entendimiento entre el Estado y los ciudadanos, principalmente en aquellos que padecen de mayores condiciones de desigualdad.

La herramienta más utilizada para coordinar los esfuerzos entre un conjunto de personas que carecen de oportunidades y el Estado son los proyectos de inversión social, entendido como un canal de comunicación.

En este contexto, la Metodología Marco Lógico (MML) es con frecuencia la metodología utilizada para formular y evaluar proyectos de inversión en el sector público, en América Latina [1].

La metodología fue desarrollada por USAID [5] en los años setenta y es usada por todos los países de Latinoamérica en el marco de la gestión pública de proyectos de inversión social [1]. En Colombia, por ejemplo, es usada por el Departamento Nacional de Planeación (DNP), organización encargada del seguimiento de las políticas públicas del país. Por ende, todas los Ministerios, Gobernaciones y Alcaldías que utilicen recursos del Presupuesto General del Estado deben formular sus proyectos de inversión usando la metodología [6].

La MML propone que para describir un problema social se debe demostrar la relación entre sus causas, problema central y consecuencias, a través de un mapa mental denominado árbol de problema [1]. Su representación gráfica es como se muestra en la figura 1, donde las casillas en blanco o no, representan oraciones relacionadas.

Por ejemplo, si previo a la construcción de la figura 1, se tiene una frase como; el rio está contaminado, y otra frase como; las empresas arrojan desperdicios en el río. Lo ideal es construir esta representación, comprendiendo que existe una relación de causalidad entre estas oraciones. Es decir, el rio está contaminado a causa de los desperdicios que arrojan las empresas.

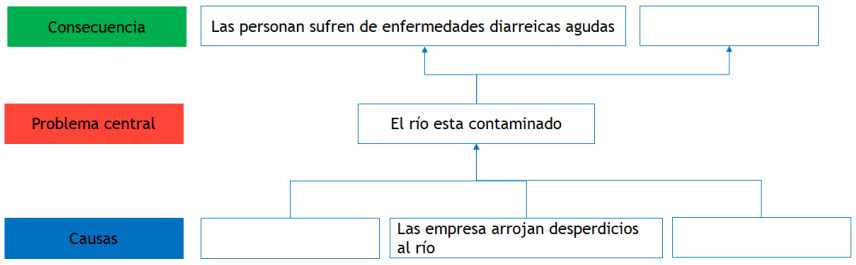


Fig. 1 Ejemplo Árbol de problema

Ilustrar la relación de causalidad en el ejemplo anterior es fácil por la reducción del árbol a tres oraciones. Sin embargo, es complicado construir dichas relaciones en un problema social real.

Es complicado porque, por definición, un problema social contiene elementos transdiciplinarios, sistémicos y múltiples intereses que dificultad su análisis [2]. Transdiciplinario significa que no existe ninguna disciplina, o combinación de estas, que permita explicar completamente el problema o su solución [7]. Sistémico hace referencia a que exhibe propiedades emergentes y requiere de soluciones de corto y largo plazo [8]. Y múltiples intereses hace referencia a cómo estos dificultan las interacciones entre los actores involucrados en el problema social [2].

De acuerdo con varios expertos en formulación del MML consultados (Ver capítulo 2.2.1), los elementos descritos hacen que sea difícil y costoso validar las relaciones, cuando se está en el proceso de construir un árbol de problemas. Se requieren observaciones multidisciplinares y amplia capacidad de análisis derivada de la experiencia. Lo que implica tener recursos para consultar a los expertos en otras disciplinas, y personas involucradas, buscando adquirir los conocimientos de los campos científicos, a los cuales se puede intentar suscribir el problema social.

Ahora bien, el desarrollo de las tecnologías de información, ligada a la disponibilidad de altos volúmenes de datos, ha impulsado el desarrollo de técnicas computacionales que permiten procesar y acceder a la información de forma eficiente, incluyendo el análisis del lenguaje humano [3].

El procesamiento de lenguaje natural (NLP) es el campo de la ciencia computacional y lingüística que se ocupa de las interacciones entre los computadores y el lenguaje humano [9].

El NLP, se descompone en seis categorías, una de ellas es el análisis semántico. La cual, consiste en el estudio del significado [10], o en otras palabras, el análisis de conceptos, episodios relevantes y experiencia sensoriales, que permiten ilustrar (graficar) la relación en el lenguaje [11]. Dicha relación podría ser la que existe entre dos oraciones, donde la primera es una causa y la segunda un problema social.

En este escenario, surgen las preguntas ¿Es posible configurar los métodos de NLP, como apoyo a la validación de árboles de problemas sociales en el marco de la MML? Y si es así ¿Cómo sería su configuración y cuáles métodos usar o descartar?

En este sentido, el objetivo de la presente investigación fue proponer un servicio web, que permitiera la validación de árboles de problemas en el marco del diseño de proyectos de inversión pública.

Si la validación de un árbol problema es vista como el análisis semántico entre las oraciones causas-problema y las oraciones problema-consecuencias. Los métodos de NLP: Part of Speech tagging POS, reconocimiento de entidades NER y similitud semántica, cuentan con madurez y cobertura científica, para realizar dicha tarea.

2. DESCRIPCIÓN DEL PROYECTO

Esta sección describe los objetivos planteados en la investigación. Dentro de los cuales estuvo explorar la base de conocimiento, configurar los métodos de NLP y construir un servicio web, como herramienta para la validación de árboles de problemas.

2.1. Objetivos

2.1.1. Objetivo general

Desarrollar un servicio web para evaluar problemas sociales en proyectos de inversión a través del procesamiento de lenguaje natural (NLP).

2.1.2. Objetivos específicos

- Definir las características que debe contener un problema social formulado dentro de un proyecto de inversión.
- Determinar la arquitectura para realizar NLP basada en las características contenidas en un problema social formulado.
- Programar un servicio web usando los métodos de NLP para evaluar problemas sociales.
- Evaluar el desempeño del servicio web a través de su aporte en la formulación de proyectos de inversión.

2.2. Metodología

El desarrollo de la investigación buscó seguir los ciclos de la investigación científica basada en el diseño (DRS): relevancia, rigor y diseño [12].

2.2.1 Ciclo de Relevancia

La relevancia de la investigación fue definida por dos pasos: 1. Revisión de la literatura asociada a la MML y 2. Aplicación de entrevistas semiestructuradas a expertos en marco lógico.

El primer paso buscó cuantificar, aproximadamente, el número de instituciones que usan la MML en el mundo. La revisión de la literatura asociada consistió en una serie de publicaciones de la Comisión Económica para América Latina (CEPAL) [1] [13] [14] y la guía para la formulación de proyectos en Colombia del DNP [6].

El segundo paso busco conocer la aplicación real de la MML en el contexto del sector público colombiano. De esta forma, se diseñó un instrumento con 19 preguntas abiertas, como guía de entrevista a expertos en formulación de proyectos del sector público (Ver anexo 1). Antes de aplicar el instrumento se realizó una simulación de la entrevista entre el asesor y el autor de la investigación, buscando el uso adecuado del lenguaje en las preguntas.

Posteriormente, se invitó a cuatro expertos en formulación de proyectos sociales a participar en la investigación, realizándoles la entrevista. Como criterio esencial los expertos fueron seleccionados por sus constantes usos de la MML en su ambiente laboral. El perfil de los expertos entrevistados es el siguiente:

1. Administradora de Empresas con dos especializaciones; gerencia de Proyectos y Economía. Siete años de experiencia en la formulación de proyectos con el uso de la MML en el sector público. Actualmente funcionaria de IDARTES, Alcaldía Mayor de Bogotá.

2. Ingeniero de sistemas, especialista en evaluación social de proyectos. Quince años de experiencia en la formulación de proyectos con el uso de la MML en el sector público. Actualmente funcionario de la Secretaría de las TIC, Gobernación de Cundinamarca.
3. Artista plástico y matemático, candidato a doctor en creatividad computacional. Tres proyectos formulados para obtener recursos de entidades públicas a través de MML. Actualmente Director de Tecnología en Contacto Móvil.
4. Administradora de empresas, especialista en evaluación social de proyectos. Diez años de experiencia en la formulación de proyectos con el uso de MML en el sector público. Actualmente funcionaria de la Presidencia de la Republica de Colombia.

Como resultado del ciclo de relevancia se determinó que actualmente no existe una métrica de evaluación de árboles de problemas, que apoye a las personas que diseñan proyectos de inversión con el uso de la MML. Los resultados del ciclo de relevancia se profundizan en el capítulo 3.1.

2.2.2 Ciclo de Rigor

Para lograr la base de conocimiento o aplicación del ciclo de rigor, se revisaron los artículos de la base de datos Web of Science[1], con la ecuación de búsqueda de la tabla 1. Es importante aclarar que este ciclo considera los resultados del ciclo de relevancia para definir las palabras claves iniciales. En este caso análisis semántico, NLP, sector público, problema social, entre otras.

Tabla 1. Aspectos de la ecuación de búsqueda

Ecuación de búsqueda	((Natural language processing or NLP or text mining) and (social or public policies or e-government or web or online) and semantic NOT (Medical or Biomedical or Biological or Radiology))
Período de tiempo	Todos los años. Fecha de la consulta 20 mayo de 2017.
Índices	SCI-EXPANDED, SSCI, A&HCI, ESCI
Número de artículos de la búsqueda	564
Número de artículos seleccionados para el estado del arte	87

La construcción de la ecuación de búsqueda fue un proceso iterativo, a la cual se le fueron agregando los términos según el aprendizaje y la lectura de los artículos. Es decir, con la intención de construir los términos asociados que respondiera a las preguntas de investigación (Ver capítulo 1).

Otro aspecto importante, consistía en evaluar el crecimiento de artículos y citas relacionadas, como parte de la existencia de una comunidad científica interesada en la investigación. La figura 2 muestra que existe un crecimiento exponencial en ambos, en artículos publicados y citas.

[1] www.webofknowledge.com

Una vez construida la ecuación de búsqueda, los artículos fueron ordenados por el número de citas de mayor a menor, y se aplicaron dos criterios de selección.

El primer criterio de selección fue que, el título y abstract del articulo expresarán la teoría o practica del análisis semántico, y de ser posible en un contexto social. Este criterio arrojo 140 artículos.

El segundo criterio consistió en el detalle y transparencia del artículo. Es decir, se realizó la lectura de los 140 artículos buscando que entre sus secciones existiera el detalle, de cómo aplicar, al menos un método de análisis semántico. Este criterio arrojó la selección de 87 artículos como base de conocimiento.

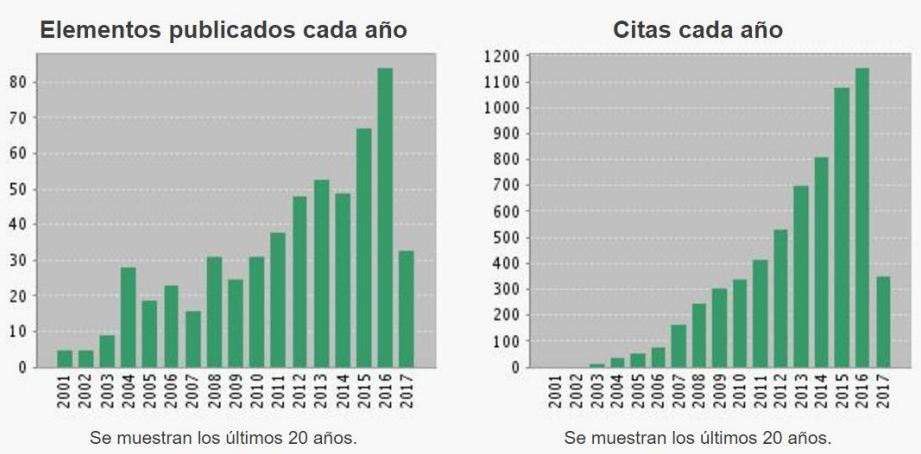


Fig. 2 Informe de publicaciones y citas. Fuente: Web of Science, consultado el 20 mayo de 2017, a través del botón "Crear informe de citas".

Como resultado del ciclo de rigor se resaltan los artículos de: D. Jurafsky and J. H. Martin [10] y A. Budanitsky and G. Hirst [15]. El primero es una explicación de toda la teoría relacionada con NLP. El segundo es un análisis detallado de las principales técnicas de similitud semántica. Ver marco capítulo 3 donde se detalla la exploración del estado del arte.

2.2.3 Ciclo de Diseño

Durante el ciclo de diseño se iteraron los métodos de NLP buscando el correcto pipeline para la validación de árboles de problemas. Este ciclo consistió en tres fases 1. Entrenamiento, 2. Verificación y 3. Escenario de uso. Estas fases guardan estrecha relación con las investigaciones en analítica de datos y texto, cuyas fases son: entrenamiento, verificación y prueba [16].

En la fase entrenamiento se buscó identificar los patrones sintácticos más significativos para detectar las entidades principales de los árboles de problemas. En la fase de verificación se evaluaron dichos patrones sintácticos de acuerdo con su precisión, cobertura y f-score. En la fase de escenario de uso se aplicó TAM a un grupo de expertos.

El grupo de expertos hace parte de la Subdirección de proyectos e información para la inversión pública SIIP del Departamento Nacional de Planeación. Esta subdirección tiene como función principal: Dirigir y administrar, conceptual, operativa y metodológicamente, la consolidación de la información relacionada con los procesos de: formulación, programación, ejecución y seguimiento de la inversión pública nacional, según las disposiciones legales, normativas, institucionales y los criterios técnicos relacionados.

Las fases se explican con más en detalle en el capítulo 4.

3. MARCO TEÓRICO

Esta sección ofrece una interpretación de la base de conocimiento consultada durante el desarrollo de la investigación.

3.1. Un árbol de problemas como la composición de oraciones

La MML es con frecuencia utilizada por las organizaciones públicas de América Latina para la gestión de proyectos sociales. La metodología propone que para el diseño de un proyecto se deben realizar 10 pasos, que se muestran la figura 3.

El paso dos de la metodología es el análisis del problema, consiste en realizar la figura de un árbol. Esto significa identificar un problema social en el centro, como tronco de un árbol, la causas como raíces y las consecuencias como ramas. Es una representación gráfica como se muestra en la figura 1.

Esta representación gráfica es el interés de la investigación por dos razones. La primera, los expertos consultados dejan entre ver que este paso es susceptible de mejoras. La segunda, esta mejora, puede estar en considerar (tomar más en serio) que un árbol es la composición de oraciones, y que validar sus relaciones, es por ende validarlo.

Lo que busca la representación, que también es el propósito[2] de la MML, es apoyar la resolución de desigualdades sociales. Es parte de un proceso de negociación entre los actores involucrados para lograr una, de muchas representaciones, de los elementos transdiciplinarios, sistémicos e intereses que conforman un problema social.

De acuerdo con las entrevistas, los expertos manifiestan que es difícil ordenar un árbol de problemas y desconocen la forma de validarlo. La dificultad se deriva del costo requerido para contar con profesionales en múltiples disciplinas y el tiempo requerido. “El cual excede las tres y cuatro semanas… Aún con profesionales idóneos esta organización exige bastante tiempo y análisis…”.

Mientras que la forma de validar los arboles de problemas es desconocido. “No existe un criterio o métrica para medir si el árbol que realice ayer está peor o mejor que la versión de hoy o la de mañana…”. “Lo que normalmente hago es enviarla a los interesados, pero a veces esto genera retroceso o mayor discusión…”

[2] Cabe aclarar que este propósito continúa siendo difícil. No existe una metodología que asegure soluciones óptimas a problema sociales. Evaluar o comparar metodologías similares está fuera del alcance de la investigación.

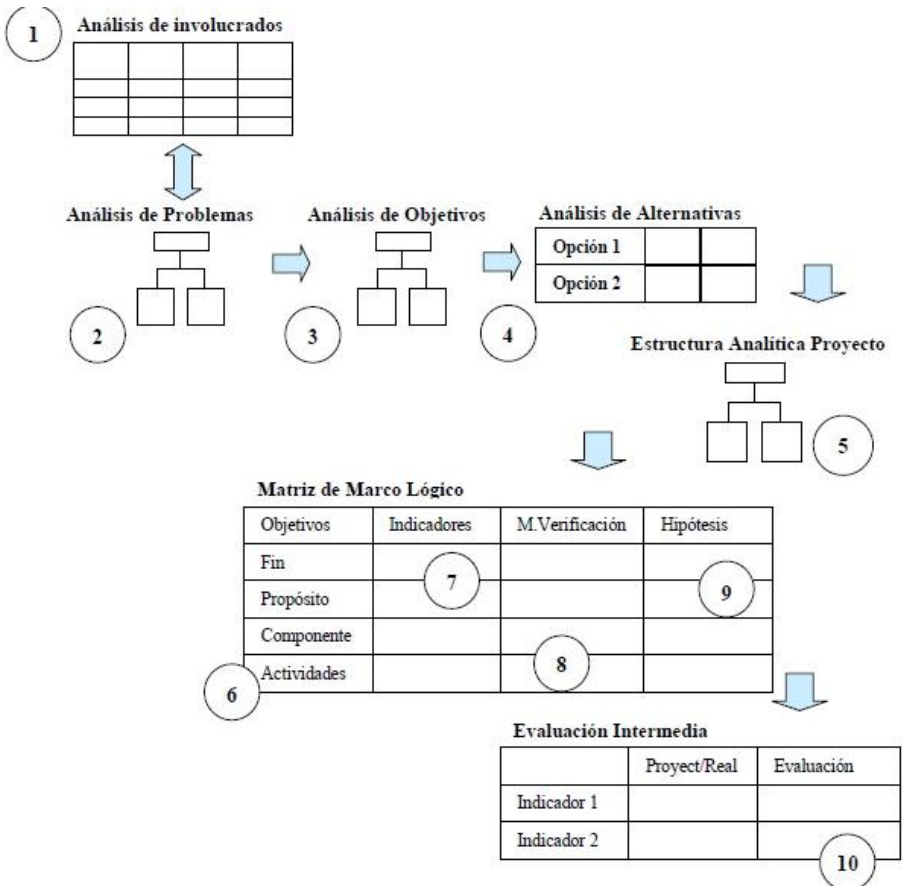


Fig. 3 Pasos de la Metodología Marco Lógico. Fuente: [1]

Así pues, el paso dos de la metodología es ineficiente y percibido de ser mejorado. La solución puede estar en utilizar el lenguaje humano para analizar un árbol de problemas[3].

La MML expresa que el árbol debe cumplir con ciertas características [1], que se muestra en la tabla 2. Sin embargo, existen ciertas características que se toman como implícitas, que no son expresadas, las cuales están asociadas con el lenguaje.

De esta forma, si se toma un árbol como la composición de oraciones de un problema social. Es decir: Oraciones causas, oración problema central, oraciones consecuencias. Dichas deben cumplir con tener correspondencia entre conceptos.

[3] Pueden existir muchas otras perspectivas de mejoras, por ejemplo, la negociación entre los actores involucrados puede ser investigada desde tópicos del pensamiento sistémico. Abarcar otras perspectivas esta fuera del alcance porque la investigación se suscribe al campo computacional.

Por ejemplo, al analizar la causa y problema de la figura 1, existe una relación entre el significado de los conceptos, aceptada por la mayoría de los seres humanos como parte del lenguaje, entre las palabras "desperdicios" y "contaminación". Esta relación se conoce como hiponimia de lenguaje. Es una correspondencia que ocurre cuando el significado de una palabra incluye a la otra[4]. En este caso, "contaminación" incluye "desperdicios".

Tabla 2. Características para realizar el árbol de problemas usando la MML

-	Formular el problema central en estado negativo.
-	Centrar el análisis de causas y efectos en torno a un solo problema central. Lo que permite acotar el análisis y ser más efectivo en recomendar soluciones.
-	No confundir el problema con la ausencia de una solución. No es lo mismo decir falta un hospital (falta de solución), que decir que existen "Altas tasas de morbilidad" en un área específica (problema).
-	Análisis de nodos críticos.

Las relaciones entre conceptos pueden ser de varios tipos: meronimia, hiponimia, funcional, entre otras. Esto es lo que se conoce como similitud semántica y se profundizan en los siguientes capítulos.

Por ende, una vez se tiene una ilustración completa de un árbol de problemas, se puede proporcionar su validación a través de la relación entre los conceptos en las oraciones causas-problema y problema-consecuencia.

Identificar la relación entre conceptos son tareas que se pueden lograr de forma automática, y que frecuentemente se buscan desde el análisis del lenguaje humano.

3.2. Breve definición de NLP

El procesamiento del lenguaje natural es el uso de técnicas computacionales para analizar el lenguaje humano hablado y escrito [17].

El lenguaje humano se puede descomponer en seis categorías de análisis [10]:

1. Fonética y fonología: el estudio de los sonidos del lenguaje.
2. Morfología: el estudio del significado que compone a las palabras.
3. Sintáctico: el estudio de las relaciones entre las palabras.
4. Semántica: el estudio del sentido de las palabras.
5. Pragmática: el estudio como el lenguaje logra las metas.
6. Discurso: el estudio de una unidad lingüística más compleja que una simple declaración.

[4] De acuerdo con la RAE

Cuatro de las categorías anteriores se pueden comprender por medio de un ejemplo en una oración escrita[5], ver figura 4. Es así como, según los aspectos presentes en el texto (oración escrita) se pueden comprender qué es NLP y la dificultad de automatizarlo.

En la oración de la figura 4 es fácil encontrar el sentido. El ser humano activa en el cerebro en milésimas de segundo las redes neuronales que permiten comprender dicha conjugación de palabras. Sin embargo, para un computador se requieren muchos pasos asociados a la descomposición del lenguaje humano.

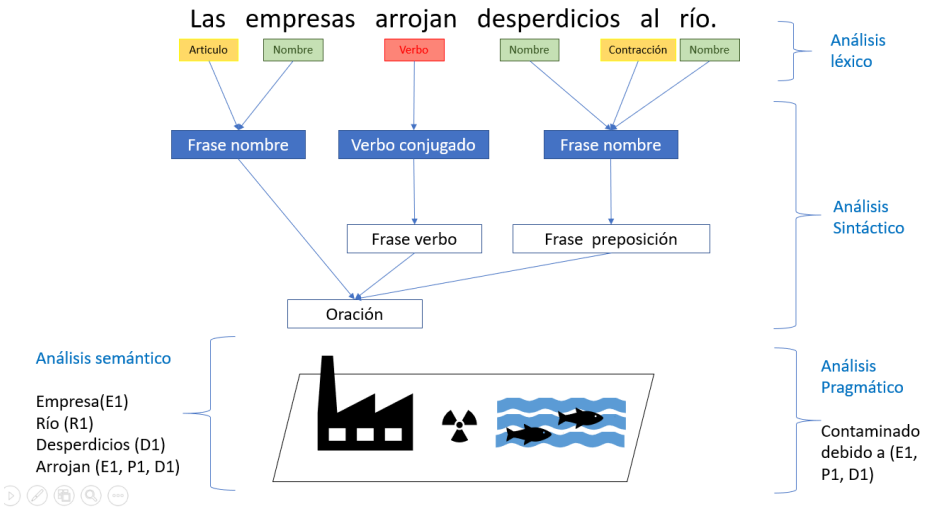


Fig. 4. Ejemplo de NLP.

Primero, se requiere determinar las palabras. En español las palabras están separadas y se pueden ver entre espacios. Luego, se requiere conocer las categorías sintácticas, las categorías de palabras: empresa es un nombre, arrojar es un verbo, río es otro nombre, entre otros. Esto es llamado análisis léxico o morfológico, donde etiquetar las palabras según su categoría sintáctica es conocido como "part-of-speech tagging" [18]. Parte superior de la figura 4.

Después se requiere conocer la relación entre las palabras. En la oración "Las empresas" es una frase compuesta de nombre, "desperdicios al río" es una frase preposición, entre otras, que están conectadas en un orden para crear un significado. En cualquier otro orden no es el mismo significado. Esto es llamado descomposición sintáctica o análisis sintáctico, cuyo resultado es un árbol de etiquetas como se muestra en centro de la figura 4.

Sin embargo, aún existen aspectos que definen el significado con mayor detalle, que no se logran con la descomposición sintáctica.

Por ende, con la intención de obtener más detalle del significado se requiere mapear las frases y su estructura en una posible representación del mundo. Es decir, "empresa" es un concepto, "desperdicio" es otro concepto, ambos conceptos están definidos y son conocidos.

[5] Los análisis de fonética y discurso implican sonido, lo cual es una restricción que los excluye del capitulo.

La representación del mundo es el dominio de conocimiento y se conoce como la representación semántica. La representación se hace a través de recursos lingüísticos o recursos léxicos. Estos últimos, pueden ser diccionarios, taxonomías u ontologías, que contienen los conceptos en una estructura que puede ser entendida por el computador.

De esta forma, la representación del mundo en la oración se puede realizar a través de símbolos que fueran previamente programados en el computador, como: E1 es el concepto de la empresa, D1 son los desperdicios, entre otros. Parte inferior izquierda de la figura 4.

Una vez se logra la representación, se puede decir que existe un nivel de comprensión que permite hacer inferencias. Esto es, por ejemplo, asegurar que el rio está contaminado dado el contexto que denota la oración y el sentido común presente en ella. Esta representación se conoce como análisis pragmático y se muestra en la parte inferior derecha de la figura 4.

Las categorías de análisis descritas se pueden usar para cualquier tipo de unidad de lenguaje requerida. Esto es: símbolos, palabras, oraciones, patrones, conjunto de palabras, documentos de todo tipo de tamaño, entre otros. Estas múltiples unidades de lenguaje hacen que los campos de aplicación sean muchos, a los que a su vez, se les puede asociar una gran variedad de tareas o métodos.

De esta forma, se puede concluir que: la unidad de lenguaje en un árbol de problemas son las oraciones, y que la búsqueda de la relación entre causas-problema y problema-consecuencia en un árbol se suscribe a la categoría del lenguaje denominado análisis semántico. Dicho de otra forma, lo que se propone como solución para mejorar el paso dos de la MML, es la validación semántica de árboles de problemas.

3.3. Métodos NLP para la validación semántica de árboles de problemas

La figura 5 muestra los principales métodos del NLP para el análisis sintáctico y semántico [11], [19], [20], [21], [22], [23], [24]. De ellos, los que pueden apoyar la validación semántica entre oraciones son: Tokenizer, Sentences Breaking, Part Of Speech Tagging, Named Entity Recognition (NER) y Semantic Similarity. A continuación, se expresa una breve definición de cada uno.

El método Tokenizer consiste en separar las palabras dentro de un texto. Es decir, etiquetar cada palabra como un token para el proceso de analisis de NLP. Algunos autores consideran este método parte del NLP, dado que en algunos idiomas las palabras no son fáciles de distinguir. Otros autores consideran que es una tarea previa al NLP.

El método Sentences Breaking o Senteces Splitter, consiste en que, dado un texto, se debe reconocer los límites que separan a las oraciones de otras. Estos límites normalmente están determinados por signos de puntuación.

Part Of Speech Tagging consiste en determinar la categoría sintáctica a la que corresponde una palabra. Usando el ejemplo de la figura 4, esto es etiquetar la palabra “arrojan” como verbo. La palabra “empresas” como nombre. Así sucesivamente para todas las palabras dentro de un texto que se está analizando.

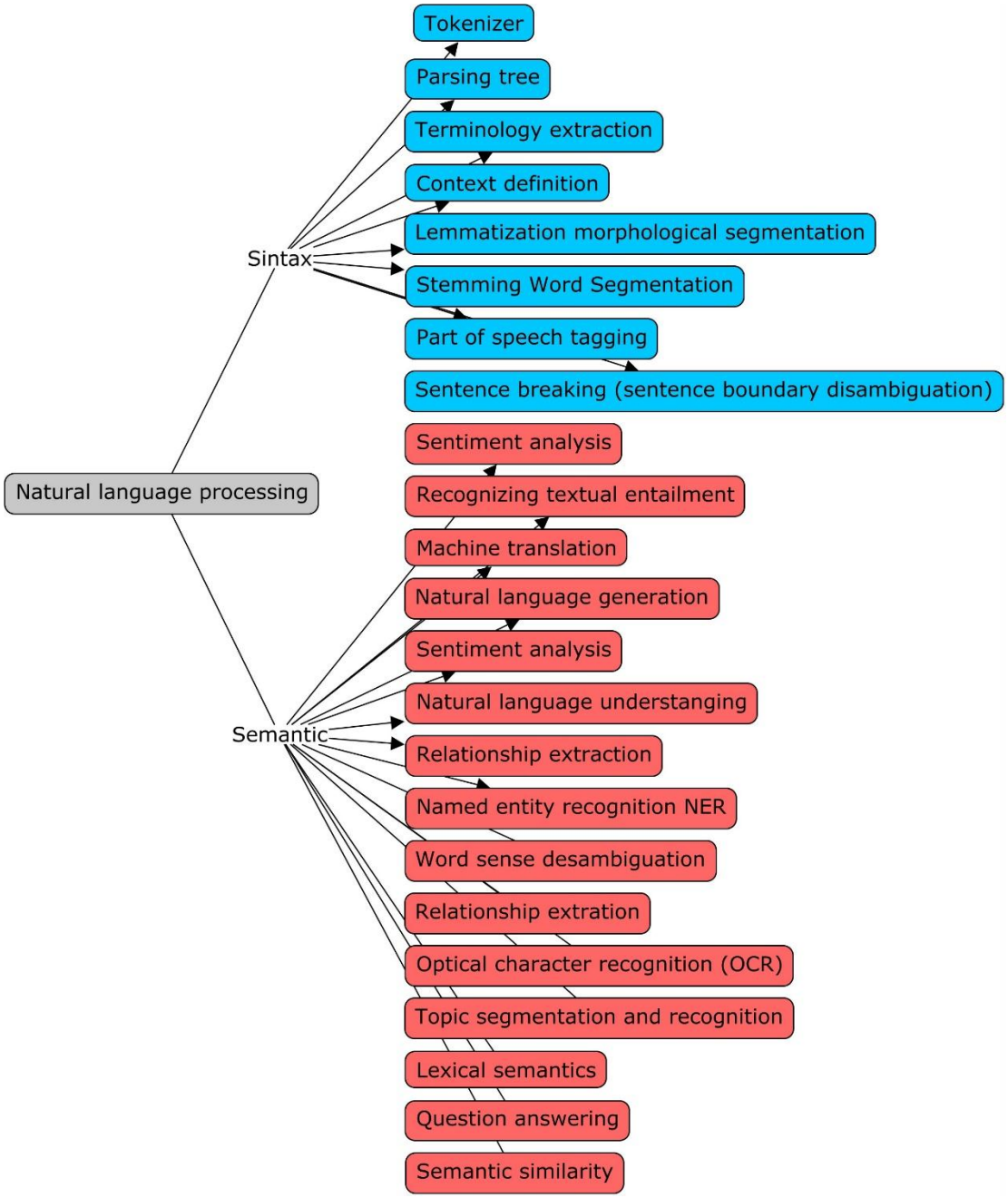


Fig. 5. Principales métodos de NLP para análisis sintáctico y semántico

Named Entity Recognition NER como su nombre lo expresa es el reconocimiento y categorización de entidades bajo el dominio del conocimiento. De forma general, es la identificación de nombres propios en el texto y la clasificación en categorías de interés [25]. Las categorías más comunes son persona, organización, localización, fecha y expresión de tiempo. La tabla 3, muestra ejemplos de cada categoría de entidad.

Tabla 3. Ejemplo de las principales de categorías de NER

Categoría de entidad	Ejemplo

Persona	Juan, él, situación de emergencia, entre otras.
Organización	Compañía, Gobernación de Cundinamarca, organización de gobierno, entre otras
Localización	Ciudad, país, ríos, entre otros
Fecha	18 mayo, 2018, 20-01, entre otros
Expresión de tiempo	8 am, ocho de la mañana, entre otros.

Una característica importante del método NER es el dominio de conocimiento, el cual es integrado como un recurso léxico. Por ejemplo, para que el método NER reconozca a la "Gobernación de Cundinamarca" como una "Organización" (ver tabla 3) deberá existir un recurso léxico que cada vez que encuentre la palabra "Gobernación", seguida de las palabras "de" y "Cundinamarca", debe etiquetar a las tres palabras como "Organización".

Un dominio de conocimiento en el NLP es la gramática aplicada al análisis de oraciones. De acuerdo con [10] las oraciones se pueden descomponer en Frases Nombres, Frases Verbo y Frases Preposiciones. De esta descomposición, la regla que se muestra en la Ecuación 1, es actualmente la más moderna y poderosa teoría de la gramática en el NLP para identificar la Frase Nombre, según los mismos autores.

$$NP \rightarrow (Det)\ (Card)\ (Ord)\ (Quant)\ (AP)\ Nominal$$

Ecuación 1. Fórmula para describir una frase nombre en una oración a través de la gramática

La Frase Nombre es la que contiene el nombre dentro de una oración. La regla se interpreta: Una Frase Nombre en una oración puede ser descrita por las categorías sintácticas siguientes: o un determinante (Det) o un numero cardinal (Card) o un numero ordinal (Ord) o un cuantificador (Quant) o una Frase Adjetivo (AP) más un Nominal. La tabla 4, muestra un ejemplo de cada una.

Un Nominal es: un nombre propio, o un determinante seguido por Nominal, o uno o más nombres.

Tabla 4. Categorías que componen una Frase Nombre

Categoría	Ejemplo
Determinantes	el, esta, su, este, qué, un, una, unos, entre otros.
Cardinal	Dos, uno, tres, entre otros.
Ordinal	Primero, segundo, siguiente, entre otros.
Cuantificador	Primera clase, sin escala, entre otros.
AP	La tarifa más barata, El actual presidente, entre otros.
Nominal	Situación de emergencia, Cundinamarca, entre otros.

Semantic Similarity se refiere a la relación entre conceptos [15]. Las principales relaciones evidentes son sinonimia, antonimia, conversión, hiponimia y meronimia [10]. La tabla 5, muestra un ejemplo de las principales relaciones entre conceptos.

Sin embargo, el método también incluye cualquier tipo de relación frecuente asociación entre conceptos. Ejemplos: Lápiz-Papel, Pingüino – Antártica, Lluvia – Flujo.

Tabla 5. Clasificación de la relación entre conceptos

Relación entre conceptos	Definición	Ejemplo
Sinonimia	Sinónimo	Calor – Caliente
Antonimia	Antónimo	Calor – Frío
Conversión	Funcional	Padrino – Ahijado o Vender – Comprar
Hiponimia	Hipónimo o relación de subtipo	Gorrión es hipónimo de pájaro.
	Hiperónimo o relación de superior	Pájaro es hiperónimo de jilguero y de gorrión.
Merónimia	Holónimo o relación es parte de	Flor es el holónimo de cáliz, corola, pistilo o estambre.
	Merónimo o relación tiene una parte de	Las palabras cáliz, corola, estambre o pistilo son merónimos de flor.

Tampoco es necesario conocer el tipo de relación para determinar que existe relación entre conceptos. De forma precisa, las investigaciones definen que Semantic Similarity consiste en conocer si existe relación o no entre dos conceptos basada en una métrica que se apoya en un recurso léxico [26].

WordNet es el recurso léxico frecuentemente más usado por el método Semantic Similarity.

El recurso léxico es tratado (comprendido) como un grafo, donde los nodos representan los conceptos y los arcos la relación entre ellos. Buscando las siguientes condiciones:

- Si existen dos nodos conectados, a través de un nodo común, significa que existe relación entre dos conceptos.
- A más alta la posición del nodo común para dos conceptos, más baja su relación.

Por lo tanto, resulta evidente en la figura 6, que “car” y “bicycle” tiene una relación entre conceptos más alta, que “car” y “fork”.

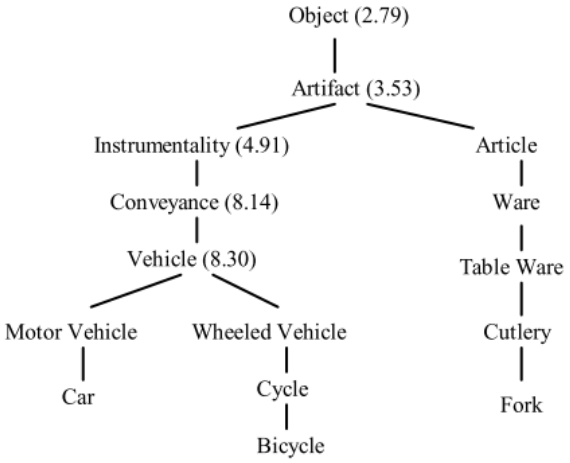


Fig. 6. Fragmento de WordNet. Tomado de Jiang and Conrath (1997)

Esta aproximación para medir la relación entre conceptos tiene dos dificultades, los antónimos y el dominio de conocimiento [27]. Pues al realizar una representación de grafo como el anterior, los antónimos estarían lejos, pero en el lenguaje humano están fuertemente relacionados. Ejemplo: Calor-Frío. El domi-

nio de conocimiento hace referencia a la complejidad que representa captar las relaciones entre conceptos de todo el lenguaje humano, incluyendo características como la polisemia, eficiencia y ambigüedad de las palabras.

Por consiguiente, es una métrica que aún está en desarrollo presentando varias aproximaciones [15][26][28][29] , donde la que mejor desempeño ha tenido es la creada por Jiang and Conrath [26].

3.4. Herramientas utilizadas para el análisis de texto

Las aplicaciones encontradas durante el ciclo de rigor, para el uso de métodos en NLP son: Stanford CoreNLP[6], OpenNLP[7], NLTK[8], Gate[9], WordNet::Similarity[10]. De ellas, se seleccionaron Gate y WordNet::Similarity para apoyar el ciclo de diseño de la investigación.

3.4.1 GATE – General architecture for text engineering

GATE es uno de los sistemas más ampliamente utilizados dentro de aquellos de su tipo, con tasas de descarga de diez miles y gran cantidad de usuarios activos en medios académicos y contextos industriales [30]. Con más de 15 años, el software se encuentra activo para su utilización en cualquier tarea computacional asociada al lenguaje humano.

Algunos los aspectos importantes:

- JAPE: Java Annotation Patterns Engine, es una versión de "CPSL – Common Pattern Specification Language". La gramática de JAPE consiste en un conjunto de fases, de las cuales cada una consiste en un conjunto de patrones/reglas. Las reglas denominadas de mano izquierda (LHS) constituyen la descripción de un patrón dentro de un grupo de anotaciones o etiquetas. Las reglas denominadas de mano derecha (RHS) consisten en declaraciones de manipulación de una anotación, definen la acción a realizar cuando un patrón se ha detectado dentro del texto gracias a lo definido por una LHS [31].
- ANNIE: Es un ejemplo de un sistema de extracción de información suministrado por GATE y que se encuentra con una aplicación pipeline ya definida, que permite facilitar el trabajo realizado con este sistema para cualquier corpus de interés.
- CREOLE Plugins: Conjunto de recursos de procesamiento que pueden ser administrados dentro de GATE para realizar tareas específicas. De acuerdo con su funcionalidad y diseño, requieren de parametrizaciones iniciales específicas para su correcta ejecución.
- GATE Developer: Interfaz gráfica de GATE que permite diseñar y ejecutar aplicaciones sobre un cuerpo de textos. La principal actividad a diseñar en una aplicación GATE consiste en generar un conjunto de anotaciones sobre textos. Cargar y ver documentos, crear y ver conjuntos de documentos, trabajar con anotaciones, utilizar CREOLE Plugins, son algunas de las tareas que pretende facilitar la utilización de GATE Developer.

[6] https://nlp.stanford.edu/
[7] https://opennlp.apache.org/
[8] http://www.nltk.org/
[9] https://gate.ac.uk/
[10] https://metacpan.org/release/WordNet-Similarity

- GATE Embedded: Es un framework orientado a objetos (librería de clases) desarrollado en java y disponible bajo GNU Lesser General Public Licence 3.0. Al igual que GATE Developer, permite realizar la debida administración de recursos de lenguaje, procesamiento y visualización.

3.4.2 WorNet::Similarity

WordNet::Similarity es un paquete de software libre diseñado para medir o aplicar el método Semantic Similarity entre conceptos. Provee seis métricas de similaridad y tres de relacionamiento, todas basadas en WordNet. Las métricas fueron implementadas en módulos Perl que toma como insumo dos conceptos, y retorna un valor numérico que representa el grado de similaridad o relacionamiento [29].

3.5. Métricas de evaluación de análisis automático de texto

Las principales métricas de clasificación texto son siete. Cuatro de ellas hacen referencia a la clasificación de las palabras, como se muestra en la tabla 6.

Tabla 6. Métricas de clasificación sobre las palabras

Métricas sobre palabras	Definición
Correctas	Palabras que son anotadas correctamente. Ejemplo: Etiquetar a "Donald Trump" como Persona.
Faltantes	Palabras que no son anotadas y que debieron serlo. Ejemplo: No etiquetar a "New York" como Locación.
Espurias	Palabras anotadas erróneamente. Ejemplo: etiquetar a "London" como locación en las palabras "Sir London"
Parcialmente correctas	Palabras anotadas correctamente, pero con otras adicionales o faltantes incorrectas. Ejemplo: etiquetar a "Trump" como Persona, es una etiqueta parcialmente correcta porque corta a "Donald". O etiquetar a "Desafortunadamente Donald Trump" como Persona es una etiqueta parcialmente correcta porque adiciona "Desafortunadamente".

Las otras tres métricas para la evaluación de análisis de texto son: Precision, Recall y F-Score, definidas de la siguiente forma.

Precisión es el número de entidades encontradas correctamente. Ver ecuación 2.

$$Precision = \frac{Correctas}{Correctas\ +\ Espurias}$$

Ecuación 2. Formula de Precision

Recall es el número de entidades que existen en el corpus que fueron encontradas. Ver ecuación 3.

$$Recall = \frac{Correctas}{Correctas\ +\ Faltantes}$$

Ecuación 3. Formula de Recall

F-Score es una combinación de Precisión y Recall conocida como la media armónica. Ver ecuación 4.

$$F - Score = 2 * \left[\frac{Precision\ *\ Recall}{Precision\ +\ Recall}\right]$$

Ecuación 4. Formula de F-Score

Con los anteriores métodos de NLP definidos, las herramientas seleccionadas y las métricas de evaluación establecidas, se construyó un servicio web que apoyará la validación semántica de árboles de problemas. El capítulo 4 resume el diseño de la investigación.

4. DISEÑO Y DESARROLLO

El diseño del servicio web consistió en tres fases principales: 1. Entrenamiento, 2. Verificación y 3. Escenario de uso. Estas fases guardan estrecha relación con las investigaciones en analítica de datos y texto, cuyas fases son: entrenamiento, verificación y prueba [16].

En el marco de la investigación basada en el diseño las tres fases mencionadas harían parte del ciclo de diseño.

La base de datos utilizada fueron los proyectos de inversión de la Gobernación de Cundinamarca de 2017. En ella se encontraron 2076 proyectos de inversión, los cuales fueron descargados en formato XML de la MGA web[11].

Un supuesto importante es que los arboles de problemas utilizados provienen de proyectos aprobados y ejecutados por la Gobernación. Esto significa que la información en ellos ha sido revisada por expertos profesionales en formulación, por lo cual las relaciones entre causas, consecuencias y problema central en los arboles de problema fueron validadas y representan un corpus adecuado para el desarrollo de proyectos de NLP.

4.1 Entrenamiento

La figura 7 muestra el pipeline utilizado durante el entrenamiento usando la notación Business Process Model and Notation (BPMN). En los cuadros se describe la tarea de forma simple, mientras que en las anotaciones de texto se mencionan los métodos NLP, con algunas excepciones.

Esta fase consistió en un pipeline de nueve pasos: limpiar los datos, crear la vista minable, traducir los problemas, identificar las palabras, identificar las oraciones, etiquetar el léxico, calcular la frecuencia de los patrones sintácticos, reconocer los patrones sintácticos y traducir los patrones de concepto principal. En esta fase solo se usó el 70% del corpus. Ver figura 7.

De los nueve pasos, siete son automáticos y dos son manuales. Los dos últimos son los pasos manuales; reconocer los patrones sintácticos y traducir los patrones de concepto principal.

Para los dos primeros pasos se construyó una aplicación en java que filtrará la información de los proyectos de inversión no requerida y obtuviera como resultado solo los árboles de problemas. Es decir, se omitió la siguiente información correspondiente a los proyectos: nombre, análisis de participantes, contribución al plan de desarrollo, localización, objetivos, análisis de alternativas, descripción de productos-servicios, análisis de beneficios y riesgos y cualquier otro tipo de etiqueta que no hiciera referencia al árbol de problemas de los XML.

[11] Aplicación utilizada por el DNP para realizar el seguimiento de los proyectos de inversión en Colombia.

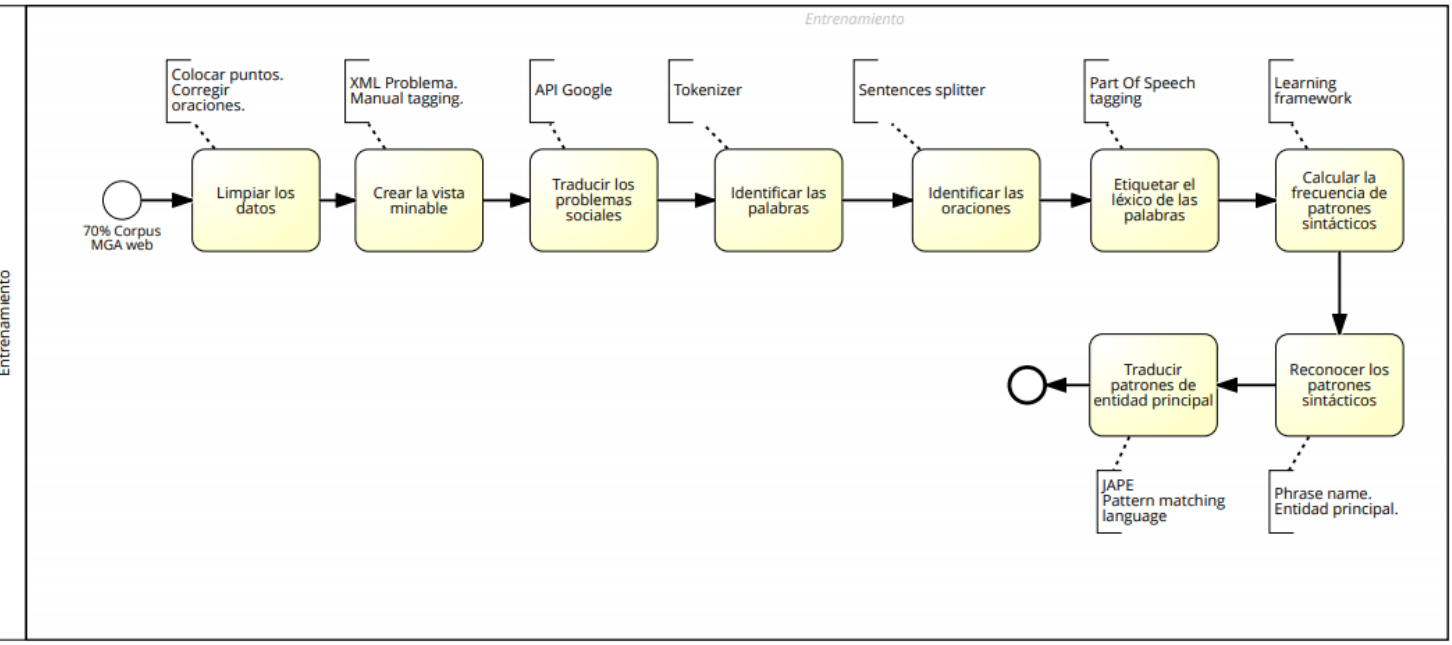


Fig. 7. Pipeline NLP fase de entrenamiento

Luego de la limpieza de datos y generación de la vista minable se obtuvieron 1794 árboles de problemas con sus respectivas oraciones causas y consecuencias. En total fueron 13037 oraciones causas, efectos y problemas analizadas en la investigación. Ver figura 8.

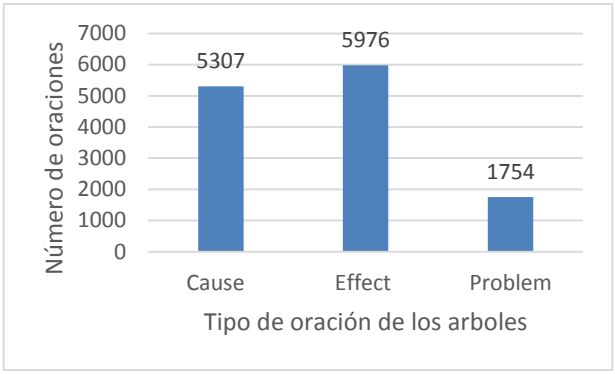


Fig. 8. Frecuencia del tipo de los tipos de oración analizadas

Adicional, durante la generación de la vista minable se decidió limitar el análisis a las causas directas, consecuencias directas y problema central del árbol de problemas porque en estas oraciones se concentran las relaciones entre conceptos. Es decir, no se consideraron causas y consecuencias indirectas.

El tercer paso consistió en conectarse con la API de Google y traducir los arboles de problemas. En este paso es importante mencionar dos aspectos de incluirlo en el pipeline. El primero, relacionado con la posibilidad de usar un recurso lingüístico para medir la similitud semántica. Dichos recursos lingüísticos están más estructurados en Ingles que en cualquier otro idioma, como es el caso de WordNet. Esta es una de las principales limitaciones a las cuales se enfrentan las investigaciones en NLP, la escasez de recursos en idiomas de origen.

El segundo, al realizar la traducción automática usando la API de Google Translate se puede agregar una distorsión del significado de las palabras y por ende a los resultados de la investigación. Sin embargo, de acuerdo con [32] [33] es la mejor aplicación para realizar traducciones de texto automáticas, al mostrar las mejores métricas de desempeño (Precision, Recall y F-Score).

Para el cuarto, quinto y sexto paso se utilizó GATE, con el plugin ANNIE. Estos pasos son métodos frecuentemente usados en el NLP, fueron definidos en el capítulo 3.3 y su aplicación no tuvo variaciones significativas.

El séptimo paso consistió en obtener la frecuencia de los patrones sintácticos en tres conjuntos asociados de acuerdo con el número de palabras a incluir. El primer conjunto se basó en la combinación de todos los patrones sintácticos en dos palabras, el segundo conjunto en tres palabras y el tercer conjunto en cuatro palabras. La figura 9 muestra el número de combinaciones por cada conjunto de patrón sintáctico.

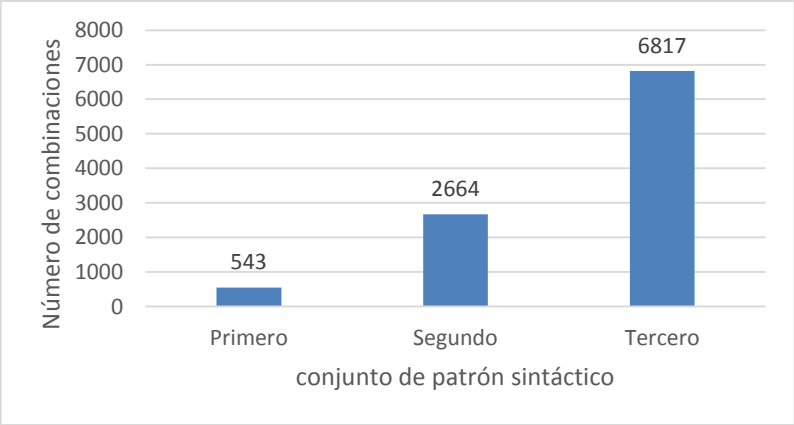


Fig. 9. Pipeline NLP fase de entrenamiento.

El octavo paso consistió en reconocer los principales patrones sintácticos de cada conjunto usando dos criterios. Un criterio visual y otro sobre la frase nombre de una oración.

En este octavo paso el supuesto es que dentro de la frase nombre se encuentra el concepto principal de las oraciones causas, problema o consecuencias. Como fue mencionado en el capítulo 3.3 la frase nombre puede ser identificada a través de la ecuación 1.

El criterio visual consistió en ordenar de mayor a menor la frecuencia los patrones de cada conjunto, y luego elegir un punto de corte con aquellos patrones que acumularan el mayor número de frecuencia. Por ejemplo, para el primer conjunto el punto de corte son 100 combinaciones como se muestra en la figura 10.

Luego de ser ordenados: el primer conjunto presentó un punto de corte en 100 patrones, el segundo conjunto en 150 patrones y el tercero en 174 patrones.

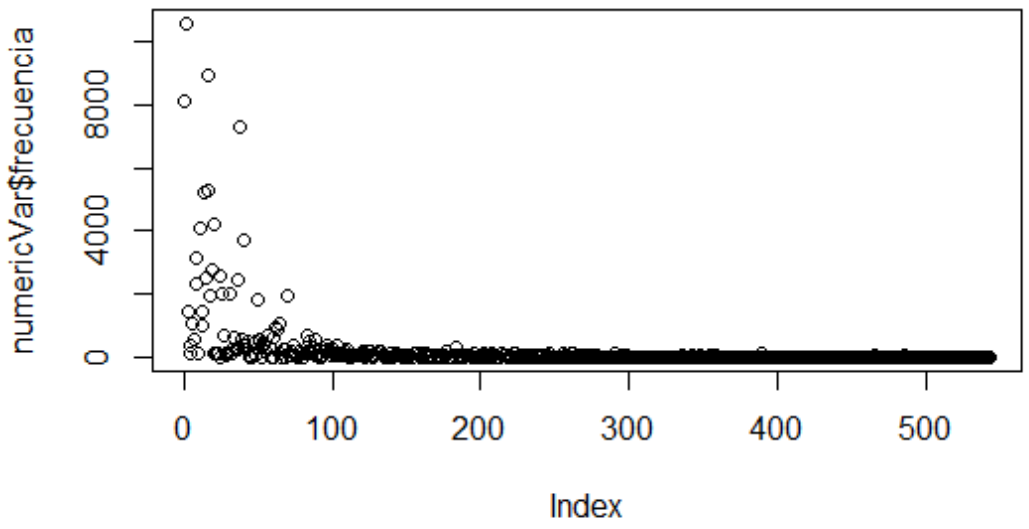


Fig. 10. Frecuencia de los patrones sintácticos del primer conjunto

El segundo criterio consistió en seleccionar las categorías sintácticas que describen una frase nombre usando la ecuación 1, descartando cualquier otro tipo de categoría sintáctica que no lo describa. Por ejemplo, en la tabla 7, se muestran algunos patrones sintácticos del primer conjunto, donde utilizando la ecuación se selecciona el patrón subrayado.

Las etiquetas utilizadas son los obtenidas por el método Part Of Speech Tagging utilizado el plugin de ANNIE en GATE [30].

Tabla 7. Ejemplo de selección de patrones para el primer conjunto

Patrones	Frecuencia
IN>DT	10588
NN>IN	8928
NNP>IN	8103
NNP>NNP	7280
DT>NN	5279
JJ>NN	5231
IN>NN	4241
DT>NNP	4105
IN>NNP	3704

El segundo criterio del octavo paso arrojó siete patrones para el primer conjunto, veintiún patrones para el segundo conjunto y nueve patrones para el tercer conjunto. Ver tabla 8.

Tabla 8. Patrones sintácticos por conjunto

Conjunto de patrones sintácticos	Numero de patrones según criterio de concepto principal
Primero	7
Segundo	21
Tercero	9

Por último, los patrones sintácticos finales fueron traducidas en reglas para identificar la entidad principal de los árboles de problemas. Esta traducción al lenguaje JAPE, es el noveno paso del pipeline en la fase de entrenamiento.

Como resultado se obtuvieron tres tipos diferentes de patrones sintácticos. Los cuales se muestran en las figuras 11, 12 y 13, traducidos como reglas JAPE para ser utilizados en la fase de verificación.

```
Phase: ReglaNGRAM2
Input: Token Sentence
Options: control = all


Rule: IdentifyMainEntity
(
 ({Token.category == "NNP"}{Token.category == "NNP"})|
({Token.category=="NN"}{Token.category=="NN"})|
({Token.category=="NN"}{Token.category=="NNS"})|
({Token.category=="NNP"}{Token.category=="NN"})|
({Token.category=="NN"}{Token.category=="NNP"})|
({Token.category=="NNS"}{Token.category=="NNS"})
)+:orgName
-->
:orgName.mainEntity = {}
```

Fig. 11. Reglas derivadas del primer conjunto de patrones sintácticos

```
Phase: ReglaNGRAM3
Input: Token Sentence
Options: control = all


Rule: IdentifyMainEntity
(
 ({Token.category == "NNS"}{Token.category == "CC"}{Token.category=="NNS"})|
({Token.category == "NN"}{Token.category == "IN"}{Token.category=="NNS"})|
({Token.category == "NN"}{Token.category == "IN"}{Token.category=="NN"})|
({Token.category == "NNP"}{Token.category == "CC"}{Token.category=="NN"})|
({Token.category == "NN"}{Token.category == "TO"}{Token.category=="NNS"})|
({Token.category == "NN"}{Token.category == "NNS"}{Token.category=="NN"})|
({Token.category == "NNS"}{Token.category == "IN"}{Token.category=="NNS"})|
({Token.category == "NNS"}{Token.category == "IN"}{Token.category=="NN"})|
({Token.category == "NNP"}{Token.category == "NN"}{Token.category=="VBZ"})|
({Token.category == "NNP"}{Token.category == "NNP"}{Token.category=="NNP"})|
({Token.category == "NNP"}{Token.category == "TO"}{Token.category=="NNP"})|
({Token.category == "NNS"}{Token.category == "IN"}{Token.category=="NNP"})|
({Token.category == "NNP"}{Token.category == "NN"}{Token.category=="NNP"})|
({Token.category == "NNP"}{Token.category == "DT"}{Token.category=="NNP"})|
({Token.category == "NN"}{Token.category == "IN"}{Token.category=="NNP"})|
({Token.category == "NNP"}{Token.category == "IN"}{Token.category=="NN"})|
({Token.category == "NNP"}{Token.category == "CC"}{Token.category=="NNP"})|
({Token.category == "NNP"}{Token.category == "IN"}{Token.category=="NNS"})|
({Token.category == "NN"}{Token.category == "CC"}{Token.category=="NN"})|
({Token.category == "NN"}{Token.category == "CC"}{Token.category=="NNS"})|
({Token.category == "NNS"}{Token.category == "CC"}{Token.category=="NN"})
)+:orgName
-->
:orgName.mainEntity = {}
```

Fig. 12. Reglas derivadas del segundo conjunto de patrones sintácticos

```
Phase: ReglaNGRAM4
Input: Token Sentence
Options: control = all


Rule: IdentifyMainEntity
(
 ({Token.category == "NNP"}{Token.category == "IN"}{Token.category=="DT"}{Token.category=="NNS"})|
({Token.category == "NNS"}{Token.category == "CC"}{Token.category=="NNS"}{Token.category=="IN"})|
({Token.category == "NNP"}{Token.category == "CC"}{Token.category=="JJ"}{Token.category=="NN"})|
({Token.category == "JJ"}{Token.category == "NN"}{Token.category=="NNS"}{Token.category=="IN"})|
({Token.category == "NN"}{Token.category == "NNS"}{Token.category=="IN"}{Token.category=="DT"})|
({Token.category == "NNS"}{Token.category == "IN"}{Token.category=="DT"}{Token.category=="NN"})|
({Token.category == "DT"}{Token.category == "NN"}{Token.category=="IN"}{Token.category=="NNS"})|
({Token.category == "NN"}{Token.category == "IN"}{Token.category=="NNS"}{Token.category=="CC"})|
({Token.category == "NNS"}{Token.category == "CC"}{Token.category=="JJ"}{Token.category=="NNS"})
)+:orgName
-->
:orgName.mainEntity = {}
```

Fig. 13. Reglas derivadas del tercer conjunto de patrones sintácticos

4.2 Verificación

La fase de verificación incluyó un pipeline de ocho pasos, buscando comparar y seleccionar el mejor conjunto de patrones sintácticos de los tres que resultaron de la fase anterior. La fase de verificación comparte los seis primeros pasos con el entrenamiento, debido a que, se debe generar la vista minable y el POS de las oraciones. En esta fase se utilizó el 30% de los árboles de problemas como corpus. Ver Figura 14.

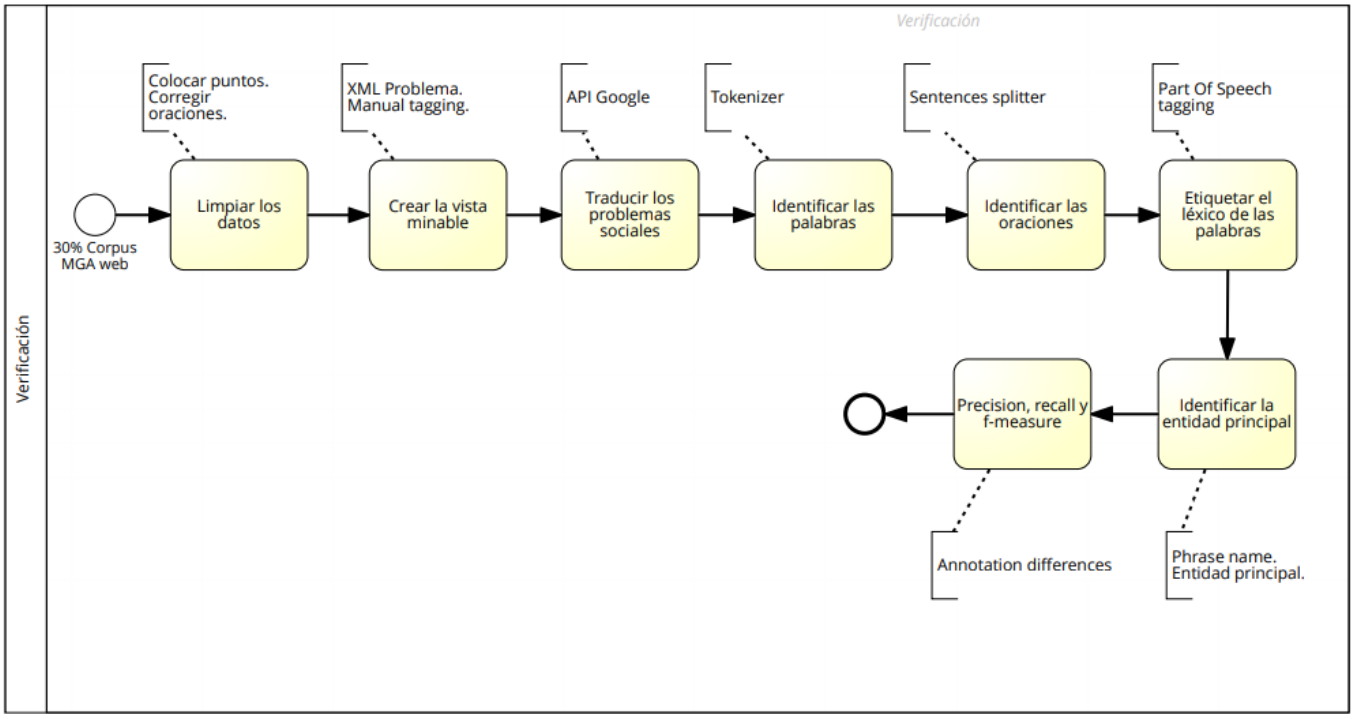


Fig. 14. Pipeline NLP fase de entrenamiento

El séptimo consistió en etiquetar las entidades principales de acuerdo con el primer conjunto de reglas. En el paso octavo se comparaban las métricas de evaluación, cada conjunto fue evaluado por los resultados obtenidos en dos formas. Primero, con las métricas de clasificación sobre las palabras. Segundo, según la Precisión, Recall y F-score cada conjunto.

La figura 15, muestra la comparación de un árbol de problemas en la primera forma de evaluación. En blanco las entidades que fueron etiquetas correctamente. En Azul, las parcialmente correctas, es decir aquellas que etiquetaron la entidad principal pero incluyeron una o más palabras adicionales. En rojo, las faltantes, aquellas entidades principales que debieron ser etiquetas. Este no contiene falsos positivos (espurias) que serían las entidades principales etiquetadas que no deberían serlo.

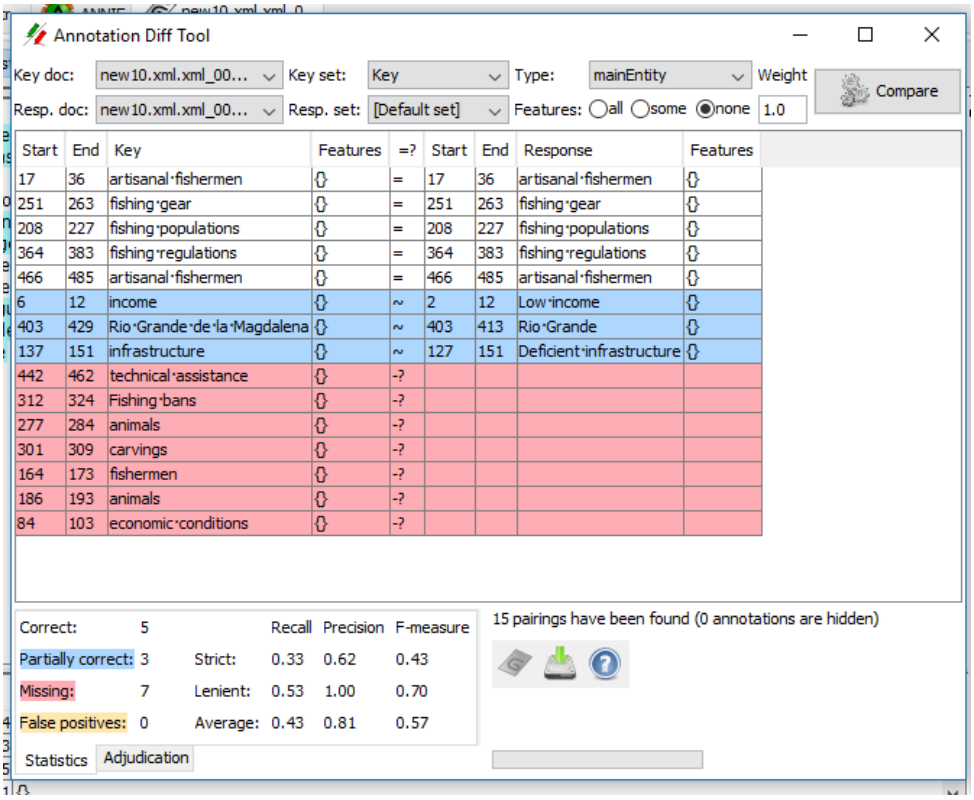
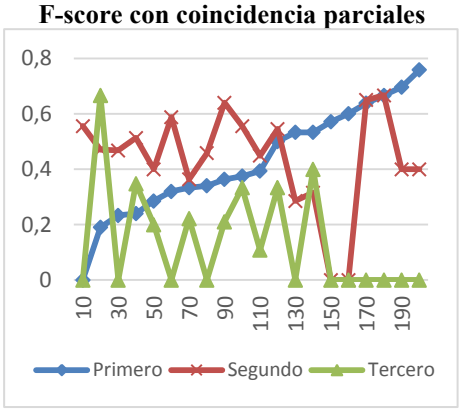


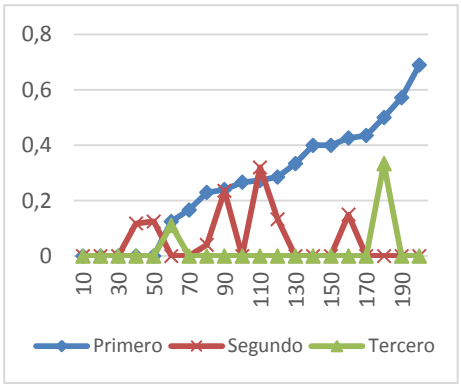
Fig. 15. Imagen comparativa de un árbol de problemas con GATE

La figura 16 muestra el comportamiento de los conjuntos de patrones sintácticos según las métricas de clasificación calculadas de manera estricta y con coincidencia parcial. La manera estricta solo usa las palabras correctas para las métricas Precision, Recall y F-Score, mientras la coincidencia parcial usa las palabras correctas más las parcialmente correctas.

F-score con coincidencia parciales



F-score estricta



Recall con coincidencias parciales

Recall estricta

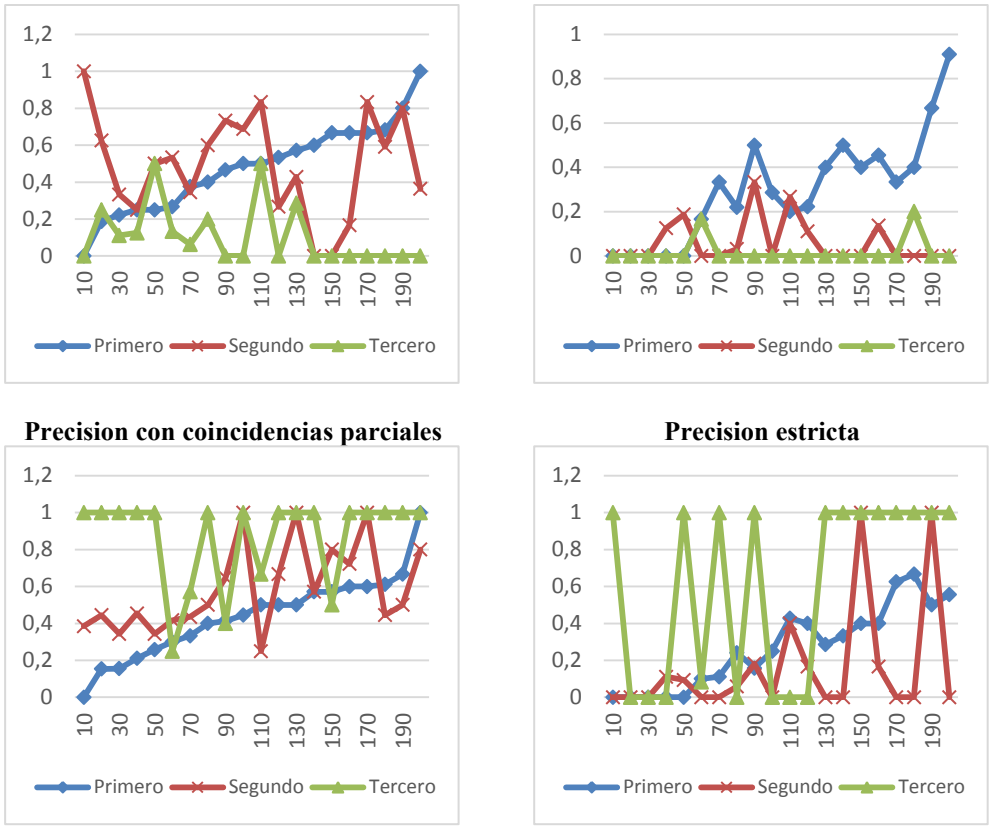


Fig. 16 comparación de las métricas de clasificación entre el conjunto de patrones sintácticos.

Al comparar los conjuntos, el tercero se destaca levemente en la Precisión con respecto a los otros conjuntos. Si la Precisión es calculada según coincidencia parciales la diferencia es menos significativa. Ver parte inferior de la figura 16.

Sin embargo, el conjunto con mejor Recall es el primero. Es decir, el primer conjunto de patrones sintácticos aumenta la diferencia sobre el segundo y tercer conjunto, entre coincidencias parciales y estrictas. Ver parte central de la figura 16.

En Recall, el tercer conjunto decae sustancialmente al compararse el cálculo con coincidencia parcial y estricta. Lo que sugiere que no son patrones sintácticos generales. Ver parte central de la figura 16.

Al comparar la metrica de F-score calculada de forma estricta es claro que el mejor conjunto de patrones sintácticos es el primero. Mantiene un buen equilibrio entre la Precision y Recall buscando identificar la entidad principal en los arboles de problemas. Ver parte superior de la figura 16.

Como resultado de la fase de verificación se tomaron los patrones sintácticos del primer conjunto para realizar el escenario de uso.

4.3 Escenario de uso

Esta sección se descompone en cuatro: la descripción del uso de los métodos NLP en los arboles de problemas, la descripción del servicio web, un prototipo de visualización y los resultados de la aplicación del TAM a los expertos del DNP.

4.3.1 Uso de los métodos NLP en los árboles de problemas

Para describir el uso propuesto en la investigación se calculará la similitud semántica en un árbol de problemas como ejemplo. De esta forma, se ilustrará cómo se enlaza la teoría de los métodos NLP expresados (capitulo 3.3), los procesos de diseño (capitulo 4.1) y los posibles usos de la investigación (capitulo 1.1).

El árbol de problemas que se usará se muestra en la figura 17, donde se describe un problema social de bajos ingresos asociados a la pesca artesanal. Contiene seis causas, dos consecuencias y un problema central.

Cada una de las causas se puede leer e interpretar de la siguiente forma: la deficiente infraestructura para la movilización de pescadores y captura de animales, es una de las causas por la cual existen bajos ingresos en los pescadores artesanales. La disminución de las poblaciones pesqueras es otra de las causas, para los bajos ingresos de los pescadores. Así para cada una de las seis causas que conforman el árbol.

Las consecuencias se pueden leer e interpretar de la siguiente forma: Los bajos ingresos para los pescadores artesanales generan como consecuencias deficientes condiciones económicas, pobreza y miseria.

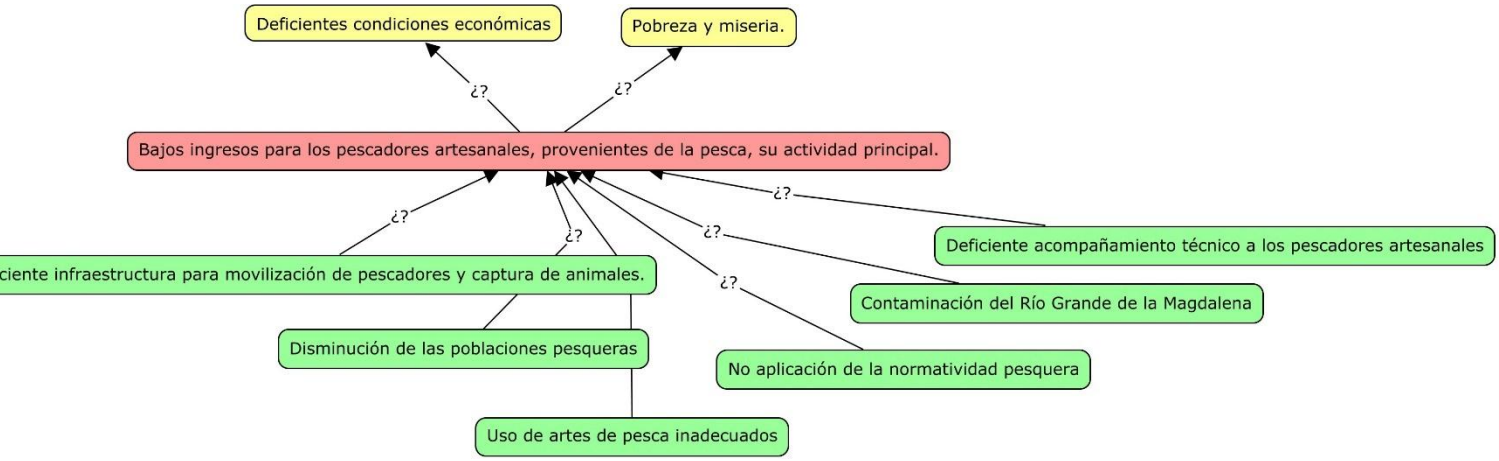


Fig. 17 Árbol de problemas de bajos ingresos derivados de la pesca artesanal

Durante un ejercicio de diseño de proyectos se debe evaluar si las relaciones construidas entre causas-problema y problema-consecuencias existen y son coherentes. Es decir, si existen relaciones entre el significado de los conceptos en las oraciones del árbol. Esta relación es la similitud semántica del lenguaje y como se expresó en el capítulo 3.1 no es fácil de lograr en los problemas sociales. Por ejemplo, no es fácil explicar la relación acerca de cómo la no aplicación de la normativa pesquera causará bajos ingresos en los pescadores. ¿Si los pescadores usaran correctamente las normas sus ingresos aumentarían?

Adicional, al evaluar la relación problema-consecuencias tampoco existe claridad acerca de cómo los bajos ingresos producen pobreza y miseria. Está relación no es directa o aceptada conceptualmente en su totalidad. La pobreza es una situación derivada de varios factores sociales, entre ellos ingreso, educación, salud. Plantear dicha relación en un árbol en el marco del diseño de proyectos está mal, porque no es clara la relación semántica directa entre el significado de los conceptos.

El análisis anterior se puede extender para todas las causas y consecuencias del árbol, donde algunas relaciones serán más evidentes que otras, pero no existirá un consenso en muchos de los escenarios reales de la aplicación de la MML. Como eje central de la investigación esta el proponer la métrica de similitud para conocer la relación entre los conceptos de las oraciones.

Los pasos para calcular la similitud semántica entre oraciones de un árbol de problemas son: Primero reconocer las palabas y oraciones, segundo etiquetar la sintaxis (léxico) de las palabras, tercero identificar los conceptos principales y cuarto calcular la similitud entre los conceptos principales. Ver pipeline en el capítulo 4.1.

Como parte de la limpieza y generación de la vista minable el árbol es reordenado y traducido al inglés. La primera oración es el problema central, las dos oraciones siguientes son las consecuencias y las seis restantes son las causas. Como se explica en la fase de entrenamiento es traducido a ingles por facilidad en el uso de los recursos lingüísticos. Ver figura 18.

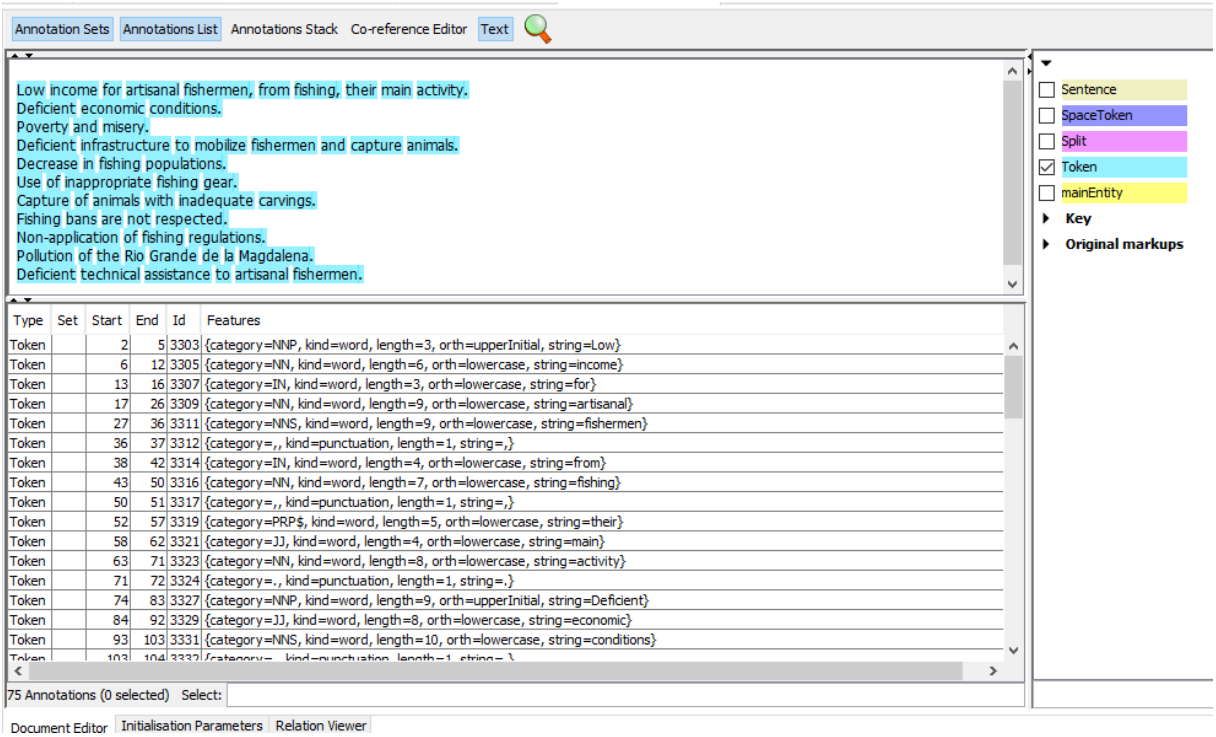


Fig. 18 Reconocimiento y etiquetado del léxico de las palabras de un árbol de problemas ejemplo

La figura 18, muestra el primer y segundo paso. Todas las palabras aparecen subrayadas porque fueron identificadas como token. En la parte inferior de la figura están etiquetadas según la sintaxis (léxico) a la que pertenecen. De esta forma: “income” es un nombre (NN), “for” es una preposición (IN), así para cada una de las palabras.

La figura 19, muestra las palabras que representan los principales conceptos en las oraciones. Para el problema central los conceptos identificados fueron “low income” y “artisanal fishermen”. Para las consecuencias no se identificó ningún concepto principal. Para las causas se identificaron “Deficient infrastructure”, “fishing populations”, fishing regulations”, “Rio Grande” y “artisanal fishermen”.

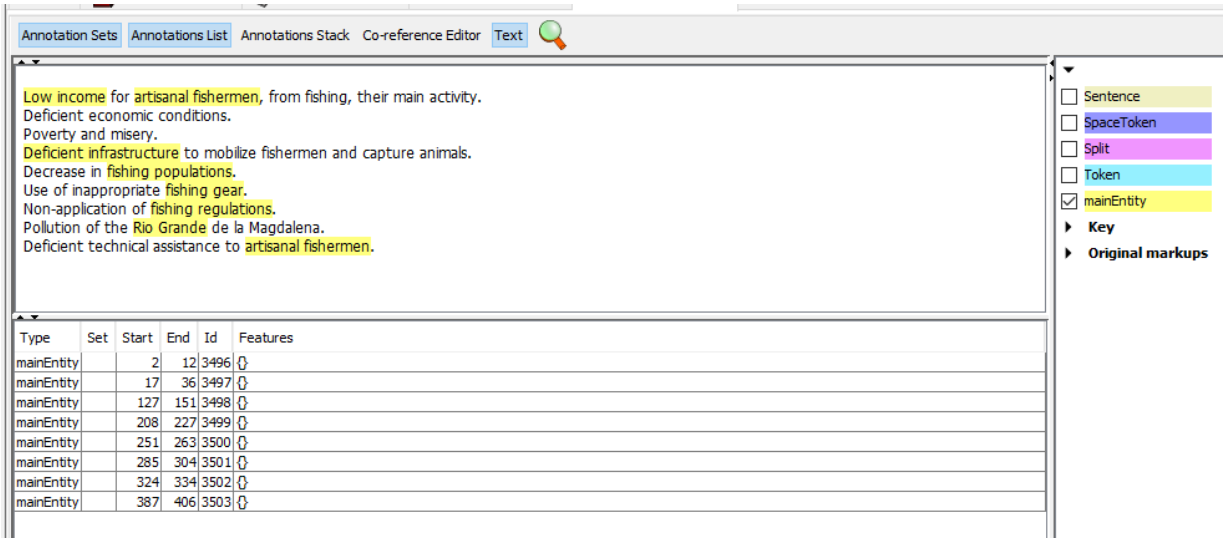


Fig. 19 Reconocimiento de los conceptos de un árbol de problemas ejemplo

Sobre estos conceptos principales se calcula la similitud semántica. La figura 20, muestra el valor derivado de aplicar el cálculo de la similitud semántica entre las oraciones causas-problema y problema-consecuencias. De acuerdo con [26] la similitud semántica arroja un valor entre 0 y 1. Donde 1 significa que existe la mayor similitud semántica y 0 que no existe similitud.

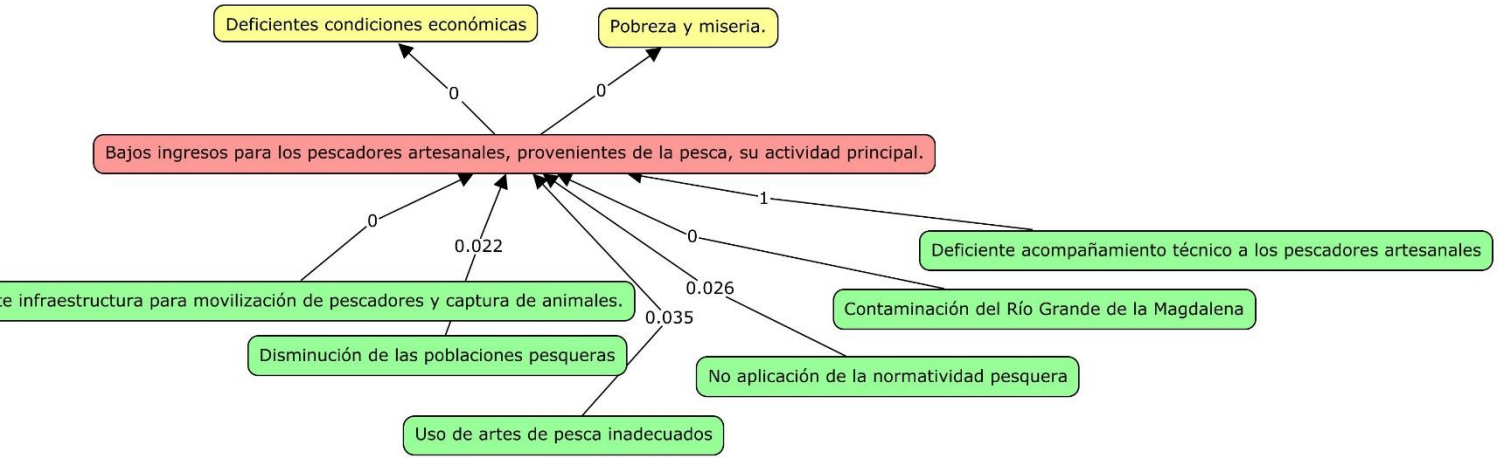


Fig. 20 Árbol de problemas de bajos ingresos derivados de la pesca artesanal

La causa con la mayor similitud semántica es el deficiente acompañamiento técnico a los pescadores con un valor de 1. Le sigue el uso de artes de pescas inadecuadas con un valor de 0.035. Seguido por la causa: no aplicación de la normatividad pesquera con 0.026 de similitud.

Las causas que no obtuvieron valor fueron: deficiencia para la movilización de pescadores y contaminación del Río Grande de la Magdalena.

Las consecuencias obtuvieron un valor de cero, dado que en ellas no hubo coincidencia en identificar un patrón sintáctico asociado al concepto principal. Aunque esto es en realidad una característica para trabajos futuros dado que la investigación no indago por los patrones sintácticos que pudieran ser prescriptivos.

Como resultado, cinco de las causas y las dos consecuencias deben ser rescritas y ajustadas buscando mejorar la relación semántica entre el árbol. La única causa que tiene relación con afectar los bajos ingresos de los pescadores es la relacionada con el deficiente acompañamiento técnico que estos reciben.

Como proceso para rescribir el problema se indagó en WordNet buscando los conceptos más cercanos. De acuerdo con WordNet los conceptos la pesca artesanal y se considera un negocio, antes que una actividad económica, por lo cual las modificaciones buscando mejorar la relación semántica deben estar en ese camino. Después de varias iteraciones, los cambios que se proponen en las oraciones son:

- Problema central: modificar “pescadores artesanales” a “pesca artesanal”.
- Consecuencia 1: modificar “económicas” a “negocios”.
- Consecuencia 2: modificar “pobreza y miseria” a “inhabilidad para acumular activos en la vivienda”.
- Causa 4: Modificar “normativos” a “conocimientos”.

La figura 21 muestra el cálculo de la similitud semántica entre las nuevas oraciones del árbol dadas las modificaciones descritas. Las consecuencias aún pueden mejorar, pero tienen una mayor relación semántica con respecto a las iniciales que se muestran en la figura 17. Adicional, cuatro causas mejoraron su relación al punto máximo. La causa “Deficiente acompañamiento técnico a los pescadores artesanales” disminuyó su relación semántica derivada de la modificación en el problema central.

Recomendaciones continuar iterando en las consecuencias, eliminar la causa relacionada con la contaminación.

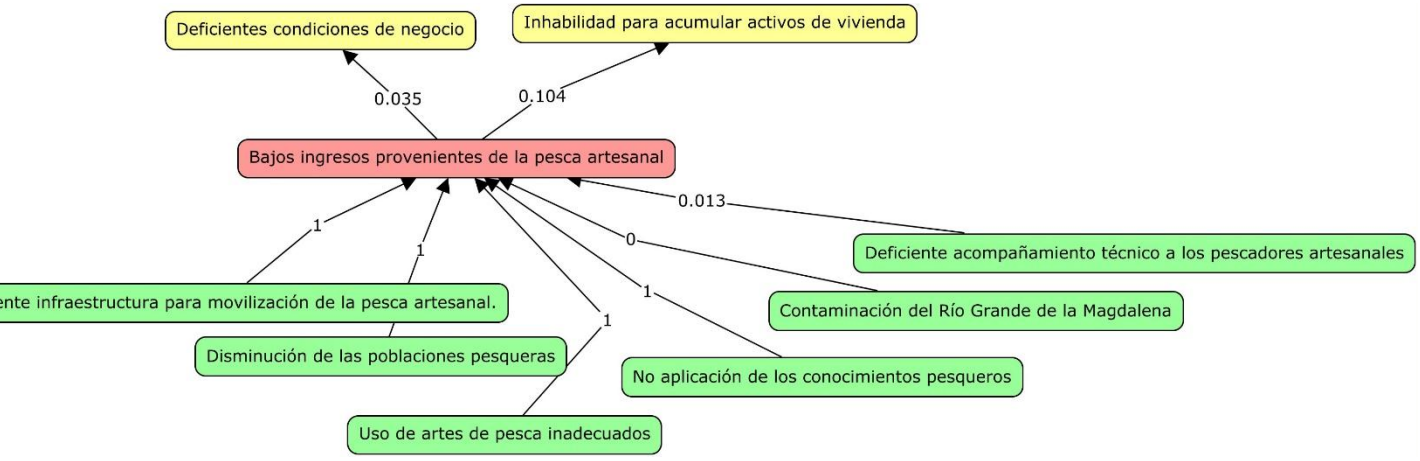


Fig. 21 Árbol de problemas de bajos ingresos derivados de la pesca artesanal modificado

Las figuras 17 y 20 representan una métrica de utilidad para los profesionales formuladores y evaluadores de proyectos, como se demuestra en el capítulo 4.3.4. A continuación se describen los aspectos computacionales del servicio web junto con un prototipo de visualización.

4.3.2 Descripción del servicio web

Un servicio web es un sistema software diseñado para soportar la interacción máquina a máquina, a través de una red, de forma interoperable. Cuenta con una interfaz descrita en un formato procesable por un equipo informático (específicamente en WSDL), a través de la que es posible interactuar con el mismo median-

te el intercambio de mensajes SOAP, típicamente transmitidos usando serialización XML sobre HTTP conjuntamente con otros estándares web [34].

El servicio web implementado tiene las siguientes características:

1. <types>, son de tipo string clasificados en el problema central, las causas y las consecuencias. Adicional también está el tipo mainEntities que es la etiqueta de las palabras consideradas las entidades principales dentro de las oraciones. Ver gráfica 16.

```
<complexType name="NodeType">
    <sequence>
        <element name="Problem" type="string" maxOccurs="1"
            minOccurs="0">
        </element>
        <element name="Effect" type="string" maxOccurs="1"
            minOccurs="0">
        </element>
        <element name="Cause" type="string" maxOccurs="1"
            minOccurs="0">
        </element>
        <element name="mainEntities"
            type="impl:MainEntitiesType" maxOccurs="1"
            minOccurs="0">
        </element>
        <element name="semanticSimilarity" type="double" maxOccurs="1" minOccurs="0"></element>
    </sequence>
</complexType>
```

Fig. 22 tipos de datos usados en el servicio web

2. < message> Existen dos mensajes que comparten el mismo elemento. El mensaje de solicitud es un árbol de problemas y el mensaje de respuesta es un árbol de problemas con el cálculo de la similitud semántica. Ver figura 17.

```
<wsdl:message name="processTreeProblemResponse">

    <wsdl:part element="impl:processTreeProblemResponse" name="parameters">

    </wsdl:part>

</wsdl:message>

<wsdl:message name="processTreeProblemRequest">

    <wsdl:part element="impl:processTreeProblem" name="parameters">

    </wsdl:part>

</wsdl:message>
```

Fig. 23. Elementos del mensaje en el servicio web

3. <portType> La operación permitida es la llamada al método “processTreeProblem” que ejecuta el cálculo de la similitud semántica. Los mensajes intercambiados es un requerimiento para un árbol de problemas y una respuesta con un árbol de problemas con el cálculo de similitud semántica.

```
<wsdl:portType name="ServiceEJB">
    <wsdl:operation name="processTreeProblem">
        <wsdl:input message="impl:processTreeProblemRequest" name="processTreeProblemRequest">
        </wsdl:input>
        <wsdl:output message="impl:processTreeProblemResponse" name="processTreeProblemResponse">
        </wsdl:output>
    </wsdl:operation>
```

4. <binding> El protocolo utilizado es Simple Object Access Protocol (Soap)

```
<wsdl:binding name="ServiceEJBSoapBinding" type="impl:ServiceEJB">
    <wsdlsoap:binding style="document" transport="http://schemas.xmlsoap.org/soap/http"/>
    <wsdl:operation name="processTreeProblem">
        <wsdlsoap:operation soapAction=""/>
        <wsdl:input name="processTreeProblemRequest">
            <wsdlsoap:body use="literal"/>
        </wsdl:input>
        <wsdl:output name="processTreeProblemResponse">
            <wsdlsoap:body use="literal"/>
        </wsdl:output>
    </wsdl:operation>
</wsdl:binding>
```

Fig. 24 Protocolo del servicio web

5. <service> el servicio web esta implementado en un puerto local.

```
<wsdl:service name="ServiceEJBService">
    <wsdl:port binding="impl:ServiceEJBSoapBinding" name="ServiceEJB">
        <wsdlsoap:address location="http://localhost:8080/similarity-web-project/services/ServiceEJB"/>
    </wsdl:port>
</wsdl:service>
```

Fig. 25 Puerto del servicio web

La descripción completa del servicio web se muestra en el Anexo 2. WSDL

4.3.3 Prototipo de visualización

Con el objetivo de recrear un escenario de uso lo más real posible, donde la experiencia de los expertos del DNP estaba relacionada más con las ciencias económicas, y no con el campo computacional, se buscó complementar el servicio web con un prototipo de visualización. Este propotipo consistió en integrar la métrica de similitud semántica con las oraciones del árbol.

De esta forma, la visualización está constituida por tres componentes como se muestra en la figura 20. El primer componente el árbol de problemas, con las causas del lado izquierdo, el problema en la parte central y las consecuencias en la derecha. Entre las líneas que unen las oraciones se visualizará la métrica de similitud semántica.

El segundo componente es un rango de la métrica de similitud semántica que permitirá al usuario identificar cuáles causas o consecuencias están bajas, bien o excelente relacionadas. En este caso, los umbrales de evaluación de similitud fueron determinados de manera provisional, pero el número en sí no tiene todavía significado real para los usuarios formuladores de proyectos. Este proceso debe ser parte de trabajos futuros.

El tercer componente es una lista de las palabras con menor similitud semántica.

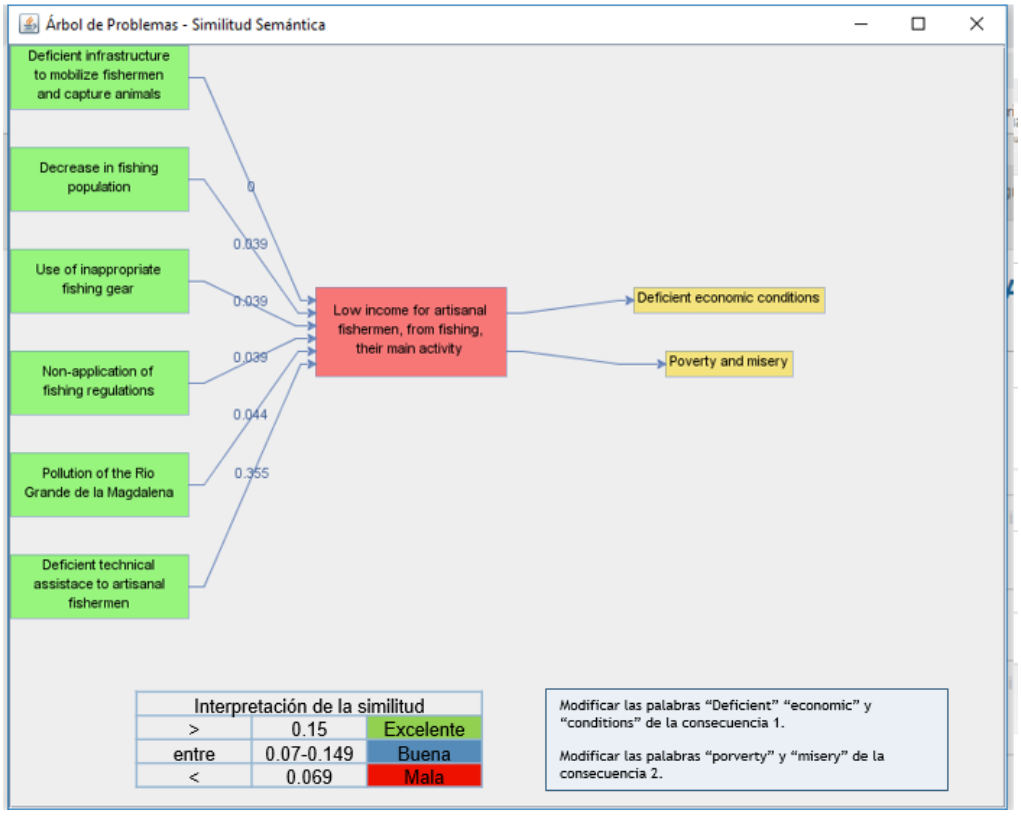


Fig. 26 Prototipo de visualización

4.3.4 Resultados TAM

El modelo de aceptación tecnológica (TAM) es una herramienta ampliamente utilizada para evaluar la utilidad potencial de los artefactos o sistemas de información [35]. La herramienta busca evaluar las percepciones de utilidad y facilidad de uso como características principales. La utilidad se entiende como la posible mejora en el desempeño del trabajo apoyado por el sistema de información propuesto. Mientras la facilidad de uso se entiende como el esfuerzo en el uso del sistema de información propuesto [36].

En los últimos años TAM ha recibido aportes para incluir otro tipo de variables que también pueden influenciar la aceptación de un sistema de información. Tales como percepción inicial, experiencia, normas subjetivas, imagen, entre otras. De esta forma las investigaciones que lo aplican realizan un análisis acerca de las variables por incluir o no de acuerdo con el contexto, teniendo en cuenta que las principales siempre deben estar. Es decir, utilidad y facilidad de uso.

En el caso de la presente investigación se incluyeron, adicional a las variables principales, normas subjetivas e intención de uso. Ambas con el objetivo de conocer por parte de los expertos del DNP la percepción sobre la posibilidad del uso del servicio. Esto es, por ejemplo, si ellos como miembros de la mayor organización encargada de la formulación y seguimiento de la política pública de Colombia y por ende de los proyectos sociales, considera que otras organizaciones públicas y profesionales relacionados con el tema percibirían el servicio web como útil.

La aplicación del TAM tuvo cuatro momentos. Un primero momento donde se realizó una presentación acerca del desarrollo de la investigación. Uno segundo, donde se realizó la demostración del servicio web. Uno tercero, donde se aplicó el TAM. Y un cuarto, donde se conversó sobre la percepción del servicio web y las potencialidades del uso de métodos de NLP en la formulación de proyectos.

La demostración del servicio web se basó en la aplicación de una consulta en tiempo real y luego su visualización a través del prototipo. La consulta en tiempo real utilizó la aplicación ReadyApi[12] como apoyo para explicar a los participantes el mensaje consulta y el mensaje respuesta. Esta aplicación es open source y su objetivo es apoyar las pruebas de APIs.

[12] https://www.soapui.org/professional/soapui-pro.html

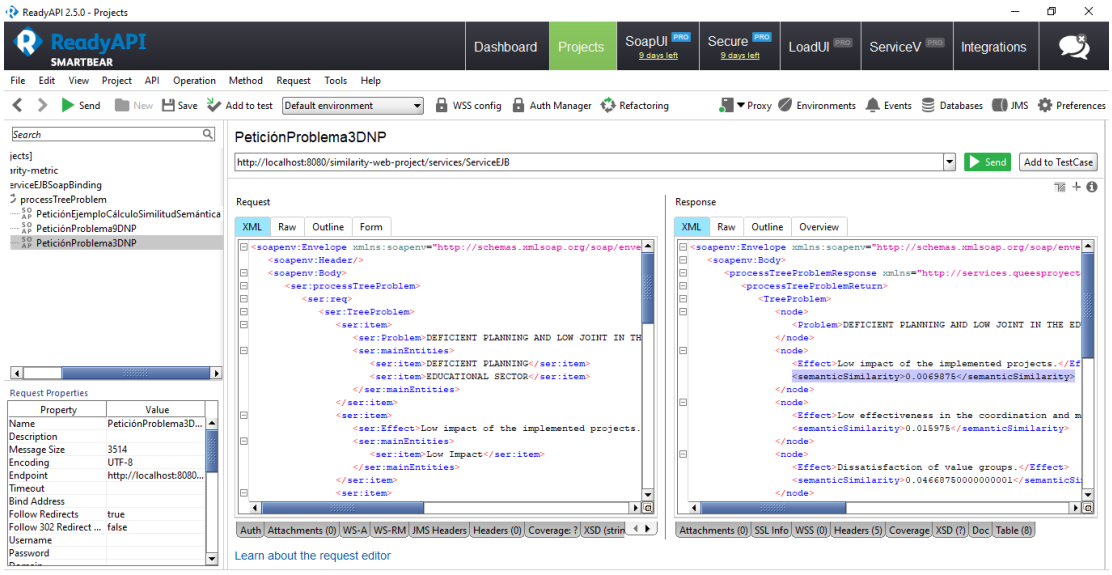


Fig. 27 Uso de la aplicación ReadyApi para probar el servicio web

La combinación de servicio web y prototipo de visualización permitió complementar el instrumento TAM, buscando de profundizar en el ciclo de diseño en la validación semántica de árboles de problemas.

Los resultados de la aplicación del TAM se muestran en la tabla 8. Subrayado se encuentran los ítems que obtuvieron baja calificación.

En general, el 83% de los entrevistados considera que el servicio mejoraría el desempeño en la evaluación de árboles de problemas en el marco del diseño de proyectos de inversión. El 60% considera que el servicio web no requiere esfuerzo en el uso.

En detalle, el ítem con menor calificación de la variable utilidad, es la percepción acerca de la facilidad para evaluar los árboles de problemas en los proyectos a cargo. Esto se debe a que los entrevistados perciben que el servicio web es un apoyo para validar arboles de problemas, pero las decisiones acerca de cómo mejorar el árbol de problemas requieren de trabajo e investigación. Es decir, cuando una causa tiene baja similitud semántica con el problema central esta debe ser modificada, pero pensar como debe ser esa modificación requiere trabajo.

Por ejemplo, uno de los expertos comentó "…como métrica propuesta era super valiosa, sobre todo porque no existen documentos sobre cómo realizar este paso desde marco lógico… pero que este era la primera pierna, hacía falta la segunda para lograr la revolución en la formulación de proyectos".

La gran percepción entre los expertos fue estar de acuerdo en que el servicio web era una métrica innovadora. Siempre existen conflictos en el proceso de realizar un árbol de problemas en el cual se necesitan aspectos objetivos para el análisis del mismo.

Por su parte, el ítem con menor calificación de la variable facilidad de uso, es la percepción de que usando el servicio no se requerirá ningún esfuerzo mental para evaluar los arboles de problemas. Evidentemente si se requiere un esfuerzo mental, el cual requiere de conocer los conceptos y elementos que componen un problema social, ya sea, para plantearlo en el servicio web y que este evalué su similitud semántica, o para corregirlo una vez se conozcan las oraciones causas-problema o problema-consecuencia que tienen baja relación.

Uno de los comentarios sobre la facilidad de uso fue el siguiente: "los proyectos de inversión tienen muchas variables asociadas al contexto, y lo que normalmente sucede es que un proyecto de un lado no puede ser usado en otro lado, tal vez la herramienta pueda ser complementada con estudios…diagnósticos y otros aspectos…".

Adicional, el 80% de los entrevistados considera que las organizaciones públicas encontrarían útil el servicio web. Mientras que la intención de uso solo es considerada por el 55% de los entrevistados.

Un aspecto adicional que surgió durante la sesión fue la revisión del componente tres de la visualización del prototipo. Uno de los expertos manifestó que "…no puede existir similitud semántica de uno… porque significaría que existen dos oraciones iguales… situación que no se presenta en los arboles de problemas nunca"

De lo anterior, se deriva que una mejora para el prototipo de visualización es definir un límite de rango superior que denote demasiada relación semántica entre las oraciones como perjudicial en el árbol de problemas.

Tabla 9. Resultados de la aplicación el TAM

ITEM	AFIRMACIÓN	MEDIANA	ACEPTACIÓN POR ÍTEM	ACEPTACIÓN POR VARIABLE
Utilidad Percibida	Usando el servicio web en mi trabajo lograré la evaluación de árboles de problemas más rápido.	4	100%	83%
	Usando el servicio web mejoraré mi desempeño evaluando los árboles de problemas.	4	100%	
	Usando el servicio web lograré incrementar mi productividad evaluando los árboles de problemas.	4	100%	
	Usando el servicio web seré más efectivo (RAE: lograr el efecto que se desea) evaluando los árboles de problemas.	4	100%	
	Usar el servicio web hará más fácil la evaluación de los árboles de problemas en los proyectos a cargo.	3	40%	
	Pienso que el servicio web es útil para apoyar la asesoría y formulación de mis proyectos.	4	60%	
Facilidad de uso percibida	Aprender a usar el servicio web me resultará fácil.	4	80%	60%
	Encuentro que será fácil hacer que el servicio web evalué los árboles de problemas.	4	60%	
	Mi interacción con el servicio web será sencilla y comprensible.	4	60%	

	Podré evaluar gran cantidad de árboles de problemas usando el servicio web.	4	60%	
	Es fácil ser un usuario experto del servicio web.	5	80%	
	Evaluar los arboles de problemas usando el servicio web no requerirá ningún esfuerzo mental.	2	20%	
Normas subjetivas	Los funcionarios públicos responsables de la formulación de proyectos, ubicados en ministerios, gobernaciones y alcaldías encontrarán útil el servicio web.	4	100%	80%
	En general, al DNP le interesaría apoyar el uso del servicio web.	4	60%	
Intención de uso	Considero que en el futuro usaré el servicio web cuando pueda	4	60%	55%
	Pienso que usaré el servicio web frecuentemente	3	40%	
	Intentaré usar el servicio web apenas esté disponible	4	60%	
	Recomendaría a mis colegas usar el servicio web en su trabajo	4	60%	

CONCLUSIONES

Para expresar las conclusiones, el análisis se descompone en tres partes. La primera la comparación de los resultados de la investigación con los objetivos propuestos. La segunda sobre la utilidad del servicio web y las mejoras posibles para otro ciclo de diseño como parte de trabajos futuros. La tercera, las preguntas que se plantearon relevantes para la investigación.

La investigación tuvo cuatro objetivos específicos alcanzados, los cuales se describen según los resultados logrados a continuación:

Los árboles de problemas contienen muchos nombres comunes y adjetivos. Se usan muy pocos los verbos y no contienen cardinales, ordinales o cuantificadores. En esta primera iteración la entidad principal puede ser identificada a través de siete patrones sintácticos, que representan el mejor conjunto debido a su desempeño en Presicion y Recall.

La investigación determinó tres pipelines, como arquitecturas para procesos de NLP en problemas sociales. Esta es una contribución significativa dado que poco existe en Colombia sobre NLP como apoyo al diseño de política pública. Específicamente, la investigación utilizó una la MGA web proponiendo una nueva forma de uso y análisis de la información que se carga en dicha aplicación.

Teniendo en cuenta lo anterior, se implementó un servicio web y un prototipo de visualización basado en NLP que apoye la evaluación de problemas sociales, que toma como insumo los XML de la MGA web. Donde el cálculo de similitud semántica se realiza entre frases con una estructura sintáctica previamente identificada que representan el concepto principal de una oración.

Los profesionales del DNP determinaron que el servicio web, el prototipo de visualización y la métrica de similitud semántica, constituyen una herramienta útil para validar semánticamente los arboles de problemas en el marco del diseño de proyectos de inversión.

Las mejoras que se pueden realizar en el diseño del servicio y prototipo de visualización son:

Debido a que los expertos perciben que el servicio web requerirá de esfuerzo mental para su uso. Adicional a determinar cuáles oraciones causas-problema y problema-consecuencia tienen baja similitud semántica, se deben incluir que conceptos adicionales pueden apoyar a mejorar las relaciones. Esto se puede lograr visualizando las secciones del grafo de WordNet sobre los conceptos comparados de las oraciones.

El prototipo de visualización debe definir un límite de rango superior que denote demasiada relación semántica entre las oraciones como perjudicial en el árbol de problemas. Que exista una similitud semántica con valor de uno entre algunas de las oraciones significa que estas son iguales, situación que iría en contra de la dinámica de un árbol de problemas y la Metodología Marco Lógico.

Sobre las preguntas que se plantearon relevantes para la investigación: ¿Es posible configurar los métodos de NLP, como apoyo a la validación de árboles de problemas sociales en el marco de la MML? Y si es así ¿Cómo sería su configuración y cuales métodos usar o descartar?

La respuesta a la primera pregunta es que sí. La correspondencia entre conceptos e identificar la entidad principal compartida en un árbol de problemas son métodos del Procesamiento del Lenguaje Natural. Una vez se tiene una ilustración completa de un árbol de problemas, los métodos de NLP pueden proporcionar una métrica de validación semántica entre las oraciones causas-problema y problema-consecuencias.

Con respecto a la segunda pregunta, los métodos NLP que permiten apoyar la validación semántica de árboles de problemas son: Tokenizer, Sentences Breaking, Part Of Speech Tagging, Named Entity Recognition (NER) y Semantic Similarity.

REFERENCIAS

[1] E. Ortegón, J. F. Pacheco, and A. Prieto, "Metodología del marco lógico para la planificación, el seguimiento y la evaluación de proyectos y programas," Santiago de Chile: Naciones Unidas, 2005, pp. 9–13.

[2] E. Aldana and A. Reyes, "Disolver Problemas: criterio para formular proyectos sociales," Bogotá: Departamento de Ingeniería Industrial, Universidad de los Andes, 2004, pp. 26–33.

[3] V. Nastase and M. Strube, "Transforming Wikipedia into a large scale multilingual concept network," *Artif. Intell.*, vol. 194, pp. 62–85, 2013.

[4] W. J. Orlikowski, "Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations," *Organ. Sci.*, vol. 11, no. 4, pp. 404–428, 2000.

[5] J. Couillard, S. Garon, and J. Riznic, "The Logical Framework Approach–Millennium.," *Proj. Manag. J.*, vol. 40, no. 4, pp. 31–44, 2009.

[6] Departamento Nacional de Planeación, "Guía del módulo de capacitación en teoría de proyectos."

[7] J. T. Klein, "Evaluation of Interdisciplinary and Transdisciplinary Research: A Literature Review," *Am. J. Prev. Med.*, vol. 35, no. 2, Supplement, pp. S116–S123, 2008.

[8] W. Ulrich, *Critical Heuristics of Social Planning: A New Approach to Practical Philosophy*. J. Wiley & Sons, 1983.

[9] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI Soc.*, vol. 30, no. 1, pp. 89–116, 2014.

[10] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," *Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit.*, vol. 21, pp. 0–934, 2009.

[11] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, 2014.

[12] A. Hevner and S. Chateerjee, *Integrated Series in Information Systems Volume 28*, vol. 28. 2012.

[13] E. Ortegon, J. F. Pacheco, and H. Roura, *Metodología general de identificación, preparación y evaluación de proyectos de inversión pública*. 2005.

[14] E. Aldunate and J. Córdoba, *Formulación de programas con la metodología de marco lógico*, vol. 68. Santiago de Chile: Serie de manuales. Instituto Latinoamericano y del Carible de Planifiación Económica y Social (ILPES)Instituto Latinoamericano y del Carible de Planificación Económica y Social (ILPES), 2011.

[15] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Comput. Linguist.*, vol. 32, no. 1, pp. 13–47, 2006.

[16] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking*, 1st ed. O'Reilly Media, Inc., 2013.

[17] E. Cambria, P. Gastaldo, F. Bisio, and R. Zunino, "An ELM-based model for affective analogical reasoning," *Neurocomputing*, vol. 149, no. Part A, pp. 443–455, 2015.

[18] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2016.

[19] R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 678–692, 2010.

[20] W. Wong, W. E. I. Liu, and M. Bennamoun, "Ontology learning from text," *ACM Comput. Surv.*, vol. 44, no. 4, pp. 1–36, 2012.

[21] D. Sánchez, M. Batet, A. Valls, and K. Gibert, *Ontology-driven web-based semantic similarity*, vol. 35, no. 3. 2010.

[22] D. Milne and I. H. Witten, "An open-source toolkit for mining Wikipedia," *Artif. Intell.*, vol. 194, pp. 222–239, 2013.

[23] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," *Proc. Natl. Conf. Artif. Intell.*, vol. 21, no. 2, p. 1419, 2006.

[24] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, 2012.

[25] J. Thomas, J. McNaught, and S. Ananiadou, "Applications of text mining within systematic reviews," *Res. Synth. Methods*, vol. 2, no. January, pp. 1–14, 2011.

[26] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," no. Rocling X, 1997.

[27] J. E. Weeds, "Measures and Applications of Lexical Distributional Similarity," *Dep. Informatics*, no. September, p. 204, 2003.

[28] Y. H. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, 2006.

[29] T. Pedersen and J. Michelizzi, "WordNet :: Similarity - Measuring the Relatedness of Concepts," *HLT-NAACL--Demonstrations '04 Demonstr. Pap. HLT-NAACL 2004*, no. July, pp. 38–41, 1998.

[30] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics," *PLoS Comput. Biol.*, vol. 9, no. 2, Feb. 2013.

[31] H. Cunningham, D. Maynard, and K. Bontcheva, *Text Processing with GATE*. Gateway Press CA, 2011.

[32] X. Wan, "Co-training for Cross-lingual Sentiment Classification," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, 2009, pp. 235–243.

[33] X. Wan, "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 553–561.

[34] W3C, "«Web Services Glossary § Web service»." [Online]. Available: https://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/#webservice. [Accessed: 12-Nov-2018].

[35] B. H. Venkatesh V., "Technology Acceptance Model 3 and a Research Agenda on Interventions," *Decis. Sci.*, vol. 39, no. 2, pp. 273–315, 2008.

[36] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information," *Inf. Technol. MIS Q.*, vol. 13, no. 3, p. 319–340., 1989.

ANEXO 1

CARTA DE AUTORIZACIÓN DE LOS AUTORES

(Licencia de uso)

Bogotá, D.C., 5 diciembre 2018

Señores
Biblioteca Alfonso Borrero Cabal S.J. Pontificia Universidad Javeriana Cuidad

Los suscritos: Juan Pablo Pajaro Hernandez con C.C No 73.207.610 de Cartagena

En mi (nuestra) calidad de autor (es) exclusivo (s) de la obra titulada: **PROCESAMIENTO DE LENGUAJE NATURAL PARA LA EVALUACIÓN DE PROBLEMAS SOCIALES.**

(por favor señale con una "x" las opciones que apliquen)

Tesis doctoral [ ] Trabajo de grado [X] Premio o distinción: **Si** [ ] **No (X)** [ ]

cual:

presentado y aprobado en el año 2018, por medio del presente escrito autorizo (autorizamos) a la Pontificia Universidad Javeriana para que, en desarrollo de la presente licencia de uso parcial, pueda ejercer sobre mi (nuestra) obra las atribuciones que se indican a continuación, teniendo en cuenta que en cualquier caso, la finalidad perseguida será facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, las atribuciones de usos temporales y parciales que por virtud de la presente licencia se autorizan a la Pontificia Universidad Javeriana, a los usuarios de la Biblioteca Alfonso Borrero Cabal S.J., así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio, son:

AUTORIZO (AUTORIZAMOS)	SI	NO
1. La conservación de los ejemplares necesarios en la sala de tesis y trabajos de grado de la Biblioteca.	X	
2. La consulta física (sólo en las instalaciones de la Biblioteca)		
3. La consulta electrónica – on line (a través del catálogo Biblos y el Repositorio Institucional)	X	
4. La reproducción por cualquier formato conocido o por conocer		
5. La comunicación pública por cualquier procedimiento o medio físico o electrónico, así como su puesta a disposición en Internet	X	
6. La inclusión en bases de datos y en sitios web sean éstos onerosos o gratuitos, existiendo con ellos previo convenio perfeccionado con la Pontificia Universidad Javeriana para efectos de satisfacer los fines previstos. En este evento, tales sitios y sus usuarios tendrán las mismas facultades que las aquí concedidas con las mismas limitaciones y condiciones	X	

De acuerdo con la naturaleza del uso concedido, la presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi

(nuestra) obra sea explotada en las condiciones aquí estipuladas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de estudiante (s) y por ende autor (es) exclusivo (s), que la Tesis o Trabajo de Grado en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la Tesis o Trabajo de Grado es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontifica Universidad Javeriana por tales aspectos.

Sin perjuicio de los usos y atribuciones otorgadas en virtud de este documento, continuaré (continuaremos) conservando los correspondientes derechos patrimoniales sin modificación o restricción alguna, puesto que de acuerdo con la legislación colombiana aplicable, el presente es un acuerdo jurídico que en ningún caso conlleva la enajenación de los derechos patrimoniales derivados del régimen del Derecho de Autor.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, *"Los derechos morales sobre el trabajo son propiedad de los autores"*, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de RESPETARLOS Y HACERLOS RESPETAR, para lo cual tomará las medidas correspondientes para garantizar su observancia.

NOTA: Información Confidencial:
Esta Tesis o Trabajo de Grado contiene información privilegiada, estratégica, secreta, confidencial y demás similar, o hace parte de una investigación que se adelanta y cuyos resultados finales no se han publicado. Si [ ] No [(X)]
En caso afirmativo expresamente indicaré (indicaremos), en carta adjunta, tal situación con el fin de que se mantenga la restricción de acceso.

NOMBRE COMPLETO	No. del documento de identidad	FIRMA
JUAN PABLO PAJARO HERNANDEZ	73207610	

FACULTAD: INGENIERÍA

PROGRAMA ACADÉMICO: MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

ANEXO 2

BIBLIOTECA ALFONSO BORRERO CABAL, S.J.
DESCRIPCIÓN DE LA TESIS O DEL TRABAJO DE GRADO
FORMULARIO

TÍTULO COMPLETO DE LA TESIS DOCTORAL O TRABAJO DE GRADO
PROCESAMIENTO DE LENGUAJE NATURAL PARA LA EVALUACIÓN DE PROBLEMAS SOCIALES
SUBTÍTULO, SI LO TIENE

AUTOR O AUTORES	
Apellidos Completos	**Nombres Completos**
PAJARO HERNANDEZ	JUAN PABLO

DIRECTOR (ES) TESIS O DEL TRABAJO DE GRADO	
Apellidos Completos	**Nombres Completos**
GONZALEZ RIVERA	RAFAEL ANDRES

FACULTAD
INGENIERIA

PROGRAMA ACADÉMICO			
Tipo de programa (seleccione con "x")			
Pregrado	Especialización	Maestría	Doctorado
		X	

Nombre del programa académico
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
Nombres y apellidos del director del programa académico
ANGELA CRISTINA CARRILLO RAMOS
TRABAJO PARA OPTAR AL TÍTULO DE:
MAESTRO EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PREMIO O DISTINCIÓN *(En caso de ser LAUREADAS o tener una mención especial)*:

CIUDAD	**AÑO DE PRESENTACIÓN DE LA TESIS O DEL TRABAJO DE GRADO**	**NÚMERO DE PÁGINAS**
BOGOTA	2018	62

TIPO DE ILUSTRACIONES (seleccione con "x")						
Dibujos	Pinturas	Tablas, gráficos y diagramas	Planos	Mapas	Fotografías	Partituras
		X				

SOFTWARE REQUERIDO O ESPECIALIZADO PARA LA LECTURA DEL DOCUMENTO
Nota: En caso de que el software (programa especializado requerido) no se encuentre licenciado por la Universidad a través de la Biblioteca (previa consulta al estudiante), el texto de la Tesis o Trabajo de Grado quedará solamente en formato PDF.

MATERIAL ACOMPAÑANTE					
TIPO	**DURACIÓN (minutos)**	**CANTIDAD**	**FORMATO**		
			CD	DVD	Otro ¿Cuál?
Vídeo					
Audio					
Multimedia					
Producción electrónica					
Otro Cuál?					

DESCRIPTORES O PALABRAS CLAVE EN ESPAÑOL E INGLÉS

Son los términos que definen los temas que identifican el contenido. (*En caso de duda para designar estos descriptores, se recomienda consultar con la Sección de Desarrollo de Colecciones de la Biblioteca* Alfonso Borrero Cabal S.J *en el correo biblioteca@javeriana.edu.co, donde se les orientará).*

ESPAÑOL	**INGLÉS**
Procesamiento de Lenguaje Natural NLP	Natural Language Processing NLP
Minería de texto	Mining text

RESUMEN DEL CONTENIDO EN ESPAÑOL E INGLÉS

(Máximo 250 palabras - 1530 caracteres)

The Logical Framework Approach (LFA) is often used to formulate and evaluate projects in the public sector in Latin America. The LFA proposes that in order to evaluate a social problem, the relationship between its causes, consequence and central problem must be demonstrated through a mental map called the problem tree. However, by definition a social problem is transdisciplinary, systemic and has multiple interests, make it difficult to evaluate its semantic analysis. The research proposes a web service, which based on the Natural Language Processing NLP, calculates a metric of semantic similarity between the sentences of the problem tree. The semantic similarity is useful for designing projects. The research was carried out following the Desing Science Research Model and ended with the validation in a use case with the application of "Technology Acceptance Model (TAM)".

La Metodología Marco Lógico (MML) es con frecuencia la metodología utilizada para formular y evaluar proyectos de inversión en el sector público, en América Latina. La MML propone que para evaluar un problema social se debe demostrar la relación entre sus causas, consecuencia y problema central, a través de un mapa mental denominado árbol de problemas. Sin embargo, por definición un problema social contiene elementos transdiciplinarios, sistémicos y múltiples intereses que dificultad evaluar su análisis semántico. La presente investigación ilustra el desarrollo de un servicio web, que basado en el Procesamiento de Lenguaje Natural NLP, calcula una métrica de similitud semántica entre las oraciones del árbol de problemas. La cual es útil durante el diseño proyectos de inversión. La investigación se desarrolló siguiendo la metodología "Desing Science Research" y finalizó con la validación del servicio web en un escenario de uso con la aplicación de "Technology Acceptance Model (TAM)" a expertos del Departamento Nacional de Planeación.