



UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA
La Universidad Católica de Loja

ÁREA TÉCNICA

**TITULACIÓN DE INGENIERÍA EN SISTEMAS INFORMÁTICOS Y
COMPUTACIÓN**

**Servicios Web para Extracción de Entidades desde Contenido HTML.
Piloto en sitios con recursos abiertos OCW**

TRABAJO DE FIN DE TITULACIÓN

AUTOR: Zaruma Sozoranga, Jhonny Alonso

DIRECTOR: Piedra Pullaguari, Nelson Oswaldo, Ing.

LOJA – ECUADOR

2014

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE FIN DE TITULACIÓN

Ingeniero.

Nelson Oswaldo Piedra Pullaguari.

DOCENTE DE LA TITULACIÓN

De mi consideración:

El presente trabajo de fin de titulación: **Servicios web para extracción de entidades desde contenido html** realizado por Zaruma SozorangaJhonny Alonso, ha sido orientado y revisado durante su ejecución, por se aprueba la presentación del mismo.

Loja, junio de 203

f)

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

“ Yo Zaruma Sozoranga Jhonny Alonso declaro ser autor (a) del presente trabajo de fin de titulación: **Servicios web para extracción de entidades desde contenido html**, de la Titulación de Ingeniero en Sistemas Informáticos y Computación, siendo Nelson Oswaldo Piedra Pullaguari director del presente trabajo; y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales. Además certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

Adicionalmente declaro conocer y aceptar la disposición del Art. 67 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad”

f.

Autor: Zaruma Sozoranga Jhonny Alonso

Cédula: 1104811482

ÍNDICE DE CONTENIDO

CARATULA.....	I
APROBACIÓN DEL DIRECTOR DEL TRABAJO DE FIN DE TITULACIÓN	II
DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS	III
ÍNDICE DE CONTENIDO	IV
ÍNDICE DE FIGURAS.....	VI
RESUMEN	1
ABSTRACT	2
INTRODUCCIÓN	3
OBJETIVOS.....	4
1. General	4
2. Específicos.....	4
MARCO TEÓRICO	5
1. La web	6
1.1. Web de datos.....	6
1.2. Web semántica.....	6
1.3. Metadatos y anotación semántica.....	8
1.4. Rdf.....	9
1.5. Sparql.....	10
1.5.1. Consultas simples.....	10
1.5.2. Sintaxis sparql.....	12
1.6. Dbpedia.....	13
1.6.1. El Dataset de dbpedia.....	14
2. Recursos educativos abiertos ocw	14
2.1. Oer (open educational resources).....	14
2.2. Ocw.....	15
3. Procesamiento del lenguaje natural.....	16
4. Servicios web	17
4.1. Arquitectura rpc (remote procedure call).....	18
4.2. Arquitectura orientada a servicios.....	19
4.3. Arquitectura rest.....	20
5. Programas y librerías.....	21
5.1. Python.....	21
5.1.1. Nltk.....	21
PROBLEMÁTICA.....	24
1. Problemática actual	25
SOLUCIÓN	28
1. Aproximación	29
2. Descripción de componentes.....	29
2.1. Vocabulario.....	29
2.2. Knowledge Base(Base de Conocimiento).....	29
2.3. Locwd.....	30
2.4. Ws tokenización.....	30
2.5. Ws Extracción de Entidades.....	30
2.6. Ws Desambiguación y Enlace.....	31
2.7. App cliente.....	31
3. Ontología y Vocabulario para descripción de Servicios Web.....	31

DESARROLLO	33
4. Ws tokenización.....	33
4.1. Arquitectura.....	33
4.1.1. Diagrama de secuencia.....	33
4.2. Implementación.....	34
4.2.1. Parámetros.....	34
4.2.2. Funciones.....	34
5. Ws desambiguación y enlace.....	43
5.1. Arquitectura.....	43
5.1.1. Diagrama de secuencia.....	44
5.2. Implementación.....	45
5.2.1. Parámetros.....	45
5.2.2. Funciones.....	45
6. App Cliente.....	50
6.1. Implementación.....	50
6.1.1. Parámetros.....	50
VALIDACIÓN Y PRUEBAS.....	52
7. Validación.....	53
7.1. Prueba de Funcionalidad.....	53
7.1.1. Objetivo.....	53
7.1.2. Escenario.....	53
7.1.3. Pruebas sobre el Servicio Web Tokenizar.....	53
7.1.1. Pruebas sobre el Servicio Web Extraer Entidades.....	53
7.1.1. Pruebas sobre el Servicio Web Desambiguar Linkear.....	54
DISCUSIÓN.....	55
8. Discusión.....	56
CONCLUSIONES.....	57
RECOMENDACIONES.....	59
BIBLIOGRAFÍA.....	61
ANEXO 1: ONTOLOGÍA PARA LA DESCRIPCIÓN DE LOS SERVICIOS WEB.....	63
ANEXO 2: ESPECIFICACIONES DE CASOS DE USO.....	74
ANEXO 3: DIAGRAMAS DE SECUENCIA.....	87

ÍNDICE DE FIGURAS

FIGURA 1 ARQUITECTURA DE LA WEB SEMÁNTICA.....	6
FIGURA 2 PARTES DE LA ORACIÓN EN RDF.....	9
FIGURA 3 REPRESENTACIÓN GRAFICA RDF	9
FIGURA 4 CONJUNTO DE ETIQUETAS	22
FIGURA 5 EJEMPLO DE EXPRESIONES REGULARES PARA ETIQUETADO CON REGEXPTAGER.....	22
FIGURA 6 SEGMENTACIÓN Y ETIQUETADO CHUNK	23
FIGURA 7 ESTRUCTURA DE LOS COMPONENTES.....	29
FIGURA 8 GRAFICO DE LA ONTOLOGÍA PARA LA DESCRIPCIÓN DE LOS SERVICIOS WEB.....	31
FIGURA 9 DIAGRAMA DE SECUENCIA DE LA FUNCIONALIDAD TOKENIZAR.....	33
FIGURA 10 DIAGRAMA DE SECUENCIA DE LA FUNCIONALIDAD DESAMBIGUAR Y ENLAZAR.....	44
FIGURA 11 APP CLIENTE.....	51
FIGURA 12 DIAGRAMA DE SECUENCIA TOKENIZAR.....	88
FIGURA 13 DIAGRAMA DE SECUENCIA EXTRAER ENTIDADES	89
FIGURA 14 DIAGRAMA DE SECUENCIA DESAMBIGUAR Y ENLAZAR.....	90

RESUMEN

El presente proyecto tiene como principal objetivo la extracción de entidades, para que sirvan como metadatos que representaran el contenido Html o texto, el piloto se lo ejecutara sobre contenidos OCW.

Para lo cual se desarrollaron servicios web cuya función serán: tokenizar, extraer entidades, desambiguar y enlazar con LOD-Cloud que en nuestro caso será Dbpedia, estos servicios web se los desarrollara y están descritos con anotaciones semánticas, para que puedan permitir la interoperabilidad entre los servicios existentes.

PALABRAS CLAVES: Extracción de entidades, Web Semántica, Anotación semántica, OCW, Dbpedia.

ABSTRACT

This project's main objective is the extraction of entities to serve as metadata to represent the Html content or text, the pilot would execute on OCW content.

To which web services are developed whose function: tokenize, extracting entities, and link to disambiguate LOD-cloud which in our case is dbpedia, these web services are the developed and are described in semantic annotations, that can enable interoperability between existing services.

KEYWORDS: Extraction of entities, Semantic Web, Semantic annotation, OCW, Dbpedia.

INTRODUCCIÓN

La web conforma un mundo de datos, información y conocimiento casi siempre nos encontramos con la dificultad de encontrar la información que realmente necesitamos. Existen muchos buscadores y algoritmos de búsqueda, pero falta mucho por recorrer para poder llegar a automatizar búsqueda y recuperación de información mediante búsquedas inteligentes.

Por lo tanto la asignación de metadatos y etiquetas es de gran importancia si se quiere tener una búsqueda inteligente en la web semántica. Con esto se encuentra la necesidad de una herramienta de extracción de meta información de calidad.

El presente trabajo se enfoca en el procesamiento de los recursos educativos abiertos OCW (Open Course Ware por sus siglas en inglés) que son una de las iniciativas educativas muy importante, ya que son libre acceso a una gran diversidad de recursos y materiales de cursos universitarios de forma gratuita e ilimitada.

En la sección de Marco Teórico se describe lo que es La web de datos, la web semántica como también que son los metadatos, las anotaciones semánticas, rdf, dbpedia, Sparql; se define también lo que son los OCW, lo que es el procesamiento de lenguaje natural, también se describe las arquitecturas de servicios web mas utilizadas. Y finalmente se describe los programas y herramientas utilizadas.

En la sección Problemática se describe en detalle el problema a resolverse que dio pie a la realización este proyecto.

En la sección Solución se detalla la aproximación de lo que se va a realizar para dar resolución al problema encontrado. Así como también se describe los componentes creados para la resolución. Y finalmente una descripción a detalle de los Servicios Web creados.

Al final se presenta las conclusiones y recomendaciones del proyecto realizado.

OBJETIVOS

1. General

- Desarrollar Servicios Web que extraigan entidades a partir del contenido HTML, el piloto se ejecutará sobre contenidos OCW.

2. Específicos

- Creación de Servicio Web para Tokenización.
- Creación de Servicio Web para Extracción de entidades.
- Creación de Servicio Web para Desambiguación y Enlace con LOD-Cloud.
- Creación de App Cliente, para integración de los Servicio Web.

MARCO TEÓRICO

1. La web

La Web se define simplemente como el universo de información accesible desde la red global. Se trata de un espacio abstracto con el cual las personas pueden interactuar, actualmente está poblado por páginas interconectadas que contienen texto, imágenes, animaciones y videos. Su existencia marca el final de una era de incompatibilidades frustrantes y debilitantes entre sistemas informáticos. (Berners-Lee, 1996)

El objetivo de la web es ser un espacio de información compartida a través del cual las personas (y máquinas) puedan comunicarse. (Berners-Lee, 1996)

1.1. Web de datos.

La Web de Datos o Linked Data permite pasar de una Web en la que los recursos son documentos HTML (en la que el usuario humano es el destinatario de la información publicada), a una Web de Datos Enlazados que están expresados en RDF.

En la actualidad la Web de Datos cuenta con información vinculada sobre organizaciones, autores, áreas de conocimiento, licencias, países, tipo de medios, etc. (Piedra, Tovar, López, Chicaiza, & Martinez, 2011)

1.2. Web semántica.

El creador del concepto, Tim Berners-Lee, define la web semántica de la siguiente manera: “no es una web separada sino una extensión de la actual, donde la información está dotada de un significado bien definido, los ordenadores están mejor capacitados y las personas trabajan en colaboración” (La Web Semántica y las Tecnologías del Lenguaje Humano)

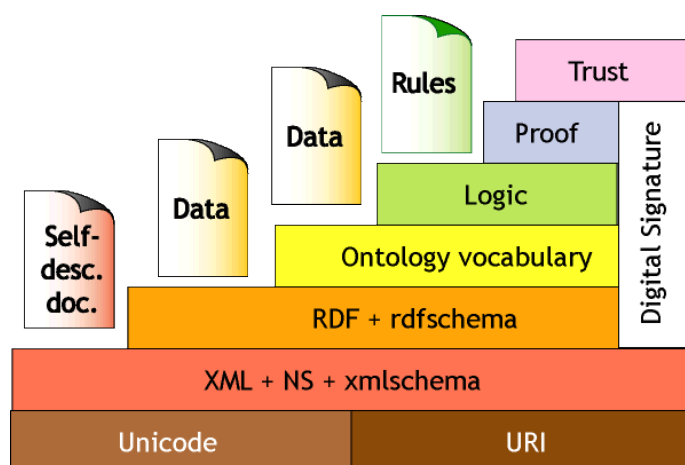


Figura 1 Arquitectura de la web semántica¹

¹ Tomado de <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

- **Unicode - URI:** Unicode es un sistema de codificación que asigna un número único para identificar cada carácter sin importar la plataforma, programa ni idioma. Este es compatible con la mayoría de sistemas operativos y con todos los exploradores actuales, además es un requerimiento para estándares modernos como XML.
- **URI** proporciona un nombre para identificar de manera única a los distintos recursos de la Web.
- **XML + NS + xmlschema:** En esta capa se integran tres tecnologías que hacen posible la comunicación entre agentes. XML ofrece un formato común para intercambiar documentos de una forma estructurada, como árboles de etiquetas con atributos.
- **XML Schema** es uno de estos lenguajes para definir su estructura, donde se describen de antemano las estructuras y tipos de datos utilizados.
- **NS** proporciona un método para cualificar elementos y atributos de nombres usados en documentos XML asociándolos con espacios de nombre identificados por referencias URI's.
- **RDF + RDFS** Schema: Es un lenguaje simple mediante el cual definimos sentencias en un formato con tres elementos: sujeto, predicado y objeto
- **RDF Schema** provee un vocabulario definido sobre RDF que permite el modelo de objetos con una semántica claramente definida. Esta capa no solo ofrece descripción de los datos, sino también cierta información semántica. Ambos corresponden a las anotaciones de la información llamados metadatos.
- **OWL:** Es uno de los lenguajes de ontologías más extendidos por la Web Semántica. Este estándar W3C fue diseñado para ser compatible con estándares web existentes. Ontology Web Language añade más vocabulario para describir propiedades, clases, relaciones entre clases, cardinalidad, igualdad, características de propiedades, clases enumeradas, etc.
- **Logic, Proof, Trust, Digital Signature:** Las capas Logic (Lógica) y Proof (Pruebas) son encargadas de aplicar reglas de inferencia con sus pruebas respectivas. En la capa Trust (Confianza) encontramos agentes que realizan un análisis completo y comprobación de las fuentes de información de la Web Semántica. Finalmente Digital Signatura (Firma Digital) garantiza que la información ofrecida proviene de sitios confiables.

La visión de la Web Semántica, defendida por Sir Tim Berners-Lee, está construida entorno al concepto de la “Web de Datos” (o LinkedData), que significa pasar de una Web actualmente centrada en Documentos a una Web centra en Datos. En esta visión la Web, los datos y sus relaciones son fundamentales. (Piedra, Tovar, López, Chicaiza, & Martinez, 2011)

Un objetivo de la Web semántica es crear un sistema de agentes inteligentes que puedan hacer deducciones de una manera automatizada con la información que esta en la Web. Este objetivo más que una realidad es una utopía incluso a medio plazo. Por otro lado, los desarrollos que se han realizado gracias a este nuevo paradigma han dado lugar a nuevos servicios ajustados con éxito en la actual Web. Como por ejemplo se logrado construir diferentes estándares para poder representar y procesar la información de una manera mas sofisticada. Estos estándares que han permitido presentar los metadatos en un formato mas lógico y controlados (por ejemplo ontologías) para que sean procesados por programas informáticos. Estos formatos ya son utilizados de manera generalizada, como por ejemplo XML, RDF, SKOS-Core y OWL. (Vallez, Rovira, Codina, & Pedraza)

1.3. Metadatos y anotación semántica.

Un elemento fundamental de la Web semántica son los metadatos, en otras palabras, información que nos describe el contenido de los documentos a los que está ligado y nos representa de una manera explícita el significado de estos.

La anotación semántica echa con metadatos ofrece contenido semántico a los documentos para logra que las máquinas interpreten la información.

Las herramientas de anotación permiten convertir en metadatos el contenido semántico extraído de las páginas web. Existen herramientas de anotación dirigidas a los autores ayudan a incorporar los metadatos dentro o fuera de las propias páginas web siguiendo los estándares (xml, rdf...). Y Las aplicaciones del herramientas de anotación externa permiten asociar meta información a páginas web, pero esta no se almacena dentro de la misma página sino que se guardada de forma externa en un repositorio.

Existen diferentes aproximaciones para realizar la anotación semántica, pero se pueden agrupar en tres grandes categorías. (Vallez, Rovira, Codina, & Pedraza)

- El primer modelo se basa en la anotación lingüística, el objetivo es etiquetar los textos a partir de los diferentes niveles de la lengua. Resulta de gran interés la identificación de los términos y saber cómo estos se relacionan

entre sí porque esta información puede incidir en el valor de un término como palabra clave. Este sistema es muy costoso computacionalmente.

- La segunda aproximación se basa en las ontologías, son utilizadas como recurso principal para extraer las conexiones entre los términos y representar su significado.
- La tercera aproximación propone el uso de un lenguaje controlado, este es un modelo que está directamente vinculado con la asignación de metadatos y la anotación semántica.

1.4. Rdf.

Marco de Descripción de Recursos(Resource Description Framework), en sus inicios fue diseñado como modelo de datos para metadatos, este modelo es similar a otros modelos como el de entidad-relación o diagrama de clases, la idea es hacer afirmaciones de los recursos en forma de triples como son conocidas en terminología RDF que son sujeto-predicado-objeto:

Sujeto: es el recurso

Predicado: propiedad que describe los rasgos del recurso y sirve de relación entre el sujeto y el **Objeto:** que puede ser un valor u otra entidad.

El uso de este modelo simple permite que los datos estructurados y semi-estructurados puedan ser mezclados, expuestos y compartidos. (W3C, RDF)

Ejemplo:

Los recursos esta identificados por un identificador de recurso(URI).

Considerando una simple oración en ingles:

Ora Lassila is the creator of the resource <http://www.w3.org/Home/Lassila>

La oración tiene las siguientes partes:

Subject (Resource)	http://www.w3.org/Home/Lassila
Predicate (Property)	Creator
Object (literal)	"Ora Lassila"

Figura 2 partes de la oración en RDF2²

Representándolo gráficamente quedaría:

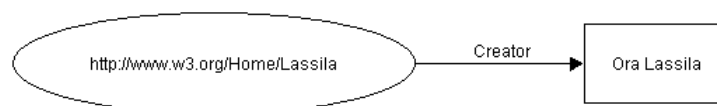


Figura 3 Representación grafica RDF2²

² Tomado de <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>

En estos diagramas, los nodos (dibujados como óvalos) representan recursos y las líneas representan las propiedades con nombre. Los nodos que representan los literales de cadena se dibujan como rectángulos.

1.5. Sparql.

SPARQL (Protocol and RDF Query Language) es un lenguaje estandarizado para consulta y de recuperación basado en RDF, que está normalizado por DAWG del World Wide Web Consortium (W3C). Algunas características son:

- Extraer información de URIs, literales y nodos vacíos
- Obtener subgrafos RDF y construir nuevos grafos a partir de la respuesta del query.

Con SPARQL se pueden expresar consultas para diversas fuentes de datos que estén almacenados como RDF o definidos mediante vistas RDF. También se pueden consultar patrones obligatorios y opcionales de un grafo, con sus conjunciones y disyunciones. Los resultados de las consultas realizadas con SPARQL pueden dar un conjunto de resultados así como también grafos RDF. (W3C, SPARQL Query Language for RDF)

1.5.1. Consultas simples.³

La mayoría de las consultas SPARQL contiene un patrón de grafo básico, estos patrones son similares a las tripletas RDF, con la diferencia que cada sujeto, predicado y objeto pueden ser una variable. (W3C, Sparql query Basicpatterns)

1.5.1.1 EJEMPLOS

1.5.1.1.1 Consulta Simple

La consulta consiste de dos partes: la condición SELECT define las variables que aparecerán en el resultado de la consulta, y la condición WHERE provee el patrón de grafo básico para la concordancia con el grafo de datos.

Consulta:

SELECT *

WHERE

{

<http://dbpedia.org/resource/Loja,_Ecuador> <http://dbpedia.org/property/officialName>

?nombre

}

³ Todos los ejemplos de las consultas están tomadas de Tomado (W3C, Sparql query Basicpatterns)

Resultado:

nombre
"Loja"@en

Consulta:

```
SELECT *
```

```
WHERE
```

```
{
```

```
  ?provincias                                <http://dbpedia.org/property/subdivisionType>
```

```
<http://dbpedia.org/resource/Provinces_of_Ecuador>.
```

```
  ?provincias <http://dbpedia.org/property/officialName> ?nombre
```

```
}
```

Resultado:

provincias	nombre
:Loja,_Ecuador	"Loja"@en
:Manta	"San Pablo de Manta"@en
...	...

1.5.1.1.2 Concordancias de literales RDF

Concordancias de literales de idioma:

Las etiquetas de idioma en SPARQL se expresan usando @ y la etiqueta de idioma. La siguiente consulta no tiene solución porque "Loja" no es el mismo literal RDF que "Loja"@en:

```
SELECT ?s WHERE { ?s ?p "Loja" }
```

Por otro lado la siguiente consulta sí, donde la variable s se relaciona, porque el idioma se especifica.

```
SELECT ?s WHERE { ?s ?p "Loja"@en }
```

Concordancias de literales con tipos numéricos:

Los enteros de una consulta SPARQL indican un literal RDF tipado con el tipo de datos xsd:integer. Por ejemplo: 21 es una forma abreviada de "21"^^<http://www.w3.org/2001/XMLSchema#integer>.

La consulta quedaría de esta manera:

```
SELECT ?v WHERE { ?v ?p 21 }
```

1.5.2. Sintaxis sparql.⁴

1.5.2.1 SINTAXIS DE EXPRESIONES RDF

1.5.2.1.1 Sintaxis para IRI

Los IRIs son tipo de URIs y son compatibles con URIs y URLs.

Los términos RDF tiene referencias RDF URI en cambio que los términos de SPARQL tienen IRIs.

Nombres prefijados: La palabra clave PREFIX enlaza una etiqueta de prefijo con un IRI. Un nombre con prefijo consta de una etiqueta y una parte local, separados por dos puntos ":".

Aquí tenemos algunos ejemplos de las diferentes formas describir la misma dirección IRI:

- <http://example.org/provinica/Loja>
- PREFIX prov: <http://example.org/provinica/>
- prov:Loja

1.5.2.1.2 Sintaxis para literales

La sintaxis para los literales es una cadena de caracteres entre comillas simples o dobles, con una etiqueta de idioma (antecedida por @) o un tipo de datos IRI o nombre prefijado (antecedido por ^^) opcionales.

Para que se aga mas fácil los facilitar los números se pueden escribirse. Los valores del tipo xsd:boolean pueden también escribirse como true o false.

Ejemplos de sintaxis de literales SPARQL:

- 'Loja'@en con la etiqueta de idioma "en"
- "xyz"^^<http://example.org/ns/userDatatype>
- "abc"^^appNS:appDataType
- ""The librarian said, "Perhaps you would enjoy 'War and Peace'.""
- 5, igual a "5"^^xsd:integer
- 1.5, igual a "1.5"^^xsd:decimal
- 1.500, igual a "1.300"^^xsd:decimal
- 1.0e5, igual a "1.0e5"^^xsd:double
- true, igual a "true"^^xsd:boolean
- false, igual a "false"^^xsd:boolean

1.5.2.1.3 Sintaxis para variables

Las variables son prefijadas con "?" o "\$";y estas no forman parte del nombre de la variable. Entonces, \$abc y ?abc son la misma variable.

⁴ Todos los ejemplos de las consultas están tomadas de Tomado (W3C, Sparql Query SparqlSyntax)

1.5.2.2 SINTAXIS PARA PATRONES DE TRIPLETA

Los Patrones de Tripletas se escriben como una lista, separada por espacios, de sujeto, predicado y objeto; hay formas abreviadas para escribir algunas construcciones de patrones comunes de tripleta.

Ejemplos:

- PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE { <http://example.org/provincia/Loja> dc:title ?titulo }
- PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://example.org/provincia/>
SELECT \$ titulo
WHERE { :Loja dc:title \$ titulo }

1.5.2.2.1 Listas de Predicado-Objeto

Los patrones de tripleta con un sujeto común pueden expresarse así:

- ?x foaf:name ?nombre ;
foaf:mbox ?mail .
- ?x foaf:name ?name .
?x foaf:mbox ? mail.

1.5.2.2.2 Listas de objetos

Si los patrones de tripleta comparten tanto el sujeto como el predicado:

- ?x foaf:nick "Loja" , "Cuenca" .
?x foaf:nick " Loja " .
?x foaf:nick " Cuenca " .
- ?x foaf:name ?nombre ; foaf:nick " Loja " , " Cuenca " .
?x foaf:name ?nombre .
?x foaf:nick "Loja" .
?x foaf:nick " Cuenca " .

1.6. Dbpedia.

La Dbpedia es un proyecto para la extracción de datos de Wikipedia para proponer una versión Web semántica . Este proyecto es realizado por la Universidad de Leipzig, Universidad Libre de Berlín y la compañía OpenLink Software. Dbpedia permite hacer consultas mas sofisticadas contra la Wikipedia, así como también enlaces diferentes conjuntos de datos de la Web con los de la Wikipedia. La información esta almacenada en RDF y se puede hacer consultas atreves de SPARQL. (DBpedia, About)

1.6.1. El Dataset de dbpedia.

La versión en Inglés de la base de conocimiento DBpedia describe actualmente 4,0 millones de cosas , de los cuales 3,22 millones se clasifican en una ontología consistente, incluyendo 832.000 personas, 639.000 plazas (incluidos 427 000 lugares poblados), 372.000 obras de creación (incluyendo 116.000 álbumes de música, 78.000 películas y 18.500 juegos de video) , 209.000 organizaciones (entre ellas 49.000 empresas y 45.000 instituciones educativas) , 226.000 especies y 5.600 enfermedades. (DBpedia, Datasets)

El conjunto de datos completo DBpedia cuenta con las etiquetas y los resúmenes de 12,6 millones de cosas únicas en 119 idiomas diferentes ; 24,6 millones de enlaces a las imágenes y los 27,6 millones de enlaces a páginas web externas; 45,0 millones de enlaces externos a otros conjuntos de datos RDF , 67,0 millones de enlaces a las categorías de Wikipedia, y 41,2 millones de categorías YAGO . El conjunto de datos consta de 2,46 mil millones de piezas de información (RDF triplica) fuera de las cuales 470 millones fueron extraídos de la edición de Inglés de Wikipedia , 1,98 millones fueron extraídos de ediciones en otros idiomas , y unos 45 millones son enlaces a bases de datos externas. (DBpedia, Datasets)

2. Recursos educativos abiertos ocw

2.1. Oer (open educational resources).

OER (Open Educational Resources) o REA (Recursos Educativos Abiertos) o son cualquier tipo de materiales educativos que se encuentran en el dominio público o introducidos con una licencia abierta. La naturaleza de estos materiales abiertos significa que cualquier persona legalmente y libremente puede copiar, usar, adaptar y re-share ellos. Gama OER de los libros de texto a los planes de estudio, programas de estudio, notas de clase, tareas, exámenes, proyectos, audio, vídeo y animación. (UNESCO)

OER nació como un concepto con gran potencial para poder sobrellevar la transformación en la educación. EL valor educativo se basa en la idea de utilizar los recursos como un método de aprendizaje, su poder se basa en la facilidad con la que los recursos ya digitalizados, se comparten a través de Internet.

Se deben tomar en cuenta tres elementos fundamentales de REA/OER:

- Contenido Educativos: Recursos educativos como libros, programas educativos completos, módulos de contenido, publicaciones, etc.

- Herramientas: Software para apoyar la creación, acceso, uso y mejoramiento de contenidos educativos abiertos.
- Recursos de Implementación: parte de legal de recursos, Diseño, adaptaciones del contenido, materiales y técnicas para apoyar al acceso del conocimiento.

Los recursos educativos abiertos frecuentemente están distribuidos bajo una licencia Creative Commons.

La tarea de aseguramiento de la calidad se ha visto complicada por la explosión de disposición contenido (tanto abierto como propietario). Esto es tanto una ventaja ya que reduce la probabilidad de necesitar para desarrollar nuevos contenidos y una maldición ya que exige mayor habilidad de nivel de búsqueda de información, selección, adaptación y evaluación, tomando en cuenta que las instituciones comparten más contenido educativo en línea, van asegurarse de que le contenido refleje una buena institución y por lo tanto podrá invertir en la mejora de su calidad antes de que este a disposición en los repositorios. (Unesco, 2011)

2.2. Ocw.

OpenCourseWare (OCW) es una publicación digital abierta y gratuita de materiales educativos de alta calidad. Están clasificados desde cursos completos hasta asignaturas. Mantiene una licencia abierta accesible a cualquiera o cualquier lugar a través de la red. (OCW Consortium, 2012)

OCW es un ejemplo de las iniciativas que desde el 2001 han emergido para promover el acceso libre y sin restricciones al conocimiento.

(UTPL, 2014)

OCW site (UPM): Es un espacio web que contiene materiales docentes creados por profesores para la formación superior. Las características que distinguen al proyecto OpenCourseWare de iniciativas similares son las siguientes:

- Los recursos didácticos publicados en un OCW site se organizan en unidades de “asignaturas” o “cursos”. Con ello se quiere indicar:
 - Los accesos se realizan por asignaturas e incluyen un conjunto significativo de todos los materiales asociados a ella.

- Los materiales se ofrecen de forma organizada por categorías: programa de la asignatura, lecturas obligatorias, materiales de clase, ejercicios, guía de aprendizaje,....
- El profesor o profesores garantizan que el material que publican en el OCW site es original o tiene los derechos, bien directamente por ser propietario o bien a través del tipo de licencia que los soporta, para ser reutilizados en “abierto” sin infringir los “copyrights” de otras personas.
- Son accesibles universalmente a través de la red:
 - Sin limitaciones geográficas.
 - Sin exclusión de usuario, ni necesidad de registrarse o utilizar palabras claves de acceso.
 - No exigen requisitos técnicos más allá de un navegador Web.

Entre los beneficios que nos brinda OCW tenemos:

- Consulta: Acceso a toda la información que mantiene.
- Colaboración: Puede obtener y brindar información OCW.
- Visibilidad colectiva: Aumentar la posibilidad colaborativa en la red.
- Sostenibilidad: asegurar que el intercambio abierto de recursos educativos tiene un futuro productivo.

3. Procesamiento del lenguaje natural

El " Procesamiento del Lenguaje Natural " (NLP) es una disciplina con una larga trayectoria. Nace en la década de 1960, como un subárea de la Inteligencia Artificial y la Lingüística, con el objeto de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural. (PNL)

El PLN investiga mecanismos eficaces computacionales para la comunicación entre personas o entre personas y máquinas por medio de lenguajes naturales. Trata de diseñar mecanismos para comunicarse que sean eficaces computacionalmente hablando. La lingüística general se relaciona con PNL, ya que esta estudia la estructura general y descubre las leyes universales de funcionalidad de los lenguajes naturales. Estas estructuras y leyes, aunadas a los métodos computacionales forman la lingüística computacional.

La lingüística computacional puede ser considerada como un sinónimo de procesamiento de lenguaje natural, ya que su tarea principal es la construcción de programas que procesen palabras y textos en lenguaje natural. (Bolshakov & Gelbukh, 2004)

Para poder realizar esa tarea, los sistemas de PLN deben tener conocimiento sobre la estructura del lenguaje, y así poder pasar de texto a significado y viceversa.

Niveles de Lenguaje

La lingüística general comprende 5 niveles principales para el análisis de la estructura del lenguaje (Bolshakov & Gelbukh, 2004) que son:

- a. **Nivel fonológico:** trata de los sonidos que comprenden el habla, permitiendo formar y distinguir palabras.
- b. **Nivel morfológico:** trata sobre la estructura de las palabras y las leyes para formar nuevas palabras a partir de unidades de significado más pequeñas llamadas morfemas.
- c. **Nivel sintáctico:** trata como las palabras pueden unirse para construir oraciones y cuál es la función que cada palabra realiza en esa oración.
- d. **Nivel semántico:** trata del significado de las palabras y de cómo se unen para dar significado a una oración.
- e. **Nivel pragmático:** estudia la intención del hablante al producir oraciones específicas o textos en una situación específica.

4. Servicios web

Los servicios web (Web Service) son componentes software que permiten intercambiar información y datos entre aplicaciones, mediante el uso de tecnologías Web basadas en estándares y protocolos. Los servicios web se diseñados para que se pueda acceder por otras aplicaciones. Los servicios web son un conjunto de herramientas que pueden ser usadas en distintas formas. (Alvarado Ruiz, Guamán Eras, & Sigcho Armijos)

Según el W3C (World Wide Web Consortium) los Servicios Web son aplicaciones de software identificadas por un URI (Uniform Resource Identifier), cuyos interfaces y vínculos tienen la capacidad de estar bien definidos, descritos y descubiertos como objetos XML. WS soporta interacciones directas con otros agentes de software usando mensajes de intercambio basados en XML vía protocolos basados en Internet. Se puede perfeccionar esta definición pidiendo que la descripción se haga a través de un documento WSDL (Web Services Description Language) y el protocolo utilizado sea SOAP. (W3C)

Puesto que cada Servicio Web puede estar implementado en una tecnología heterogénea es necesario cumplir una serie de estándares para hacer posible la comunicación entre ellos. Los más utilizados son los siguientes (González):

Web Services Protocol Stack: conjunto de servicios y protocolos de los servicios Web.

- XML: Es un lenguaje de marcas capaz de describir distintos tipos de datos. Es un estándar aceptado y utilizado como medio de descripción de datos.
- SOAP: Es un protocolo de comunicación entre procesos basado en el intercambio de mensajes en formato XML dentro de una red. A su vez SOAP está basado en XML y es completamente independiente de la plataforma y del lenguaje en el que estén implementados los procesos que se comunican.
- WSDL: Es un lenguaje basado en XML que permite describir servicios web (como su nombre indica). Un documento WSDL especifica, entre otras cosas, dónde se encuentra el servicio así como las operaciones que pone accesibles a otros servicios.
- UDDI: Es un directorio, basado en XML, en el que las distintas empresas dan de alta servicios web que ponen al servicio de otras empresas.
- WS-Security (Web Service Security): Protocolo de seguridad aceptado como estándar por OASIS (Organization for the Advancement of Structured Information Standards). Garantiza la autenticación de los “actores” y la confidencialidad de los mensajes enviados.

4.1. Arquitectura rpc (remote procedure call).

En el paper de la idea fue permitir que los programas llamaran a procedimientos localizados en otras máquinas. Cuando un proceso en la máquina A llama a un procedimiento en la máquina B, el proceso llamado en A es suspendido, y la ejecución del procedimiento llamado se lleva a cabo en el B. Información puede ser transportada de la persona que llama al destinatario de la llamada en los parámetros y puede volver a aparecer en el resultado del procedimiento. (Birrell & Nelson)

No hay ningún mensaje que pasa visible para el programador. Este método se conoce como Remote Procedure Call, a menudo solo RPC.

La idea básica parece simple y elegante, los problemas existen. Para empezar, porque los procedimientos de llamada y la llamada se ejecutan en máquinas diferentes, que se

ejecutan en espacios de direcciones diferentes, lo que provoca complicaciones. Parámetros y resultados también tienen que ser pasados, que puede ser complicado, especialmente si las máquinas no son idénticas. Por último, ambas máquinas pueden fallar y cada uno de los posibles fallos causa diferentes problemas. Sin embargo, la mayoría de estos pueden ser tratados, y RPC es una técnica ampliamente utilizada que subyace a muchos de los sistemas distribuidos.

El RPC sigue los siguientes pasos:

- El procedimiento en el cliente llama al cliente stub de manera normal
- El cliente stub construye un mensaje y llama al sistema operativo local
- El Sistema operativo cliente envía un mensaje para el OS remoto
- El OS remoto da un mensaje al servidor stub
- El servidor stub desempaqueta los parámetros y llama al servidor
- El servidor hace el trabajo y retorna el resultado para el stub
- El servidor stub empaqueta en un mensaje para el OS cliente
- El OS servidor envía el mensaje al OS cliente
- El OS cliente da un mensaje al cliente stub
- El stub desempaqueta el resultado y retorna al cliente

4.2. Arquitectura orientada a servicios.

Arquitecturas orientadas a servicios se han utilizado durante muchos años. SOA es de una articulación flexible por lo que la distingue otras arquitecturas. Acoplamiento débil significa que el cliente de un servicio es esencialmente independiente del servicio. La forma en que un cliente se comunica con el servicio no depende de la aplicación del servicio. De manera significativa, por lo cual el cliente no tiene que saber mucho sobre el servicio para usarlo. El acoplamiento flexible permite a los servicios ser un documento-orientado. El cliente se comunica con el servicio de acuerdo a una interfaz específica, bien definida, y luego se deja en manos de la implementación del servicio para realizar el procesamiento necesario. Si la implementación del servicio cambia, el cliente se comunica con ella en la misma forma que antes, siempre que la interfaz sigue siendo el mismo.

Sin embargo lo que es relativamente nuevo es la aparición de servicios web basados SOAs. Un servicio web es un servicio que se comunica con los clientes a través de un conjunto de protocolos y tecnologías estándar. Estos estándares de servicios web se ejecutan en plataformas y productos de todos los principales proveedores de software, lo

que permite a los clientes y servicios para comunicarse de una manera consistente a través de un amplio espectro de plataformas y entornos operativos. Esta universalidad ha hecho que los servicios web de la forma más extendida implementen SOA. (SunMicrosystems, 2005)

4.3. Arquitectura rest.

REST (Transferencia de estado representacional) es un estilo de arquitectura de software para sistemas hipermedia distribuidos como la World Wide Web. El término fue introducido en la tesis doctoral en el año 2000 por Roy Fielding , y ha entrado en uso generalizado en la comunidad de redes. (Fielding)

REST se refiere estrictamente a una colección de principios de la arquitectura de red que describen cómo los recursos son definidos y tratados. El término se utiliza a menudo en un sentido más amplio para describir cualquier interfaz simple que transmite datos específicos del dominio a través de HTTP y sin capa de mensajería adicionales, tales como SOAP o una sesión de seguimiento a través de cookies HTTP. (Alvarado Ruiz, Guamán Eras, & Sigcho Armijos)

- Principios de diseño:
- El estado de aplicación y la funcionalidad se dividen en recursos.
- Cada recurso es direccionable únicamente utilizando una sintaxis universal para su uso en enlaces hipermedia
- Todos los recursos comparten una interfaz uniforme para la transferencia de estado entre el cliente y los recursos, constituido por
 - Un conjunto limitado de operaciones bien definidas
 - Un conjunto limitado de tipos de contenido, opcionalmente soporte de código on-deman
- Un protocolo que es:
 - cliente / servidor
 - sin estado
 - Cacheable
 - en capas

La separación de cliente-servidor REST simplifica la implementación de componentes, reduce la complejidad de la semántica de los conectores, mejora la eficacia de la

optimización del rendimiento, y aumenta la escalabilidad de los componentes de servidor puros. Limitaciones en capas del sistema permiten a los intermediarios - los proxies, gateways y servidores de seguridad - que se presentó en varios puntos de la comunicación sin necesidad de cambiar las interfaces entre los componentes, lo que les permite ayudar en la traducción de la comunicación o mejorar el rendimiento a través de gran escala, el almacenamiento en caché compartida.

5. Programas y librerías

5.1. Python.

Python es un lenguaje de programación dinámico y potente que es utilizado en varios dominios de aplicación . A Python se lo compara con Tcl, Perl, Ruby, Scheme o Java. Sus características mas relevantes son (Python):

- Su sintaxis es muy clara y legible
- Fuerte Capacidad de introspección
- Orientado a objetos
- Expresión natural del código procedimental
- Completamente modular, Soporte para paquetes jerárquicos
- Manejo de errores basado en excepciones
- Los tipos de datos son dinámicos de muy alto nivel
- Bibliotecas estándar y módulos de terceros para prácticamente todas las tareas
- Puede ser integrada en aplicaciones mediante interfaz de script

Python es un lenguaje de programación que te permite trabajar con mayor rapidez e integrar sus sistemas con mayor eficacia. También se puede ejecutar en Windows, Linux / Unix, Mac OS X, y ha sido adaptada a la de Java y máquinas virtuales. NET.

Python es libre de usar, incluso para los productos comerciales, por su aprobada por OSI licencia de código abierto. (Python)

5.1.1. Nltk.

NLTK es una plataforma para python que trabajar con datos de lenguaje humano. Proporciona interfaces fáciles de usar, junto con un conjunto de bibliotecas de procesamiento de textos para la clasificación, tokenización , derivado, el etiquetado , el análisis y el razonamiento semántico. (Nltk)

NLTK se lo puede trabajar en Windows, Mac OS X y Linux. NLTK también es código abierto, impulsado por la comunidad.

El público objetivo de NLTK consiste en lingüistas y científicos de computación, y es accesible y desafiante en muchos niveles de las aptitudes computacionales. Y se basa en un lenguaje de programación orientado a objetos apoyándose prototipado rápido y programación literaria.. (Loper & Bird)

Librerías de NLTK utilizadas:

Word_tokenize: es una funcionalidad de NLTK, que sirve para la clasificación de la palabras, lo clasifica mediante etiquetas.

Tag	Meaning	Examples
ADJ	adjective	<i>new, good, high, special, big, local</i>
ADV	adverb	<i>really, already, still, early, now</i>
CNJ	conjunction	<i>and, or, but, if, while, although</i>
DET	determiner	<i>the, a, some, most, every, no</i>
EX	existential	<i>there, there's</i>
FW	foreign word	<i>dolce, ersatz, esprit, quo, maitre</i>
MOD	modal verb	<i>will, can, would, may, must, should</i>
N	noun	<i>year, home, costs, time, education</i>
NP	proper noun	<i>Alison, Africa, April, Washington</i>
NUM	number	<i>twenty-four, fourth, 1991, 14:24</i>
PRO	pronoun	<i>he, their, her, its, my, I, us</i>
P	preposition	<i>on, of, at, with, by, into, under</i>
TO	the word <i>to</i>	<i>to</i>
UH	interjection	<i>ah, bang, ha, whee, hmpf, oops</i>
V	verb	<i>is, has, get, do, make, see, run</i>
VD	past tense	<i>said, took, told, made, asked</i>
VG	present participle	<i>making, going, playing, working</i>
VN	past participle	<i>given, taken, begun, sung</i>
WH	<i>wh</i> determiner	<i>who, which, when, what, where, how</i>

Figura 4 Conjunto de etiquetas⁵

RegexpTagger: sirve para hacer una clasificación y etiquetado de las palabras pero con expresiones regulares en base ha patrones. Por ejemplo se puede decir que todas las palabras en ingles que termina en ed son verbos en pasado participio, así como también las palabras que terminen en 's son nombres posesivos.

```
patterns = [
    (r'.*ing$', 'VBG'),          # gerunds
    (r'.*ed$', 'VBD'),          # simple past
    (r'.*es$', 'VBZ'),          # 3rd singular present
    (r'.*ould$', 'MD'),         # modals
    (r'.*'s$', 'NN$'),          # possessive nouns
    (r'.*s$', 'NNS'),           # plural nouns
    (r'^-?[0-9]+(.[0-9]+)?$', 'CD'), # cardinal numbers
    (r'.*', 'NN')               # nouns (default)
]
```

Figura 5 Ejemplo de expresiones regulares para etiquetado con RegexpTager⁵

⁵ Tomado de <http://www.nltk.org/book/ch05.html>

Chunk: es un paso preliminar y útil para la extracción de información, sé que crea árboles de análisis de texto no estructurado con un Chunker. Una vez obtenido un árbol de análisis sintáctico de una oración, se puede hacer la extracción de información más específica, como el reconocimiento de entidades y extracción de relaciones.

Fragmentación es básicamente un proceso de 3 pasos:

Se etiqueta una sentencia

Se hace un Chunk de la sentencia etiquetada

Analizar el árbol obtenido para extraer información

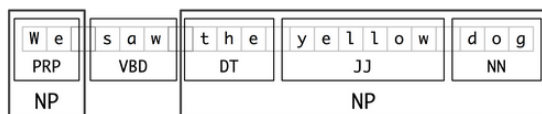


Figura 6 Segmentación y Etiquetado Chunk⁶

⁶ Tomado de <http://www.nltk.org/book/ch07.html>

PROBLEMÁTICA

1. Problemática actual

Actualmente se puede encontrar una gran cantidad de información bastante detallada, que se encuentra representada en un enorme conjunto de documentos electrónicos que están localizados en distintos lugares y en diferentes formatos.

Por ejemplo, en la web se dispone de una amplia variedad de datos relacionados a un sinnúmero de temas específicos: medicina, educación, matemáticas, entretenimiento, noticias, entre otros; en lo que concierne a la educación se dispone de bastantes recursos como son Wikis, blogs, tutoriales, presentaciones, papers, noticias, videos, infografías, bases de datos, entre otros.

La educación superior dispone de los OER y OCW, estos manejan una gran cantidad de los recursos anteriormente mencionados, dichos contenidos están presentados en su mayoría en lenguaje natural, es decir, aquel que puede ser entendido y discernido fácilmente por las personas pero no es entendible para las computadoras.

Para que una máquina pueda llevar a cabo la inferencia de conocimiento a partir de unos datos entregados, dichos datos deberán estar definidos en un lenguaje que se conoce como entendible por máquina.

En el caso de los cursos OCW se dispone de una gran cantidad de información que puede ser entregada a un usuario para su consumo, manipulación y almacenamiento, pero existen ciertas preguntas con respecto a todo este conjunto de información:

¿Cuál información de todo este conjunto de datos es relevante?

¿Cómo seleccionar aquellos cursos que sean más idóneos para la búsqueda planteada?

¿Cómo clasificar los datos/recursos/autores/artículos que manejan los cursos OCW?

Esto lleva a plantearse una pregunta más general:

¿Qué entidades se puede extraer de un curso OCW?

Puesto que un OCW presenta un conjunto de datos bastante amplio como:

- Organización a la que pertenece
- Autor
- Áreas de conocimiento
- Países

- Contenido, dentro de este existe todavía más datos que pueden ser relevantes:
 - Menciones a papers
 - Autores relacionados
 - URLs, URIs
 - Software
 - Conceptos
 - Organizaciones (universidades, empresas, consorcios)

Con esta amplia variedad de datos se puede crear etiquetas que nos permitirán de alguna forma resumir o detallar el contenido de un curso OCW.

Por ejemplo, supongamos que se realiza una búsqueda sobre aplicaciones de la Web Semántica en ámbito médico, al llevar a cabo una búsqueda basada en ciertas palabras claves las resultantes de dicha búsqueda sería bastante amplia; en cambio que si cada curso OCW dispone de etiquetas relevantes como: Web Semántica, aplicaciones de Web Semántica, medicina, la búsqueda resultaría más eficaz, al basar la misma en etiquetas o al realizarla mediante palabras claves y luego filtrar la misma mediante las etiquetas.

La extracción de entidades no sólo resultaría en un enriquecimiento de la información disponible del curso, al extraer todos aquellos datos que son relevantes, sino también en una forma de pre-clasificación del curso OCW, mediante la generación de etiquetas a partir del contenido.

En la web toda la información está estrechamente relacionada una con otra, así mismo al extraer entidades de un curso OCW, dichas entidades pueden usarse para interrelacionar un curso con otro.

Por ejemplo, una entidad de tipo autor que ha sido extraída de un curso OCW que habla sobre ontologías biológicas, puede ser usada para relacionar con algún paper, recurso, libro o publicación que dicho autor haya hecho anteriormente.

Esto permite enriquecer o delimitar la información sobre un OCW:

- enriquecer porque se puede relacionar un OCW con otros mediante las distintas entidades extraídas;

- delimitar, por que mediante la generación de etiquetas se puede llevar a cabo una búsqueda más exacta, precisa y eficaz sobre un tema o curso en específico.

SOLUCIÓN

1. Aproximación

Se planea realizar la extracción de entidades como: lugares, países, ciudades, áreas de conocimientos, organizaciones académicas, autores.

Las entidades extraídas además servirán como metadatos que representaran el contenido de documentos.

Para lo cual se desarrollaran servicios web cuya función serán: tokenizar, extraer entidades, desambiguar y enlazar con LOD-Cloud, estos servicios web se los desarrollara con anotaciones semánticas, para que puedan permitir la interoperabilidad entre los servicios existentes.

En la siguiente grafica se muestra la estructura mencionada:

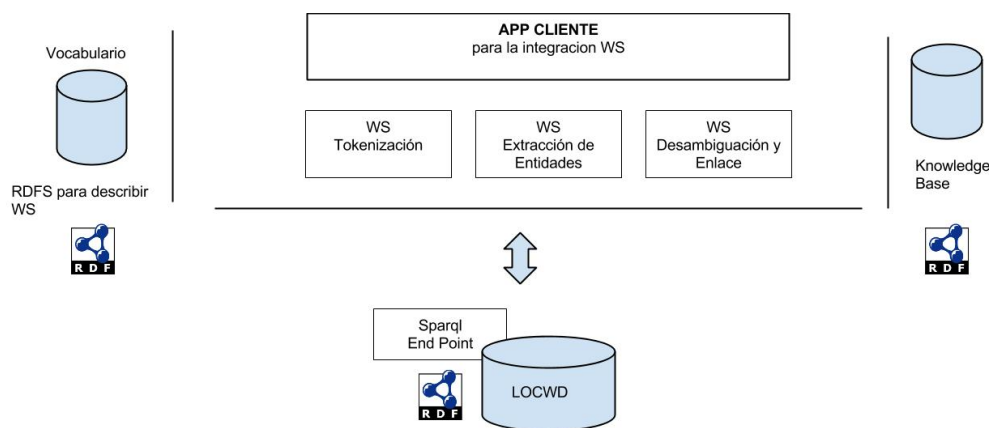


Figura 7 Estructura de los Componentes

2. Descripción de componentes.

La solución estará modelada en base a distintos componentes y recursos como son:

2.1. Vocabulario.

Se definirá un vocabulario para describir los servicios web semánticamente, puesto que dichos servicios deben llevar anotaciones semánticas para facilitar y optimizar la interoperabilidad entre los servicios existentes, el vocabulario estará definido en RDFs.

2.2. Knowledge Base(Base de Conocimiento).

Se necesita definir una data para llevar a cabo la extracción de entidades en base a conceptos que hayamos definido anteriormente; debemos definirla puesto que existen datos que son propios del ámbito universitario; entre estos conceptos tenemos:

- a. **Entidades**, entre estas tenemos universidades, organizaciones académicas, empresas, organizaciones gubernamentales.
- b. **Autores**, sean conferencistas, autores de publicaciones, sean libros, papers, revistas, artículos, de cursos, de talleres, de toda persona que sea expositora de un tema.
- c. **Áreas de conocimiento** .
- d. **Lugares**, como son países, ciudades, provincias, localizaciones.
- e. **Eventos**, como conferencias, simposios, congresos, encuentros .

2.3. Locwd.

Son las siglas de Linked OpenCourseWare Data que es un vocabulario que implementa datos enlazados en contenidos abiertos educativos, dicho vocabulario reutiliza un conjunto de vocabularios RDF. Cada vocabulario incluye un conjunto de términos y clases que son comunes a nuestro dominio particular de conocimiento. (Piedra , Chicaiza, & Lopez, Combining Linked Data and Mobiles Devices to improve access to OCW)

2.4. Ws tokenización.

Se definirá este servicio web para el manejo del texto, sea:

- **Codificación**, mediante esta funcionalidad se definirá el formato estándar con el cual trabajar el texto, sea este utf-8 o ascii, así como se extraerá el idioma en el cual se encuentra el texto introducido.
- **Tokenización**, mediante este servicio se tokenizará cada palabra perteneciente al texto, mediante tokenización de oraciones completas para no perder el sentido ni la sintaxis.
- **Tageo**, se establecerá una etiqueta por cada palabra dentro de cada oración del texto introducido, esto con el fin de determinar el tipo de palabra pudiendo ser: un artículo, un verbo, un sustantivo, un adjetivo, un adverbio, un pronombre o nombres propios.

2.5. Ws Extracción de Entidades.

Mediante este servicio se analizará y procesará el texto para encontrar aquella información relevante dentro del mismo, como principal enfoque se buscará extraer entidades como autores, organizaciones, lugares, datos informáticos, entre otros.

Dicho servicio utilizará procesamiento de lenguaje natural, puesto que en la actualidad existen librerías, documentación, programas y recursos que nos facilitan el PLN; así mismo se emplearán algoritmos, métodos y técnicas para extraer conocimiento útil de datos de texto no estructurados a través de la identificación de conceptos básicos.

2.6. Ws Desambiguación y Enlace.

Este servicio web se encargará de 2 tareas:

- Desambiguación, se entiende por esto que teniendo varios conceptos relacionados a una palabra específica se definirá un sentido del conjunto de posibilidades predefinidas, de acuerdo al contexto de la palabra.
- Se enlazara mediante este servicio se creará un enlace entre la data extraída, definida y almacenada con el LOD-Cloud que son servicios en la nube para Linked Open Data, por medio de este enlace la información se enriquecerá, puesto que al estar disponible en la web se crearán relaciones entre las entidades que han sido extraídas y la información que la web tiene sobre la misma entidad.

2.7. App cliente.

Es la interfaz que permitirá integrar todos los servicios definidos anteriormente, para hacer más fácil y sencillo el consumo de los mismos.

3. Ontología y Vocabulario para descripción de Servicios Web.

Para la descripción semántica de los Servicios Web, se a creado una ontología reutilizando términos de otros vocabularios.

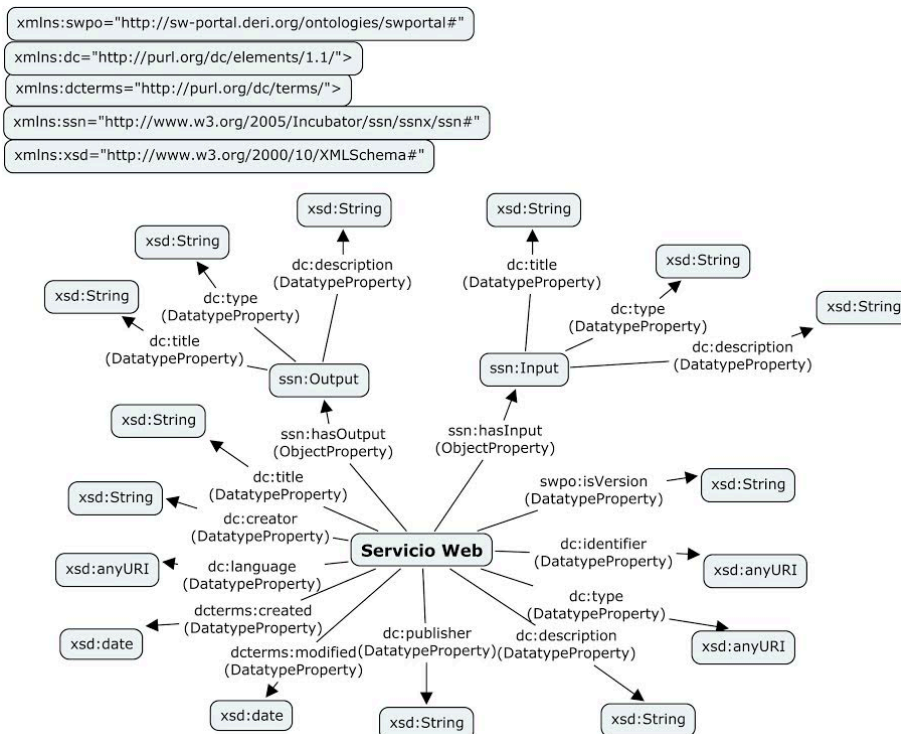


Figura 8 Grafico de la Ontología para la descripción de los Servicios Web

En la Figura 8 esta descrito gráficamente la ontología creada poder anotarlos a los servicios web semánticamente. Para esta ontología se utilizo varios términos de distintos vocabularios, a continuación se lista los vocabularios reutilizados:

- swpo="http://sw-portal.deri.org/ontologies/swportal#"
- dc="http://purl.org/dc/elements/1.1/">
- dcterms="http://purl.org/dc/terms/">
- ssn="http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#"
- xsd="http://www.w3.org/2000/10/XMLSchema#"

De los vocabularios reutilizados se utilizo ciertos términos, que para mejor entendimiento se los describirá, para dejar claro cual es su función en esta ontología:

- dc:title: Se lo utiliza para describir el nombre el titulo del Servicio Web o de la entrada o salida del Servicio.
- dc:creator: Es para definir el autor que creo el Servicio Web.
- dc:language: Describe el tipo de lenguaje del Servicio Web.
- dcterms:created: Sera para colocar la fecha en que se creo el Servicio Web.
- dcterms:modified: Se lo utiliza para agregar las fechas de las ultimas modificaciones del Servicio Web.
- dc:publisher: Identifica quien o donde se lo publico.
- dc:description: Sirve para dar una descripción mas a detalle de el Servicio Web o de la entrada o salida del servicio.
- dc:type: Se lo utiliza tanto para describir el tipo de Servicio Web asi como también el tipo de entrada o salida del Servicio.
- dc:identifier: Es para identificar la dirección donde se encuentra el servicio web.
- swpo:isVersion: Identifica la versión en la que esta el Servicio Web.
- ssn:Input: Clase para describir las entradas del Servicio Web
- ssn:Output: Clase para describir las salidas del Servicio Web

En el Anexo 1 se encuentra la ontología así como también la descripción de los servicios web en formato rdf/xml.

DESARROLLO

4. Ws tokenización.

El diseño de este prototipo se realiza la tokenización de texto. Para la elaboración de este prototipo se realizó lo siguiente:

- El texto se lo separa en oraciones utilizando la función *sent_tokenize* de la librería NLTK , para luego esas oraciones se la separe en tokens con la misma librería con la función *word_tokenize* .
- Para el tagueo de los tokens antes mencionados se establece etiquetas para cada palabra del texto con *RegexTagger de NLTK* y se taguea con la función *pos_tag*.
- Este prototipo nos devuelve una lista con los tokens tagueados

4.1. Arquitectura.

4.1.1. Diagrama de secuencia.

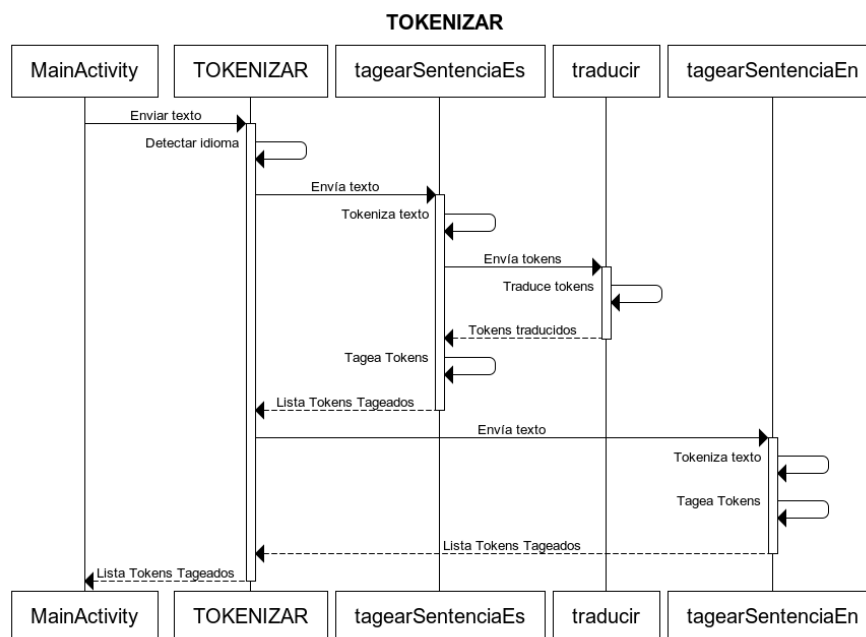


Figura 9 Diagrama de secuencia de la funcionalidad Tokenizar

A continuación se describe su comportamiento:

1. Se llama al servicio web *TOKENIZAR* y se le envía el texto.
2. *TOKENIZAR* recibe el texto y se procede a detectar el idioma y dependiendo si esta en español o ingles se lo envía a las otras funciones.

3. Si el idioma es español se envía el texto a *tagearSentenciaEs*, para q esta función tokenice el texto
4. Luego se lo envía los tokens a la función *traducir* y nos devuelve los tokens traducidos
5. Con los tokens traducidos la función *tagearSentenciaEs* los taguea y devuelve una lista con los token tagueados
6. En cambio si el idioma esta en ingles envía el texto a la función *tagearSentenciaEn*
7. La función *tagearSentenciaEn* con el texto recibido procede a tokenizar y luego a taguear los tokens y nos devuelve una lista con los token tagueados.
8. El servicio Web no devuelve una Lista con los tokens tagueados.

4.2. Implementación.

4.2.1. Parámetros.

Como parámetro de entrada se tiene :

- *Sentencia*, consiste en una variable de tipo carácter o string que almacena el texto a ser analizado, procesado y etiquetado.

Como parámetro de salida:

- *Lista tokens*, es una lista desglosada mediante PLN, en la cual cada palabra se encuentra separada y clasificada mediante una etiqueta que denota su funcionalidad dentro del texto:

```
( 'El', 'DT' )
( 'Comit\xc3\xa9', 'NN' )
( 'del', 'DT' )
( 'Patrimonio', 'NNP' )
( 'Mundial', 'NN' )
( 'reunido', 'VBD' )
( 'en', 'IN' )
( 'Phnom', 'NNP' )
( 'Penh', 'NNP' )
```

Figura 10 Salida del servicio web de tokenización

4.2.2. Funciones.

Esta clase implementa funciones para:

- Tokenizar texto
- Taggear texto tokenizado

Para acceder al servicio de esta clase se lo hace mediante la función:

```
def tagear(self,sentencia):
    idioma=gs.detect(sentencia)
    lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Albanian","ar":"Arabic","
    idiomaCompleto=lenguajes[idioma]

    if idioma == 'es':
        Lista=self.tagearSentenciaEs(sentencia)
    elif idioma == 'en':
        Lista=self.tagearSentenciaEn(sentencia)
    else:
        mensaje='El idioma %s no es soportado'%idiomaCompleto
        print mensaje
        mensaje=gs.translate(mensaje, idioma)
        print mensaje
        Lista=[]
    return Lista
```

Figura 11. Script para tagear del servicio web de tokenizar

la misma recibe como parámetro el texto ingresado que se almacena en la variable **sentencia**.

Se procede a detectar el idioma del texto mediante la función

```
def tagear(self,sentencia):
    idioma=gs.detect(sentencia)
    lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Albanian","ar":"Arabic","
    idiomaCompleto=lenguajes[idioma]
```

Figura 12. Script para detectar idioma del servicio web de tokenizar

de esta manera se determina el lenguaje en que se encuentra escrito el texto ingresado para posteriormente llevar a cabo la tokenización y taggeo del mismo.

Pueden darse 3 casos dentro de la tokenización y taggeo:

A. Idioma Español.

Si el idioma está en español se procede a llamar a la función:

```

def taggearSentenciaEs(self, tags):
    tags=word_tokenize(tags)
    patterns = [
        (r'^@w+', 'NNP'),
        (r'^d+$', 'CD'),
        (r'^ing$', 'VBG'),
        (r'^ment$', 'NN'),
        (r'^ful$', 'JJ'),
        (r'^la', 'IN'),
        (r'^La', 'IN'),
        (r'^de', 'IN'),
        (r'^el', 'IN'),
        #(r'^.*', 'NN')
    ]
    patterns = [
        (r'^$', ' '),
        (r'^$', ' '),
        (r'^$', ' '),
        (r'^$', ' '),
        (r'^.$', ' '),
        (r'^-$', ' '),
        (r'^/$', ' '),
        (r'^\($', ' '),
        (r'^\)$', ' '),
        (r'^-$', ' '),
        (r'^-$', ' '),
        (r'^de|De|con|Con|En|en|xe2\x80\x93$', 'IN'),
        (r'^A|A|(a|A)nte|(b|B)ajo|(c|C)on|(c|C)ontra|(d|D)esde|(e|E)n|(e|E)ntre|(h|H)acia|(h|H)a', 'IN'),
        (r'^el|EL|EL|Los|los|la|La|Las|las|del$', 'DT'),
        (r'^éste|Este|ésta|Esta|esto|Esto|éstos|Estos|éstas|Éstas|ése|Ése|ésa|Ésa|eso|Eso|ésos|Esos|estas|Estas|ese|Ese|esa|Esa|eso|Eso|esos', 'IN'),
        (r'^este|Este|esta|Esta|esto|Esto|estos|Estos|estas|Estas|ese|Ese|esa|Esa|eso|Eso|esos', 'IN'),
        (r'^yo|Yo|tú|tú|él|Él|Ella|ella|nosotros|Nosotros$', 'PRP'),
        (r'^Así|Así|así|así|más|Más$', 'RB'),
        (r'^Así|Así|así|así|más|Más$', 'RB'),
        (r'^Y|y|o|O|u|u|ni|Ni|ya|Ya$', 'CC'),
        (r'^véase$', 'VR'),
        (r'^admitir|afectar|estar|apuntar|permitir|responder|aparecer|aplicar|discutir|arreglar|', 'IN')
    ]
    regex_tagger = nltk.RegexpTagger(patterns)
    for index,t in enumerate(tags):
        newtag=regex_tagger.tag(nltk.word_tokenize(t))
        if newtag[0][1]!=None:
            t=newtag[0]
            tags[index]=newtag[0]
        else:
            if index+1<len(tags):
                tTra=self.traducir(t, tags[index+1])
            else:
                tTra=self.traducir(t, tags[index])
            nuevotag=pos_tag(nltk.word_tokenize(tTra))[0]
            nuevotag=(t, nuevotag[1])
            tags[index]=nuevotag
    return tags

```

Figura 13. Script para taggear del servicio web de tokenizar

esta función recibe el texto ingresado y lo tokeniza mediante el método:
word_tokenize(tags)

Se define patrones para el etiquetado mediante NLTK (en la variable **patterns**), dicho etiquetado se realiza mediante el método:

nltk.RegexpTagger(patterns)

En el cual se clasifica cada palabra encontrada en el texto ingresado, con una etiqueta que se encuentra definida según la variable **patterns**.

Luego de este método, si existen palabras que no han podido ser etiquetadas, las mismas serán traducidas mediante la función:

```
def traducir(self, token, token2):
    #print token
    if len(token)>2:
        tokenTra=gs.translate(token, 'en')
    else:
        #tokenTra= gs.translate(token+' '+token2, 'en')
        tokenTra= gs.translate(token+' '+token+' '+token+' '+token, 'en')
        #print tokenTra
        tokenTra=word_tokenize(tokenTra)
        tokenTra=tokenTra[0]
        tokenTra=word_tokenize(tokenTra)
        tokenTra=tokenTra[len(tokenTra)-1]
        #print tokenTra
        if token.islower()==True:
            tokenTra=tokenTra.lower()
    return tokenTra
```

Figura 14. Script para traducir los tokens del servicio web de tokenizar

con esta función se reciben los tokens, si se encuentran en español se los traduce en inglés y devuelve la misma mediante la variable **tokenTra**, una vez que se regresa esta palabra a la función principal se procede a tokenizarla mediante el corpus en inglés de NLTK, y se almacena la misma palabra en español con el token extraído de su versión traducida.

Luego de todos estos procesos se regresa una lista (la variable **tags**) que contiene todas las palabras del texto separadas y clasificadas según etiquetas o tags.

B. Idioma inglés.

En este caso se llama al método:

```
def tagearSentenciaEn(self, tags):
    tags=word_tokenize(tags)
    tags=pos_tag(tags)
    return tags
```

Figura 15. Script para tagear los tokens del servicio web de tokenizar

y puesto que el texto está en inglés y el corpus del NLTK está especificado completamente en este idioma, resulta más sencillo y simple llevar a cabo el taggeo de las palabras; igual que el método anterior se devuelve una lista de tags.

C. Idioma no encontrado.

Se presentará un mensaje en el cual se especifica que el idioma no se encuentra soportado por la clase.

5. Ws extracción de entidades

El diseño de este prototipo se realiza la extracción de entidades. Para la elaboración de este prototipo se realizó lo siguiente:

- El texto introducido se lo envía a descomponer con la función `sent_tokenize` de la librería NLTK, luego se lo envía al prototipo anterior (WS tokenización) obteniendo una lista con los tokens tageados.

- Utilizando la función *RegexpParser*(función de NLTK) se realiza un conjunto de expresiones regulares para la extracción de las entidades en base a los tags recuperados.
- Con la función *chunker* en fusión con expresiones regulares realizadas, se extraen las entidades.

5.1. Arquitectura

5.1.1. Diagrama de secuencia

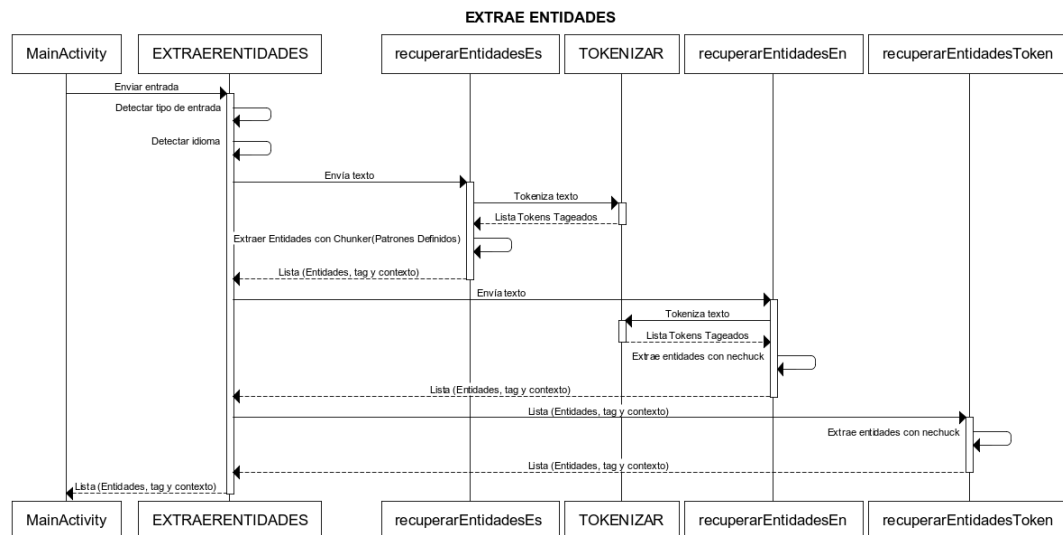


Figura 16 Diagrama de secuencia de la funcionalidad Extrae Entidades

A continuación se describe su comportamiento:

1. Se llama el servicio web EXTRAERENTIDADES y se le envía una entrada, esta puede ser un texto o una lista de tokens con sus tags.
2. Al recibir la entrada se detecta el tipo y si es un texto se detecta el idioma(pueden ser español o ingles)
3. Si el idioma detectado es el español se envía el texto ala función recuperarEntidadeEs.
4. La función recuperarEntidadesEs recibe el texto y lo envía a la funcionalidad TOKENIZAR la cual nos devuelve una lista de tokens tagueados.
5. Luego de recibir los la lista de tokens tagueados se procede a extraer las entidades con chunker una funcionalidad de NLTK definiendo patrones para la extracción. Y devolvemos una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.

6. Si el idioma detectado fue el ingles se procede a enviar el texto a la función `recuperarEntidadesEn`, y esta la envía a la funcionalidad `TOKENIZAR`
7. Luego de recibir la lista con los tokens tageados de la funcionalidad `TOKENIZAR`, extrae las entidades utilizando la función de NLTK `nechunk`. Y se devuelve una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.
8. El servicio Web devuelve una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.

5.2. Implementación

5.2.1. Parámetros.

Como parámetro de entrada se tiene :

- *Entrada*, esta variable puede ser de tipo string o list que almacena el texto o una lista de tokens tageados a ser analizados y procesados .

Como parámetro de salida:

- *Lista entidades*, es una lista desglosada mediante PLN, en la cual se encuentra las entidades extraídas:

```
[['Jap\xc3\x3n', 'NNP']], 'El volc\nc3\nc3\x3n Fuji, 13 de Enero del 2014
[['Educaci\nc3\x3n', 'NNP']], 'El volc\nc3\nc3\x3n Fuji, 13 de Enero del 2
[['Unesco', 'NNP']], 'El volc\nc3\nc3\x3n Fuji, 13 de Enero del 2014 en es
[['Patrimonio', 'NNP'], ('Mundial', 'NN')], 'El Comit\nc3\nc3\x3n del Patrin
[['Phnom', 'NNP'], ('Penh', 'NNP')], 'El Comit\nc3\nc3\x3n del Patrimonio M
[['Fujiyama', 'NNP']], 'Los japoneses consideran sagrado el Fujiyama (mc
[['Fuyi', 'NNP']], 'Los japoneses consideran sagrado el Fujiyama (monte
[['Unesco', 'NNP']], 'Los expertos de la Unesco tambi\nc3\nc3\x3n inscribie
[['Honghe', 'NNP'], ('Hani', 'NNP')], 'Los expertos de la Unesco tambi\nc3\nc3\x3n
[['China', 'NNP']], 'Los expertos de la Unesco tambi\nc3\nc3\x3n inscribier
[['Parque', 'NNP'], ('Nacional', 'NNP'), ('Sehlabathebe', 'NNP')], 'Los
[['Lesoto', 'NNP']], 'Los expertos de la Unesco tambi\nc3\nc3\x3n inscribie
[['Hongye', 'NNP'], ('Hani', 'NNP'), ('cubren', 'NN'), ('16.603', 'CD')]
```

Figura 17 Salida del servicio web de extracción de entidades

5.2.2. Funciones.

Esta clase implementa funciones para:

- Recuperar Entidades a partir de:
 - Texto
- Lista de tokens de entidades extraídas con sus respectivos tags

Para acceder al servicio de esta clase se lo hace mediante la función:

```

def ExtEntidades(self,Entrada):
    if type(Entrada) is str:
        gs= goslate.Goslate()
        idioma=gs.detect(Entrada)
        lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Albanian"}
        idiomaCompleto=lenguajes[idioma]

        if idioma == 'es':
            Lista=self.recuperarEntidadesEs(Entrada)
        elif idioma == 'en':
            Lista=self.recuperarEntidadesEn(Entrada)
        else:
            mensaje='El idioma %s no es soportado'%idiomaCompleto
            print mensaje
            mensaje=gs.translate(mensaje, idioma)
            print mensaje
            Lista=[]
    elif type(Entrada) is list:
        Lista=self.recuperarEntidadesToken(Entrada)
    else:
        mensaje='Tipo de datos no soportados'
        print mensaje
        Lista=[]
    return Lista

```

Figura 18. Script principal para extraer entidades del servicio web de extraer entidades la misma recibe como parámetro el texto ingresado que se almacena en la variable Entrada. Se procede a detectar el tipo de entrada mediante la función

```

elif type(Entrada) is list:
if type(Entrada) is str:

```

Figura 19. Script para detectar tipo de entrada del servicio web de extraer entidades

Pueden darse 3 casos de tipo de entrada:

1. Entrada tipo String

Si el tipo es String se procede a detectar el idioma mediante:

```

gs= goslate.Goslate()
idioma=gs.detect(Entrada)
lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Albanian"}
idiomaCompleto=lenguajes[idioma]

```

Figura 20. Script para traducir del servicio web de tokenizar de esta manera se determina el lenguaje en que se encuentra escrito el texto ingresado para posteriormente llevar a cabo la tokenización y taggeo del mismo.

Pueden darse 3 casos dentro de la extracción de entidades:

A. Idioma Español.

Si el idioma del string de entrada esta en español llama al siguiente método:

```

def recuperarEntidadesEs(self, texto):
    chunker = RegexpParser("""
    ENTI:
        {<NNP|NNPS>+<NNP|NNPS|NN|NNS>} |
        {<NN|NNS>+<NN|NNS><JJ>}
        {<NNP|NNPS><IN|DT><NNP|NNPS|NN|NNS>}
        {<NN|NNS><JJ>|<JJ><NN|NNS>}
        {<NNP|NNPS>}
    ENTIDACOMP:
        {<DT><NN|NNS><ENTI>}
        {<DT><NN|NNS><IN><ENTI>}
        {<ENTI>(<IN>|<IN><DT>)<ENTI|NN|NNS>}
        {<ENTI|ENTIDACOMP><JJ><IN><ENTI|ENTIDACOMP>}
        {<ENTI|ENTIDACOMP><IN><ENTI|ENTIDACOMP>}
        {<ENTI|ENTIDACOMP><IN><ENTI|ENTIDACOMP><IN><ENTI|ENTIDACOMP>}
    ENTIDACOMP2:
        {<ENTI|ENTIDACOMP><IN><ENTI|ENTIDACOMP>}
    FECHA:
        {<LS|CD><IN><ENTI><DT><LS|CD>}
        {<LS|CD><IN><ENTI>}
        {<ENTI><DT><LS|CD>}
        {<ENTI><LS|CD>}
    """)
    ObjTag = Tokenizar()
    Lista = []
    Lista2 = []
    for sentence in sent_tokenize(texto):
        tags=ObjTag.tagear(sentence)
        parsed = chunker.parse(tags)
        for chunk in parsed:
            if hasattr(chunk, 'node'):
                Lista2.append([chunk.leaves(),sentence])
                Lista.append(' '.join(c[0] for c in chunk.leaves()))
    return Lista2

```

Figura 21. Script para extraer entidades en español del servicio web de extraer entidades

Esta función recibe el texto y lo envía a la case de Tokenizacion para ser procesada y nos devuelve los tokens son los tags de texto enviado, mediante la siguiente método:

```

ObjTag = Tokenizar()
Lista = []
Lista2 = []
for sentence in sent_tokenize(texto):
    tags=ObjTag.tagear(sentence)

```

Figura 22. Script para definir patrones del servicio web de extraer entidades

Se define patrones para la extracción mediante NLTK(en la variable **chunker**), dicha extracción se la realiza mediante el método :

```

parsed = chunker.parse(tags)
for chunk in parsed:
    if hasattr(chunk, 'node'):
        Lista2.append([chunk.leaves(),sentence])
        Lista.append(' '.join(c[0] for c in chunk.leaves()))

```

Figura 23. Script para extraer entidades en ingles del servicio web de extraer entidades

en el cual se clasifica y extrae las entidades mediante los tokens con sus tags en contraste a los patrones definidos en la variable **chunker**.

Luego de todos estos procesos se regresa una lista (la variable **Lista2**) que contiene todas las entidades extraídas con sus etiquetas y con el contexto de donde se las extrajo.

B. Idioma inglés

En este caso se llama al método:

```
def recuperarEntidadesEn(self, texto):
    ObjTag = Tokenizar()
    Lista = []
    Lista2= []
    for sentence in sent_tokenize(texto):
        tags=ObjTag.tagear(sentence)
        parsed = ne_chunk(tags)
        for chunk in parsed:
            if hasattr(chunk, 'node'):
                Lista2.append([chunk.leaves(), sentence])
                Lista.append(' '.join(c[0] for c in chunk.leaves()))
    return Lista2
```

Figura 24. Script para extraer entidades en español del servicio web de extraer entidades

Esta función recibe el texto y lo envía a la case de Tokenizacion para ser procesada y nos devuelve los tokens son los tags de texto enviado, mediante la siguiente método:

```
ObjTag = Tokenizar()
Lista = []
Lista2= []
for sentence in sent_tokenize(texto):
    tags=ObjTag.tagear(sentence)
```

Figura 25. Script para tokenizar del servicio web de extraer entidades

Y como el texto esta en ingles y el corpus de NLTK esta especificado para ese idioma, es mas sencillo la extracción, y se lo hace con el siguiente método:

```
parsed = ne_chunk(tags)
for chunk in parsed:
    if hasattr(chunk, 'node'):
        Lista2.append([chunk.leaves(), sentence])
        Lista.append(' '.join(c[0] for c in chunk.leaves()))
```

Figura 26. Script para extraer entidades en español del servicio web de extraer entidades

se clasifica y extrae las entidades mediante los tokens con sus tags en contraste a los patrones definidos por NLTK en la variable **nechunk**.

Luego de todos estos procesos se regresa una lista (la variable **Lista2**) que contiene todas las entidades extraídas con sus etiquetas y con el contexto de donde se las extrajo.

C. Idioma no encontrado.

Se presentará un mensaje en el cual se especifica que el idioma no se encuentra soportado por la clase.

2. Entrada tipo Lista

Si el tipo es un tipo list(lista) llama a este método:

```
def recuperarEntidadesToken(self,tokens):
    Lista = []
    Lista2 = []
    sentence= ' '.join(n[0] for n in tokens)
    parsed = ne_chunk(tokens)
    for chunk in parsed:
        if hasattr(chunk, 'node'):
            Lista2.append([chunk.leaves(),sentence])
            Lista.append(' '.join(c[0] for c in chunk.leaves()))
    return Lista2
```

Figura 27. Script para extraer tokens de una lista del servicio web de extraer entidades

Y como ya la entrada ya es una lista de tokens con sus respectivos tags, para corpus de NLTK le resulta mas sencillo para poder extraer las entidades; este método devuelve la lista de las entidades extraídas.

3. Tipo no encontrado

Se presentará un mensaje en el cual se especifica que el tipo de entrada no se encuentra soportado por la clase.

6. Ws desambiguación y enlace.

El diseño de este prototipo se realiza la extracción de entidades. Para la elaboración de este prototipo se realizo lo siguiente:

- El texto introducido se lo envía al prototipo anterior(WS tokenización) obteniendo una lista con las entidades extraídas.
- Utilizando consultas SPARQL se realiza consultas ala Dbpedia para desambiguar y extraer Uris para enlazarlos.

6.1. Arquitectura.

6.1.1. Diagrama de secuencia

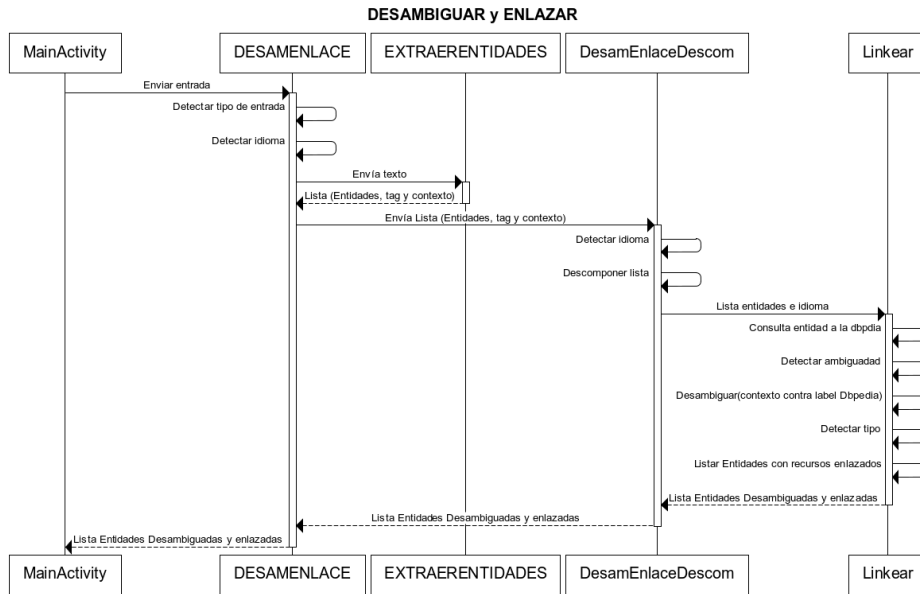


Figura 28 Diagrama de secuencia de la funcionalidad Desambiguar y Enlazar

A continuación se describe su comportamiento:

1. Se llama el servicio web *DESAMENLACE* y se le envía una entrada, esta puede ser un texto o una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.
2. Al recibir la entrada se detecta el tipo y si es un texto se detecta el idioma(pueden ser español o ingles)
3. Si el idioma detectado es el español o ingles se envía el texto a la funcionalidad *EXTRAERENTIDADES*.
4. La funcionalidad *EXTRAERENTIDADES* recibe el texto y nos devuelve una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.
5. Una vez teniendo la lista de la entidades se la envía a DesamEnlaceDescom y esta función detecta el idioma y descompone la lista para enviarla a las función *Linkear*.
6. La función linkear recibe las lista y realiza consultas a la Dbpedia de la entidades.
7. Luego detecta si hay ambigüedad en la consulta de la entidad y desambigua con el contexto de la entidad contra los datos de la Dbpedia. Y procede a enlazar los datos con los recursos de la Dbpedia.
8. Posteriormente detecta el tipo de entidad y lista los enlaces a los recursos ordenándolos desacuerdo a la relevancia.

9. Luego de recibir los la lista de tokens tageados se procede a extraer las entidades con chunker una funcionalidad de NLTK definiendo patrones para la extracción. Y devolvemos una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad.
10. Y se devuelve la lista de la entidades Desambiguadas y enlazadas.
11. El servicio web devuelve dicha lista que contiene las entidades desambiguados con los enlaces a los recursos de la Dbpedia.

6.2. Implementación.

6.2.1. Parámetros.

Como parámetro de entrada se tiene :

- Entrada, esta variable puede ser de tipo string que almacena el texto o una lista con las entidades y su contexto a ser analizados y procesados .

Como parámetro de salida:

Lista entidades enlazadas, es una lista desglosada, en la cual se encuentra las entidades extraídas, desambiguadas y enlazadas:

```
[u'Quito', [[1, u'http://dbpedia.org/resource/Quito', 'GEO']]]
[u'Ecuador', [[1, u'http://dbpedia.org/resource/Ecuador', 'GEO']]]
[u'Peru', [[1, u'http://dbpedia.org/resource/Peru', 'GEO']]]
[u'Argentina', [[1, u'http://dbpedia.org/resource/Argentina', 'GEO']]]
[u'Israel', [[1, u'http://dbpedia.org/resource/Israel', 'GEO']]]
[u'Korea', [[1, u'http://dbpedia.org/resource/Korea', 'ORG']]]
[u'China', [[1, u'http://dbpedia.org/resource/China', 'GEO']]]
[u'North Korea', [[1, u'http://dbpedia.org/resource/North_Korea', 'GEO']]]
[u'Loja', [[1, u'http://dbpedia.org/resource/Loja', 'GEO']]]
[u'Ambato', [[2, u'http://dbpedia.org/resource/Ambato,_Ecuador', 'GEO'], [2, u'http://dbpedia.org/resource/Ambato-Madagascar', 'Sin Tipo']]]
```

Figura 29 Salida del servicio web de desambiguar y enlazar

6.2.2. Funciones.

Esta clase implementa funciones para:

- Desambigua y Enlaza entidades a partir de:
 - Texto
 - Lista con entidades
- Lista de entidades desambiguadas con su enlace a la Dbpedia.

Para acceder al servicio de esta clase se lo hace mediante la función:

```
def DesamEnlace(self,Entrada):
    DesamEnl=[]
    if type(Entrada) is str:
        gs= goslate.Goslate()
        idioma=gs.detect(Entrada)
        lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Alb
        idiomaCompleto=lenguajes[idioma]

        if idioma == 'es' or idioma == 'en':
            Entrada=ObjExt.recuperarEntidadesEs(Entrada)
            DesamEnl=DesamEnlaceDescom(Entrada)
        else:
            mensaje='El idioma %s no es soportado'%idiomaCompleto
            #print mensaje
            mensaje=gs.translate(mensaje, idioma)
            #print mensaje
            DesamEnl=[]
    elif type(Entrada) is list:
        DesamEnl=DesamEnlaceDescom(Entrada)
    else:
        mensaje='Tipo de datos no soportados'
        #print mensaje
        DesamEnl=[]

    return DesamEnl
```

Figura 30. Script principal del servicio web de desambiguar y enlazar

la misma recibe como parámetro el texto ingresado que se almacena en la variable **Entrada**. Se procede a detectar el tipo de entrada mediante la función

```
if type(Entrada) is str:
elif type(Entrada) is list:
```

Figura 31. Script para detectar tipo de entrada del servicio web de desambiguar y enlazar

Pueden darse 3 casos de tipo de entrada:

1. Entrada tipo String

Si el tipo es String se procede a detectar el idioma mediante:

```
gs= goslate.Goslate()
idioma=gs.detect(Entrada)
lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Al
idiomaCompleto=lenguajes[idioma]
```

Figura 32. Script para detectar idioma del servicio web de desambiguar y enlazar

de esta manera se determina el lenguaje en que se encuentra escrito el texto ingresado para posteriormente llevar a cabo la tokenización y taggeo del mismo.

Pueden darse 2 casos dentro de la desambiguación y enlace de entidades:

A. Idioma Español o Ingles.

Si el idioma del string de entrada esta en español o ingles llama al servicio web de extracción de entidades:

```

def ExtEntidades(self,Entrada):
    if type(Entrada) is str:
        gs= goslate.Goslate()
        idioma=gs.detect(Entrada)
        lenguajes={"auto":"Detect language","af":"Afrikaans","sq":"Albanian","ar":"Arabic","az":"Azerbaijani","be":"Belarusian","bg":"Bulgarian","bn":"Bengali","bs":"Bosnian","ca":"Catalan","cs":"Czech","da":"Danish","de":"German","el":"Greek","en":"English","es":"Spanish","et":"Estonian","eu":"Basque","fa":"Persian","fi":"Finnish","fr":"French","ga":"Irish","he":"Hebrew","hi":"Hindi","hr":"Croatian","hu":"Hungarian","id":"Indonesian","is":"Icelandic","it":"Italian","ja":"Japanese","ka":"Georgian","kk":"Kazakh","km":"Khmer","kn":"Kannada","ko":"Korean","lt":"Lithuanian","lv":"Latvian","mk":"Macedonian","ml":"Malayalam","mn":"Mongolian","mr":"Marathi","ms":"Malay","mt":"Maltese","ne":"Nepali","nl":"Dutch","no":"Norwegian","or":"Oriya","pa":"Punjabi","pl":"Polish","pt":"Portuguese","ro":"Romanian","ru":"Russian","sk":"Slovak","sl":"Slovenian","sq":"Albanian","sr":"Serbian","sv":"Swedish","sw":"Swahili","ta":"Tamil","te":"Telugu","th":"Thai","tl":"Tagalog","tr":"Turkish","uk":"Ukrainian","ur":"Urdu","vi":"Vietnamese","wa":"Walloon","xh":"Xhosa","yi":"Yiddish","yo":"Yoruba","zh":"Chinese","zu":"Zulu"}
        idiomaCompleto=lenguajes[idioma]

        if idioma == 'es':
            Lista=self.recuperarEntidadesEs(Entrada)
        elif idioma == 'en':
            Lista=self.recuperarEntidadesEn(Entrada)
        else:
            mensaje='El idioma %s no es soportado'%idiomaCompleto
            print mensaje
            mensaje=gs.translate(mensaje, idioma)
            print mensaje
            Lista=[]
    elif type(Entrada) is List:
        Lista=self.recuperarEntidadesToken(Entrada)
    else:
        mensaje='Tipo de datos no soportados'
        print mensaje
        Lista=[]
    return Lista

```

Figura 33. Script del servicio web extraer entidades

Este servicio web nos devolverá una lista desglosada mediante PLN, en la cual se encuentra las entidades extraídas con su contexto.

```

[[('Japón', 'NNP'), ('El volcán Fuji', '13 de Enero del 2014'), ('Educación', 'NNP'), ('El volcán Fuji', '13 de Enero del 2014'), ('Unesco', 'NNP'), ('El volcán Fuji', '13 de Enero del 2014'), ('Patrimonio', 'NNP'), ('Mundial', 'NN'), ('El Comité del Patrimonio', 'NNP'), ('Phnom', 'NNP'), ('Penh', 'NNP'), ('El Comité del Patrimonio', 'NNP'), ('Fujiyama', 'NNP'), ('Los japoneses consideran sagrado el Fujiyama (monte)', 'NNP'), ('Fuji', 'NNP'), ('Los japoneses consideran sagrado el Fujiyama (monte)', 'NNP'), ('Unesco', 'NNP'), ('Los expertos de la Unesco también inscriben', 'NNP'), ('Honghe', 'NNP'), ('Hani', 'NNP'), ('Los expertos de la Unesco también inscriben', 'NNP'), ('China', 'NNP'), ('Los expertos de la Unesco también inscriben', 'NNP'), ('Parque', 'NNP'), ('Nacional', 'NNP'), ('Sehlabathebe', 'NNP'), ('Los', 'NNP'), ('Lesoto', 'NNP'), ('Los expertos de la Unesco también inscriben', 'NNP'), ('Hongye', 'NNP'), ('Hani', 'NNP'), ('cubren', 'NN'), ('16.603', 'CD')]

```

Figura 34 Salida del servicio web de extraer entidades

De este modo la entrada la convertimos en una lista.

B. Idioma no encontrado.

Se presentará un mensaje en el cual se especifica que el idioma no se encuentra soportado por la clase.

2. Entrada tipo Lista

Si el tipo es un tipo list(lista) llama a este método:

```

def DesamEnlaceDescom(self,Entrada):
    DesamEnl=[]
    paraidioma=' '.join(entidad[1] for entidad in Entrada)
    gs = goslate.Goslate()
    idioma=gs.detect(paraidioma)
    for entidad in Entrada:
        entidades=' '.join(c[0] for c in entidad[0])
        contexto= entidad[1]
        link=self.Linkea(entidades,contexto,idioma)
        DesamEnl.append(link)
    return DesamEnl

```

Figura 35 Script extraer entidades de una lista del servicio web de desambiguar y enlazar

Este método separa la lista y envía las entidades separadas con el contexto a este otro método para ser procesadas:

```
def Linkear(self,entidad,contexto,idioma):
    linkear=[]
    entidad=entidad.decode('utf-8')
    try:
        results = self.ConsuDbpedia(entidad,'rdfs:label',idioma)
        #results = self.ConsuDbpedia(entidad,'rdfs:label','en')
        if results["results"]["bindings"]==[]:
            if idioma=='en':
                results = self.ConsuDbpedia(entidad,'rdfs:label','es')
            else:
                results = self.ConsuDbpedia(entidad,'rdfs:label','en')
    except Exception, e:
        print e
        return
    for result in results["results"]["bindings"]:
        link=(result["label"]["value"])
        if 'Category' in link:
            continue
        try:
            results2 = self.ConsuDbpedia2(link,'rdf:type')
        except Exception, e:
            continue

        results2 = results2["results"]["bindings"]
        if results2 ==[]:
            UrlyTipo=[]
            results3=self.ConsuDbpedia3(link,'dbpedia-owl:wikiPageDisambiguates')
            results3=results3["results"]["bindings"]
            if results3!=[]:
                for result3 in results3:
                    link=(result3["label"]["value"])
                    if 'Category:' in link:
                        continue
                    try:
                        results4 = self.ConsuDbpedia2(link,'rdf:type')
                    except Exception, e:
                        #print e
                        continue

                    results4 = results4["results"]["bindings"]
                    TipoFinal=self.TipoEntidad(results4)
                    UrlyTipo.append([link,TipoFinal])
                    linkear=self.SeleccionaTipo(UrlyTipo,contexto)
            else:
                TipoFinal=self.TipoEntidad(results2)
                linkear=[[1,link,TipoFinal]]
        else:
            TipoFinal=self.TipoEntidad(results2)
            linkear=[[1,link,TipoFinal]]
        #break
    LisDesyEnlase=[entidad,linkear]
    return LisDesyEnlase
```

Figura 36 Script para enlazar del servicio web de desambiguar y enlazar

Esta función recibe la entidad con su contexto y el idioma en el que esta, una vez obtenido esto se envía la entidad para ser procesada en este otro método:

```
def ConsuDbpedia(self,entidades,predicado,idioma):
    sparql = SPARQLWrapper("http://dbpedia.org/sparql")
    sparql.setQuery("""
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT *
WHERE {
?label %s "%s"@%s
FILTER regex(str(?label), "http://dbpedia.org/resource/", "i")
}
""")%(predicado,entidades,idioma)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()
```

Figura 37 Script de consulta sparql del servicio web de desambiguar y enlazar

Este método recibe la entidad a ser procesada, así como también su idioma y el valor de predicado para poder realizar una consulta SPARQL a la Dbpedia, este método nos devolverá una lista con todos los resultados encontrados en la consulta.

Dependiendo de lo q nos devuelva el método anterior se pueden dar dos casos:

A. Entidad Ambigua.

Si la entidad es ambigua, es decir la lista obtenida del método ConsuDbppedia devolvió un enlace a un recurso que contiene un predicado wikiPageDisambiguates, se procede a realizar otra consulta para extraer todos los recursos relacionados ha esa entidad:

```
def ConsuDbppedia3(self, link, predicado):
    sparql = SPARQLWrapper("http://dbpedia.org/sparql")
    sparql.setQuery("""
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/20002-rdf-syntax-ns#>
SELECT *
WHERE {
  <%s> %s ?label
  FILTER regex(str(?label), "http://dbpedia.org/resource/", "i")
}
""")%(link, predicado)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()
```

Figura 38 Script de consulta sparql del servicio web de desambiguar y enlazar

Ese método contiene otra consulta SPARQL que nos devuelve la lista con todos los recursos posibles para la entidad extraída. Esta lista se lo envía a este método:

```
def SeleccionaTipo(self, listaTipos, contexto):
    listafrec=[]
    for link in listaTipos:
        try:
            LabelsExtraidos=self.ConsuDbppediaExtraLabel(link[0], 'rdfs:label')
        except Exception, e:
            #print e
            continue
        LabelsExtraidos=LabelsExtraidos["results"]["bindings"]
        labels=(' '.join(self.eliminasignos(c["label"]["value"]) for c in LabelsExtraidos)).lower()
        labelsSinDuplicados=list(set(labels))
        count=0
        for label in labelsSinDuplicados:
            if label in self.eliminasignos(contexto.lower()).split(' '):
                count=count+1
        listafrec.append([count, link[0], link[1]])
    ListaOrdenado= sorted(listafrec, key=lambda x: x[0], reverse=True)
    return ListaOrdenado
```

Figura 39 Script para definir el tipo de entidad del servicio web de desambiguar y enlazar

A este método llegan como paramentos de entrada la lista de los recursos de la entidad ambigua y el contexto de dicha entidad, para procede a determinar cual es el recurso valido o mas cercano, para esto nos servimos de otra consulta para extraer los rdfs:label de los recursos, para proceder ha hacer una comparación con el contexto de la entidad y los labels

extraídos con la consulta ordenándolos en una lista de acuerdo a al que tenga mas concordancia en dicha comparación. El método devuelve la lista ordenada de los recursos encontrados y las entidades.

B. Entidad no Ambigua

Si la entidad no es ambigua nos devuelve la uri del recursos con el tipo de recurso

3. Tipo no encontrado

Se presentará un mensaje en el cual se especifica que el tipo de entrada no se encuentra soportado por la clase.

7. App Cliente.

La app cliente es el frontal para la integración de los Servicios Web. Para la elaboración de este frontal o app cliente se realizo lo siguiente:

- El frontal se lo hizo con JavaScript y hml, cosumiendo los Servicios Web realizados.
- Este frontal envía un texto uo url a los Servicios Web y presenta en tablas los resultados asi como también tiene un botón para poder descargar los resultados en un archivo json.

7.1. Implementación.

7.1.1. Parámetros.

Como parámetro de entrada se tiene :

- Una caja de texto que se puede introducir ya sea un texto o una url.
- Un combo box para seleccionar li se quiere tokenizar, extraer entidades o desambiguar y linkear.

Como parámetro de salida:

- El resultado se presenta en una tabla, también se puede descargar el resultado en un archivo json.

En el siguiente figura se muestra el frontal o App cliente:

Servicios Web

Desambiguar y Enlazar +

This module deals with the mathematical elements essential to understanding physics courses, namely the study of real functions, derivation and integration of a function with one and several real variables, the development of a function, some elements of numerical calculations and, finally, solving a system of linear equations. Learning activities of different difficulty levels are developed with formal assessments. Moreover, online word and useful links enable students to study certain topics in detail. Finally, the students will also be able to use software such as "Microsoft Excel 2000" and Maxima.

Extraer

ENLACE

Entidad	Enlace	Tipo
Microsoft Excel	http://dbpedia.org/resource/Microsoft_Excel	Software-Work
Maxima	http://dbpedia.org/resource/Maxima_(software)	Software-Work

Descargar JSON

Figura 40 App cliente

VALIDACIÓN Y PRUEBAS

Se ha llevado a cabo algunas pruebas sobre los Servicios Web, entre las mismas tenemos:

8. Validación.

Se las llevara acabo sobre todos los Servicios Web, asi como también al app Cliente:

- Servicio Web Tokenizar
- Servicio Web Extraer Entidades
- Servicio Web Desambiguar y Enlazar
- App Cliente

La prueba que se llevara acabo será la Prueba de Funcionalidad.

8.1. Prueba de Funcionalidad

8.1.1. Objetivo

El objetivo de esta prueba será comprobar un desempeño correcto de los servicios, lo que se busca verificar es:

- Los servicios Web funcionan de manera adecuada
- No presenta errores en ejecución
- No se paralizan los servicios
- Responde como fue programada, devuelve el resultado lo que cada servicio debe devolver.

8.1.2. Escenario

Las pruebas se llevaran acabo en un servidor local.

8.1.3. Pruebas sobre el Servicio Web Tokenizar

Tabla 1 Pruebas sobre el Servicio Web Tokenizar

Texto introducido	Errores	Soluciones
Cadena vacía	Ninguno	No necesita
Solo un espacio o espacios vacíos	Ninguno	No necesita
Caracteres especiales	Ninguno	No necesita
Caracteres "extranjeros"	Ninguno	No necesita
Texto en idioma diferente a español o ingles	Ninguno	No necesita

8.1.1. Pruebas sobre el Servicio Web Extraer Entidades

Tabla 2 Pruebas sobre el Servicio Web Extraer Entidades

Texto introducido	Errores	Soluciones
Cadena vacía	Ninguno	No necesita
Solo un espacio o espacios vacíos	Ninguno	No necesita
Caracteres especiales	Ninguno	No necesita
Caracteres "extranjeros"	Ninguno	No necesita
Texto en idioma diferente a español o ingles	Ninguno	No necesita

8.1.1. Pruebas sobre el Servicio Web Desambiguar Linkear

Tabla 3 Pruebas sobre el Servicio Web Desambiguar Linkear

Texto introducido	Errores	Soluciones
Cadena vacía	Ninguno	No necesita
Solo un espacio o espacios vacíos	Ninguno	No necesita
Caracteres especiales	Ninguno	No necesita
Caracteres "extranjeros"	Ninguno	No necesita
Texto en idioma diferente a español o ingles	Ninguno	No necesita

DISCUSIÓN

Para dar solución al problema encontrado, se a desarrollado Servicios web, los cuales dan solución al problema.

9. Discusión

- Con la creación del servicio web de tokenizar se a podido taggear o etiquetar cada palabra un texto o del contenido de una pagina web, este etiquetado describe la función de la palabra con respecto a su contexto, y nos sirve para poder extraer las entidades localizadas en ese contenido.
- Mediante la creación del servicio web extracción de entidades, se logro encontrar las entidades relevantes que componen un texto o el contenido de una pagina web, para así etiquetar de mejor manera ese contenido, con las entidades extraídas, esto ayudara a clasificar mejor dicha información, como también se facilitara búsquedas futuras de ese contenido.
- Al tener el servicio de desambiguar y enlazar, se dará mas relevancia a las entidades encontradas ya que también se enlaza la entidad con un recurso de la Dbpedia, si dicha entidad se encuentra en la base de conocimientos (Dbpedia). La entidad será desambiguada con la base de conocimientos, ya que el recurso será escogido de entre muchos similares pero el que tenga mas relación con el contexto de la entidad encontrada será el q se asignara a la entidad. Esto ayuda a enriquecer las entidades y a su vez el contenido, ya que con estas relaciones se podrán establecer conexiones entre los contenidos relacionados. Con la ayuda de esto se podrían realizar búsquedas mas exactas y precisas sobre un tema.

CONCLUSIONES

- Se logro cumplir con los objetivos planteados para el proyecto y se a dado solución a la necesidad de extracción de entidades.
- Mediante PNL (procesamiento de lenguaje natural) se pueden extraer información adicional a partir de un contenido, dicha información es útil cuando se desea ampliar la información o enriquecer los datos que se haya obtenido.
- Los Servicios Web creados facilita no sólo la gestión de etiquetas, tags, entidades contenido dentro de un texto, sino que además al ser implementados como servicios a ser consumido brinda la oportunidad de a futuro ser parte de un sistema mucho más amplio y versátil para el manejo de la información, así como también a una migración hacia la web semántica.

RECOMENDACIONES

- Para el procesamiento de texto el lenguaje python es muy bueno, por ese motivo se lo utilizo como programa principal, ya que este facilita mucho en trabajo con texto y además tiene muchas librerías para el procesamiento de lenguaje natural siendo NLTK la mejor.

BIBLIOGRAFÍA

- W3C. (n.d.). Web Services Architecture. Retrieved 19 de 01 de 2014 from W3C:
<http://www.w3.org/TR/ws-arch/>
- Python. (n.d.). About Python. Retrieved 19 de 01 de 2014 from Python:
<http://www.python.org/about/>
- González, M. (n.d.). ESTUDIO DE ARQUITECTURAS DE REDES ORIENTADAS A SERVICIO. Retrieved 19 de 01 de 2014 from UpCommons:
[http://upcommons.upc.edu/pfc/bitstream/2099.1/12312/1/ESTUDIO_DE_ARQUITEC
TURAS_DE_REDES_ORIENTADAS_A_SERVICIO.pdf](http://upcommons.upc.edu/pfc/bitstream/2099.1/12312/1/ESTUDIO_DE_ARQUITEC_TURAS_DE_REDES_ORIENTADAS_A_SERVICIO.pdf)
- UPM. (n.d.). ¿Qué es OCW? Retrieved 20 de 01 de 2014 from OpenCourseWare de la Universidad Politécnica de Madrid: <http://ocw.upm.es/bfque-es-ocw>
- Loper, E., & Bird, S. (n.d.). NLTK: The Natural Language Toolkit. Retrieved 21 de 01 de 2014 from Cornell University: <http://arxiv.org/pdf/cs/0205028v1.pdf>
- Vallez, M., Rovira, C., Codina, L., & Pedraza, R. (n.d.). Procedimientos para la extracción de palabras clave de páginas web basados en criterios de posicionamiento en buscadores. Retrieved 21 de 01 de 2014 from UPF:
http://www.upf.edu/hipertextnet/numero-8/extraccion_keywords.html
- La Web Semántica y las Tecnologías del Lenguaje Humano. (n.d.). Retrieved 23 de 01 de 2014 from e-Lis:
[http://eprints.rclis.org/15586/1/La%20Web%20Sem%C3%A1ntica%20y%20las%20T
ecnolog%C3%ADas%20del%20Lenguaje%20Humano%20-%20Preprint.pdf](http://eprints.rclis.org/15586/1/La%20Web%20Sem%C3%A1ntica%20y%20las%20Tecnolog%C3%ADas%20del%20Lenguaje%20Humano%20-%20Preprint.pdf)
- Unesco. (2011). A Basic Guide to Open Educational Resources. Retrieved 22 de 2 de 2014 from www.unesco.org/education
- Bolshakov, I., & Gelbukh, A. (2004). Computational Linguistics. Models, Resources, Applications. Ciencia de la Computacion Primera Edición .
- Alvarado Ruiz, P. A., Guamán Eras, D. E., & Sigcho Armijos, J. P. (n.d.). Aplicación de tecnologías móviles para la búsqueda de recursos educativos abiertos. Retrieved 20 de 01 de 2014 from Bibliotec UTPL:
[http://dspace.utpl.edu.ec/jspui/bitstream/123456789/4938/1/Pablo%20Antonio%20Alv
arado%20Ruiz.pdf](http://dspace.utpl.edu.ec/jspui/bitstream/123456789/4938/1/Pablo%20Antonio%20Alvarado%20Ruiz.pdf)
- UNESCO. (n.d.). UNESCO. Retrieved 20 de 02 de 2014 from What are Open Educational Resources (OERs)?: [http://www.unesco.org/new/en/communication-and-
information/access-to-knowledge/open-educational-resources/what-are-open-
educational-resources-oers/](http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-educational-resources/what-are-open-educational-resources-oers/)
- Piedra, N., Tovar, E., López, J., Chicaiza, J., & Martinez, O. (4 de abril de 2011). www.ocw.org. Retrieved 18 de 02 de 2014 from
<http://conference.ocwconsortium.org/index.php/2011/cambridge/paper/view/162>
- UTPL. (20 de 01 de 2014). OpenCourseWare UTPL. Retrieved 21 de 02 de 2014 from UTPL OCW: <http://ocw.utpl.edu.ec/>
- PNL. (n.d.). Retrieved 23 de 02 de 2014 from El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines:
<http://www.upf.edu/hipertextnet/numero-5/pln.html>
- Birrell, & Nelson. (n.d.). RCP. From Implementing Remote Procedure Calls:
<http://www.cs.princeton.edu/courses/archive/fall03/cs518/papers/rpc.pdf>

- SunMicrosystems, I. (04 de 2005). RESTfulWeb Services. Retrieved 03 de 02 de 2014 from <http://docs.huihoo.com/glassfish/v3/820-4867.pdf>
- Berners-Lee, T. (agosto de 1996). www.w3c.org. Retrieved 13 de 02 de 2014 from Actas de la V Conferencia Internacional World Wide Web: www.w3c.org
- Nltk. (n.d.). Retrieved 27 de 01 de 2014 from NLTK: <http://nltk.org/>
- OCW Consortium. (2012). OpenCourseWare Consortium. Retrieved 7 de 02 de 2014 from <http://www.ocwconsortium.org/>
- Fielding, R. T. (n.d.). Architectural Styles and the Design of Network-based Software Architectures. Retrieved 02 de 2014 from https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf
- Piedra , N., Chicaiza, J., & Lopez, J. (2012). Combining Linked Data and Mobiles Devices to improve access to OCW. Retrieved 18 de 02 de 2014 from Mendeley: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6201202&contentType=Conference+Publications&searchWithin%3Dpiedra%2C+n%26queryText%3Dpiedra>
- Tovar, E., Piedra, N., Chicaiza, J., Lopez , J., & Martínez, O. (2012). Development and promotion of OERs. Outcomes of an international research project under OpenCourseWare model. . Journal of Universal Computer Science .
- Piedra , N., Chicaiza, J., & Lopez, J. (2012). Combining Linked Data and Mobiles Devices to improve access to OCW. Retrieved 25 de Julio de 2012 from Mendeley: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6201202&contentType=Conference+Publications&searchWithin%3Dpiedra%2C+n%26queryText%3Dpiedra>
- W3C. (n.d.). RDF. Retrieved 25 de 04 de 14 from RDF: <http://www.w3.org/RDF/>
- W3C. (n.d.). SPARQL Query Language for RDF. Retrieved 28 de 04 de 2014 from RDF sparql-query: <http://www.w3.org/TR/rdf-sparql-query/>
- DBpedia. (n.d.). About. Retrieved 05 de 05 de 2014 from DBpedia: <http://wiki.dbpedia.org/About>
- DBpedia. (n.d.). Datasets. Retrieved 05 de 05 de 2014 from DBpedia Datasets: <http://wiki.dbpedia.org/Datasets>

ANEXO 1: ONTOLOGÍA PARA LA DESCRIPCIÓN DE LOS SERVICIOS WEB

```

<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY dcterms "http://purl.org/dc/terms/" >
  <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
  <!ENTITY dc "http://purl.org/dc/elements/1.1/" >
  <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
  <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <!ENTITY swpo "http://sw-portal.deri.org/ontologies/swportal#" >
  <!ENTITY ssn "http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#" >
]>
<rdf:RDF xmlns="http://utpl.edu.ec/WebServices#"
  xml:base="http://utpl.edu.ec/WebServices"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:swpo="http://sw-portal.deri.org/ontologies/swportal#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:ssn="http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/">
  <owl:Ontology rdf:about="http://utpl.edu.ec/WebServices"/>
  <!--
  //////////////////////////////////////
  //
  // Anotaciones de las Propiedades
  //
  //////////////////////////////////////
  -->
  <!-- http://purl.org/dc/elements/1.1/contributor -->
  <owl:AnnotationProperty rdf:about="&dc;contributor"/>

  <!-- http://purl.org/dc/elements/1.1/date -->
  <owl:AnnotationProperty rdf:about="&dc;date"/>

  <!-- http://purl.org/dc/elements/1.1/description -->

```

```
<owl:AnnotationProperty rdf:about="&dc:description"/>
```

```
<!-- http://purl.org/dc/elements/1.1/identifier -->
```

```
<owl:AnnotationProperty rdf:about="&dc;identifier"/>
```

```
<!-- http://purl.org/dc/elements/1.1/language -->
```

```
<owl:AnnotationProperty rdf:about="&dc;language"/>
```

```
<!-- http://purl.org/dc/elements/1.1/publisher -->
```

```
<owl:AnnotationProperty rdf:about="&dc;publisher"/>
```

```
<!-- http://purl.org/dc/elements/1.1/source -->
```

```
<owl:AnnotationProperty rdf:about="&dc;source"/>
```

```
<!-- http://purl.org/dc/elements/1.1/title -->
```

```
<owl:AnnotationProperty rdf:about="&dc;title"/>
```

```
<!-- http://purl.org/dc/elements/1.1/type -->
```

```
<owl:AnnotationProperty rdf:about="&dc;type"/>
```

```
<!-- http://purl.org/dc/terms/created -->
```

```
<owl:AnnotationProperty rdf:about="&dcterms;created"/>
```

```
<!-- http://purl.org/dc/terms/modified -->
```

```
<owl:AnnotationProperty rdf:about="&dcterms;modified"/>
```

```
<!--
```

```
////////////////////////////////////////////////////////////////
```

```
//
```

```
// Propiedades de Objeto
```

```
//
```

```
////////////////////////////////////////////////////////////////
```

```
-->
```

```

<!-- http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#hasInput -->
<owl:ObjectProperty rdf:about="&ssn;hasInput">
  <rdfs:domain rdf:resource="http://utpl.edu.ec/WebServices#WebService"/>
  <rdfs:range rdf:resource="&ssn;Input"/>
</owl:ObjectProperty>

```

```

<!-- http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#hasOutput -->
<owl:ObjectProperty rdf:about="&ssn;hasOutput">
  <rdfs:domain rdf:resource="http://utpl.edu.ec/WebServices#WebService"/>
  <rdfs:range rdf:resource="&ssn;Output"/>
</owl:ObjectProperty>

```

```

<!--
////////////////////////////////////
//
// Datos de Propiedades
//
////////////////////////////////////
-->

```

```

<!-- http://purl.org/dc/elements/1.1/creator -->
<owl:DatatypeProperty rdf:about="&dc;creator"/>

```

```

<!-- http://purl.org/dc/elements/1.1/description -->
<owl:DatatypeProperty rdf:about="&dc;description"/>

```

```

<!-- http://purl.org/dc/elements/1.1/identifier -->
<owl:DatatypeProperty rdf:about="&dc;identifier"/>

```

```

<!-- http://purl.org/dc/elements/1.1/language -->
<owl:DatatypeProperty rdf:about="&dc;language"/>

```

```

<!-- http://purl.org/dc/elements/1.1/publisher -->
<owl:DatatypeProperty rdf:about="&dc;publisher"/>

```



```

<!-- http://purl.org/dc/elements/1.1/title -->
<owl:DatatypeProperty rdf:about="&dc;title"/>


<!-- http://purl.org/dc/elements/1.1/type -->
<owl:DatatypeProperty rdf:about="&dc;type"/>


<!-- http://purl.org/dc/terms/created -->
<owl:DatatypeProperty rdf:about="&dcterms;created"/>


<!-- http://purl.org/dc/terms/modified -->
<owl:DatatypeProperty rdf:about="&dcterms;modified"/>


<!-- http://sw-portal.deri.org/ontologies/swportal#isVersion -->
<owl:DatatypeProperty rdf:about="&swpo;isVersion"/>


<!--
////////////////////////////////////
//
// Clases
//
////////////////////////////////////
-->


<!-- http://utpl.edu.ec/WebServices#WebService -->
<owl:Class rdf:about="http://utpl.edu.ec/WebServices#WebService"/>


<!-- http://www.w3.org/2002/07/owl#Thing -->
<owl:Class rdf:about="&owl;Thing"/>


<!-- http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#Input -->
<owl:Class rdf:about="&ssn;Input"/>

```

```

<!-- http://www.w3.org/2005/Incubator/ssn/ssnx/ssn#Output -->
<owl:Class rdf:about="&ssn;Output"/>

<!--
////////////////////////////////////
//
// Servicios Web Descritos
//
////////////////////////////////////
-->

<!-- http://localhost:5000/WebServices/desambiguarenlazar/ -->
<owl:NamedIndividual
rdf:about="http://localhost:5000/WebServices/desambiguarenlazar/">
  <rdf:type rdf:resource="http://utpl.edu.ec/WebServices#WebService"/>
  <dcterms:modified rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:modified>
  <dcterms:created rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:created>
  <swpo:isVersion rdf:datatype="&xsd:string">1.0</swpo:isVersion>
  <dc:creator rdf:datatype="&xsd:string">Jhonny Zaruma</dc:creator>
  <dc:description rdf:datatype="&xsd:string">Servicio Web que desambigua y enlaza con
la dbpeda a partir de entidades q son extraídas del texto q se le envié a este servicio, y este
servicio devuelve como resultado un json con una lista de las entidades con un enlace a la
Dbpedia</dc:description>
  <dc:publisher rdf:datatype="&xsd:string">TAW-SBC</dc:publisher>
  <dc:title rdf:datatype="&xsd:string">Web Service Desambiguar y Enlazar</dc:title>
  <dc:language
rdf:datatype="&xsd:anyURI">http://dbpedia.org/resource/English_language</dc:language>
  <dc:language
rdf:datatype="&xsd:anyURI">http://es.dbpedia.org/resource/Idioma_español</dc:language>
  <dc:identifier
rdf:datatype="&xsd:anyURI">http://localhost:5000/WebServices/desambiguarenlazar/</dc:ident
entifier>
  <dc:type
rdf:datatype="&xsd:anyURI">http://localhost:5000/WebServices/extrarentidades/</dc:type>
  <ssn:hasInput rdf:resource="http://utpl.edu.ec/WebServices#IWSD01"/>
  <ssn:hasOutput rdf:resource="http://utpl.edu.ec/WebServices#OWSD01"/>

```

```

</owl:NamedIndividual>

<!-- http://localhost:5000/WebServices/extrarentidades/ -->
<owl:NamedIndividual rdf:about="http://localhost:5000/WebServices/extrarentidades/">
  <rdf:type rdf:resource="http://utpl.edu.ec/WebServices#WebService"/>
  <dcterms:created rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:created>
  <dcterms:modified rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:modified>
  <swpo:isVersion rdf:datatype="&xsd:string">1.0</swpo:isVersion>
  <dc:creator rdf:datatype="&xsd:string">Jhonny Zaruma</dc:creator>
  <dc:description rdf:datatype="&xsd:string">Servicio web que extrae entidades a partir
de un texto que se le envíe, devuelve un json con una lista de entidades con etiquetas y
contexto.</dc:description>
  <dc:publisher rdf:datatype="&xsd:string">TAW-SBC</dc:publisher>
  <dc:title rdf:datatype="&xsd:string">Web Service Extrae Entidades</dc:title>
  <dc:language
rdf:datatype="&xsd:anyURI">http://dbpedia.org/resource/English_language</dc:language>
  <dc:type
rdf:datatype="&xsd:anyURI">http://dbpedia.org/resource/Web_service</dc:type>
  <dc:language
rdf:datatype="&xsd:anyURI">http://es.dbpedia.org/resource/Idioma_español</dc:language>
  <dc:identifier
rdf:datatype="&xsd:anyURI">http://localhost:5000/WebServices/extrarentidades/</dc:identifie
r>
  <ssn:hasInput rdf:resource="http://utpl.edu.ec/WebServices#IWSE01"/>
  <ssn:hasOutput rdf:resource="http://utpl.edu.ec/WebServices#OWSE01"/>
</owl:NamedIndividual>

```

```

<!-- http://localhost:5000/WebServices/tokenizar/ -->
<owl:NamedIndividual rdf:about="http://localhost:5000/WebServices/tokenizar/">
  <rdf:type rdf:resource="http://utpl.edu.ec/WebServices#WebService"/>
  <dcterms:modified rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:modified>
  <dcterms:created rdf:datatype="&xsd;dateTime">03-05-2014</dcterms:created>
  <swpo:isVersion rdf:datatype="&xsd:string">1.0</swpo:isVersion>
  <dc:creator rdf:datatype="&xsd:string">Jhonny Zaruma</dc:creator>

```

```

    <dc:description rdf:datatype="&xsd:string">Servicio Web que tokeniza y etiqueta el
    texto q se le envia, y de vuelve un json con una lista de los tokens
    tageados</dc:description>
    <dc:publisher rdf:datatype="&xsd:string">TAW-SBC</dc:publisher>
    <dc:title rdf:datatype="&xsd:string">Web Service Tokenizar</dc:title>
    <dc:language
    rdf:datatype="&xsd:anyURI">http://dbpedia.org/resource/English_language</dc:language>
    <dc:type
    rdf:datatype="&xsd:anyURI">http://dbpedia.org/resource/Web_service</dc:type>
    <dc:language
    rdf:datatype="&xsd:anyURI">http://es.dbpedia.org/resource/Idioma_español</dc:language>
    <dc:identifier
    rdf:datatype="&xsd:anyURI">http://localhost:5000/WebServices/tokenizar/</dc:identifier>
    <ssn:hasInput rdf:resource="http://utpl.edu.ec/WebServices#IWST01"/>
    <ssn:hasOutput rdf:resource="http://utpl.edu.ec/WebServices#OWST01"/>
</owl:NamedIndividual>

```

```

<!-- http://utpl.edu.ec/WebServices#IWSD01 -->
<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#IWSD01">
    <rdf:type rdf:resource="&ssn;Input"/>
    <dc:description rdf:datatype="&xsd:string">La entrada puede ser un texto o
    string</dc:description>
    <dc:type rdf:datatype="&xsd:string">String</dc:type>
    <dc:title rdf:datatype="&xsd:string">sentence</dc:title>
</owl:NamedIndividual>

```

```

<!-- http://utpl.edu.ec/WebServices#IWSE01 -->
<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#IWSE01">
    <rdf:type rdf:resource="&ssn;Input"/>
    <dc:description rdf:datatype="&xsd:string">La entrada puede ser un texto o
    string</dc:description>
    <dc:type rdf:datatype="&xsd:string">String</dc:type>

```

```
<dc:title rdf:datatype="&xsd:string">sentence</dc:title>
</owl:NamedIndividual>
```

```
<!-- http://utpl.edu.ec/WebServices#IWS01 -->
<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#IWS01">
  <rdf:type rdf:resource="&ssn;Input"/>
  <dc:description rdf:datatype="&xsd:string">La entrada puede ser un texto o
string</dc:description>
  <dc:type rdf:datatype="&xsd:string">String</dc:type>
  <dc:title rdf:datatype="&xsd:string">sentence</dc:title>
</owl:NamedIndividual>
```

```
<!-- http://utpl.edu.ec/WebServices#OWSD01 -->
<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#OWSD01">
  <rdf:type rdf:resource="&ssn;Output"/>
  <dc:type rdf:datatype="&xsd:string">Json</dc:type>
  <dc:title rdf:datatype="&xsd:string">enlace</dc:title>
</owl:NamedIndividual>
```

```
<!-- http://utpl.edu.ec/WebServices#OWSE01 -->
<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#OWSE01">
  <rdf:type rdf:resource="&ssn;Output"/>
  <dc:type rdf:datatype="&xsd:string">Json</dc:type>
  <dc:description rdf:datatype="&xsd:string">devuelve un json con una lista de entidades
con etiquetas y contexto.
```

Ejemplo:

```
{
  "entidades": [
```

```

[[["&quot;United&quot;,&quot;NNP&quot;],[&quot;States&quot;,&quot;NNPS&quot;]],&quot;One year ago , several hours cities across ...&quot;],
[["&quot;European&quot;,&quot;NNP&quot;],[&quot;Center&quot;,&quot;NNP&quot;]],&quot;One year ago , several
hours cities...&quot;],

[[["&quot;Nuclear&quot;,&quot;NNP&quot;],[&quot;Research&quot;,&quot;NNP&quot;]],&quot;
One year ago , several hours cities ...&quot;]
]
}</dc:description>
<dc:title rdf:datatype="&xsd:string">entidades</dc:title>
</owl:NamedIndividual>

```

```

<!-- http://utpl.edu.ec/WebServices#OWST01 -->

```

```

<owl:NamedIndividual rdf:about="http://utpl.edu.ec/WebServices#OWST01">
  <rdf:type rdf:resource="&ssn;Output"/>
  <dc:type rdf:datatype="&xsd:string">Json</dc:type>
  <dc:description rdf:datatype="&xsd:string">Salida es un json con una lista de los tokens
tagueados

```

Ejemplo:

```

{
  &quot;tokens&quot;: [
    [&quot;cities&quot;,&quot;NNS&quot;],
    [&quot;across&quot;,&quot;IN&quot; ],
    [&quot;the&quot;,&quot;DT&quot;],
    [&quot;United&quot;,&quot;NNP&quot;],
    [&quot;States&quot;,&quot;NNPS&quot;]
  ]
}</dc:description>

```

```

  <dc:description rdf:datatype="&xsd:string">servicio devuelve como resultado un json
con una lista de las entidades con un enlace a la Dbpedia.

```

Ejemplo:

```

{

```

```

    &quot;enlace&quot;; [
      [&quot;United
States&quot;,[1,&quot;http://dbpedia.org/resource/United_States&quot;,&quot;GEO&quot;]],
      [&quot;European Center&quot;,[],
[&quot;CERN&quot;,[1,&quot;http://dbpedia.org/resource/CERN&quot;,&quot;ORG&quot;]]
    ]
  }</dc:description>
    <dc:title rdf:datatype="&xsd:string">tokens</dc:title>
    </owl:NamedIndividual>
</rdf:RDF>

```

```

<!-- Generated by the OWL API (version 3.4.2) http://owlapi.sourceforge.net -->

```

ANEXO 2: ESPECIFICACIONES DE CASOS DE USO

Especificación de Caso de Uso (ECS)

Tokenizar

Versión [1.0.0]

Información del Documento

TÍTULO: Especificación de Caso de Uso (ECS)
SUBTÍTULO: Tokenizar
VERSIÓN: [1.0.0]
ARCHIVO: 01 - Tokenizar
AUTOR: Jhonny Alonso Zaruma
ESTADO: Borrador

Lista de Cambios

VERSIÓN	FECHA	AUTOR	DESCRIPCIÓN
1.0.0	2014-05-13	JAZ	Emisión Inicial

Firmas y Aprobaciones

ELABORADO POR: Jhonny Alonso Zaruma
Analista - Desarrollador

FECHA: 2014-05-13 Firma: _____

REVISADO POR: Nelson Piedra

Especificación de Caso de Uso (ECS)

TOKENIZAR

Especificación de Casos de Uso

Número:	01	
Nombre:	Tokenizar	
Actores:	Usuario, WebService	
Descripción:	Su función es Separar por tokens el texto que se le envíe, y dichos tokens también tagarlos.	
Precondiciones:	El usuario a enviado el texto al WebService	
Pos condiciones:	Recibe una lista con los tokens tageados	
Flujo Normal:	Actor:	Sistema:
	<ol style="list-style-type: none"> 1. Accede al Webservice y envía una entrada(texto) 	<ol style="list-style-type: none"> 2. El sistema recibe el texto 3. Detecta el idioma FA 1 4. Si el idioma del texto esta en español lo traduce, tokeniza y tagea 5. Si el idioma del texto esta ingles lo tokeniza y tagea 6. El Webservice devuelve una lista con los tokens tageados.
Sub Flujo		
Flujo Alternativo:	FA 1 Idioma ingresado no soportado Si el idioma del texto q se envía no es español ni ingles nos devuelve una lista vacía.	
Excepciones:		
Prioridad:	Alta	
Referencias Cruzadas:		
Requerimientos Especiales:		
Asunciones y Dependencias:		
Notas adicionales:		

Glosario

Actor	Usuario
WebService	Servicio web

Especificación de Caso de Uso (ECS)

Extraer Entidades

Versión [1.0.0]

Información del Documento

TÍTULO:	Especificación de Caso de Uso (ECS)
SUBTÍTULO:	Extraer Entidades
VERSIÓN:	[1.0.0]
ARCHIVO:	02 – Extraer Entidades
AUTOR:	Jhonny Alonso Zaruma
ESTADO:	Borrador

Lista de Cambios

VERSIÓN	FECHA	AUTOR	DESCRIPCIÓN
1.0.0	2014-05-13	JAZ	Emisión Inicial

Firmas y Aprobaciones

ELABORADO	Jhonny Alonso Zaruma
POR:	Analista - Desarrollador
FECHA:	2014-05-13

Firma: _____

REVISADO POR: Nelson Piedra

Especificación de Caso de Uso (ECS)

EXTRAER ENTIDADES

Especificación de Casos de Uso

Número:	02	
Nombre:	Extraer Entidades	
Actores:	Usuario, WebService	
Descripción:	Su función es extraer entidades del texto que se le envíe o la lista de tokens tageados que se le envíe.	
Precondiciones:	El usuario a enviado el texto o la lista con los tokens tageados al WebService	
Pos condiciones:	Recibe una lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad	
Flujo Normal:	Actor:	Sistema:
	<ol style="list-style-type: none">1. Accede al Webservice y envía una entrada(texto o lista)	<ol style="list-style-type: none">2. El sistema recibe la entrada3. Detecta el tipo de entrada FA 14. Detecta el idioma FA 25. Si la entrada es texto y el idioma esta en español lo va a tokenizar FA 3. Luego extrae entidades definiendo patrones con el método Chunker de NLTK, y se crea una lista con las entidades extraídas.6. Si la entrada es texto y el idioma esta en ingles se lo envía tokenizar FA 3, y se procede a extraer entidades con nechunk de NLTK7. Si la entrada es una lista(que contiene tokens tageados) se hace la extracción de la entidades con nechunk8. El Webservice devuelve una

		lista con las entidades , los tags y con el contexto de donde se la extrajo.
Sub Flujo		
Flujo Alternativo:	<p>FA 1 Tipo de entrada no soportado Si el tipo de entrada no es un texto ni tampoco una lista con tokens tageados, el Webservice nos devolverá una lista vacía.</p> <p>FA 2 Idioma ingresado no soportado Si el idioma del texto q se envía no es español ni ingles nos devuelve una lista vacía.</p> <p>FA 3 Enviar texto a tokenizar Se envía el texto al webservice Tokenizar y nos devuelve una lista con los tokens tageados.</p>	
Excepciones:		
Prioridad:	Alta	
Referencias Cruzadas:		
Requerimientos Especiales:		
Asunciones y Dependencias:		
Notas adicionales:		

Glosario

Actor	Usuario
WebService	Servicio web
OCW	OpenCourseWare

UTPL-ECC

Especificación de Caso de Uso (ECS) Desambiguar y Enlazar

Versión [1.0.0]

Información del Documento

TÍTULO: Especificación de Caso de Uso (ECS)
SUBTÍTULO: Desambiguar y Enlazar
VERSIÓN: [1.0.0]
ARCHIVO: 01 - Tokenizar
AUTOR: Jhonny Alonso Zaruma
ESTADO: Borrador

Lista de Cambios

VERSIÓN	FECHA	AUTOR	DESCRIPCIÓN
1.0.0	2014-05-13	JAZ	Emisión Inicial

Firmas y Aprobaciones

ELABORADO Jhonny Alonso Zaruma
POR: Analistas - Desarrolladores

FECHA: 2014-05-13 Firma: _____

REVISADO POR: Nelson Piedra

Especificación de Caso de Uso (ECS)

DESAMBIGUAR Y ENLAZAR

Especificación de Casos de Uso

Número:	03	
Nombre:	Desambiguar y Enlazar	
Actores:	Usuario, WebService	
Descripción:	Su función es Desambiguar y Enlazar con recursos de la Dbpedia las entidades recibidas en un tipo de lista o el texto recibido.	
Precondiciones:	El usuario a enviado el texto o la lista de las entidades extraídas con respectivos tags y con su contexto de donde se extrajo la entidad al WebService	
Pos condiciones:	Recibe una lista que contiene las entidades desambiguados con los enlaces a los recursos de la Dbpedia.	
Flujo Normal:	Actor:	Sistema:
	<ol style="list-style-type: none">1. Acede al Webservice y envía una entrada(texto o lista)	<ol style="list-style-type: none">2. El sistema recibe la entrada3. Detecta el tipo de entrada FA 14. Detecta el idioma FA 25. Si la entrada es texto y el idioma esta en español o ingles lo envía a tokenizar FA 3. Y contenemos una lisa con las entidades, tags y contexto .6. Si la entrada es una lista que contiene entidades con tags y contexto, se procede a detectar el idioma FA 2 , luego se descompone la lista y mediante consultas SPARQL a la Dbpedia se desambigua y se enlaza con recursos de la Dbpedia. Para crear una lista con las entidades desambiguadas y enlazadas.7. El Webservice devuelve una lista con las entidades desambiguadas y enlazadas a los recursos de la Dbpedia.

Sub Flujo	
Flujo Alternativo:	<p>FA 1 Tipo de entrada no soportado Si el tipo de entrada no es un texto ni tampoco una lista con las entidades sus tokens y contexto, el Webservice nos devolverá una lista vacía.</p> <p>FA 2 Idioma ingresado no soportado Si el idioma del texto q se envía no es español ni ingles nos devuelve una lista vacía.</p> <p>FA 3 Enviar texto a tokenizar Se envía el texto al webservice Tokenizar y nos devuelve una lista con los tokens tageados.</p>
Excepciones:	
Prioridad:	Alta
Referencias Cruzadas:	
Requerimientos Especiales:	
Asunciones y Dependencias:	
Notas adicionales:	

Glosario

Actor	Usuario
WebService	Servicio web

ANEXO 3: DIAGRAMAS DE SECUENCIA

Secuencia 01 - Tokenizar

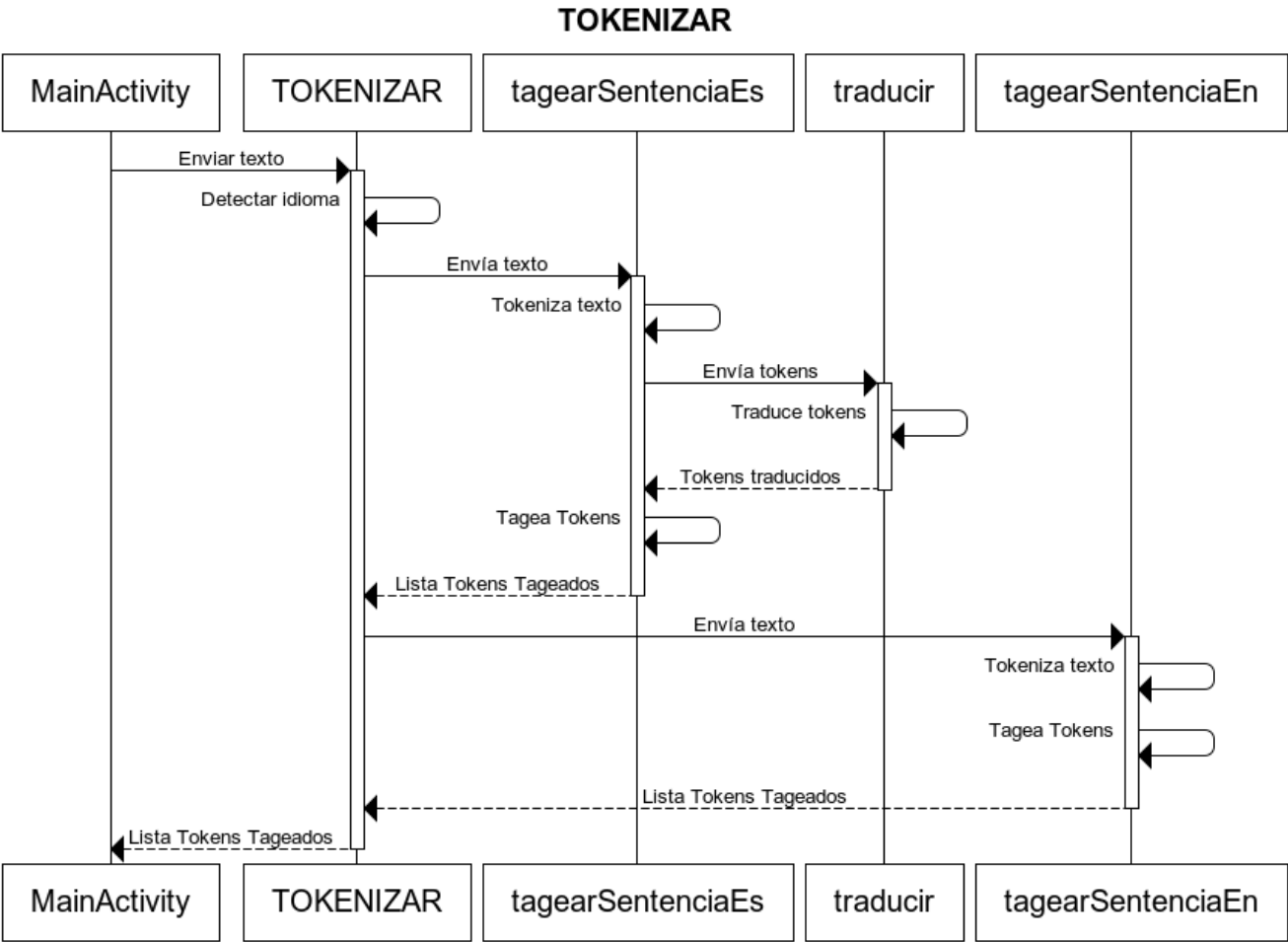


Figura 41 Diagrama de Secuencia Tokenizar

Secuencia 02 – Extraer Entidades

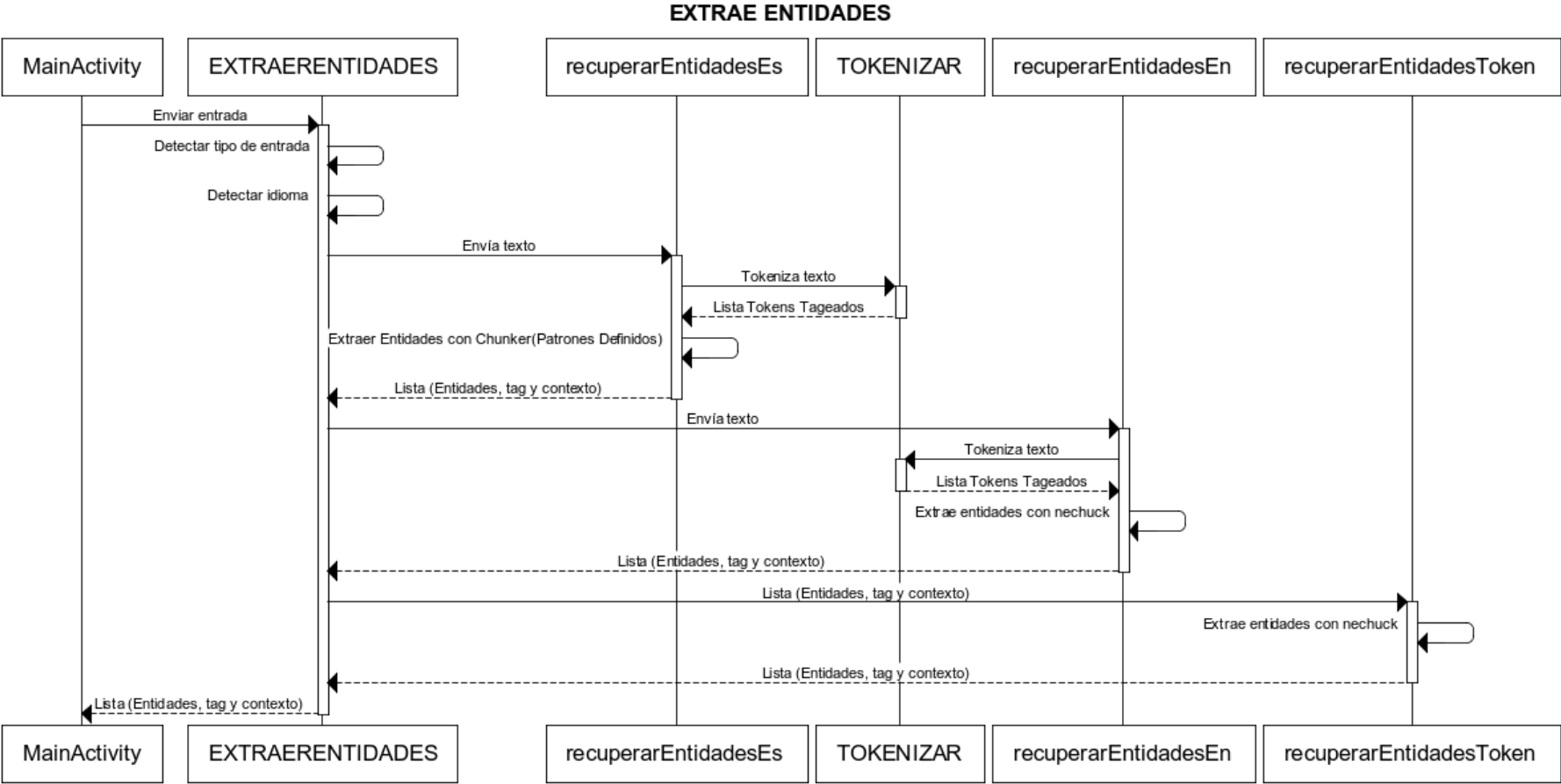


Figura 42 Diagrama de Secuencia Extrae Entidades

Secuencia 03 – Desambiguar y Enlazar

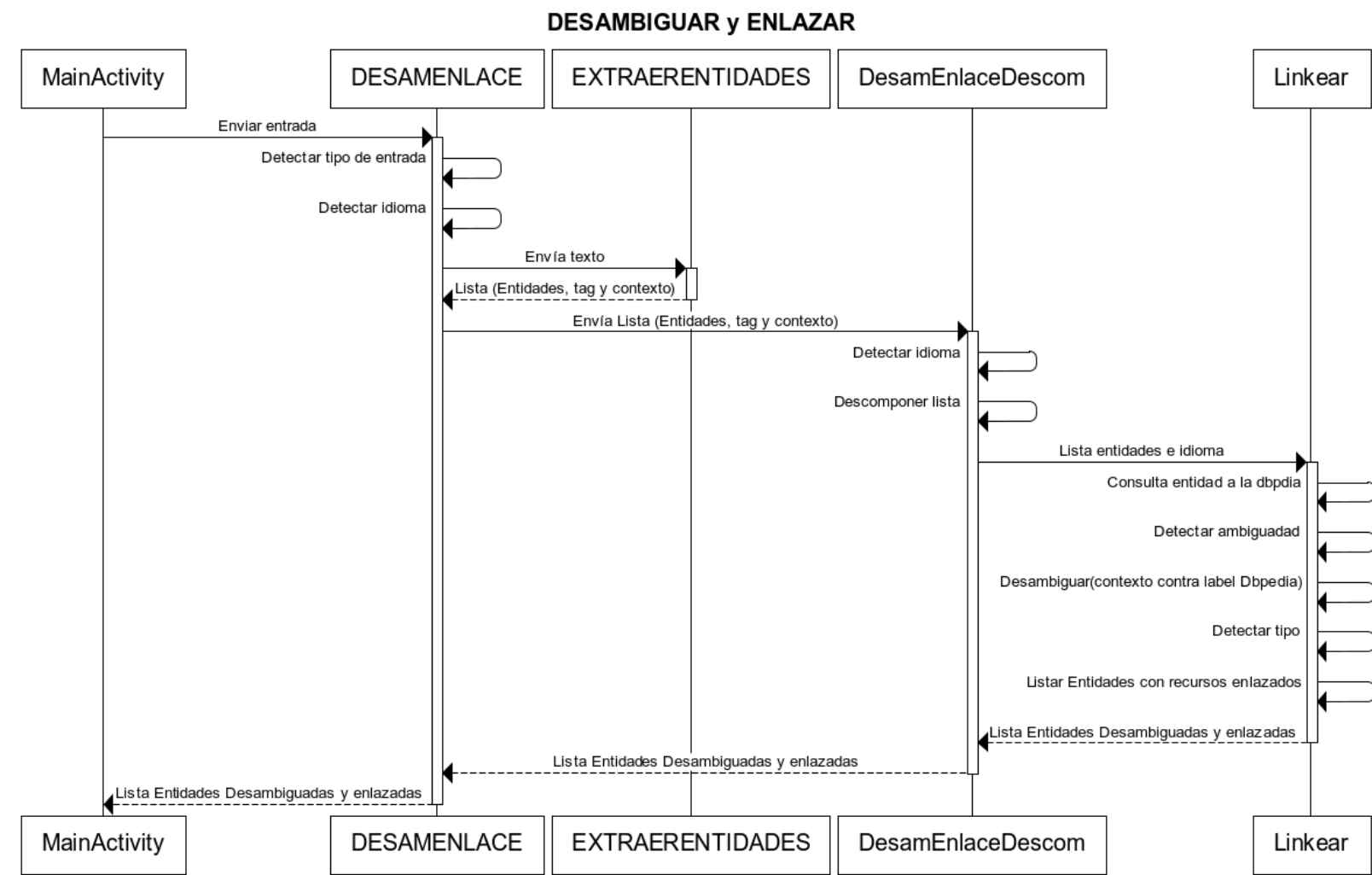


Figura 43 Diagrama de Secuencia Desambiguar y Enlazar