PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

SISTEMA DE EXTRACCIÓN DE INFORMACIÓN BASADO EN ONTOLOGÍAS PARA COMENTARIOS DE UN FORO DE DISCUSIÓN EN LÍNEA EN EL DOMINIO DE CURSOS BRINDADOS POR UNA ENTIDAD DE EDUCACIÓN SUPERIOR

Tesis para optar el Título de **Ingeniero Informático,** que presenta el bachiller**:**

Willy Alexis Peña Vilca

ASESOR: Héctor Andrés Melgar Sasieta

Lima, febrero del 2015

RESUMEN

En la actualidad, las empresas necesitan estar informadas acerca de la opinión que tienen sus principales clientes respecto a los productos o servicios que ofrecen. Esto se debe a que estas dependen de esa información para poder tomar decisiones estratégicas al respecto. Para poder lograr esto, muchas de ellas optan por contratar servicios de empresas consultoras que realicen una encuesta tradicional o un focus group para poder obtener la información requerida; sin embargo, la parcialidad que guardan este tipo de estudios hace que en algunos casos se tomen decisiones estratégicas a partir de información no del todo fiable o representativa.

En base a lo mencionado anteriormente, el presente proyecto de fin de carrera brindará una herramienta para poder aprovechar la información contenida dentro de los comentarios hechos en foros de discusión en línea. Estas fuentes de conocimiento muchas veces no son procesadas para ningún fin; sin embargo, por medio de la herramienta propuesta se podrá extraer información relevante para ser utilizada como base de conocimiento por una empresa del rubro educativo. A partir de la cual se podrá contar con una alternativa confiable para obtener información sobre a la opinión directa de sus alumnos respecto a los cursos y profesores pertenecientes a la organización. Por último, dicha información podrá ser utilizada como base durante la toma de decisiones estratégicas de dicha organización educativa.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO: **SISTEMA DE EXTRACCIÓN DE INFORMACIÓN BASADO EN ONTOLOGÍAS PARA COMENTARIOS DE UN FORO DE DISCUSIÓN EN LÍNEA EN EL DOMINIO DE CURSOS BRINDADOS POR UNA ENTIDAD DE EDUCACIÓN SUPERIOR**

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Facultad de Ciencias e Ingeniería
06 ABR. 2015
MIGUEL MEJIA PUENTE
DECANO

ÁREA: CIENCIAS DE LA COMPUTACIÓN
PROPONENTE: Dr. Héctor Andrés Melgar Sasieta
ASESOR: Dr. Héctor Andrés Melgar Sasieta
ALUMNO: Willy Alexis Peña Vilca
CÓDIGO: 20072329
TEMA N°: 555
FECHA: San Miguel, 14 de diciembre de 2014

DESCRIPCIÓN

En la actualidad, las empresas necesitan estar informadas de la opinión que tienen sus principales clientes respecto a los productos o servicios que ofrecen con el objetivo de tomar decisiones estratégicas. Para lograr esto, las empresas generalmente optan por contratar servicios de empresas consultoras que realicen una encuesta tradicional o un *focus group* para poder obtener la información requerida; sin embargo, la parcialidad que guardan este tipo de estudios hace que en algunos casos se tomen decisiones a partir de información que no es representativa o no es fiable.

En este contexto, el presente proyecto de fin de carrera propone el desarrollo de una herramienta para poder aprovechar la información contenida dentro de los comentarios hechos en foros de discusión en línea. Estas fuentes de conocimiento muchas veces no son procesadas para ningún fin; sin embargo, por medio de la herramienta propuesta se podrá extraer información relevante para ser utilizada como base de conocimiento por una empresa del rubro educativo. A partir de la cual se podrá contar con una alternativa confiable para obtener información sobre a la opinión directa de sus alumnos respecto a los cursos y profesores pertenecientes a la organización. Por último, dicha información podrá ser utilizada como base durante la toma de decisiones estratégicas de dicha organización educativa.

OBJETIVO GENERAL

Desarrollar un sistema de extracción de información basado en ontologías que permita extraer información relevante respecto al dominio de cursos brindados por una entidad de educación superior, a partir de los comentarios hechos por alumnos en un foro de discusión en línea.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos del presente proyecto son:

- Diseñar un componente pre-procesador de texto que permita convertir los comentarios hechos en un foro de discusión en un formato que permita su procesamiento.

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

- Diseñar un componente de desambiguación lexical que permita resolver la ambigüedad de las palabras en un dominio específico.
- Diseñar un componente de extracción de información que permita obtener información relevante a partir de comentarios contenidos en un foro de discusión.

ALCANCE

El proyecto construirá una base de conocimientos a partir de datos no estructurados, por lo que se realizará bajo el enfoque de la Ingeniería del Conocimiento, para lo cual se desarrollará un análisis de correferencias con el fin de encontrar todas las menciones respecto a un dominio especificado a partir de la base de datos de comentarios del foro de discusión. Para lograr este objetivo, se desarrollará un sistema de extracción de información basado en ontologías para una organización del rubro educativo, bajo el dominio de cursos brindados por una entidad de educación superior. Para tal fin, se contará con una base de datos de comentarios en español de un foro de discusión en línea de alumnos de la Facultad de Estudios Generales Ciencias de la Pontificia Universidad Católica del Perú. Asimismo, se reusará y adaptará una ontología ya existente para guiar el proceso de extracción de información y para mostrar los resultados obtenidos por el sistema que realizará dicha tarea. En otras palabras, el proyecto no se enfocará en la construcción de una ontología nueva para su desarrollo; así como también, no clasificará los resultados de tal manera que se pueda identificar si un comentario es positivo o negativo.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

DEDICATORIA

A mis padres Willy y Ana María, quienes con su gran experiencia y lucidez han guíado mi camino desde el día en que nací, permitiendome ver y entender aquello que no alcanzo a ver.

A Glenda, maravillosa y brillante consejera, por su incondicional apoyo, su enorme paciencia y por compartir su alegría conmigo en cada momento.

AGRADECIMIENTOS

Al Dr. Andrés Melgar, por brindarme su asesoría y experiencia en el campo de investigación , enriqueciendo el contenido de este proyecto.

Al Mag. Cesar Aguilera y Mag. Johan Baldeon, por su orientación y mentoría profesional a lo largo de mi carrera universitaria.

A todos aquellos que en su momento me ayudaron con sus valiosos comentarios a mejorar este proyecto.

ÍNDICE

CAPÍTULO 1: Generalidades

1 Problemática

Las empresas que ofrecen servicios o productos para un determinado mercado necesitan conocer la opinión de los clientes respecto al servicio o producto que ofrecen, de forma tal que con esta información puedan tomar decisiones estratégicas. Para tal fin, muchas organizaciones optan por encuestas tradicionales para recopilar información. Por lo general, este tipo de estudio busca obtener la opinión del cliente a través de cuestionarios o *focus group*, los cuales se encuentran limitados a la cantidad de personas que eligen participar del estudio; así como también, pueden presentar cierta parcialidad, lo cual depende de cómo fueron redactadas las preguntas o la interpretación de las mismas (Kongthon, Angkawattanawit et al. 2010).

En la actualidad, la Internet ha permitido que emerjan otras opciones a las encuestas tradicionales (Bodendorf and Kaiser 2010). Las discusiones de las personas en la Web 2.0 son una fuente de información relevante para las empresas debido a la gran cantidad de datos que es generada a partir de fotos, textos y videos que son publicados cada día (Hongjiang, Dan et al. 2010). Un caso en particular son los foros de discusión en línea, los cuales han hecho posible que las personas puedan compartir sus opiniones libremente sobre un tema en específico. Esto se logra gracias a los comentarios hechos en las publicaciones que se encuentran en ellos; a partir de los cuales, información útil puede ser extraída para ser utilizada en diferentes propósitos (Castellanos, Hsu et al. 2012) .

Por esta razón, las empresas necesitan herramientas que faciliten la extracción y procesamiento de la información que brindan estas fuentes no tradicionales. En los últimos años, se han construido soluciones para satisfacer esta necesidad; sin embargo, gran cantidad de ellas se concentran en procesar data estructurada o semi-estructurada y no se enfocan en la información generada a partir de los foros de discusión en línea que en gran proporción generan información no estructurada (Shu, Wen-Jie et al. 2009). Para tal fin, existen sistemas de extracción de información basados en ontologías, los cuales realizan dicha tarea (Wimalasuriya and Dou 2010). De forma general, un sistema de este tipo se apoya en el proceso de extracción de información, el cual consiste en extraer automáticamente cierto tipo de información según plantillas predefinidas, a partir de un texto escrito en lenguaje natural. Para lograrlo, también hace uso de ontologías, las cuales brindan de manera formal y

explícita un esquema conceptual dentro de un dominio específico, lo cual permite representar el conocimiento que se busca extraer (Wimalasuriya and Dou 2010). En conclusión, este tipo de sistemas se diferencian de los sistemas tradicionales en los siguientes puntos: i) Procesar datos no estructurados y/o semi-estructurados contenidos en textos escritos usando lenguaje natural; ii) Presentar el resultado usando ontologías; y iii) Usar los procesos de extracción de información guiados por una ontología (Wimalasuriya and Dou 2010).

Aunque este tipo de sistemas presenten una confiable solución al problema no existen muchas herramientas creadas que utilicen esta tecnología, debido a que es un nuevo campo de estudio (ver capítulo 2, sección 2). La falta de soluciones de este tipo que se enfoquen en el problema de extracción de información relevante a partir de comentarios hechos en un foro de discusión en línea, se puede ver reflejada en cuatro puntos principales. En primer lugar, existe un desperdicio de la información contenida en los comentarios hechos en un foro de discusión en línea, ya que cada persona tiene una opinión formada, pensamientos negativos o positivos y otros tipos de información respecto a un tema, las cuales son compartidas dentro de los foros de discusión en línea (Manuel, Indukuri et al. 2010); sin embargo, una vez que se ha llenado el espacio disponible para almacenar la información contenida en los comentarios, se opta por crear un *back up* de los datos y resguardarlos sin ser procesados previamente (Yang, Ng et al. 2007). En segundo lugar, gran parte de las organizaciones optan por contratar los servicios de consultoras o utilizar sus propios recursos para poder conocer las opiniones de sus clientes por medio de encuestas tradicionales o grupos de enfoque; no obstante, un sistema de este tipo puede obtener dicha información a partir de foros de discusión en línea, en cualquier momento y presentarla de manera estructurada. (Kongthon, Angkawattanawit et al. 2010). En tercer lugar, las encuestas tradicionales dependen de muchos factores para poder ser totalmente fiables; sin embargo, la información contenida en los comentarios hechos en foros de discusión en línea presentan opiniones directas de las personas respecto a un producto, servicio o tema en particular, lo cual permite que sea una fuente directa de este tipo de información (Kongthon, Angkawattanawit et al. 2010). Por último, debido al gran tamaño y cantidad de comentarios que pueden existir en un foro de discusión en línea, la extracción de información de estas fuentes de manera manual es costosa y requiere mucho tiempo (Hariharan, Srimathi et al. 2010, Galvis Carreno and Winbladh 2013).

Por consiguiente, los sistemas de extracción de información basados en ontologías son aceptados, ya que tiene gran potencial para la solución del problema de extracción

de información contenida en los comentarios realizados en los foros de discusión en línea. Esto se debe a que realiza un procesamiento automático de la información contenida en textos escritos en lenguaje natural. Así como también, los sistemas de este tipo permiten crear metada de manera automática que servirá como datos de entrada para sistemas que quieran utilizar la información de la web semántica. Finalmente, este tipo de sistemas también puede ser usado para evaluar la calidad de una ontología; es decir, si una ontología puede ser utilizado exitosamente por un sistema de extracción de información de este tipo indicara que es una buena representación del dominio (Wimalasuriya and Dou 2010).

Por esta razón, el presente proyecto de fin de carrera busca brindar una alternativa de solución al problema de extracción de información contenida en los comentarios realizados en los foros de discusión en línea. En este caso en particular, se enfocará en una organización del rubro educativo; es decir, se buscará extraer información que se encuentren bajo el dominio de cursos brindados por una universidad. Esto se debe a que las universidades como organización les interesa saber cuál es la opinión de sus estudiantes respecto al servicio que brindan, con el fin de poder tener un mejor estudio del mercado en el cual se desarrollan (Soutar and Turner 2002). Asimismo, los datos extraídos serán representados mediante un modelo de conocimiento, lo que permitirá facilitar la comprensión de la información extraída para que pueda ser utilizada como base de conocimiento para la toma de decisiones estratégicas de la universidad. Para lograrlo se plantea el desarrollo de un sistema de extracción de información basado en ontologías que facilite el entendimiento de los textos escritos en lenguaje natural dentro de los comentarios.

1.1 Objetivo general

Desarrollar un sistema de extracción de información basado en ontologías que permita extraer información relevante respecto al dominio de cursos brindados por una universidad, a partir de los comentarios hechos por alumnos en un foro de discusión en línea.

1.2 Objetivos específicos

1. Diseñar un componente pre-procesador de texto que permita convertir los comentarios hechos en un foro de discusión en un formato que permita su procesamiento.

2. Diseñar un componente de desambiguación lexical que permita resolver la ambigüedad de las palabras en un dominio específico.
3. Diseñar un componente de extracción de información que permita obtener información relevante a partir de comentarios contenidos en un foro de discusión.

1.3 Resultados esperados

- Para Objetivo Específico 1:

  - ✓ Análisis cualitativo de la información de los comentarios del foro de discusión en línea.
  - ✓ Análisis cualitativo del formato de los comentarios del foro de discusión en línea.
  - ✓ Módulo de pre-procesamiento de los comentarios en lenguaje natural que permita normalizar el texto en los comentarios.

- Para Objetivo Específico 2:

  - ✓ Estructura para almacenar el conocimiento contenido en los comentarios.
  - ✓ Módulo de desambiguación lexical de palabras en lenguaje natural dentro de un dominio usando ontologías.

- Para Objetivo Específico 3:

  - ✓ Módulo de análisis de correferencias que permita identificar una misma entidad en distintas partes del comentario.
  - ✓ Análisis de los resultados de las métricas estadísticas de precisión, *recall* y valor-f aplicadas al sistema de extracción de información.

2 Herramientas, métodos, metodologías y procedimientos

En el cuadro 1 se presentan las herramientas utilizadas en cada uno de los resultados esperados del presente proyecto de fin de carrera:

Cuadro 1: Tabla de herramientas, métodos y procedimientos

Resultado Esperado	Herramienta a usarse
OE1-RE2: Análisis cualitativo de la información de los comentarios del foro de discusión en línea.	**MySQL** es un sistema de administración de datos relacional.
OE1-RE2: Análisis cualitativo del formato de los comentarios del foro de discusión en línea.	**CommonKADS** es una metodología apoyar la ingeniería del conocimiento estructurado.
OE1-RE3: Módulo de pre-procesamiento de los comentarios en lenguaje natural que permita normalizar el texto en los comentarios.	**Freeling** es una librería que provee servicios de análisis de lenguaje.
	Java es un lenguaje de programación orientado a objetos.
	Eclipse es un entorno de desarrollo de programación
	Apache Lucene es una librería que provee servicios para sistemas de recuperación de información.
OE2-RE1: Estructura para almacenar el conocimiento contenido en los comentarios.	**CommonKADS** es una metodología apoyar la ingeniería del conocimiento estructurado.
	OWL es un lenguaje de marcado que permite mostrar datos mediante ontologías.
	Protege es un software que permite modelar una ontología.
OE2-RE2: Módulo de desambiguación de palabras en lenguaje natural dentro de un dominio usando ontologías.	**Apache Jena** es un marco de trabajo que permite procesar ontologías.
	OWL es un lenguaje de marcado que permite mostrar datos mediante ontologías.
	Java es un lenguaje de programación orientado a objetos.
	Eclipse es un entorno de desarrollo de programación
	SparQL es un lenguaje de consulta a estructuras RDF
OE3-RE1: Módulo de análisis de correferencias que permita	**Análisis de correferencias** es un procedimiento que permite extraer todas las menciones a una entidad determinada en un texto.
	CommonKADS es una metodología apoyar la

identificar una misma entidad en distintas partes del comentario.	ingeniería del conocimiento estructurado.
	Java es un lenguaje de programación orientado a objetos.
	Eclipse es un entorno de desarrollo de programación
	Apache Jena es un marco de trabajo que permite procesar ontologías.
	Freeling es una librería que provee servicios de análisis de lenguaje.
	SparQL es un lenguaje de consulta a estructuras RDF.
OE3-RE2: Análisis de los resultados de las mediciones de precisión, *recall* y valor-f del sistema de extracción de información.	**Precisión y *Recall*** son unas métricas que permiten medir el rendimiento de un sistema de extracción de información.
	Valor-F es una métrica que permite establecer una relación entre la precisión y el *recall*.

2.1 Herramientas

2.1.1 Lenguaje de marcado OWL

OWL es un lenguaje de marcado utilizado para el modelar ontologías. Está diseñado para ser usado en aplicaciones que en lugar de únicamente representar información para los humanos, necesitan procesar el contenido de la información. OWL facilita un mejor mecanismo de interpretabilidad de contenido Web, proporcionando vocabulario adicional junto con una semántica formal (W3C 2012).

En el presente proyecto se decidió seleccionar esta herramienta debido a que facilita el manejo y edición de una ontología.

2.1.2 Lenguaje de programación Java

Java es un lenguaje de programación orientado a objetos. Su sintaxis está basada principalmente en C y C++ pero posee menos facilidades de bajo nivel a diferencias de ellos.

El objetivo principal de este lenguaje de programación es permitir a los desarrolladores implementar una aplicación que pueda ser ejecutada en cualquier dispositivo. En otras

palabras, el código que es ejecutado en una plataforma no tiene que ser recompilado para correr en otra (Oracle 2014).

La principal razón de su elección en el presente proyecto de fin de carrera se debe a que es un lenguaje de programación sencillo y tiene gran integración con las otras herramientas que se usarán.

2.1.3 Protégé

Protégé es un entorno de edición de ontologías con un soporte completo de OWL. Esta herramienta es compatible con la creación y edición de una o varias ontologías en un único espacio de trabajo a través de una interfaz de usuario completamente personalizable. Las herramientas de visualización permiten una navegación interactiva de las relaciones de la ontología (Standford 2014).

El principal motivo de la elección de esta herramienta para el desarrollo del presente proyecto es que servirá para poder adaptar la ontología que permitirá la realización de la extracción de información, ya que cuenta con un sólido soporte para el lenguaje de marcado OWL.

2.1.4 IDE Eclipse

Eclipse es un entorno de desarrollo realizado principalmente para el lenguaje de programación Java. Esta herramienta permite a los programadores escribir, compilar, depurar y ejecutar programas.

Su principal característica es la modularidad. Todas las funciones del IDE son provistas por módulos. Cada módulo ofrece una función bien definida como el soporte de Java o el sistema de control de versiones (Eclipse 2014).

El principal motivo para su elección en el presente proyecto de fin de carrera se debe a que es una herramienta que permite una gestión sencilla y practica de sus proyectos. Esto permite optimizar los tiempos de desarrollo del código y enfocarse en alcanzar el objetivo general.

2.1.5 MySql

MySql es un sistema administración de base de datos relacional, multiplataforma, mutihilo y multiusuario que cuenta con un tipo de licenciamiento dual; es decir, cuenta con una versión de código libre y otra de código privativo. Esta herramienta fue escrita en C y C++ y destaca por su gran adaptación a diferentes entornos de desarrollo,

permitiendo su interacción con los lenguajes de programación más utilizados y su integración con distintos sistemas operativos (Oracle 2014).

La principal razón por la que se ha seleccionado esta herramienta para el desarrollo del proyecto es que es rápida, segura y fácil de usar. Además, la base de datos de los comentarios del foro de discusión en línea con la que se cuenta para realizar la extracción de información se encuentra almacenada en este motor de base de datos.

2.1.6 Freeling

Freeling es una librería de código abierto que provee servicios de análisis de lenguaje para desarrolladores, ya que está diseñado para ser utilizado como una biblioteca externa de cualquier aplicación. En la actualidad, soporta el análisis para el idioma español (Mambo 2004).

La elección de esta herramienta se debe que los servicios que ofrece brindan soporte al idioma español, lo cual facilitará el procesamiento y análisis de los textos escritos en lenguaje natural que se encuentran dentro de los comentarios.

2.1.7 SparQL

SparQL es un lenguaje estandarizado de consulta que es orientado a datos donde solo se obtiene la información contenida dentro del modelo representado con el marco de descripción de recursos (RDF). El RDF es un método para descomponer el conocimiento en piezas pequeñas con algunas reglas acerca de la semántica de dichas piezas (Apache 2014).

La principal razón por la que se ha seleccionado esta herramienta es para recuperar términos dentro de la ontología, los cuales se encuentran representados bajo modelos en RDF. De tal forma, dichos términos puedan ser analizados y guíen el proceso de extracción de información.

2.1.8 Apache Jena

Apache Jena es una librería de código libre para Java que permite extraer datos y escribirlos en gráficos RDF. Los gráficos son representados como un modelo abstracto, el cual puede tener como fuente de datos un archivo, una base de datos, url's o una combinación de las mismas (Apache 2014).

A diferencia de otros software de este tipo, Jena provee soporte a OWL esto facilitará del manejo de ontologías a lo largo del proyecto; por esta razón ha sido considerada en esta lista de herramientas.

2.1.9 Apache Lucene

Apache Lucene es una librería de código abierto creada para ser realizar tareas de recuperación de información. Esta herramienta permite realizar indexado y búsqueda de texto completo por lo que ha sido ampliamente usado para la implementación de motores de búsqueda (Apache 2014).

La elección de esta herramienta se debe a que provee los componentes necesarios para poder realizar algunos de los mecanismos para el pre-procesamiento de los comentarios (ver capítulo 3, sección 4).

2.2 Métodos y Procedimientos

2.2.1 Análisis de correferencias

La extracción de información consiste en extraer de manera automática la información contenida en un texto escrito en lenguaje natural (Wimalasuriya and Dou 2010). Para esto, una de las tareas que se realiza es el análisis de correferencias que permite conocer todas las relaciones que existen de una misma entidad dentro de un texto, es decir, permite reconocer todas las frases que hablan de una misma entidad del mundo real. En este contexto, estas frases son llamadas menciones, las cuales pueden ser de los siguientes tipos (Cunningham 1997):

- Nombradas: Prof. Chávez, Jorge Chávez.
- Nominales: El señor del abrigo gris.
- Pronominales: él.

La importancia de este procedimiento radica en que permite la asociación de información descriptiva que se encuentra por medio de referencias a las entidades principales encontradas. Esto permite encontrar y extraer mayor cantidad información relevante acerca de la entidad que se tiene como objetivo.

La elección de este procedimiento para el desarrollo del presente proyecto se debe a que permitirá tener un mayor alcance al momento de extraer la información. Con el fin de rescatar información relevante a partir de los textos que procesará el sistema.

2.2.2 Precisión, Recall y Valor-F

En primero lugar, la precisión y *recall* son métricas estadísticas que permiten medir el rendimiento de un sistema de extracción de información. Para este fin, se denomina precisión a las instancias recuperadas que son relevantes y *recall* instancias relevantes recuperadas. Por ejemplo, si se tienen en un comentario 15 menciones a un determinado servicio que brinda una organización y el sistema solo recupera 7 menciones de los cuales solo 3 eran de la empresa en cuestión, se tendrían los siguientes resultados:

- Precisión: 3/7
- *Recall*: 3/15

Por otro lado, el valor-F es un método que combina las métricas de precisión y *recall*, la cual se define de la siguiente manera:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Imagen 1: Relación entre precisión y *recall* (Zechner 1997).

Donde ß es el parámetro usado para representar el peso de la precisión sobre el de *recall* o viceversa.

La principal razón para la utilización de estos métodos es poder contar con un indicador que muestre la calidad del sistema de extracción de información que se desarrollará. De esta manera, se podrá asegurar que se ha podido cumplir con los objetivos contando con una calidad adecuada en el sistema.

2.3 Metodologías

2.3.1 CommonKADS

Es una metodología que apoya a la ingeniería del conocimiento estructurado, mediante un conjunto de modelos de ingeniería construidas con la organización y aplicación en mente; es decir, son elaboradas a partir del conocimiento humano (Schreiber, Wielinga et al. 1994). Esta estrategia será utilizada para todo el sistema, en especial durante el análisis de patrones en el dominio. Dicho enfoque es preferido cuando se trata de generar reglas para el lenguaje natural donde se deben determinar

el léxico que se usará y los lineamientos que se tienen al escribir (Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

En el presente proyecto de fin de carrera se ha decidido realizar la extracción de información en base al enfoque de la ingeniería del conocimiento. La razón principal de dicha elección es que se busca aplicar ontologías para extraer la información requerida por lo que esta metodología aportará de manera significativa a los objetivos del proyecto.

3 Alcance

El proyecto construirá una base de conocimientos a partir de datos no estructurados, por lo que se realizará bajo el enfoque de la ingeniería del conocimiento, para lo cual se desarrollará un análisis de correferencias con el fin de encontrar todas las menciones respecto a un dominio especificado a partir de la base de datos de comentarios del foro de discusión. Para lograr este objetivo, se desarrollará un sistema de extracción de información basado en ontologías para una organización del rubro educativo, bajo el dominio de cursos brindados por una universidad. Para tal fin, se contará con una base de datos de comentarios en español de un foro de discusión en línea de alumnos de la Facultad de Estudios Generales Ciencias de la Pontificia Universidad Católica del Perú. Asimismo, se reusará y adaptará una ontología ya existente para guiar el proceso de extracción de información y para mostrar los resultados obtenidos por el sistema que realizará dicha tarea. En otras palabras, el proyecto no se enfocará en la construcción de una ontología nueva para su desarrollo; así como también, no clasificará los resultados de tal manera que se pueda identificar si un comentario es positivo o negativo

4 Justificación

En la actualidad, las empresas necesitan estar informadas acerca de la opinión que tienen sus principales clientes respecto a los productos o servicios que ofrecen. Esto se debe a que estas dependen de esa información para poder tomar decisiones estratégicas al respecto. Para poder lograr esto, muchas de ellas optan por contratar servicios de empresas consultoras que realicen una encuesta tradicional o un *focus group* para poder obtener la información requerida; sin embargo, la parcialidad que guardan este tipo de estudios hace que en algunos casos se tomen decisiones estratégicas a partir de información no del todo fiable o representativa.

En base a lo mencionado anteriormente, el presente proyecto de fin de carrera brindará una herramienta para poder aprovechar la información contenida dentro de los comentarios hechos en foros de discusión en línea. Estas fuentes de conocimiento muchas veces no son procesadas para ningún fin; sin embargo, por medio de la herramienta propuesta se podrá extraer información relevante para ser utilizada como base de conocimiento por una empresa del rubro educativo. A partir de la cual se podrá contar con una alternativa confiable para obtener información sobre a la opinión directa de sus alumnos respecto a los cursos y profesores pertenecientes a la organización. Por último, dicha información podrá ser utilizada como base durante la toma de decisiones estratégicas de dicha organización educativa.

CAPÍTULO 2: Marco conceptual y Estado del arte

1 Marco conceptual

En esta sección se detallarán los conceptos necesarios para entender el presente proyecto de fin de carrera. Esto comprenderá el problema respecto a la extracción de información de datos no estructurados generados a partir de comentarios hechos en un foro de discusión en línea; así como también, los principales conceptos que apoyarán el entendimiento de la extracción de información basada en ontologías como solución al problema propuesto. Para esto se busca responder las siguientes preguntas:

- ¿Qué es extracción de información?
- ¿Cuáles son las tareas de la extracción de información?
- ¿Qué tipos de metodologías pueden ser usadas para construir sistemas de extracción de información?
- ¿Cuáles son los principales componentes de un sistema de extracción de información?
- ¿Cuáles son las principales métricas que se usan para evaluar sistemas de extracción de información?
- ¿Qué desempeño consiguen los sistemas de extracción de información?
- ¿Qué es la extracción de información basada en ontologías?, Especialmente:
    - ¿Qué es una ontología? ¿Cómo son usadas las ontologías en los sistemas de extracción de información basados en ellas?

1.1 Extracción de información

La extracción de información es un proceso que a partir de textos logra extraer automáticamente datos ordenados y poco ambiguos, respecto a un campo en específico, lo cual incluye entidades, relaciones y eventos de los mismos (Cunningham 1997, Abolhassani, Fuhr et al. 2003).

1.2 Tareas de la extracción de información

Existen cinco tareas de la extracción de información las cuales se detallan a continuación (Cunningham 1997):

1.2.1 Reconocimiento de entidades

Consiste en identificar todos los nombres de personas, lugares, organizaciones, fechas y montos de dinero (Cunningham 1997). Por ejemplo, según el siguiente comentario respecto a quién es el mejor profesor de una facultad universitaria; extraído de un foro de discusión en línea:

> *"Para mí el mejor profesor es Rubén Agapito, no regala nota ni te la quita, es justo, enseña todo lo que se debe aprender, te proyecta al futuro y te dice todo lo que la matemática te va a servir en otros cursos, da su buen descanso en medio tiempo, envía problemas propuestos, envía prácticas y exámenes de ciclos pasados con solucionario, siempre después de cada practica o examen da el solucionario. Yo he llevado Calculo 1 y Calculo 4 con ese profesor y en los dos casos me gustaba como enseñaba aprendí mucho"* (Clasesmas 2014).

Se puede encontrar los siguientes nombres de entidades: Rubén Agapito, Calculo 1 y Calculo 4. A partir de este resultado, se puede deducir que este proceso depende del dominio, especialmente cuando los textos que se procesarán contienen gran cantidad de lenguaje informal. En otras palabras, si los comentarios que tratáramos de analizar fueran respecto a opiniones sobre un producto de belleza involucraría algunos ajustes en el sistema respecto al dominio para el cual ha sido construido.

1.2.2 Resolución de correferencias

Esta tarea involucra identificar relaciones entre las entidades encontradas en la tarea anterior. De esta manera siguiendo con el ejemplo mencionado anteriormente; una correlación encontrada en el texto sería entre *"Rubén Agapito"* y *"ese profesor"* (Cunningham 1997).

Aunque de cara a los usuarios esta tarea no aporta tanto como las otras en un proceso de extracción de información; es muy útil al momento de intentar desarrollar un sistema de este tipo. La mayor importancia radica en que es una pieza clave para las siguientes tareas, ya que permite la asociación de información descriptiva que se encuentra por medio de referencias a las entidades principales encontradas.

1.2.3 Construcción de plantilla de elementos

La presente tarea se basa en el reconocimiento de nombres de entidades y en la resolución de correferencias obtenido a partir de las tareas adicionales. A partir de los cuales busca información descriptiva para las entidades encontradas. Por ejemplo, siguiendo con el caso antes mencionado, para la entidad encontrada *"Rubén Agapito"* y a su correferente *"ese profesor"* se obtiene que Rubén Agapito es un profesor como información descriptiva de la entidad. (Cunningham 1997).

1.2.4 Construcción de plantilla de relaciones

Esta tarea requiere la identificación de un número pequeño de posibles relaciones entre las plantilla de elementos identificadas en la tarea anterior. Por ejemplo, en el caso planteado se obtienen las relaciones entre las entidades *"Rubén Agapito", "Cálculo 1"* y *"Cálculo 2",* donde cada una de ellas cuentan con propiedades particulares que se relacionan entre sí; es decir, se obtiene que Rubén Agapito es profesor de Cálculo 1 y Cálculo 2. Por esta razón, la construcción de plantilla de relaciones brinda una de las principales características para cualquier tipo de extracción de información (Cunningham 1997).

1.2.5 Producción de plantilla de escenarios

Las plantillas de escenarios muestran prototipos de los resultados obtenidos por un sistema de extracción de información. Ellos logran mostrar las plantillas de elementos como entidades con descripciones de eventos y relaciones. Por ejemplo, podrá identificar como Pedro Pérez está cursando el segundo ciclo de la carrera de ingeniería informática y se encuentra llevando el curso de Calculo 2 con el profesor Rubén Agapito.

Esta tarea depende del dominio y se encuentra relacionado al escenario de interés de los usuarios. Esto permite evitar cualquier error en el desarrollo de esta tarea, dejando de lado algunas de las ocurrencias en escenarios importantes (Cunningham 1997).

1.3 Principales componentes de un sistema de extracción de información

Los principales componentes de un sistema de extracción de información son los siguientes (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006):

1.3.1 Zonificador de Texto

Este componente transforma los textos en segmentos de texto. Al menos este componente separa las regiones formateadas de las no formateadas; incluso, algunos sistemas pueden llegar a segmentar los textos no formateados en temas específicos a través de frases en el texto o significados estadísticos (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

1.3.2 Preprocesador

Toma los textos como una secuencia de caracteres y logra localizar los enlaces de las oraciones y produce una secuencia de elementos léxicos. Dichos elementos, son palabras juntas con atributos léxicos. Este componente, determina las partes del texto para cada palabra y escoge una parte para trabajar (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

1.3.3 Filtro

Este componente filtra las oraciones que son poco importantes, reduciendo de esta manera la longitud del texto original para que pueda ser procesado mucho más rápido (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

1.3.4 Pre analizador

Gran parte de los sistemas recientes buscan incluir este módulo ya que permite reconocer estructuras de menor escala que son muy comunes en textos en lenguaje natural y puede ser reconocida fácilmente. Por ejemplo: preposiciones, complementos, entre otros (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

1.3.5 Analizador

Este componente toma los elementos lexicales y frases para intentar producir un árbol analizado de la oración completa.

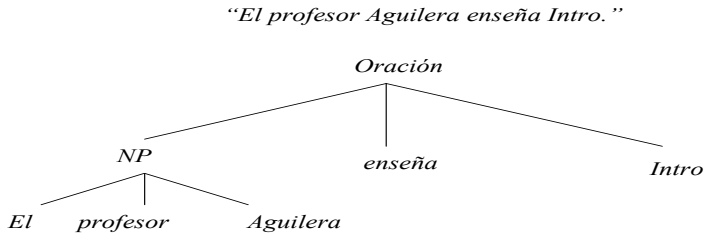

Imagen 2: Ejemplo estructura de arbol producido por el analizador

Como e puede observar en la imagen 2, la oración es transformada en un árbol donde los se detectan ciertos grupos de palabras para tener distintos fragmentos que puedan ser procesados. Recientemente, más sistemas están abandonando el análisis de la oración completa en aplicaciones de extracción de información, ya que vocabulario incluido es limitado. (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006).

1.3.6 Combinador de fragmentos

Convierte los grupos de árboles analizados en una sola forma lógica para toda la oración. Por ejemplo, un método usado para este fin es agrupar todos los fragmentos encontrados por medio de conjunciones para formar una sola forma lógica (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006). Por ejemplo:

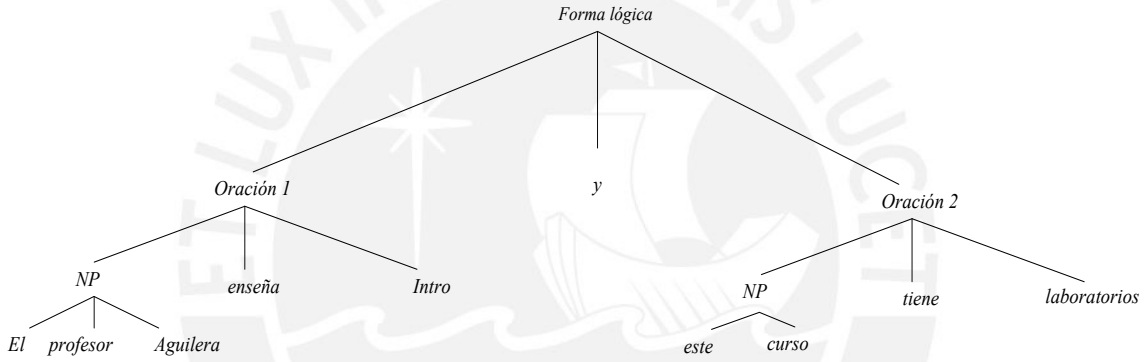

Imagen 3: Ejemplo estructura de forma lógica producida por el combinador de fragmentos

Como se puede observar en la imagen 3, los grupos de árboles originados a partir de las oraciones son relacionados mediante conjunciones de tal forma que se vuelven una sola forma lógica.

1.3.7 Interpretador semántico

Traduce la forma lógica producido en una estructura semántica; es decir, básicamente representa las acciones y las relaciones que se encuentran implícitas dentro de las oraciones (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006). Según los diagramas anteriormente mostrados, este componente busca relaciones entre los miembros de la forma lógica según el significado de las palabras que contiene; es decir, en este caso detectará que *"el curso"* cuenta con *"laboratorios"*.

1.3.8 Desambiguación lexical

A través de reglas desarrolladas manualmente, logra resolver ambigüedades que se encuentran en una estructura semántica (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006). Por ejemplo, según el caso mostrado anteriormente *"Intro"* dentro del texto puede referirse a los cursos *"Introducción a la computación"* o *"Introducción a la Ingeniería Informática"*; sin embargo, a partir de las reglas creadas por el contexto se deduce que *"Intro"* se refiere al curso *"Introducción a la computación"*, debido a que este curso lo dicta el profesor *"Aguilera"* y cuenta con *"laboratorios".*

1.3.9 Resolución de Correferencias

Este componente convierte las estructuras semánticas construidas que se encuentran en nodos separados en una sola entidad; en otras palabras, logra que se conviertan en estructuras interrelacionadas (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006). Por ejemplo, según el caso mostrado anteriormente, las entidades identificadas como *"este curso"* y *"Intro"* son correferentes debido a que hablan de una misma entidad que este caso es el curso *"Introducción a la computación".*

1.3.10 Generador de plantillas

Toma las estructuras semánticas generadas por el procesamiento del lenguaje natural y produce plantillas oficiales según las reglas establecidas. En otros términos, eventos identificados como poco relevantes son dejados de lados; así como también, las comas son removidas, los porcentajes son redondeados, entre otros; para de esta manera obtener el resultado final esperado, el cual dependerá de las reglas establecidas (Hobbs 1993, Abolhassani, Fuhr et al. 2003, Turmo, Ageno et al. 2006). Por ejemplo, para el curso *"Introducción a la computación"* se tiene como profesor a *"Aguilera"* y el curso cuenta con *"laboratorios"*; de tal forma que para otros cursos encontrados en otros textos se buscará completar dicha plantilla con los datos extraídos.

1.4 Principales métricas para evaluar sistemas de extracción de información

En términos de calidad de los sistemas de extracción de información, los siguientes puntos deben ser considerados (Zechner 1997, Turmo, Ageno et al. 2006):

- **Interfaz de usuario:** Evalúa que tan fácil es el uso del sistema y qué tan rápido puede llegar a ser la curva de aprendizaje.

- **Velocidad del sistema:** Tiempo de espera para que el sistema muestre un resultado al usuario.

- ***Recall* del sistema:** El resultado obtenido contiene elementos relevantes para el usuario.

- **Precisión del sistema:** El resultado obtenido es relevante según lo que el usuario intentaba buscar.

Respecto a la precisión y *recall* existe una combinación para comparar estos dos parámetros llamada valor-f (ver capítulo 1, sección 2.2.2).

1.5 Desempeño de los sistemas de extracción de información

Cada una de las tareas del proceso de extracción de información ha sido rigurosamente evaluada por las diferentes ediciones del *Message Understanding Conference* (MUC). A partir de dichas evaluaciones, se ha logrado determinar los siguientes factores que afectan el desempeño de cada una de las tareas de extracción (Cunningham 1997):

- Tipo de texto: Las clases de texto con los que se está trabajando. Por ejemplo, en el presente proyecto de fin de carrera se trata de comentarios hecho en un foro de discusión. Este factor determinará el nivel de aplicación las reglas establecidas, ya que no será lo mismo aplicar dichas reglas a textos provenientes de un portal de noticias, debido a que los formatos de los textos son diferentes.

- Dominio: El tema del cual hablan dichos textos y la forma en que son escritos. Por ejemplo, los comentarios que se analizarán se encuentran escritos en lenguaje informal y están dentro del dominio de servicios que brinda una universidad a sus alumnos; específicamente los cursos y profesores de la organización educativa. Este factor permitirá identificar de manera general cual es el tema que se aborda dentro de los textos, permitiendo establecer las reglas para dicho dominio, las cuales deberán ser cambiadas en caso se cambiará el mismo.

- Escenario: Los eventos en los que el usuario está interesado. Por ejemplo, los mejores profesores de una facultad, la mejor cafetería en una universidad, entre otras. Este factor permitirá ser más preciso al momento de obtener información e influirá directamente con el objetivo del sistema de extracción de información; es decir, determinará cuál es la información que se considera relevante.

Estos tres factores hacen que cada sistema de extracción de información sea único; es decir, que el desempeño de cada sistema será predecible solo para la conjunción de dichos factores.

1.6 Extracción de información basada en ontologías

En primer lugar, una ontología es una conceptualización de un dominio de conocimiento donde se describe conceptos y relaciones, las que son comprensibles tanto para los humanos como para las máquinas (Gruber 1995). Esta conceptualización se encuentra representada por los siguientes componentes: clases, propiedades, tipos de datos, objetos, valores de las propiedades de los objetos y las relaciones.

Por otro lado, un sistema de extracción de información basado en ontologías permite procesar datos no estructurados y semi-estructurados a partir de texto que se encuentran en lenguaje natural a través de mecanismos guiados por ontologías. Adicionalmente, permite presentar los resultados mediante ontologías lo que facilita el entendimiento de la información extraída (Wimalasuriya and Dou 2010).

En un sistema de este tipo, existen dos enfoques para la utilización de las ontologías. En primer lugar, uno de ellos es cuando se usan como datos de entradas del sistema, lo cual genera que la ontología sea construida de forma manual para que pueda ser usada. Por otro lado, existe otro enfoque que se centra en la construcción de una ontología como parte de las tareas realizadas por el sistema (Wimalasuriya and Dou 2010).

1.6.1 Arquitectura general

En la imagen 4, se puede observar que inicialmente el usuario realiza una consulta a la base de conocimientos para obtener la información resultante de la extracción. Para obtener dicha base, previamente se han ejecutado los siguientes mecanismos de extracción de información:

- **Preprocesador:** Los textos que contienen información relevante son enviados a este componente para que pueda convertirlos a un formato que el módulo de extracción pueda procesar.

- **Ontología, editor de ontologías y generador de ontologías:** Estos tres componentes son utilizados como datos de entrada para el procesador; así como también, directamente para el módulo de extracción de información. De esta manera, se puede observar ambos enfoques de la utilización de ontologías, ya sea como dato de entrada directo al módulo principal guiando de esta manera el proceso de extracción de información o la generación de la ontología como como una tarea más del sistema de extracción de información a partir de los textos de entrada.

- **Léxico semántico:** Este componente contiene el origen y formas de las palabras según el idioma que se está tratando en el sistema. Para esto se basa en un análisis en los textos de entrada del sistema.

- **Módulo de extracción de información:** Este componente recibe los textos pre-procesados, el léxico semántico definido y la ontología para realizar la tarea de extracción de información.

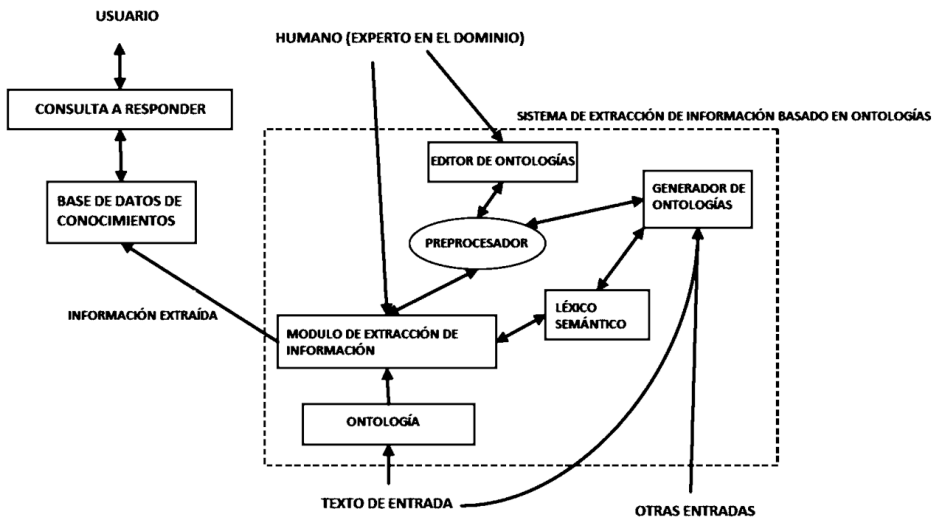

Imagen 4: Arquitectura general de un sistema de extracción de información basado en ontologías (Wimalasuriya and Dou 2010).

1.7 Conclusiones sobre el marco conceptual

En base a los conceptos presentados se puede afirmar que no existirá una solución única al problema, debido a la existencia de una gran variedad de escenarios y la extensión de metodologías aplicables a cada una de ellas. Sin embargo, basados en el potencial que guardan las ontologías para los sistemas de extracción de información y a que se encuentran predispuestas para la aplicación en la web semántica; se puede deducir que el sistema de extracción de información basado en ontologías, permitirá obtener datos relevantes a partir de comentarios hechos en un foro de discusión en línea. Para este fin, en el presente proyecto se ha optado por tomar como referencia el enfoque, en el cual la ontología es tomada como un dato de entrada del sistema de extracción de información. De tal forma, permitirá guiar y obtener información respecto a los profesores y cursos de una organización educativa.

2 Estado del arte

2.1 Introducción

El presente estado del arte se realizó en base a una revisión sistemática. Esto se debe a la ventaja que ofrece el método para sintetizar las investigaciones realizadas, minimizando la parcialidad gracias al proceso estructurado que sigue (Kitchenham 2004).

El objetivo de la presente revisión es recuperar y extender el conocimiento acumulado acerca de la extracción de información en comentarios realizados en un foro de discusión en línea. Para tal fin, se planteó la siguiente pregunta de investigación: ¿Cuáles son las iniciativas que han sido llevadas a cabo para la extracción de información a partir de datos no estructurados y que al mismo tiempo han sido extraídos a partir de los comentarios realizados en foros de discusión u otras páginas web?

2.2 Método usado en la revisión del estado del arte

En base a la pregunta planteada, se elaboró una lista de cadenas que fueron usadas para la búsqueda de fuentes primarias. Esta está conformada por los siguientes elementos:

Cuadro 2: Cadenas generales básicas de búsqueda

	Cadenas generales básicas de búsqueda
1	"extracción" Y ("datos" O "información") Y ("foros de discusión" O "página web") Y ("comentarios" O "opiniones" O "texto")
2	"procesamiento" Y ("datos no estructurados" O "información" O "lenguaje natural") Y ("foros de discusión" O "página web") Y ("comentarios" O "opiniones" O "texto")

En la presente revisión sistemática, estas cadenas fueron usadas directamente en páginas de las librerías digitales de IEEE, ACM, Scopus, Science Direct y Springer. Una vez encontrada una gran cantidad de estudios primarios relacionados con el tema; el análisis de su inclusión dentro del estudio estuvo basado en el título, el resumen y palabras claves de los artículos. Además, los objetivos principales de estos artículos debían ser la extracción de información a través de la comprensión de la misma.

Por último, se encontraron muchos artículos sobre el procesamiento y la extracción de información a partir de datos no estructurado, sin embargo algunos se centraban en la elaboración de algoritmos para la tarea mencionada; por lo que, no aportaban de manera correcta al proyecto, ya que no se trata de plantear un método que permita manejar este tipo de datos, sino que se busca establecer un marco de trabajo que permita realizar el entendimiento de los datos no estructurados para luego puedan ser utilizados para diferentes fines.

2.3 Estudio N°1: Extracción de información de foros basado en expresiones regulares

Este proyecto tiene como objetivo analizar un grupo de foros analizando cada una de sus caracteristas particulares, para poder obtener un sistema que identifique otros foros del mismo tipo y así extraer información de los mismos. Para lograr este propósito, utiliza expresiones regulares.

La idea principal del método es utilizar una muesta de páginas web de algunos tipos de foros escogidos para el estudio, como Discuz!, Phpwind, entre otros; y analizar los diferentes campos de los paquetes de protocolos HTTP capturados para extraer las características de cada uno de los foros pertenecientes a los grupos determinados. A partir de esta muestra, se generá un conjunto de reglas y patrones que permitirán extraer información de otros foros que pertenezcan al mismo tipo.

Aunque el enfoque utilizado brinda una forma efectiva de extraer información de un foro de discusión en línea, debido a que se enfoca en solo algunos tipos de foros específicos y solo se toma una pequeña muestra de cada uno de ellos, todavía existen algunos tipos; como vBulletin, phpBB, entre otros; que no han sido utilizados dentro del estudio y que tienen caracteristicas particulares que no podran ser identificadas por el sistema planteado (Gang, Yingwei et al. 2013).

2.4 Estudio N°2: Extracción de características de un API desde un foro

Este proyecto busca extraer información de los foros que abordan temas sobre librerías de software y API's para desarrolladores. Para tal fin, se busca utilizar técnicas de procesamiento del lenguaje natural y análisis de los sentimientos para extraer los problemas respecto a los API's de manera automática.

Para lograr este objetivo, han elaborado una herramienta llamada Haystack que sigue los hilos de todos los sentimientos negativos de los usuarios en las discusiones, elaborando de esta manera un único texto que les sirve como dato de entrada para su proceso de extracción. A partir de esto se analiza y se extraen las principales características del API que se intenta investigar.

El proceso seguido por este estudio requiere una transformación previa de la información, lo cual genera que su aplicación se encuentre limitada a una cierta cantidad de foros de dicusión online; asi como también, a herramientas de código libre utilizadas para la identificación de la orientación de los sentimientos y el análisis de las oraciones; como lo son Sentiment140 API y Stanford CoreNLP respectivamente (Yingying and Daqing 2013).

2.5 Estudio N°3: Navegación en foros de discusión sobre salud usando extracción relacional y ontologías médicas

Este proyecto se enfoca en foros de discusión que abordan temas de la medicina y la salud. Esto permite diseñar unos métodos de extracción basados en ontologías que son más exactos y poderosos debido al dominio del conocimiento existente de los investigadores.

Para la implementación de este sistema se desarrollaron tres métodos de extracción basados en ontologías diferentes y se aplicaron en los foros de discusión para extraer relaciones entre entidades médicas.

Cada uno de los métodos fue evaluado de manera manual y presentaron un desempeño aceptable para los datos de entrada considerados; sin embargo, solo se tomaron en cuenta términos médicos de una sola palabra, es decir otros términos más complejos fueron dejados de lado. A pesar de ello, el método aplicado para el reconocimiento de patrones y la matriz de correferencia pueden ser adaptados para también aceptar dichos términos (Mohajeri, Esteki et al. 2013).

2.6 Estudio N°4: Un enfoque integrado para la extracción de información

En este proyecto, se utiliza un método que utiliza el algoritmo de alineación de árboles y el método de transferencia de aprendizaje, los cuales son propuestos para generar las capas de páginas web de foros de discusión en línea, blogs y páginas de noticas. Para tal fin se consideran los siguientes puntos:

- El algoritmo de alineación de árboles es utilizado para encontrar la estructura más similar de las páginas web usadas. Luego, se generan regresiones lineales para darle un peso a dichas comparaciones. Basado en esto, se mezcla los árboles en uno solo que almacena información estadística obtenida de múltiples páginas web.

- El método de transferencia de aprendizaje es utilizado para obtener el contenido de bloque más parecido y aplicar el algoritmo mencionado anteriormente para encontrar patrones repetidos en el árbol encontrado.

- Por último el método de generación de capas mide la similitud entre las capas y las páginas web; en caso se encontrará algún cambio drástico en aquella similitud se genera una nueva capa según datos estadísticos.

-

Los resultados experimentales muestran que el método tiene un 93% de precisión y un 96% *recall* aplicado a foros de discusión (YingJu, YuHang et al. 2011).

2.7 Estudio N°5: La extracción de datos de los foros web basados en la similitud de diseño de página

En este proyecto se resuelven los problemas de extracción de datos de foros de discusión siguiendo dos pasos. En el primero de ellos se utiliza la estructura de la página web para identificar el tema del foro de discusión. En el segundo se realiza la

extracción de metadatos utilizando la regularidad estadística de los metadatos. Este método propuesto se realiza sin intervención manual, realizando la extracción en dos fases permite obtener un 98% de precisión y 97% de *recall* (Yun, Bicheng et al. 2009).

2.8 Estudio N°6: Extractor automático de datos de foros de discusión en línea

En este proyecto se presenta un enfoque para la extracción automática de los datos del autor, la fecha y el contenido de una publicación realizada en un foro de discusión. Los resultados del experimento muestran que el enfoque es efectivo; sin embargo, no puede ser utilizado en páginas web que generan dinámicamente datos a través de JavaScript.

Es importante mencionar que este estudio no realiza la tarea de extracción información de las entidades de los datos de entrada mencionados (Suke, Liyong et al. 2009).

2.9 Estudio N°7: Marco de trabajo orientado hacia el contenido para el análisis de discusión en línea

En este documento se propone un nuevo marco de trabajo que plantea modelar discusiones en línea a través de grafos basados en mensajes. Esto permite que la extracción sea orientada al contenido, así como la identificación de las partes de la discusión que son más interesantes. Adicionalmente, facilita la clasificación de la discusión desde el punto de vista de los temas que están siendo discutidos.

Para lograr este objetivo, los grafos son utilizados para representar entidades y las relaciones entre ellas. En primer lugar, las entidades encontradas fueron dos: los usuarios que participaban en las discusiones y los mensajes o respuestas a un mensaje ya publicado. Por otro lado, las relaciones entre mensajes intercambiados y los usuarios eran las respuestas que estos últimos realizaban a los primeros.

Este tipo de enfoque también identifica cuál es el mensaje más popular; es decir, el que ha causado muchas reacciones y permite la concentración del análisis en el contenido más que en los participantes de la discusión. En otras palabras, un mensaje que ha recibido muchas respuestas es más atractivo a otro que no ha recibido respuesta alguna (Stavrianou, Chauchat et al. 2009).

2.10 Estudio N°8: Foro de extracción de datos sin reglas explícitas

En este proyecto se discuten las características de los datos de un foro y se analiza la extracción de estos en diferentes foros. Para este fin, se propone un método efectivo basado en la redundancia de la información explotando la estructura para extraer de manera automática la información del usuario y los contenidos generados por el mismo desde diferentes foros de discusión.

En primer lugar, se resumió las características de los foros:

- Todas las áreas de usuario utilizan una plantilla similar.

- Todas las áreas de usuario incluyen dos partes: una parte estática y una parte dinámica. La primera describe información del usuario y la segunda almacena el contenido generado por el usuario.

Por esta razón, como se mencionó en un primer momento, este marco de trabajo consta de dos fases:

- Extracción de la información del usuario

- Extracción del contenido generado por el usuario

Para la primera fase se hace uso de un algoritmo, el cual en una primera etapa, a través de unas páginas de publicaciones transformadas en árboles DOM, construye los nodos del texto como etiquetas DOM que son consideradas como términos, luego todas las áreas de usuarios válidas son usadas para generar un nuevo árbol. En la segunda etapa de este algoritmo, se organiza el contenido de todas las áreas de usuario y a partir de ellos son extraídos los principales atributos del usuario.

Por último, para la segunda fase, se utiliza el método de computación combinada, donde se usan los datos de usuario extraídos para localizar sus publicaciones (Jingwei, Cheqing et al. 2012).

2.11 Estudio N°9: Extracción de los textos de discusión web para el análisis de opinión

Este documento intenta resolver el problema de extracción de información de comentarios hechos en un foro de discusión. Para esto utiliza la técnica de alineación parcial de un árbol que es basada en el supuesto de que una página web contiene

notas estructuradas para los objetos; es decir, contiene una estructura de HTML estándar.

En general, la estructura de una discusión dentro de un foro es generalmente creada con etiquetas de HTML como TABLE o DIV. En el caso del uso de una tabla, cada contribución es representada como filas de la misma. Dentro de esas filas se encuentra datos relevantes y datos que funcionan como ruido que no deben ser considerados.

Por esta razón, para la extracción el proceso se dividió en dos pasos:

- Devolución de bloques con textos.
- Limpieza de estos bloques de ruido.

Para realizar esta tarea se aplicó un algoritmo, a partir del cual se obtuvo un 77% de *recall* (Machova and Penzes 2012).

2.12 Estudio N°10: Encontrar respuestas relevantes en los foros de software

En este documento, se presenta un enfoque para encontrar respuestas relevantes hechas en foros de software, utilizando un mecanismo que automáticamente infiere las etiquetas de las publicaciones en los hilos de discusión hechos en un foro.

Como se puede observar en la imagen 5, el mecanismo asigna siete diferentes etiquetas a la publicación: pregunta, respuesta, respuesta aclaradora, pregunta aclaradora, comentario positivo, comentario negativo y basura. Asimismo, se entiende que todo parte de una pregunta, la cual puede tener dos flujos. El primero de ellos incluye otras preguntas que permiten esclarecer el motivo de la pregunta y el segundo obtiene directamente la respuesta a la pregunta planteada inicialmente; así mismo, una vez encontrada la respuesta, se producen comentarios positivos y negativos al respecto. Por último, lo que no es considerado como información relevante para el análisis es etiquetado como basura.

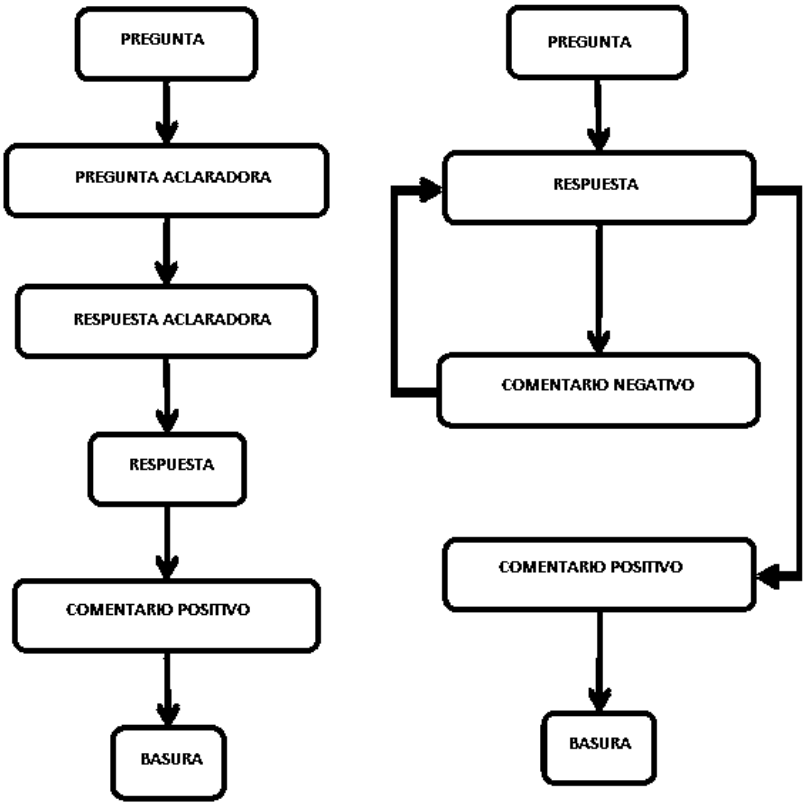

Imagen 5: Cadena de conversaciones en un foro (Gottipati, Lo et al. 2011).

A partir de esto se construyó un mecanismo de búsqueda semántica cuyos resultados mostraron que podía llegar a un 67% de precisión, 71% de *recall* y a un 69% valor-f. La eficacia de este método es validada al compararlo con un sistema de extracción de información estándar confirmando que en promedio se puede mejorar considerablemente la precisión (Gottipati, Lo et al. 2011)

2.13 Conclusiones sobre el estado del arte

Una vez presentados los proyectos encontrados para resolver el problema de extracción de información de datos no estructurados generados a partir de comentarios realizados en un foro; se observa que se deben considerar muchos factores para el desarrollo de una solución del mismo. Además, es importante resaltar que la complejidad del problema varía según el dominio elegido para realizar la extracción. Por lo tanto, los métodos mostrados en el desarrollo de esta sección, si bien es cierto nos brindan un panorama general, no se aplican a cada uno de los escenarios posibles; por lo que es necesario utilizar un método diferente dependiendo si se trata de comentarios hechos en un foro de discusión en línea respecto a una organización del rubro educativo

Cuadro 3: Resumen de estudios revisados

Nombre de Estudio	Dominio de aplicación	Método utilizado	Resumen
Estudio N°1 (Gang, Yingwei et al. 2013)	Foros de discusión en línea que abordan diferentes temas.	-Sistema de huellas digitales.	Analiza diferentes tipos de foros incorporando al sistema de extracción de información un sistema de huellas digitales para detectar las diferentes estructuras de cada uno.
Estudio N°2 (Yingying and Daqing 2013)	Foros de discusión en línea que abordan el tema de desarrollo de software.	-Procesamiento de lenguaje natural. -Técnicas de análisis de sentimiento.	Investiga métodos de extraer problemas ocurridos en los diferentes API's utilizados para el desarrollo de software.
Estudio N°3 (Mohajeri, Esteki et al. 2013)	Foros de discusión en línea que abordan los temas relacionado a la salud y la medicina.	-Sistema basado en ontologías.	Simplifica la exploración en foros de discusión extrayendo las entidades de los comentarios hechos en lenguaje natural a través de la utilización de ontologías.
Estudio N°4 (YingJu, YuHang et al. 2011).	Foros de discusión en línea, blogs y sitios web de noticia.	-Transferencia de aprendizaje. -Algoritmo de alineamiento de árbol.	Propone un método para la extracción automática de información a partir de capas, los cuales son construidos utilizando los métodos mencionados.
Estudio N°5 (Yun, Bicheng et al. 2009)	Foros de discusión en línea que abordan diferentes temas.	-Extracción de metadatos.	Resuelve el problema de extracción de información mediante la presentación de un método de dos pasos, el primero que reconoce el tema a través de la estructura de la página web y luego extrae metadatos.

Estudio N°6 (Suke, Liyong et al. 2009)	Foros de discusión en línea que abordan diferentes temas.	-Análisis de estructura HTML de las páginas web	Propone un enfoque que permite extraer información automáticamente de los foros de discusión a partir de la estructura HTML de los mismos.
Estudio N°7 (Stavrianou, Chauchat et al. 2009)	Foros de discusión en línea que abordan diferentes temas.	-Gráficos basado en mensajes y usuarios	Utiliza la representación mediante gráficos de los mensajes y usuarios para extraer y analizar los datos contenidos en un foro de discusión en línea.
Estudio N°8(Jingwei, Cheqing et al. 2012)	Foros de discusión en línea que abordan diferentes temas.	-Proceso de inducción estructural. -Proceso de combinación de términos.	Propone un método para la extraer información de los foros de discusión usando las características estructurales y visuales de los mismos.
Estudio N°9 (Machova and Penzes 2012)	Foros de discusión en línea que abordan diferentes temas.	-Técnica de alineación parcial de un árbol.	Resuelve el problema de extracción de información basándose en la asunción que la página web contiene notas estructuradas para los objetos; es decir, contiene una estructura de HTML.
Estudio N°10 (Gottipati, Lo et al. 2011)	Foros de discusión en línea que abordan el tema de desarrollo de software.	-Inferencia de etiquetas semánticas.	Brinda un mecanismo de búsqueda semántica que permite encontrar respuestas relevantes hechas en foros de discusión en línea respecto a desarrollo de software.

En conclusión, luego analizar los estudios primarios obtenidos como resultado de la revisión sistemática se puede apreciar que la mayoría de investigaciones que están dedicadas a la extracción de información, utiliza métodos como la aplicación de capas, aprendizaje de reglas automático, aprendizaje de modelos estadísticos, entre otros; mientras que, en menor proporción, se tienen investigaciones que buscan realizar la

extracción de información de este tipo basado en ontologías mediante un enfoque de la ingeniería del conocimiento. A partir de esto, el presente proyecto de fin de carrera se enfocará en aportar un sistema de extracción de información basado en ontologías, el cual permitirá el entendimiento e interpretación del contenido de los comentarios hechos en un foro de discusión. De este modo, se aprovechara al máximo los datos no estructurados; con lo que la organización educativa podrá hacer uso de ellos para generar estadísticas de opinión de los alumnos respecto a sus cursos y profesores.

CAPÍTULO 3: Pre-procesamiento de texto escrito en lenguaje natural

En el presente capítulo se desarrollará el primer objetivo específico del proyecto. Se presentará un análisis cualitativo de los comentarios que se procesarán dentro del sistema y el diseño de un componente que permitirá normalizarlos de su forma original a un formato que el sistema podrá procesar.

Todos los comentarios realizados en los foros de discusión en línea que serán analizados en el desarrollo del presente proyecto se encuentran escritos en lenguaje natural por los usuarios, los cuales utilizan un lenguaje informal para redactarlos. Por consiguiente, para el procesamiento de los mismos se deberá resolver los problemas presentes a consecuencia de esta redacción como errores ortográficos, lenguaje ofensivo, entre otros.

Por esta razón, teniendo en cuenta lo mencionado anteriormente, este objetivo está enfocado en preparar los comentarios para poder realizar la tarea de extracción de manera eficiente.

1 Resultado Esperado 1: Análisis cualitativo de la información de los comentarios del foro de discusión en línea.

1.1 Introducción

Este resultado tiene como objetivo brindar una visión general respecto a la información que se encuentra dentro de los comentarios hechos en el foro de discusión escogido para el análisis; de esta forma se podrá entender el dominio de cursos brindados por una universidad que se representara para los siguientes resultados; así como también, los textos en lenguaje natural que servirán como datos de entrada al sistema. Para este fin se ha elegido el foro de discusión en línea *Enlacato.com* (Clasesmas 2014), en el cual alumnos de la Pontificia Universidad Católica del Perú; comparten dudas, opiniones, ideas y archivos respecto los servicios que brinda dicha universidad.

1.2 Resultado Alcanzado

Para realizar el análisis se utilizó una base de datos en MySql, la cual contiene los comentarios de los usuarios del foro de discusión en línea.

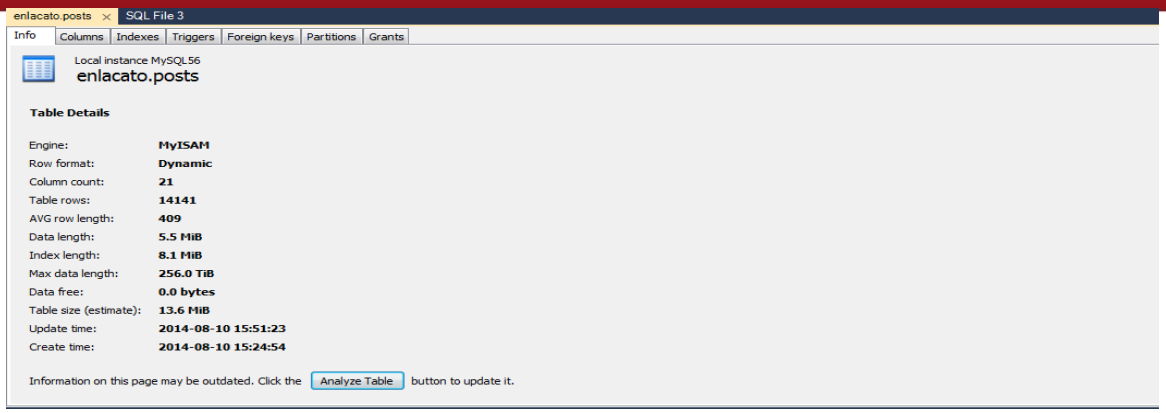

Imagen 6: Reporte extraído de la base de datos de comentarios Enlacato.com al 07 de agosto del 2014 (Clasesmas 2014)

Como se puede observar en la imagen 6, el foro contaba con 4988 miembros registrados, los cuales han realizado un total de 14,141 comentarios que se encuentran dentro de un total de 2798 publicaciones, las cuales representan 13.6 MiB en datos no estructurados.

Sin embargo, como se puede observar en el Cuadro 4, se identificaron distintas secciones predefinidas dentro del foro, donde se detectaron que algunas que eran muy similares, por lo que fueron agrupadas.

El resultado final de esta agrupación se puede observar en la Cuadro 5, donde se tiene la cantidad efectiva de comentarios y publicaciones que serán procesados para cada grupo dentro del sistema.

Cuadro 4: Secciones dentro de Enlacato.com (Clasesmas 2014)

Secciones	
Estudios Generales Ciencias	Recomienda Audiovisuales
Cálculo	Ing. Minas - Cursos y dudas
Cálculo 1	Estructuras Discretas
Cálculo 2	Tecnicas de Programacion
Cálculo 3	Todo sobre Enlacato
Cálculo 4	Ciencias e Ingeniería
Física	Ingeniería Informática
Física 1	Ingeniería Industrial
Física 2	Ingeniería Mecánica
Física 3	Ingeniería Electrónica
Dibujo en Ingeniería	Ingeniería Civil
Profes EEGGCC	Quinto Ciclo
Rotonda EEGGCC	Sexto Ciclo
Puros Electivos	Séptimo Ciclo
Intro. Matemáticas Universitarias	Octavo Ciclo
Intro. Física Universitaria	Noveno Ciclo
Introducción a la Computación	Décimo Ciclo
Química	General
Matemáticas Básicas	Ingenierí a de Minas
Ing. Informática - Cursos y Dudas	Ingeniería de Telecomuniaciones
Ing. Electrónica e Ing. Teleco- Cursos y Dudas	Compra, vende, intercambia
Ing. Industrial - Cursos y Dudas	Estática
Ing. Mecánica - Cursos y Dudas	Ingeniería Mecatrónica
Ing. Civil - Cursos y Dudas	Estadística
Cursos de Matemáticas, Química y Física	Lenguajes de Programación
Ing. Mecatrónica - Cursos y Dudas	Zona Staff
Estudios Generales Letras	Todo Cine, TV, Radio y Espectáculos
Profes EEGGLL	Filosofía, Teología, RyC e ICOE
Rotonda EEGGLL	Recomienda un libro
Cursos EEGGLL	Libros disponibles
Comunidad - No te aburras	Matemáticas, Física , Química.
Discusiones Generales - Recontra Off Topic	Letras
Tu primer tema - Preséntate aquí	Informática
Química 1	Miscelánea
Química 2	Biblioteca Enlacato

Cuadro 5: Grupos de publicaciones y comentarios

Grupos	Cant. Publicaciones	Cant. Comentarios
Profesores	372	1611
Cálculo	255	1106
Física	196	1031
Otros	173	904
Química	135	453
Electivos	129	338
Introducción a la Computación	91	384
Dibujo en Ingeniería	77	633
Matemáticas Básicas	67	358
Ingeniería Civil - Otros	66	285
Ingeniería Industrial - Otros	65	130
Filosofía, Teología, Redacción y Comunicación	64	265
Estática	52	311
Introducción a la Física Universitaria	50	240
Ingeniería Industrial - Otros	48	124
Ingeniería Electrónica y Telecomunicaciones - Otros	45	124
Técnicas de Programación	38	130
Estadística	33	102
Introducción a la Mátematica Universitaria	33	78
Ingeniería Mecánica - Otros	31	93
Ingeniería Minas - Otros	21	14
Ingeniería Mecatrónica - Otros	17	56
Estructuras Discretas	15	44
Lenguajes de Programación	14	40
Ingeniería Informática - Otros	12	65
Cursos de Ciencias	11	16
Ingenieria Informática - Otros	6	5
Ingeniería de Minas - Otros	5	21
Total general	**2121**	**8961**

A partir de lo mencionado, se obtuvo que las principales facultades que se encuentran representadas en los comentarios son la Facultad de Ciencias e Ingeniería, Facultad de Estudios Generales Letras y Facultad de Estudios Generales Ciencias de dicha universidad. Es importante resaltar que la mayor parte de usuarios pertenecen a esta última facultad mencionada, por lo que la mayoría de comentarios son respecto a cursos que se dictan en la misma. Por ejemplo, introducción a la física universitaria, física, cálculo, introducción a la matemática universitaria, dibujo en ingeniería, entre otros. Asimismo, como se puede observar en el Cuadro 4, la mayor parte de comentarios son respecto a los profesores de dichos cursos.

1.3 Pruebas

A continuación, se presentarán algunos comentarios donde se muestra la información que se encuentra dentro de los mismos.

Input 1 (Comentario Inicial):

Estando en mi ultimo ciclo de generales les dejo mis comentarios sobre mis profesores pasados:

 Norberto Chau - **MB y Calculo 3** - Buen profesor, excelente pizarra. Si eres ordenado copiando, con una leida a tu cuaderno sera suficiente para pasar las evaluaciones. Muy recomendable

 Jose Phan - **IFU** - A pesar de su mala fama fue muy bueno para IFU, llegue a pasar el curso con 19.

 Sergio Pavletich - **Calculo 2 y estadistica** - Me gusto como profe de estadistica, a pesar de su desorden hacia pensar mucho a los alumnos. Para cal2 no lo recomiendo ya que es un curso donde la practica es escencial y el profe no es de resolver problemas.

 Israel Cabrera - Estatica - Buen profe, corrige medio-alto y si observa que te esfuerzas al final te recompensa con ese puntito que falto para aprobar. Recomendable

 Rosa Jabo - Calculo 1 - Excelente profesora, quizas la mejor para calculo 1. Explica muy bien y es ordenada. Corrige normal. Muy recomendable

 Patrizia Pereyra - Fisica 2 - Se preocupa porque sus alumnos aprendan pero en ocaciones no se explica muy bien. La dificultad de sus practicas oscila bastante en el curso dependiendo de como vaya el salon. Recomendable.

 Espero sea de ayuda para las nuevas generaciones

Input 2 (Comentario Inicial):

HI! Yo aun no llevo **Fisica 2** pero llevare el curso con **Pablo Vilela**.
Porque diablos hare eso?
Bueno, el profesor me ensenio **IFU y Fa1** y ambos los pase con buena nota.
Mira las Stats! tiene el mayor promedio en **Fa1 y Fa2** del ciclo 20091

Opino que su manera de enseniar es pasable y sus jps te ayudan bastante!

Y.....eso es todo jeje
bye

Como se puede observar dentro del input 1 y el input 2, las palabras resaltadas mencionan cursos que se encuentran dentro de la Facultad de Estudios Generales Ciencias. Así como también, se mencionan profesores que pertenecen a la misma unidad académica donde se indica para cada uno la metodología de enseñanza, calificación y la asistencia a sus clases. Por otro lado, se pudo observar que en el dominio se utilizan numerales y adverbios que son considerados como fuentes; por ejemplo, *"Fa1"*, *"Fa2"* y *"cal2"* que representan a los curso de *"Física"* y *"Calculo"*

Input 3 (Comentario Inicial):

[quote name="Alek"]Todo el material que tengo de **cal4** ![/quote] **[attachment=526:Ex1-6.rar]
**
**[attachment=525:Ex1-5.rar]
**
**[attachment=524:Ex1-4.rar]
**
**[attachment=523:Ex1-3.rar]
**
**[attachment=522:Ex1-2.rar]
**
[attachment=527:Ex1-1.rar]. he dicho.
Bye

Por último, en el input 3 se puede observar como el último comentario menciona el curso Calculo 4, perteneciente a la facultad que se mencionó en los anteriores inputs y los materiales que se comparten en el comentario, en este caso exámenes pasados.

2 Resultado Esperado 2: Análisis cualitativo del formato de los comentarios del foro de discusión en línea.

2.1 Introducción

El presente resultado tiene como objetivo analizar los formatos de los comentarios para describir los problemas de procesamiento de lenguaje natural (NLP) que se enfrentarán para que la tarea de extracción de información tenga un mejor desempeño. Cada texto que se analizará tendrá formatos y contenidos diferentes, por

lo que para realizar la tarea de extracción se necesita que estos textos se encuentren estandarizados. En otros términos, los comentarios deben de cumplir un formato establecido, en el cual se hayan superado los NLP que pudieran existir.

A partir de esto se podrá determinar el diagrama de arquitectura que permitirá procesamiento de los comentarios brindando un mayor entendimiento al lector respecto al funcionamiento general del sistema y los módulos con los que se contarán.

2.2 Resultado Alcanzado

a) Formato de los comentarios

El formato de los comentarios, como se puede observar en la sección 2.3 del presente capítulo, se encuentra representado por texto en lenguaje natural y etiquetas básicas de HTML. Si bien es cierto el procesamiento en lenguaje natural tiene varias limitantes o dificultades, muchas de ellas podrán ser resueltas antes de realizar la tarea de extracción de información. Por ejemplo, los comentarios que serán usados como texto de entrada para el sistema, usan lenguaje informal por lo que contienen lenguaje ofensivo o errores ortográficos.

En primer lugar, el comentario deberá pasar por un proceso de estandarización, el cual consiste en reemplazar cualquier palabra que se considere ofensiva con una etiqueta predefinida; asimismo, se descartarán caracteres que no brinden información relevante o se reemplazarán si fueran especiales. De igual manera se intentará resolver algunos errores ortográficos del texto, para poder procesarlos en el siguiente paso. Según lo mencionado anteriormente, una vez que el comentario ha sido estandarizado, se realizará un proceso de *lematización,* el cual consiste en volver a su forma base las palabras contenidas en el comentario; con el fin de tener todas las palabras en una sola forma. Por ejemplo, si se tiene la palabra “practicando” o “practicaremos”; ambas se convertirán en “practicar” luego de pasar por dicho proceso. Esto permitirá que se pueda realizar una adecuada búsqueda de palabras o patrones al realizar las siguientes tareas del sistema.

b) Diagrama de arquitectura

Como se puede observar en la imagen 7, la arquitectura propuesta para el presente proyecto contará con los siguientes módulos y componentes:

- **Módulo de pre-procesamiento en lenguaje natural:** Permitirá realizar un análisis previo al comentario, para lograr establecer una estandarización

inicial que permita realizar los procesamiento de lenguaje natural necesarios para los siguientes módulos. Para este módulo se hará uso de las librerías de clases de Freeling y Lucene.

- **Ontología de dominio:** Permitirá la representación del conocimiento en el dominio de cursos brindados por una universidad. Para este módulo se utilizará la herramienta de diseño Protege para incluir la información del dominio que se desarrollará.
- **Módulo de desambiguación de palabras:** Tendrá como objetivo resolver las palabras ambiguas encontradas en los comentarios ya estandarizados con la información contenida en la ontología de dominio. Para este módulo, se utilizará la librería Jena para poder manejar la ontología de dominio dentro de la aplicación.
- **Módulo de extracción de información:** Permitirá extraer la información especificada según la ontología de dominio. Para este módulo se hará uso de la librería Jena para poder manejar la ontología de dominio dentro de la aplicación.

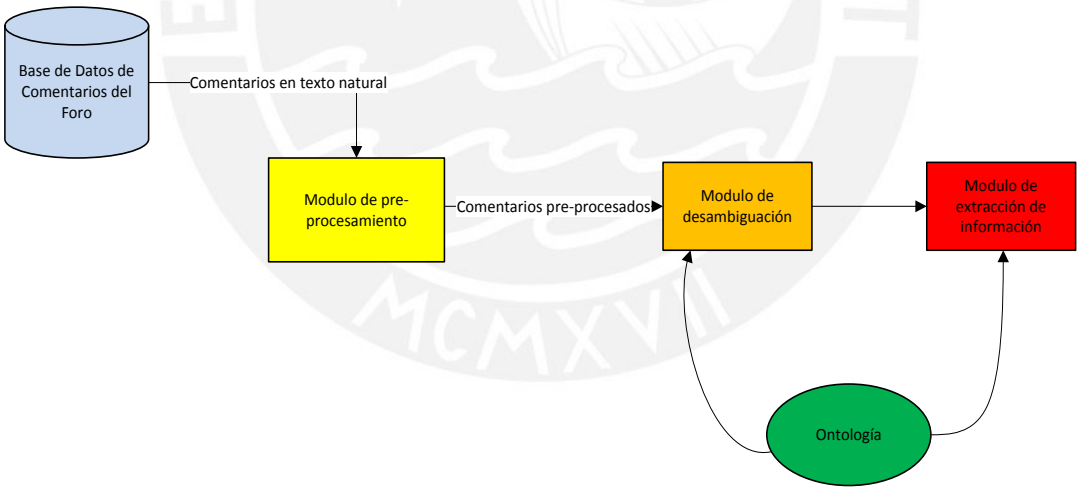

Imagen 7: Diagrama de arquitectura del sistema de extracción.

En los siguientes capítulos, se detallará el funcionamiento y contenido de cada uno de los componentes mencionados.

2.3 Pruebas

A continuación, se presentan los distintos tipos de comentarios que se encuentran dentro del foro de discusión analizado:

Input 1 (Comentario Inicial):

Estando en mi ultimo ciclo de generales les dejo mis comentarios sobre mis profesores pasados**:
**
**
** Norberto Chau **-** MB y Calculo 3 - Buen profesor, excelente pizarra. Si eres ordenado copiando, con una leida a tu cuaderno sera suficiente para pasar las evaluaciones. Muy recomendable**
**
**
** Jose Phan **- IFU -** A pesar de su mala fama fue muy bueno para **IFU**, llegue a pasar el curso con 19.**
**
**
** Sergio Pavletich **-** Calculo 2 y estadistica **-** Me gusto como profe de estadistica, a pesar de su desorden hacia pensar mucho a los alumnos. Para cal2 no lo recomiendo ya que es un curso donde la practica es escencial y el profe no es de resolver problemas.**
**
**
** Israel Cabrera **-** Estatica **-** Buen profe, corrige medio-alto y si observa que te esfuerzas al final te recompensa con ese puntito que falto para aprobar. Recomendable**
**
**
** Rosa Jabo - Calculo 1 - Excelente profesora, quizas la mejor para calculo 1. Explica muy bien y es ordenada. Corrige normal. Muy recomendable**
**
**
** Patrizia Pereyra **-** FÃsica 2 **-** Se preocupa porque sus alumnos aprendan pero en ocaciones no se explica muy bien. La dificultad de sus practicas oscila bastante en el curso dependiendo de como vaya el salon. Recomendable.**
**
**
** Espero sea de ayuda para las nuevas generaciones

Como se puede observar, muchos de los comentarios no se encuentran estandarizados; es decir, cuentan con etiquetas *"
"* que no aportan información relevante para extraer; así como también, caracteres especiales que pueden presentar una dificultad durante el análisis de correferencias. En el cuadro 6 y 7, se pueden observar otros casos similares que han sido considerados para el desarrollo del sistema.

Cuadro 6: Caracteres especiales

Caracter especial	reemplazo
Ã¡	á
Ã©	é
Ã³	ó
Ãº	ú
Ã±	ñ
Ã	í

Cuadro 7: Texto sin información relevante

Textos descartados
\
"
[
[
<
>
(
)
{
}
/
!
,
?
¿

Por otro lado, se puede observar que se presentan abreviaturas dentro del texto; por ejemplo, *"IFU"* que representa el curso de *"Introducción a la Física Universitaria"*. Según lo mencionado anteriormente, las palabras que cuentan con una longitud de caracteres menor o igual a cinco, serán consideradas como abreviaturas de una palabra, debido a que, en el dominio, esa fue la cantidad máxima de caracteres para representar una abreviatura.

Input 2 (Comentario Inicial):

Medina deja sus pcs regaladas eso si no **enseañ** ni **mierda.**

Por otro lado, en este último comentario se muestran un error ortográfico donde la palabra resaltada **"enseañ"** se encuentra mal escrita, para esto será necesario corregir estos errores para poder estandarizar la mayoría de palabras del comentario. De igual manera, la palabra resaltada "**mierda"** representa el uso del lenguaje ofensivo en los comentarios los cuales serán etiquetados para un análisis formal**.** En el cuadro 8, se presenta la lista de palabras consideradas ofensivas y la cantidad de veces que aparecen en los comentarios.

Cuadro 8: Grupos de publicaciones y comentarios

Lenguaje ofensivo	Cant. Publicaciones
puta	102
ptm	40
mierda	27
concha	25
mare	16
carajo	15
idiota	13
marica	7
gay	6
chucha	5
xuxa	2
maricon	1
ctm	1
cojudo	1

Por otro lado, una vez resueltas dichas dificultades, se deberá estandarizar cada una de las palabras *lemmatizandolas* y realizando un análisis morfosintáctico.

Para poder realizar la tarea de extracción de información será necesario estandarizar estos comentarios de tal forma que se pueda realizar un análisis de correferencias con partes del texto que contengan información relevante considerando el texto en lenguaje natural pre-procesado.

3 Resultado Esperado 3: Módulo de pre-procesamiento de los comentarios en lenguaje natural que permita normalizar el texto en los comentarios.

3.1 Introducción

Este mecanismo de pre-procesamiento de los comentarios permitirá estandarizarlos y así prepararlos para su futuro análisis dentro de la tarea de extracción. Para tal fin, se está considerando como entrada a este módulo un comentario escrito en lenguaje natural, el cual está compuesto por la concatenación de todos los comentarios que se han realizado para una publicación.

En primer lugar, para la entrada, se realizará un etiquetado de las palabras que contengan lenguaje ofensivo; asimismo, se reemplazarán los caracteres especiales y los textos que no tengan información relevante dentro del comentario, luego se corregirán algunos errores ortográficos que se presenten dentro los comentarios; a continuación se realizará el proceso de *lematización,* el cual permitirá reducir las palabras a su forma más básica (*lemma*) y; finalmente el etiquetado morfosintáctico que permitirá identificar del tipo, género y número de las palabras dentro del comentario.

3.2 Resultado Alcanzado

A partir de la imagen 8, se puede observar que cada comentario pasará por el siguiente proceso:

- En el primer nivel, se buscaran y reemplazarán, dentro del comentario, caracteres especiales y texto que ha sido considerado como poco relevante (ver sección 2.3).

- En el segundo nivel, Se procederá a *tokenizar* las palabras contenidas en el comentario, de tal forma que el comentario sea separado en un conjunto de *tokens* (palabras); es decir, cada palabra recibirá un identificador para que pueda ser procesado. Para tal fin, se hizo uso de la herramienta Freeling.

- En el tercer nivel, se selecciona un *token*, el cual será buscado dentro de la lista de palabras consideradas ofensivas. Si se cumpliera la evaluación presentada en el cuarto nivel de la imagen; el *token* será reemplazado por la etiqueta "*lenguaje_ofensivo*" (ver sección 2.3).

- A continuación, si el *token* no fue considerado como lenguaje ofensivo; en el quinto nivel, será analizado para poder ser transformado a su forma base de la siguiente manera.

    - el *token* pasará por una búsqueda en el diccionario de *lemmas,* sin considerar la etiqueta de "*lenguaje_ofensivo*" utilizada previamente.

    - Si se encontrará, dentro del diccionario, en el octavo nivel, se obtendrá el *lemma* (forma base de la palabra). Así como también, en el noveno nivel, se realizará un análisis morfosintáctico para determinar si el *token* es un verbo, adjetivo, nombre, pronombre, entre otros; dentro del comentario. Para realizar esta tarea se utilizó el diccionario en español por defecto que brinda la herramienta Freeling, donde por medio de las librerías que ofrece, se logró *lemmatizar* las palabras y etiquetarlas morfosintácticamente cada palabra; es decir, al ser encontrada en el diccionario en español utilizado fue reemplazado por su forma base y analizada para determinar si era un verbo, nombre, pronombre, entre otros; se puede encontrar una descripción más detalladas de dichas etiquetas en (Freeling 2014).

    - De no ser encontrada, en el séptimo nivel, dicha palabra será considerada como un error ortográfico, la cual será procesada para determinar que palabra puede ser la más óptima a reemplazar. Este análisis se realizará solo si no se tratará de un nombre, verbo, numeral, pronombre o adverbio, los cuales, según el dominio, podrían ser considerados como posibles fuentes en los siguientes módulos del sistema de extracción de información; asimismo, si el *token* contará con un *lemma* con longitud menor o igual a 5 caracteres, tampoco será procesado, debido a que es considerada una abreviatura. Una vez corregida la palabra, esta pasa nuevamente por el proceso de *lemmatizacion* y etiquetado morfosintáctico de palabras mencionado en los puntos anteriores.

      Para la realización se esta tarea se utilizó la herramienta Apache Lucene, la cual por medio de las librerías que ofrece permitió analizar mediante un diccionario en español, las posibles palabras que podían reemplazar a un error ortográfico; el desarrollo de este módulo se basó

en el algoritmo supervisado planteado en (van Delden, Bracewell et al. 2004), el cual plantea la implementación de la distancia de Levenshtein, para calcular el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra, para esto se utilizó la implementación del algoritmo en Apache Lucene y un diccionario con palabras en español para obtener la mejor opción para reemplazar una palabra con error ortográfico.

- Por último, en el décimo nivel, se evaluará si es el último token de no serlo se volverá al tercer nivel para realizar el análisis del siguiente *token*.

Una vez culminado este proceso, se obtendrá el texto del comentario con las palabras relevantes con la información de su forma base o diccionario (*lemma*), etiquetadas indicando el resultado del análisis morfosintáctico, la cantidad de errores ortográficos reducida y la etiqueta de lenguaje ofensivo, según sea el caso.

Por otro lado, es importante mencionar que los diccionarios utilizados contienen términos del idioma español de España, lo cual puede llevar a que no todas las palabras sean *lematizadas* o corregidas ortográficamente a nivel léxico, ya que los comentarios que se están analizando se encuentran redactados por personas latinoamericanas. Sin embargo, no se ha enfatizado mucho en este punto debido a que no se trata de un objetivo principal dentro de la tarea de extracción.

Por último, se puede considerar, para futuras investigaciones, reemplazar los diccionarios mencionados anteriormente; con el objetivo de realizar el pre-procesamiento de comentarios con mayor exactitud.

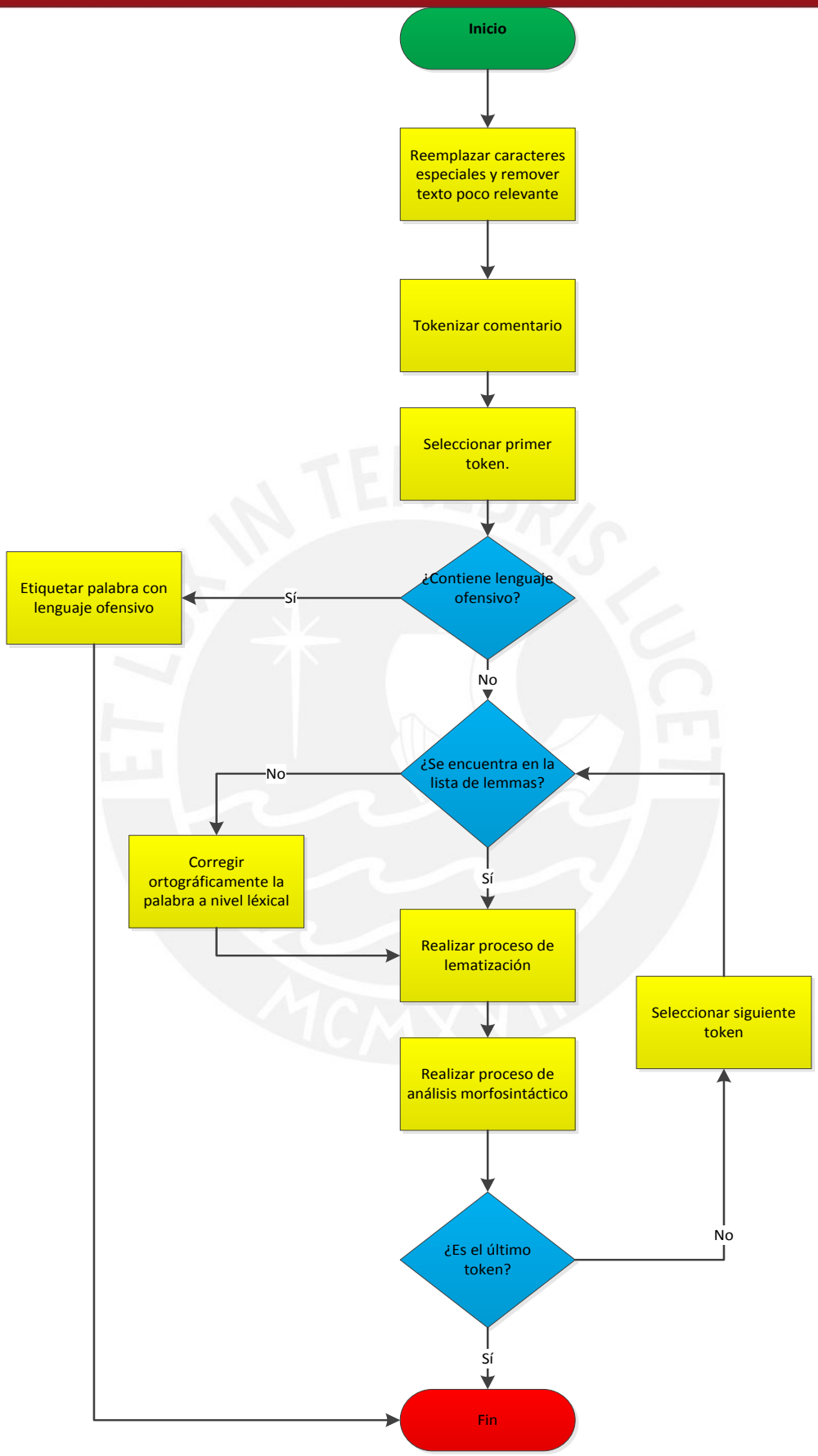

Imagen 8: Mecanismo de pre-procesamiento de comentarios.

3.3 Pruebas

Para la verificación del mecanismo se realizaron pruebas con el 20% del total de las publicaciones obteniendo una precisión de 79% y un *recall* del 71%. Estos resultados se deben principalmente al corrector ortográfico, ya que muchos de los errores ortográficos se encontraban dentro del comentario eran representados por nombres, numerales o adverbios, los cuales no están considerados para la aplicación de la corrección ortográfica; sin embargo, al momento de analizar las palabras que si estaban consideradas para la aplicación del corrector ortográfico este realizaba el proceso de corrección de manera aceptable. A continuación se muestra el resultado a partir del siguiente comentario:

Input (Comentario):

Comentario: SÃ¡nchez es un buen profesor!!, explicr bien, no se apura pero tampoco se atrasr y te acepta cualquier duda. Por otro lado, el profeosr Medina es una mierda.

A continuación se analizará paso a paso el proceso por el que pasa este comentario inicial hasta el resultado final:

Output 1: Reemplaza lenguaje ofensivo, caracteres especiales y texto no relevante.

Comentario: Sánchez es un buen profesor explicr bien no se apura pero tampoco se atrasr y te acepta cualquier duda. Por otro lado el profesor Medina es una lenguaje_ofensivo.

En este primer resultado del módulo de pre-procesamiento, se puede observar como la palabra considerada ofensiva ha sido reemplazada con la etiqueta *"lenguaje_ofensivo"*. Asimismo, el carácter *"Ã¡"* fue reemplazado por *"á"* y el texto *"!!"* fue descartado por ser poco relevante para el análisis.

Output 2: Corrige ortográficamente a nivel lexical

Comentario: Sánchez es un buen profesor explicr bien no se apura pero tampoco se atrasar y te acepta cualquier duda. Por otro lado el profesor Medina es una lenguaje_ofensivo.

En el segundo resultado, se puede observar que la palabra *"atrasr"* fue reemplazada por *"atrasar"*, ya que la palabra fue considerada como un verbo. Una vez realizado dicho cambio, la palabra pasará el proceso de *lemmatización* y análisis morfosintáctico. Por otro lado, la palabra *"explicr"* al ser considerada un nombre y no un verbo, no es analizada por el corrector ortográfico; sin embargo, también pasará por el proceso de *lemmatización* y análisis morfosintáctico.

Output 3: *Lematiza* y realiza análisis morfosintáctico

Comentario:

Sánchez sánchez NP00000
es ser VSIP3S0
un uno DI0MS0
buen bueno AQ0MS0
profesor profesor NCMS000
explicr explicr NCMS000
bien bien RG
no no RN
se se P00CN000
apura apurar VMIP3S0
pero pero CC
tampoco tampoco RG
se se P00CN000
atrasar atrasar VMN0000
y y CC
te te PP2CS000
acepta acepto AQ0FS0
cualquier cualquiera DI0CS0
duda duda NCFS000
. . Fp
Por por SPS00
otro otro DI0MS0
lado lado NCMS000
el el DA0MS0
profesor profesor NCMS000

```
Medina medina NP00000
es ser VSIP3S0
una uno DI0FS0
lenguaje_ofensivo lenguaje_ofensivo AQ0MS0
. . Fp
```

Como se puede observar en el resultado, se realiza el proceso de *lemmatización* y análisis morfosintáctico de las palabras que se encuentran dentro del comentario. A partir de dicha información se procede a continuar la ejecución de los siguientes módulos dentro del sistema de extracción de información.

4 Conclusión

Según lo observado, los comentarios que se encuentran en el foro de discusión contienen, en gran medida, la opinión de alumnos que pertenecen a la etapa inicial de carreras universitarias relacionadas a las ciencias e ingeniería. Por esta razón, la información más relevante que se podrá extraer de esta base de conocimiento será relacionada a los cursos brindados para esta clase de alumnos y los profesores que dictan estos cursos. Por consiguiente, esta información será representada en el dominio de cursos de una universidad dentro de una ontología que se planteará en los siguientes resultados.

Por otro lado, la tarea de extracción de información aplicada a los comentarios en su formato original, tendría un bajo desempeño de no ser por el aporte del conocimiento de dominio y el entrenamiento del sistema a partir de una muestra de los comentarios que recibió como entrada este mecanismo. Dicha muestra brindó un mayor alcance respecto al comportamiento inicial del mecanismo, lo cual permitió realizar modificaciones pertinentes para que se obtenga un mejor desempeño.

En conclusión, los siguientes módulos dependerán tanto de la ontología de dominio diseñada como de la muestra de datos que se ha usado como entrenamiento para este mecanismo; es decir, la información que se logré obtener será en base al conocimiento presentado que dependerá de la persona que diseñe la ontología y analice el dominio.

CAPÍTULO 4: Desambiguación lexical de palabras

En el presente capítulo se desarrollará el segundo objetivo específico del proyecto. Se buscará resolver el problema de ambigüedad de palabras presentadas en un comentario. Por ejemplo, si se tuviera la palabra *"intro"* dentro del comentario, se podrían tener distintas interpretaciones de la misma, según el contexto; es decir, puede ser considerado como el curso de *"Introducción a la computación"* o el curso de *"Introducción a la ingeniería informática".*

Por lo mencionado anteriormente, este objetivo está enfocado en resolver este problema del procesamiento de lenguaje natural. Con el fin de identificar las fuentes dentro del comentario de manera natural; en otros términos, poder identificar los cursos y profesores que son considerados como fuentes dentro del dominio, lo cual permitirá que el análisis de correferencias sea realizado de manera óptima. Para esto se busca que se pueda identificar el concepto que intenta representar cada palabra del comentario. Por esta razón, se planteará una estructura que permitirá identificar conceptos ambiguos y resolverlos dentro del comentario para luego realizar la tarea de extracción de información.

1 Resultado Esperado 1: Estructura para almacenar el conocimiento contenido en los comentarios

1.1 Introducción

Este resultado busca superar una de las causas identificadas como raíz del problema planteado: No se conoce del todo el dominio sobre el cual se realiza la extracción.

Por esta razón, se brindará una base que permita un posterior procesamiento del conocimiento del dominio, otorgando a los usuarios finales la facilidad del manejo del mismo. Para tal fin, según lo especificado en el alcance del presente proyecto se utilizó para el desarrollo de este mecanismo una ontología de dominio desarrollada anteriormente (Carranza 2014).

1.2 Resultado Alcanzado

a) Ontología base:

Para fines del desarrollo de este resultado, se modificó una ontología en el dominio de los conceptos estudiados en la rama de ciencias de la computación de la

especialidad de ingeniería informática de la Pontificia Universidad Católica del Perú, en formato OWL/RDF(Carranza 2014). Esta ontología cuenta con:

- Una propiedad “nombrePreferente” que contiene el nombre principal de cada concepto. Cada nodo de la ontología tiene exactamente un nombre preferente asociado a él. Este elemento existe con la única finalidad de ser mostrado al usuario final para su fácil lectura y comprensión del concepto al que se quiere hacer referencia.
- Una propiedad “sinonimos” que relaciona cada nodo con sus respectivos términos equivalentes en versiones *lematizadas*. Cada nodo de la ontología puede tener uno o varios sinónimos asociados a él.
- Una propiedad “lemma” que relaciona cada concepto con su respectiva denominación en forma base (*lemma*). Cada nodo de la ontología tiene exactamente un *lemma* asociado a él.

Si bien el valor asociado a la propiedad “nombrePreferente” de cada nodo es completado en base al conocimiento especializado propio de quien diseña la ontología, las formas *lematizadas* a colocar en las otras dos propiedades se generan empleando los mecanismos propuestos en este proyecto (ver capítulo 3, sección 3). Luego de obtener dichos *lemmas*, estos deben ser introducidos manualmente en la ontología como parte del proceso de modificación de la ontología.

b) Modificaciones:

Según lo mencionado anteriormente, se agregaron las clases relacionadas al dominio que se busca atacar en el presente proyecto:

- La clase “Comentario” que contiene los tipos de comentarios a los cuales se refieren cada uno de los alumnos que opinan en el foro de discusión.
- La clase “Profesor” que permitirá relacionar la información planteada para cada profesor que se mencione.
- La clase “Material” que contiene los tipos de materiales que son compartirdos diariamente por los alumnos que opinen en el foro de discusión.

Además, se agregó una propiedad llamada “nombreFuente”, la cual permitirá, dentro de la resolución de correferencias, identificar las fuentes dentro de un comentario analizado. Esta propiedad almacena la versión *lematizada* de las clases, pero con los espacios reemplazados por el carácter “_”.

Adicionalmente, se añadieron propiedades para la relación entre las clases para de esta forma poder identificar al momento de realizar la tarea de extracción las referencias de cada uno de los nodos.

- La propiedad “tieneComentario” relaciona directamente a la clase “Profesor” y la clase “Comentario”, ya que según lo revisado en la sección 1 del capítulo 1, mucha de la información contenida en el foro son opiniones o pequeñas reseñas respecto a un profesor de un curso específico.
- La propiedad “tieneProfesor” relaciona las clases “Curso” y la clase “Profesor” para lograr identificar que profesor enseña un determinado curso.
- La propiedad “tieneMaterial” relaciona las clases “Curso” (la cual ya se encontraba en la ontología actual) y la clase “Material” para de tal forma representar e identificar que material se ha compartido para cada curso especificado.

Por último, se incluyeron algunos nodos adicionales para poder cubrir la información que se busca extraer referente a los cursos, profesores y conceptos relacionados al dominio, lo cual será detallado en la siguiente sección.

c) Clases y Sub-Clases:

En esta sección se muestran los diagramas y cuadros que detallan la estructura de objetos y propiedades de la ontología:

En primer lugar, se plantearon las clases principales que representan a entidades de la realidad de la universidad; así como también, las entidades principales dentro del foro como lo son la clase Comentario y Material.

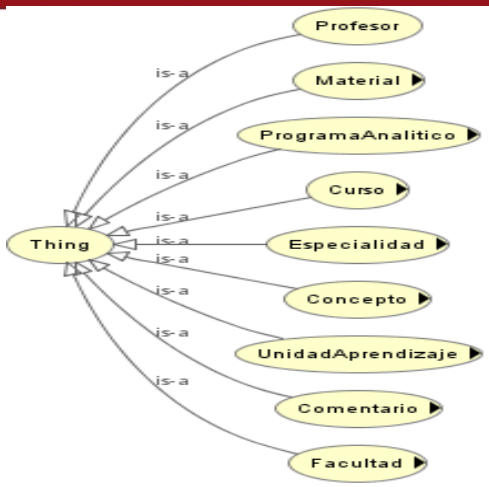

Imagen 9: Estructura global con las clases principales

En el cuadro 9, se procedió a definir las categorías de los comentarios que tiene cada profesor: metodología, enseñanza y asistencia (ver capítulo 3, sección 1.3).

Cuadro 9: Estructura clase Comentario

CLASE COMENTARIO	Asistencia
	Calificacion
	Enseñanza

De igual manera, en el cuadro 10, se definieron los materiales que se comparten dentro del foro para cada uno de los cursos dentro del dominio: exámenes, libros y prácticas.

Cuadro 10: Estructura clase Material

CLASE MATERIAL	Examen
	Libro
	Practica

Con estas categorías y materiales identificados, se podrán clasificar algunas partes del comentario que sean considerados dentro de alguna de ellos para poder tener mayor conocimiento de la información que se logró obtener.

Como siguiente paso, en los cuadros 11 y 12, se definieron los cursos y profesores, respectivamente, pertenecientes a la Facultad de Estudios Generales Ciencias, los cuales representan las fuentes principales que se buscarán identificar dentro de los comentarios para la realización del análisis de correferencias. Asimismo, en el cuadro 13, se definieron algunos conceptos referentes al dominio. Dichos cuadros, presentan términos ambiguos que serán resueltos por el mecanismo que se desarrollará en la siguiente sección.

Cuadro 11: Estructura clase Curso

CLASE CURSO	Algoritmia
	Calculo
	CienciaFilosofia
	CircuitosDigitales
	CircuitosElectricos
	DesarrolloHabilidades
	DibujoIngenieria
	Estadistica
	Estatica
	EstrategiaEstudioEficaz
	EstructurasDiscretas
	EstudioTrabajo
	Filosofia
	MatematicasBasicas
	MecanicaIngenieros
	Quimica
	RedaccionComunicacion
	IntroduccionIngenieriaElectronicaTelecomunicaciones
	Fisica
	FundamentosProgramacion
	IntroduccionComputacion
	IntroduccionComunicacionEscrita
	IntroduccionFisicaUniversitaria
	IntroduccionIngenieriaCivil
	TecnicasProgramacion
	IntroduccionIngenieriaGeologicaMinas
	IntroduccionIngenieriaIndustrial
	IntroduccionIngenieriaInformatica
	IntroduccionIngenieriaMecanicaMecatronica
	IntroduccionMatematicasUnivervistarias
	LeguajesProgramacion
	LiderazgoGestionEmpresarial
	MotivacionLiderazgoPersonal
	SistemasOperativos
	Sociologia

Cuadro 12: Estructura clase Profesor

CLASE PROFESOR	AccostupaJuan	GomezMaria	OrozcoRichard
	AgapitoRuben	GomezSophia	PavletichSergio
	AguileraCesar	GonzagaMiguel	PereyraPatrizia
	AllasiDavid	GonzalesMariao	PhanJose
	AltunaMartin	GuaniraJuan	PiaggioMiguel
	AlvaFernando	GuerraCesar	PizarroCarlos
	AngelesLuis	GuerraJorge	PortillaZalatiel
	AtocheWilmer	GuimarayRosa	RamirezJuan
	AñiAdriana	GutierrezJulio	RamirezVictoria
	BaldeonJohan	GuzmanAbimael	RauJose
	BancesDiana	HadzichMiguel	RenwickRicardo
	BancesRicardo	HenostrozaJose	RiquerosJose
	BarMarco	HinojosaHilmar	RiveraJose
	BelloAlejandro	HirshLayla	RobertsonKarem
	BeltranAndres	HuaragEduardo	RoblesJuana
	BenavidesJorge	IriarteLuis	RodriguezSusana
	BerrocalJorge	JaureguiHernan	RomeroSilvana
	BossioStefano	JaureguiJohn	RoncalAna
	BrahimDaniella	JimenezJuan	RosalesEmiliano
	BravoLuis	JonesJoel	RosalesJose
	BringasLuis	KhlebnikovViktor	RubioManuel
	CaldasIvan	LandaIsabel	RubioNorma
	CaluaLuis	LeonRuben	RuedaDandy
	CamposJavier	LeyvaVanessa	RuisJenniel
	CarlosAlexander	LiraJuan	SalazarJorge
	CaroArnulfo	LunaAna	SanchezEder
	CastilloHernan	LunaMaritza	SanchezRoy
	ChamorroRoberta	LunaWalter	SantosDennis
	ChauNorberto	MadridEricka	SaraviaNancy
	ChavezNoelia	MancillaCesar	SichaAlberto
	ChongMiguel	MassoniEduardo	SosaCarlos
	ChuquinFrank	MatosMariela	TapiaCarlos
	CisnerosVictor	MauchiBeatriz	TaveraMaria
	CorralesCesar	MazaIly	TorresLuis
	CorteganaHumberto	MedinaNelida	ValderramaAna
	DiazEdwar	MelgarHector	ValdiviaAna
	DiazWilson	MendozaAldo	VassalloEttore
	DonofrioSandro	MendozaCesar	VelaJulio
	EspinozaNancy	MestanzaAdalberto	VelardeLuis
	EstevesMaria	MontealegreJuan	VerasteguiTeodulo
	EzcurraAlvaro	MontenegroFlor	VilcaFernando
	FarfanJonathan	MonteroGualberto	VilcapomaLuis
	FerreyraAna	MontesHernan	VilelaPablo
	FigueroaChristiam	MontesinosEnrique	VillagomezDiego
	FloresBerry	MoralesEmma	VillogasEdwin
	FloresDonato	MoscosoRichard	ZapataClaudia
	FloresJose	MurguiaDanny	ZapataJesus
	FloresMaria	MurilloBarulio	ZarateJennifer
	FourmentKatherine	NeciosupHernan	ZegarraKatia
	FrancoRosendo	ObregonJose	ZeladaRosio
	GalvezGonzalo	OrihuelaFreri	ZevallosRodrigo

Cuadro 13: Estructura clase Concepto

CLASE CONCEPTO	AplicacionesArboles
	AplicacionesArbolesBinariosBusquedaGrafos
	AplicacionesColas
	AplicacionesListas
	AplicacionesPilas
	AplicacionesRecorridosMetodosEspecialesArbolesGrafos
	ArchivosBinarios
	ArchivosDescriptor
	ArchivosTexto
	AsistenciaClases
	AsistenciaExamen
	AsistenciaLaboratorios
	AsistenciaPractica
	AYPCadenasCaracteres
	AYPPunterosFunciones
	AYPPunterosGenericos
	BusquedaBinaria
	BusquedaFuncionesHash
	BusquedaSecuencial
	BusquedaTablasHash
	BusquedaTransformacionLlaves
	BusquedaTratamientoColisiones
	CalificacionExamen
	CalificacionLaboratorio
	CalificacionPractica
	DerivacionCalculoPredicados
	DerivacionLenguajeOrdenesConGuarda
	EnseñanzaDinamica
	EnseñanzaMetodología
	ExamenFinal
	ExamenParcial
	OrdenacionArchivos
	OrdenacionInsercion
	OrdenacionIntercambio
	OdenacionMonticulos
	OdenacionSeleccion
	OrdenamientoRapido
	POOAtributos
	POOClases
	POOMetodos
	POOSobrecarga
	PracticaCalificada
	PracticaEjercicio
	PracticaLaboratorio
	PracticaProblema
	ProcesosExclusionMutua
	ProcesosMonitor
	ProcesosSemaforos
	TADColas
	TADConjuntos
	TADGrafos
	TADListas
	TADPilas
	TADVectores

Por último, dentro del dominio de conocimiento se presentan otras estructuras que ya existían dentro de la ontología y no han sido modificadas, las cuales complementarán el dominio desarrollado.

En primer lugar, se plantean los programas analíticos identificados por el código que tiene cada curso, donde se describen el contenido curricular de dichas asignaturas.

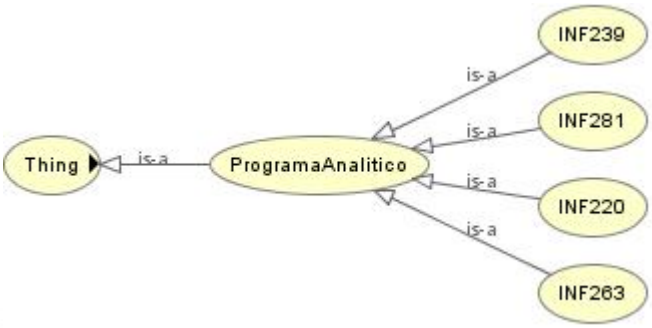

Imagen 10: Estructura clase ProgramaAnalitico

Por otro lado, se definieron algunas facultades que pertenecen al dominio de la universidad, las cuales a través de propiedades permitirá identificar a que facultad pertenece un curso.

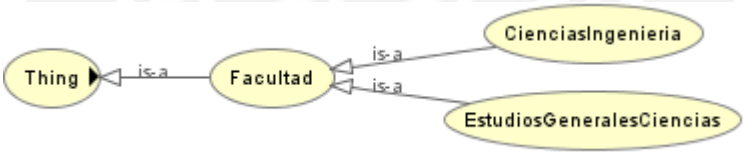

Imagen 11: Estructura clase Facultad

Asimismo, se definieron las especialidades que pertenecen a cada una de las facultades definidas previamente. De tal forma que se podrán identificar los cursos que pertenecen a cada una de ellas.

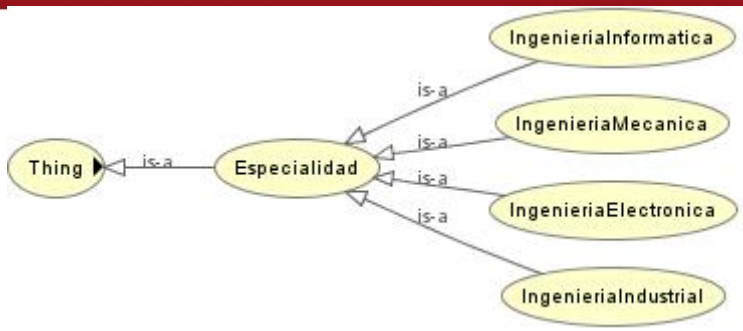

Imagen 12: Estructura clase Especialidad

Por último, se definió la siguiente clase que contiene el conocimiento respecto a las unidades aprendizaje de cada uno de los conceptos definidos en el cuadro 13. Asimismo, estas unidades de aprendizaje estan relacionadas a los programas análiticos de cada curso.

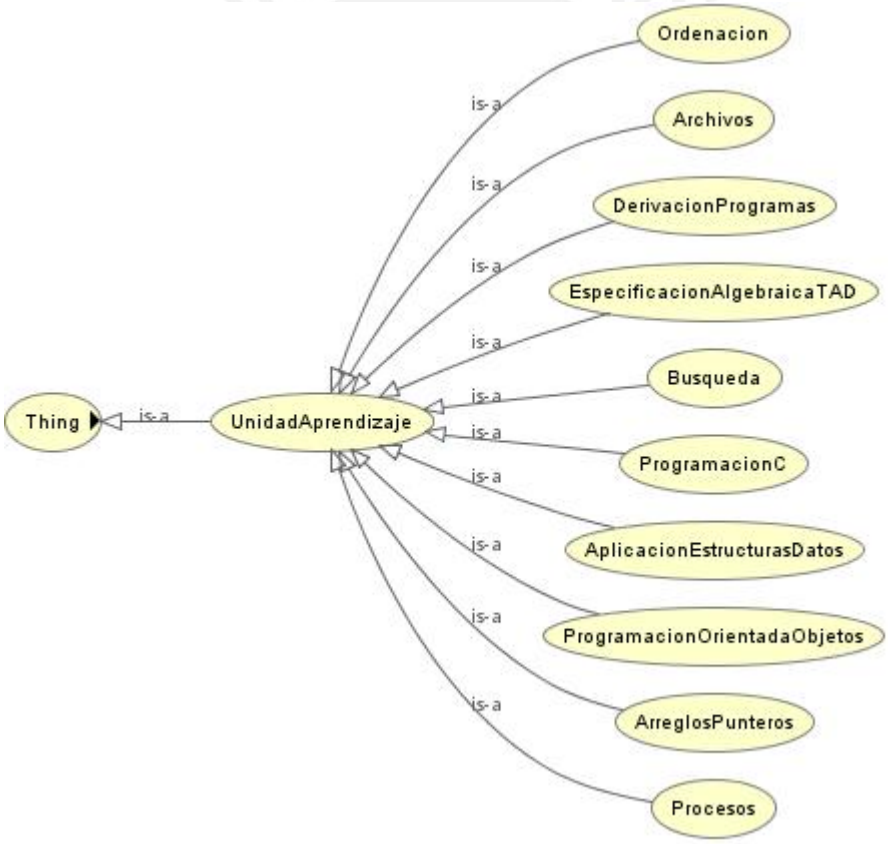

Imagen 13: Estructura clase UnidadAprendizaje

2 Resultado Esperado 2: Módulo de desambiguación lexical de palabras en lenguaje natural dentro de un dominio usando ontologías.

2.1 Introducción

Este objetivo está orientado a atacar otra de las causas identificadas como raíz de la problemática, la cual es la dificultad de los procesamientos en lenguaje natural para lidiar con la ambigüedad de las palabras.

Un caso concreto de ambigüedad se puede identificar en la homonimia; por ejemplo, al tener el término “Bances” dentro del comentario. Esta palabra puede hacer referencia al concepto de la acción de “Introducción”, al concepto de evaluación dentro del dominio de cursos de una universidad. Inclusive, dentro del dominio mencionado, el término “Bances”, puede hacer referencia a distintos términos según el dominio; por ejemplo, “BancesDiana” o “BancesRicardo”.

Lo que se espera con el desarrollo de este objetivo es poder identificar el concepto al que está asociada cada palabra del comentario de un usuario en caso este sea ambiguo.

2.2 Resultado Alcanzado

El módulo de desambiguación permitirá determinar los conceptos a los que hace referencia los términos que puedan considerarse ambiguos dentro de cada comentario,. Para ello se basó el desarrollo de este mecanismo según lo planteado en (Zhao, Zhixian et al. 2013), donde a partir de una ontología y las relaciones planteadas en ella, se resuelve el problema de desambiguación léxical.

Para este fin, se ha hecho uso de los términos complementarios (términos denominados no-ambiguos respecto a los términos ambiguos analizados) del comentario. Por ejemplo:

“Se debe ir estudiando para el curso de cálculo 1, porque El profesor Bances califica bajo”

Como se puede observar en el ejemplo, el término ambiguo identificado sería la palabra “Bances” y sus términos complementarios serían todas las otras palabras que componen el comentario.

Para este caso, se utilizará una ontología de dominio sobre la cual se identificará los posibles conceptos a los que puede estar haciendo referencia el término ambiguo en análisis, y con apoyo de los términos complementarios se encontrará término que representará el concepto más apropiado para la palabra ambigua.

A partir de la imagen 15, se puede observar que cada comentario pasará por el siguiente proceso:

- En el primer nivel, cada palabra del comentario pre-procesado pasará por una primera evaluación que verifica si dicho término es ambiguo y si no es considerado como una abreviatura. En otras palabras, es considerado como ambiguo si se encuentra más de una vez en nodos diferentes de la ontología de dominio y es considerado como abreviatura si la cantidad de caracteres de su versión *lemmatizada* no es menor o igual a 5. Si es que dicho término solo es encontrado una vez, no es necesario que dicha palabra pase por el proceso de desambiguación, caso contrario, se procede a calcular la similitud de cada una de esas coincidencias retornadas.

- A continuación, en el segundo nivel, se obtienen los nodos con los cuales el término ha sido considerado ambiguo.

- Luego, en el tercer nivel, para cada uno de los términos complementarios se realiza lo siguiente:

    - el cálculo de la similitud se realiza con apoyo de los términos complementarios, los cuales se encuentran representados por todas las palabras que rodean al término ambiguo, dentro del comentario. Dicho cálculo consiste en realizar una evaluación dentro de cada nodo y los nodos padres de cada uno de ellos; con sus sinónimos, nombrePreferente, nombreFuente y *lemmas*. Además, para poder tener una mejor precisión de este mecanismo, se considerará que mientras más alejado este el término complementario del nodo ambiguo meno será el puntaje; es decir, si el término complementario fuera encontrado en el mismo nodo ambiguo tendrá un peso de 1; en caso sea encontrado en el nodo padre tendrá un peso de 0.5.

- Una vez obtenido el puntaje de similitud para el término complementario a nivel de la ontología, se procede a asignarle un peso debido a la proximidad con que se encuentra el término complementario del término ambiguo. Por ejemplo, como se puede observar en la imagen 14, se asigna un peso según la posición del término complementario; considerando a *"ti"* como el término ambiguo y a *"tk"* como el término complementario.

$$[\, t_k, t_{i-4}, t_{i-3}, t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}, t_{i+3}, t_{i+4}, t_k]$$

$$f(x) = \begin{cases} 1/(i-k), & k < i \\ 1/(k-i), & k > i \end{cases}$$

Imagen 14: Peso asignado según la posición de los términos

- Por último, en el cuarto nivel, en caso se pudiera detectar un puntaje mayor dentro de los candidatos para resolver la ambigüedad, se reemplazará la forma *lemmatizada* de dicho termino en el comentario.

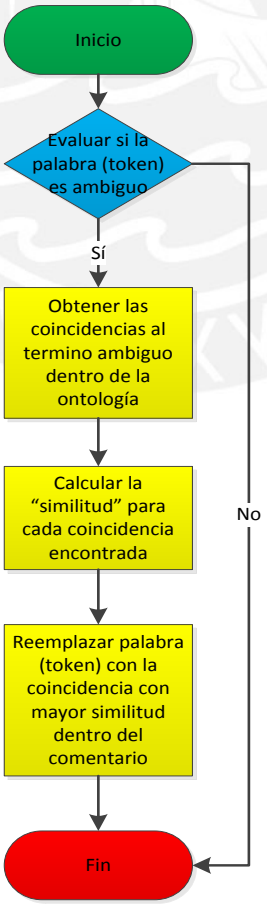

Imagen 15: Mecanismo de desambiguación de palabras.

2.3 Pruebas

Para la verificación del mecanismo se realizaron pruebas con el 20% de la totalidad de los comentarios donde se obtuvo una precisión y *recall* del 76%, debido a que si bien es cierto se lograba identificar todos los términos ambiguos; el mecanismo no podía resolver la ambigüedad para términos donde el puntaje obtenido por los candidatos era el mismo. Por ejemplo, esto ocurría cuando dos términos ambiguos compartían las mismas relaciones, lo cual puede ocurrir cuando dos profesores dictan el mismo curso o dos cursos distintos es dictado por el mismo profesor. En estos casos, los términos complementarios no aportan de manera significativa a la resolución del problema, por lo que no se resuelve la ambigüedad. A continuación se muestra el resultado de una de las pruebas:

Input (Comentario lematizado):

Comentario: deber ir estudiar para el curso calcular 1, porque el profesor bances calificar bajo.

Output (Comentario final):

Comentario: deber ir estudiar para el curso calcular 1, porque el profesor bances_ricardo calificar bajo

Resultado obtenido de evaluación

Comentario: Bances bances bances_ricardo 0.058169014084507043 bances_diana 0.04169014084507043

Como se puede observar en la el resultado mostrado, el término *"bances_ricardo"* obtiene mejor puntaje. Esto se debe a que, según lo mostrado en la imagen 16, se encuentra asociado al término complementario *"Calculo"* cuya versión *lemmatizada* es *"calcular",* por lo que se le brindará mayor puntaje al momento de realizar la evaluación

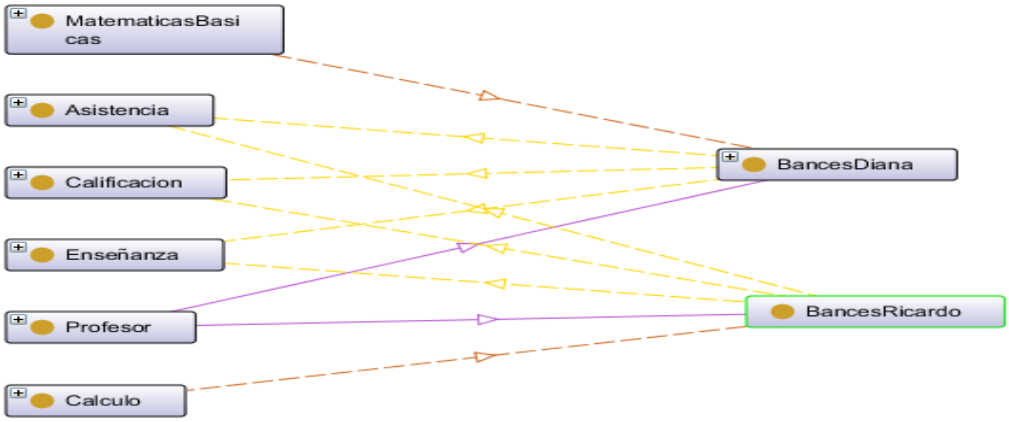

Imagen 16: Diagrama de relaciones según la ontología de dominio

3 Conclusión

En conclusión, con el resultado obtenido se logró representar el conocimiento del dominio de cursos brindados por una universidad, el cual será utilizado durante las siguientes secciones. Además, a partir ello se pudo observar que el alcance y representación del conocimiento depende directamente de la persona que diseña la ontología, lo cual hace que dicha representación del conocimiento pueda ser mejorada o modificada para futuros trabajos.

Asimismo, según lo expuesto se ha podido observar que los resultados cumplen con el objetivo específico planteado, ya que dadas las condiciones limitadas por el comentario se ha podido encontrar un concepto al que puede hacer referencia el término ambiguo, con lo cual se realiza la desambiguación, descartando conceptos que podrían generar ruido innecesario si fueran considerados al momento de realizar la tarea de extracción de información; sin embargo, se debe recalcar que esta tarea depende de la ontología de dominio utilizada; por lo que para futuros trabajos podría aumentarse el número de conceptos ambiguos para el dominio, de tal forma que se logre resolver las ambigüedades que existan para otros conceptos.

CAPÍTULO 5: Resolución de correferencias

En el presente capítulo se desarrollará el tercer objetivo específico del proyecto. Se buscará desarrollar la tarea de extracción de información conocida como resolución de correferencias para poder encontrar todas las menciones en un texto de las fuentes que se buscan extraer. De igual manera, una vez resuelta estas menciones se procederán a almacenar la información obtenida en un modelo estructurado de datos. Además, mediante las métricas precisión, *recall* y valor-f se mostrará la eficiencia del sistema desarrollado.

A continuación se mostrarán los resultados obtenidos, resaltando el método encontrado para la implementación de la resolución de correferencias en el presente proyecto y se presentarán los resultados del desarrollo de las métricas planteadas para la evaluación del rendimiento del sistema.

1 Resultado Esperado 1: Módulo de resolución de correferencias que permita identificar una misma entidad en distintas partes del comentario.

1.1 Introducción

El objetivo del presente resultado es encontrar todas las referencias de una fuente específica dentro de un comentario, de tal forma que se pueda identificar dichas menciones para realizar la extracción de información. Para lograr esta tarea, se realizó una recuperación de las fuentes dentro del texto que se busca analizar, lo cual permite identificar cuáles son las fuentes primarias y secundarias dentro de un texto escrito en lenguaje natural. Una fuente primaria se puede definir como las entidades dentro del mundo real que se busca extraer las cuales pueden tener diferentes menciones dentro de un texto. Por otro lado, las fuentes secundarias son algunas categorías que permitirán clasificar la información extraída a través del proceso de extracción de información. Además, se identifican posibles fuentes que son consideradas omitidas, debido a que hacen referencia a alguna fuente primaria o secundaria en el texto, pero la oración no contiene un término que se pueda evaluar directamente para recuperar la fuente. Por último, las otras fuentes son consideradas como fuentes de referencia, las cuales se analizarán para determinar la correferencia con las fuentes primarias y secundarias. Por ejemplo, en el siguiente comentario:

"[Calculo 1] ***fuente primaria*** *es muy sencillo si practicas al menos media hora al día. []* ***fuente omitida de cálculo 1*** *Se está dictando en el 2014-1. [Este curso]* ***fuente***

***referencia**, puede ser llevado junto con [Física 1] **fuente primaria** sin problemas. A continuación les comparto dos [practicas] **fuente secundaria** del ciclo 2013-2"*

Por otro lado, una vez identificadas las fuentes a analizar el siguiente paso es realizar la resolución de correferencias, mediante la cual se buscará las relaciones entre dichas fuentes, con el fin de obtener mayor cantidad de información de la fuente primaria. Una vez obtenido este resultado se almacenará la información obtenida en un modelo de datos que permitirá utilizar estos datos para otros fines.

1.2 Resultado Alcanzado

Para lograr el objetivo planteado, se utiliza como base lo propuesto en (Acerenza, Rabosto et al. 2012), donde se presenta una aplicación heurística a partir de la información morfosintáctica y semántica para determinar la relación entre las fuentes de un texto. Sin embargo, a diferencia de esta solución, este módulo resolverá las relaciones entre las entidades en base a la ontología desarrollada en el capítulo 4.

Como se puede observar en la imagen 15, el análisis de correferencias consta de dos fases para poder realizar el análisis de correferencias:

a) Identificación y recuperación de fuentes

En el primer nivel, se realiza la identificación de las fuentes que se buscan extraer de cada uno de los comentarios. Para esto, se definieron los siguientes tipos de fuentes:

- Fuente primaria: Son los cursos y profesores dentro de la ontología los cuáles serán las fuentes de las que se buscará extraer información.
- Fuente secundaria: Son los materiales y comentarios dentro de la ontología los cuales permitirán calificar algunas partes de los comentarios.
- Fuente omitida: Son fuentes que en principio no se encuentran en el texto, pero son recuperadas a través del sentido que tiene la oración.
- Fuente referencia: Son nombres, verbos, adjetivos, numerales, adverbios y fuentes omitidas que son candidatos a correferir con la fuente primaria y secundaria.

En caso la fuente primaria o secundaria sea encontrada mediante un numeral o un adverbio, esta serán calificadas como especiales, debido a que en el dominio, pueden representar una fuente de manera directa (ver capítulo 3, sección 1.3).

Según lo mencionado anteriormente, el funcionamiento de esta fase es el siguiente.

- A partir de la lista de palabras que se tiene como resultado el módulo de pre-procesamiento, para cada nombre (propio o común), verbo, adjetivo, adverbio y numeral dentro de dicha lista; se verifica si alguna de las palabras en su forma *lemma* se encuentran dentro de la ontología en alguna de las propiedades de las clases de profesores o cursos.

- Si se encuentra es calificada como una fuente primaria. En caso no encontrarse, se verifica nuevamente si se trata de información que se busca extraer; es decir, si se encuentra en las propiedades de las clases comentario o material. De ocurrir esto será calificada como una fuente secundaria.

- Por último, si no se encontrase en ninguno de los casos mencionados anteriormente la palabra será calificada como fuente de referencia.

- Una vez realizada dicha clasificación se pasa a identificar posibles fuentes omitidas dentro de las opiniones; las candidatas a ser este tipo de fuente son las fuentes de referencia. Para lograr esta tarea, se tiene en cuenta el cuadro 14, asignando una puntación positiva o negativa según cada caso.

Cuadro 14: Criterios aplicados para encontrar fuentes omitidas

Puntuación postiva	Puntuación negativa
fuente referencia coincide en genero o numero con la fuente principal	fuente referencia no coincide en número o genero con la fuente principal
fuente referencia es una persona	
fuente referencia se encuentra en alguna fuente principal	
fuente referencia se encuentra en la misma oración que alguna fuente principal	

En este caso el candidato con mayor puntaje será el que sea reemplazado con la mención de la fuente primaria y será etiquetado como *"fuente omitida"*. Para poder recuperar una fuente se estableció

un puntaje mínimo que debe tener el candidato para considerarse como una fuente omitida; el cual para fines de este proyecto se estableció con un valor de dos, debido a que esta tipo de fuente al ser omitida debe de tener una referencia de género y número con alguna fuente ya existente, de no ser el caso al menos deberá cumplir todas las otras reglas.

b) Resolución de correferencias

En esta etapa se buscará analizar los comentarios buscando correferencias entre sus fuentes; es decir, cuando las fuentes identificadas previamente hacen referencia a la misma entidad del mundo real.

Se utiliza un enfoque de asignación de puntajes de forma de penalizar o premiar a los candidatos de ser antecedentes de una fuente dependiendo sus características particulares. Para esto se busca la correferencia entre las siguientes fuentes:

- Las fuentes de referencia o fuentes omitidas serán candidatas para correferir con una fuente primaria o fuente secundaria mencionadas antes de dicha fuente.
- Las fuentes secundarias serán candidatas para correferir con una fuente primaria mencionadas antes de dicha fuente.

Como se puede observar, en la imagen 17, a partir del segundo nivel, para cada una de las fuentes candidatas se siguen los siguientes pasos según lo planteado en (Acerenza, Rabosto et al. 2012):

- Se asignan puntajes a cada una de las fuentes con las que puede correferir. Para este análisis, se aplican criterios sintácticos y semánticos de forma de otorgar puntajes a cada fuente candidata. Como se puede observar en el cuadro 15, se establecen ciertos criterios a partir de lo planteando en (Acerenza, Rabosto et al. 2012), donde se muestra las categorías de las fuentes y los criterios utilizados para cada uno, asimismo se establece un impacto según ocurrencia: ocurre / no ocurre

- En base a los puntajes obtenidos se determina con cuál fuente correfiere la fuente candidata. Para esto, se considera como correferente la fuente que tenga mayor puntaje, solo si, este puntaje es mayor a cero; caso contrario la fuente candidata no tiene correferencia con ninguna de las fuentes que fueron analizadas.

Cuadro 15: Criterios aplicados para resolución de correferencias

Criterio	Fuente especial	Fuente Omitida	Pronombre	Adjetivo	Verbo	Nombre
Proximidad*	+4	+3	+2			
Concordancia género			+1/-3	+1/-2	+1/-2	+1/-2
Concordancia número			+1/-3	+1/-2	+1/-2	+1/-2
Concordancia persona			+1/-3	+1/-2	+1/-1	+1/-1
Concordancia clase semántica				+1/-2	+1/-1	+1/-1
Concordancia lemmas					+2	+2

* La proximidad se mide según la cantidad de palabras que se encuentren entre dos fuentes; es decir, serán consideradas como próximas aquellas se encuentren separadas por menos palabras

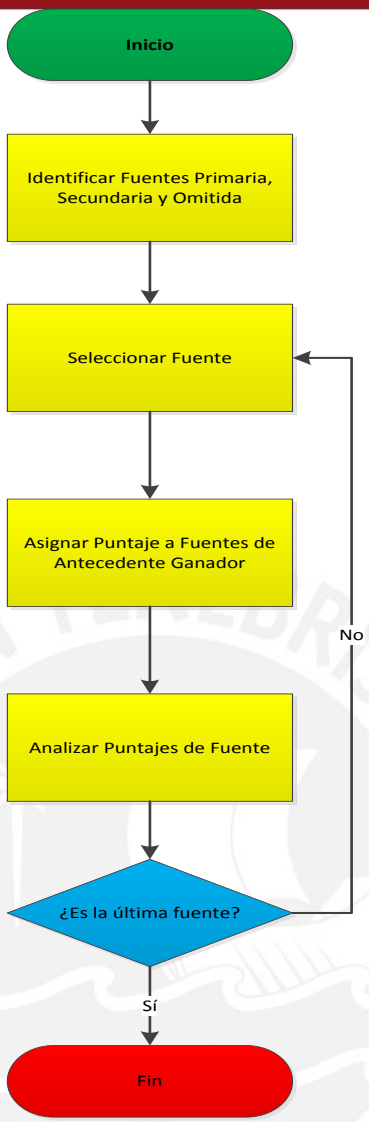

Imagen 17: Mecanismo de resolución de correferencias

c) Almacenamiento de la información

A partir de la información obtenida en las secciones anteriores se procede a almacenar el resultado, en una base de datos estructurada. La cual contiene la información de los materiales brindados por los cursos y las opiniones respecto a cada uno de los profesores que se establecieron en secciones anteriores.

En la estructura que se presenta a continuación se almacenarán las opiniones respecto a los profesores y el material que se ha compartido de cada curso. Es importante mencionar que una opinión es una parte del comentario original mas no el comentario completo.

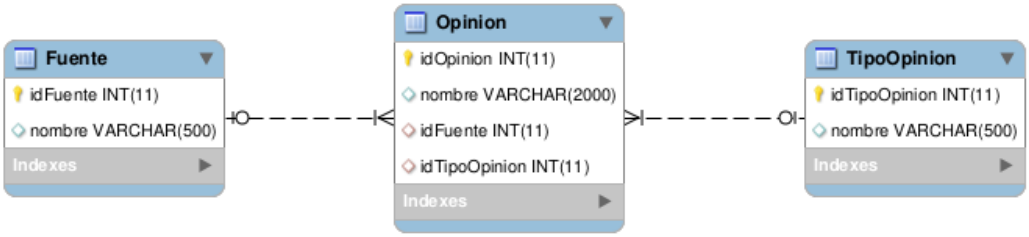

Imagen 18: Modelo de datos para el almacenamiento de la información.

1.3 Pruebas

Para la verificación del mecanismo se realizaron pruebas con el 20% de la totalidad de los comentarios donde se obtuvo una precisión de 74% y *recall* de 77%. Esto se debe a que si bien es cierto, se recuperaban la gran mayoría las fuentes primarias y secundarias en el texto, se encontró que se identificaban fuentes que no estaban dentro del texto, debido a que se trataban de algunos errores ortográficos que no habían sido resueltos en los módulos anteriores. Esto ocasiona que se encuentren correferencia erróneas para las fuentes calificadas incorrectamente como primarias, por consiguiente la precisión es menor a lo esperado. Sin embargo, al ser menor la cantidad de fuentes que en realidad existen, el *recall* que se obtuvo para este mecanismo fue aceptable. A continuación se muestra el resultado de una de las pruebas:

Input (Comentario original):

Comentario:

es te material es del qerido profesor pavletich espero les sirva atoa lagent chkenlooooo
attachment=637:Diapos Pavletich.rar.
gracias, me servirá bastante para el ciclo!!!.
siii muchas graciass.
genial justo lo que buscaba gracias! img src= http: enlacato.com public style_emoticons #EMO_DIR# wink.png class= bbc_emoticon alt= ;

Output (correferencias):

Comentario:

<fuente principal> pavletich 1 sergio_pavletich </fuente principal>
<opinion> </opinion>
<fuente referencia> es </fuente referencia>
<opinion> es </opinion>
<fuente referencia> te </fuente referencia>
<opinion> te </opinion>
<fuente referencia> material </fuente referencia>
<opinion> material </opinion>
<fuente referencia> es </fuente referencia>
<opinion> es de </opinion>
<fuente referencia> él </fuente referencia>
<opinion> él </opinion>
<fuente referencia> qerido </fuente referencia>
<opinion> qerido </opinion>
<fuente referencia> profesor </fuente referencia>
<opinion> profesor </opinion>
<fuente referencia> espero </fuente referencia>
<opinion> espero </opinion>
<fuente referencia> les </fuente referencia>
<opinion> les </opinion>
<fuente referencia> sirva </fuente referencia>
<opinion> sirva </opinion>
<fuente referencia> atoa </fuente referencia>
<opinion> atoa </opinion>
<fuente referencia> lagent </fuente referencia>
<opinion> lagent </opinion>
<fuente referencia> chkenlooooo </fuente referencia>
<opinion> chkenlooooo attachment=637:Diapos </opinion>
<fuente referencia> rar </fuente referencia>
<opinion> rar . . </opinion>
<fuente referencia> gracias </fuente referencia>
<opinion> gracias </opinion>

<fuente referencia> me </fuente referencia>
<opinion> me </opinion>
<fuente referencia> servirá </fuente referencia>
<opinion> servirá bastante para el ciclo&#33 &#33 &#33 . siii muchas </opinion>
<fuente referencia> gracia </fuente referencia>
<opinion> gracia . </opinion>
<fuente referencia> genial </fuente referencia>
<opinion> genial </opinion>
<fuente referencia> justo </fuente referencia>
<opinion> justo lo </opinion>
<fuente referencia> que </fuente referencia>
<opinion> que </opinion>
<fuente referencia> buscaba </fuente referencia>
<opinion> buscaba gracias&#33 </opinion>
<fuente referencia> img </fuente referencia>
<opinion> img </opinion>
<fuente referencia> src </fuente referencia>
<opinion> src </opinion>
<fuente referencia> http </fuente referencia>
<opinion> http </opinion>
<fuente referencia> enlacato </fuente referencia>
<opinion> enlacato . </opinion>
<fuente referencia> com </fuente referencia>
<opinion> com public </opinion>
<fuente referencia> style_emoticons </fuente referencia>
<opinion> style_emoticons </opinion>
<fuente referencia> EMO_DIR </fuente referencia>
<opinion> EMO_DIR </opinion>
<fuente referencia> wink </fuente referencia>
<opinion> wink . </opinion>
<fuente referencia> png </fuente referencia>
<opinion> png </opinion>
<fuente referencia> clase </fuente referencia>
<opinion> clase </opinion>
<fuente referencia> bbc_emoticon </fuente referencia>
<opinion> bbc_emoticon </opinion>
<fuente referencia> alt </fuente referencia>

<opinion> alt </opinion>
<fuente principal> Pavletich 1 sergio_pavletich </fuente principal>
<opinion> . </opinion>

Como se puede observar en el resultado, se identifican las menciones a la fuente primaria *"sergio_pavletich",* lo cual permite encontrar las correferencias para dicha fuente, a partir de las fuentes referencias identificadas. De igual manera se puede observar que existen algunas palabras que generan ruido al momento de realizar este análisis como lo son los emoticones y algunos caracteres especiales que no han sido identificados en módulos anteriores, debido a que se encuentran escritos de una manera diferente a lo esperado.

2 Resultado Esperado 2: Análisis de los resultados de las métricas estadísticas de precisión, recall y valor-f aplicadas al sistema de extracción de información.

2.1 Introducción

El objetivo del presente resultado es mostrar el desempeño que tiene el sistema desarrollado a partir de los valores resultantes obtenidos para las métricas precisión, *recall* y valor-f. Asimismo, se describirán las consideraciones que se tuvieron para obtenerlas y la interpretación de los resultados obtenidos.

2.2 Resultado Alcanzado

Para lograr el objetivo de este resultado se consideraron las métricas precisión y *recall.* Además, tal como se muestra en la imagen 19, se consideró la métrica valor-f, en especial el valor-f1, donde tanto la precisión como el *recall* tienen la misma relevancia para evaluar el desempeño del sistema.

$$F_1 = \frac{2PR}{P + R}$$

Imagen 19: Valor-f1 considerado para la medición

A partir de lo planteado anteriormente se obtuvieron los siguientes resultados mostrados en el cuadro 16.

Cuadro 16: Cuadro de métricas para el sistema

	Precisión	Recall	Valor-f
Módulo Preprocesamiento	79%	71%	**75%**
Módulo Desambiguación	76%	76%	**76%**
Módulo Correferencias	74%	77%	**75%**
Sistema Extracción Información	**76%**	**75%**	**75%**

En dicho cuadro se detallan los resultados para cada una de las métricas, los cuales fueron calculados de la siguiente manera:

- **Módulo Preprocesamiento:** Para la obtención de las métricas de este módulo se consideró el desempeño del corrector ortográfico a nivel lexical, para esto se contabilizó la cantidad de errores ortográficos totales en el comentario original, la cantidad de errores que fueron considerados por el mecanismo y la cantidad de errores ortográficos resueltos correctamente por el mecanismo. A partir de esto se obtuvo que la mayor parte de los errores ortográficos que debían ser procesados fueron corregidos exitosamente; sin embargo, dentro del texto original existían mayor cantidad de errores a los que fueron considerados durante la elaboración del mecanismo. Esto último se puede ver reflejado en el cuadro 14, el resultado final mostro que la precisión resultaba mayor al *recall* obtenido.

- **Módulo Desambiguación:** Para la obtención de las métricas de este módulo se consideró el desempeño de la desambiguación de palabras, para esto se contabilizó la cantidad de palabras ambiguas para el dominio en el texto original, la cantidad de palabras ambiguas detectadas por el mecanismo y la cantidad de palabras ambiguas resueltas correctamente por el mecanismo. Debido a que la decisión si un término es ambiguo o no depende enteramente del dominio, las cantidades de palabras ambiguas en texto original y las detectadas por el mecanismo resultaron ser las mismas. Por esta razón, los valores obtenidos para las métricas precisión y *recall*, para los comentarios

analizados también fueron los mismos para la gran mayoría de los comentarios.

- **Módulo Correferencias:** Para la obtención de las métricas de este módulo se consideró el desempeño del análisis de correferencias, para esto se contabilizó la cantidad de referencias a las fuentes primarias en el texto original, la cantidad de referencias a las fuentes primarias detectadas por el mecanismo y la cantidad de referencias que fueron correctamente asignadas a una fuente primaria por el mecanismo. Debido a que dentro del módulo se detectaban algunas fuentes primarias que no existían en el comentario original, causo que la precisión sea baja en algunos comentarios, ya que a estas fuentes se les asignaba referencias de otras fuentes primarias que eran relevantes para el dominio; sin embargo, debido a que se detectaban mayor cantidad de referencias el *recall* obtenido fue mayor en algunos comentarios, lo cual se ve reflejado en el cuadro 14.
- **Sistema Extracción de Información:** Para la obtención de las métricas en general del sistema se realizó un promedio simple de los resultados obtenidos para los módulos descritos anteriormente

Por último, como se puede observar en el cuadro 14, en promedio el sistema tiene un desempeño aceptable, esto depende directamente de cada uno de los componentes del sistema, los cuales aportan de manera significativa a esta métrica.

3 Conclusión

Según lo expuesto, mediante este módulo se podrá obtener y almacenar los datos extraídos de manera estructurada, lo cual permitirá que puedan ser aprovechados por una universidad para un futuro análisis de los resultados obtenidos. Es importante mencionar que algunas fuentes podrán no ser recuperadas debido a que algunas de las reglas no se puedan aplicar al no encontrarse dentro del dominio planteado en la ontología; sin embargo, para futuros trabajos se podría ampliar el dominio abarcado en este proyecto, agregando más sinónimos o *lemmas* para las palabras analizadas u otros miembros a las clases principales asegurando de esta manera que se abarquen nuevos términos.

En primer lugar, se observa que para realizar un análisis de correferencias es necesario que se resuelvan las ambigüedades del dominio, debido a que esto permitirá diferenciar distintos conceptos dentro del dominio, lo cual mejorará el desempeño al momento de identificar fuentes, por tal motivo reducirá los errores por correferencias asignadas a fuentes principales mal identificadas.

Además, se detectó que los errores ortográficos influyen negativamente al desempeño del sistema debido a que dichas palabras no siempre son las mismas y no se aplica al léxico del dominio. Por esta razón, se recomienda que para futuros trabajos se pueda considerar como datos de entrada comentarios que no cuenten con este tipo de errores; es decir, se podrían considerar como datos de entrada artículos o noticias que se encuentren en formato de texto para probar los mecanismos y evaluar nuevamente las métricas que evalúan el desempeño.

Por último, se puede concluir que el análisis del dominio tiene gran relevancia en el desarrollo de este tipo de sistemas. Esto se debe a que, el léxico y el contexto, pueden determinar el sentido que se le da a una palabra dentro del dominio especificado; es decir, sin un conocimiento claro del dominio existiría gran dificultad para poder realizar una tarea de extracción de información. De igual manera, se debe de contar con datos que permitan revisar y evaluar las reglas generadas a partir del análisis inicial del dominio; de tal forma que estas puedan ser modificadas y mejoradas según los resultados obtenidos.

Referencias bibliográficas

Abolhassani, M., et al. (2003). Information extraction and automatic markup for XML documents. Intelligent Search on XML Data, Springer**:** 159-174.

Acerenza, F., et al. (2012). Coreference resolution between sources of opinions in Spanish texts. Informatica (CLEI), 2012 XXXVIII Conferencia Latinoamericana En.

Apache (2014). "Apache Jena." Retrieved 03-06-2014, from https://jena.apache.org/.

Apache (2014). "Apache Lucene." Retrieved 20/08/2014.

Bodendorf, F. and C. Kaiser (2010). Mining Customer Opinions on the Internet - A Case Study in the Automotive Industry. Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on.

Carranza, B. (2014). Diseñar un modelo de recuperación de información usando expansión de consultas basadas en ontologías en el dominio de la currícula en el área de ciencias de la computación para la especialidad de ingeniería informática de una universidad., Pontificia Universidad Católica del Perú.

Castellanos, M., et al. (2012). Intention insider: Discovering people's intentions in the social channel.

Clasesmas (2014). "Enlacato." Retrieved 02-05-2014, from http://enlacato.com/.

Cunningham, H. (1997). "Information extraction-a user guide." arXiv preprint cmp-lg/9702006.

Eclipse (2014). "Eclipse." Retrieved 02-05-2014, from https://www.eclipse.org/home/index.php.

Freeling (2014). "Etiquetas Eagles." Retrieved 22/09/2014, from http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html.

Galvis Carreno, L. V. and K. Winbladh (2013). Analysis of user comments: An approach for software requirements evolution. Software Engineering (ICSE), 2013 35th International Conference on.

Gang, H., et al. (2013). Information Extraction of Forum Based on Regular Expression. Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on.

Gottipati, S., et al. (2011). Finding relevant answers in software forums. Automated Software Engineering (ASE), 2011 26th IEEE/ACM International Conference on.

Gruber, T. R. (1995). "Toward principles for the design of ontologies used for knowledge sharing?" International Journal of Human-Computer Studies **43**(5–6): 907-928.

Hariharan, S., et al. (2010). Opinion mining and summarization of reviews in web forums. Proceedings of the Third Annual ACM Bangalore Conference. Bangalore, India, ACM**:** 1-4.

Hobbs, J. R. (1993). The generic information extraction system. MUC.

Hongjiang, W., et al. (2010). Network online comments and sentiment features analysis. Intelligent Control and Automation (WCICA), 2010 8th World Congress on.

Jingwei, Z., et al. (2012). Forum Data Extraction without Explicit Rules. Cloud and Green Computing (CGC), 2012 Second International Conference on.

Kitchenham, B. A. (2004). Systematic reviews. Software Metrics, 2004. Proceedings. 10th International Symposium on.

Kongthon, A., et al. (2010). Using an opinion mining approach to exploit Web content in order to improve customer relationship management. Technology Management for Global Economic Growth (PICMET), 2010 Proceedings of PICMET '10:.

Machova, K. and T. Penzes (2012). Extraction of web discussion texts for opinion analysis. Applied Machine Intelligence and Informatics (SAMI), 2012 IEEE 10th International Symposium on.

Mambo (2004). "Freeling." Retrieved 02/08/2014, from http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=12&Itemid=41.

Manuel, K., et al. (2010). Analyzing internet slang for sentiment mining.

Mohajeri, S., et al. (2013). Innovative navigation of health discussion forums based on relationship extraction and medical ontologies. Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on.

Oracle (2014). "MySQL :: The world's most popular open source database." Retrieved 28-05-2014, from http://www.mysql.com/.

Oracle (2014). "Oracle Technology Network for Java Developers." Retrieved 30-05-2014, from http://www.oracle.com/technetwork/java/index.html.

Schreiber, G., et al. (1994). "CommonKADS: A Comprehensive Methodology for KBS Development." IEEE Expert: Intelligent Systems and Their Applications **9**(6): 28-37.

Shu, Z., et al. (2009). Opinion Analysis of Product Reviews. Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on.

Soutar, G. N. and J. P. Turner (2002). "Students' preferences for university: a conjoint analysis." International Journal of Educational Management **16**(1): 40-45.

Standford (2014). "Protégé." Retrieved 01-06-2014, from http://protege.stanford.edu/.

Stavrianou, A., et al. (2009). A Content-Oriented Framework for Online Discussion Analysis. Advanced Information Networking and Applications Workshops, 2009. WAINA '09. International Conference on.

Suke, L., et al. (2009). Automatic Data Extraction from Web Discussion Forums. Frontier of Computer Science and Technology, 2009. FCST '09. Fourth International Conference on.

Turmo, J., et al. (2006). "Adaptive information extraction." ACM Computing Surveys (CSUR) **38**(2): 4.

van Delden, S., et al. (2004). Supervised and unsupervised automatic spelling correction algorithms. Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on.

W3C (2012). "OWL 2 Web Ontology Language Document Overview (Second Edition)." Retrieved 03-06-2014, from http://www.w3.org/TR/owl2-overview/.

Wimalasuriya, D. C. and D. Dou (2010). "Ontology-based information extraction: An introduction and a survey of current approaches." Journal of Information Science **36**(3): 306-323.

Yang, C. C., et al. (2007). Analyzing and visualizing gray Web forum structure. **4430 LNCS:** 21-33.

YingJu, X., et al. (2011). An integrated approach for information extraction. Information Science and Service Science (NISS), 2011 5th International Conference on New Trends in.

Yingying, Z. and H. Daqing (2013). Extracting problematic API features from forum discussions. Program Comprehension (ICPC), 2013 IEEE 21st International Conference on.

Yun, W., et al. (2009). Data extraction from Web forums based on similarity of page layout. Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on.

Zechner, K. (1997). "A literature survey on information extraction and text summarization." Computational Linguistics Program, Carnegie Mellon University.

Zhao, L., et al. (2013). OnPerDis: Ontology-Based Personal Name Disambiguation on the Web. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on.