



Titulación de Ingeniería en Sistemas Informáticos y Computación

Desarrollo de Servicios Web para el proceso de Enlace y Enriquecimiento de Datos Enlazados.

(Prototipo en datos universitarios)

Fabrizio Montaña

Ing. Nelson Piedra

05/06/2014

Agenda

- Propósito y Resultados Esperados
- Fases del Proyecto
- Detalle de Fases del Proyecto
- Avance General del Proyecto
- Riesgos del Proyecto
- Proyecto
- Problemas encontrados
- Aprendizajes

Datos del proyecto

Propósito del Proyecto

- Propósito:
 - **Extracción** de entidades.
 - Proceso de **Desambiguación**.
 - **Enlace** de Datos con fuentes externas.
 - Levantar Servicio Web - Rest
- Fecha de inicio del Proyecto:
 - Noviembre 2013
- Fecha de finalización del Proyecto:
 - Agosto 2014 (9 meses)

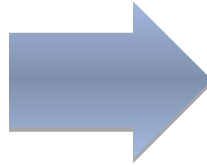
Resultados Esperados

- Servicio Web (REST) – Desambiguación y Enlace
 - Módulo de extracción de entidades
 - Modulo desambiguación
 - Modulo de Enlace

Fases del Proyecto (Componentes)

Desarrollo de la
propuesta formal

- Propuesta
- Objetivos

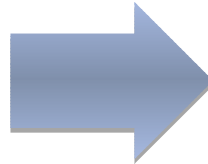


Investigación
Preliminar

- Marco teórico

Implementación
de la solución

- Módulos
- Servicio - Cliente



Análisis de
Resultados,
Conclusiones y
Recomendaciones

- Resultados

Parte II

Estado del Proyecto

Detalle de componentes

Implementación de la solución

- Construcción de los módulos y servicios necesarios para que el sistema cumpla con su objetivo
- **Entregables propuestos**
 - Modulo de etiquetado
 - Module Extracción Entidades y Keywords
 - Modulo de Desambiguación
 - Servicio Web - Rest
- **Actividades realizadas**
- **Actividades pendientes**
 - Corregir errores en resultados de servicio web

Avances del proyecto

% Avance Real: %68


% Avance Esperado: %80

% Avance Desvío: **%12**

Observaciones: retrasos en documentación

Nombre Fase	Fecha fin real	% Avance Estimado	% Avance Real	% Retraso
Desarrollo de la propuesta formal	12/12/2013	100%	100%	0,0%
Investigación Preliminar	30/01/2014	90%	90%	10%
Implementación de la solución	---	95%	85%	10%
Análisis de Resultados, Conclusiones y Recomendaciones	---	40%	40%	10%

Riesgos del Proyecto

Resp.	Riesgo	Fecha Creación	Fecha Cierre	Mitigación	Estado/ Impac.
Tesista	Falta de conociendo den conceptos, herramientas y tecnologías	03/12/20 13	---	-Dedicación tiempo extra investigación	Abierta 

Marco Teórico

- Datos enlazados
 - Principios de Datos Enlazados
 - Tecnologías
 - URI, HTTP, RDF & SPARQL
- Procesamiento de Lenguaje Natural (PLN)
 - Part of Speech Tagger
 - Chunking
 - Desambiguación (WSD)
- RESTful Web Service

Marco Teórico (Linkend Data)

- W3C: “Linked Data se refiere a la utilización de las mejores prácticas para publicación, estructuración de los datos en la web, de tal forma que puedan ser enlazados entre sí, utilizando tecnología propias de web semántica como RDF, OCW, SPARQL, etc
- Principios :
 - Usar URIs como nombre de las cosas
 - Usar URIs HTTP para que esas cosas puedan ser referenciadas
 - Representar los datos en RDF y SPARQL como lenguaje de consulta
 - Incluir enlaces hacia otras cosas, para descubrir más cosas

Marco Teórico (Lingüística Computacional o PLN)

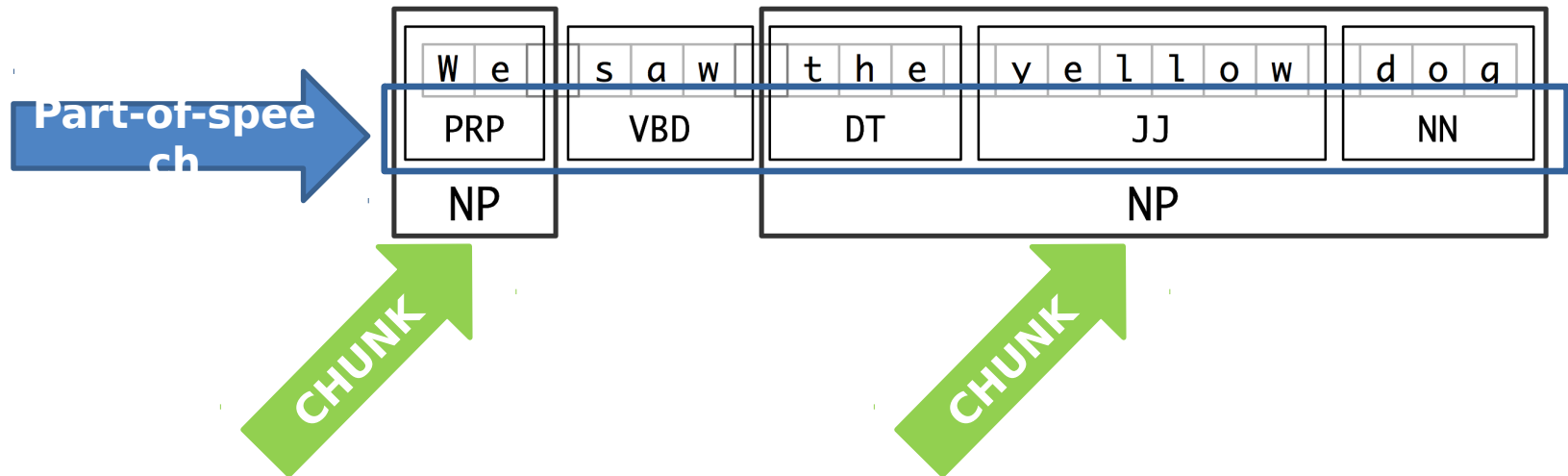
- Entender el lenguaje humano, una tarea que para las personas e inclusive animales es tan natural y cotidiana se vuelve un reto al tratar de interpretarlo mediante procesos computacionales a fin de comprenderlo y poder replicarlo.

PLN – Part of Speech Tagging

➤ Penn Treebank (Penn Treebank - Universidad de Pennsylvania)

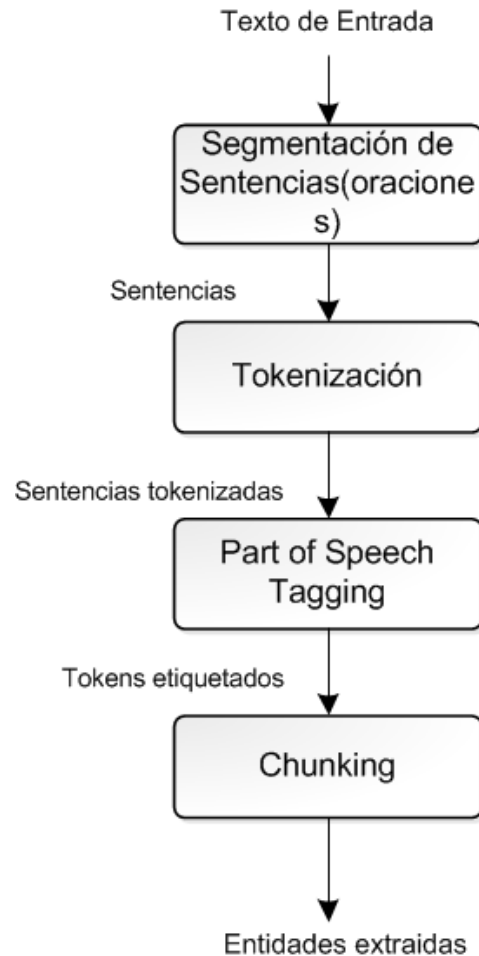
Tag	Meaning	Examples
ADJ	adjective	<i>new, good, high, special, big, local</i>
ADV	adverb	<i>really, already, still, early, now</i>
CNJ	conjunction	<i>and, or, but, if, while, although</i>
DET	determiner	<i>the, a, some, most, every, no</i>
EX	existential	<i>there, there's</i>
FW	foreign word	<i>dolce, 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>

PLN - Chunking



- **Entidades:**
 - We
 - The yellow dog

PLN - Proceso de extracción de entidades



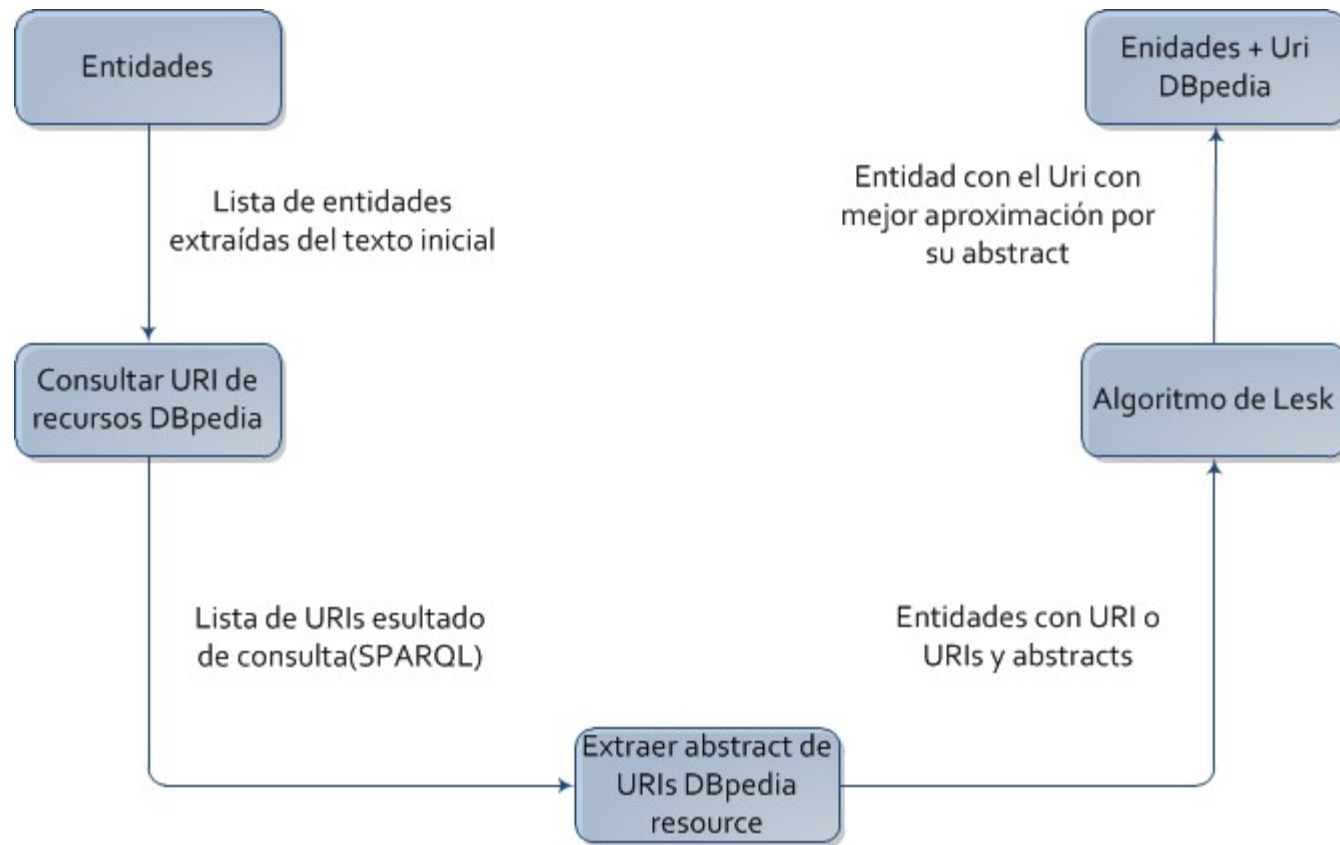
PLN – Desambiguación WSD

- Métodos basados en conocimiento
 - Algoritmo de Lesk 1986
 - En base a los sentidos de las palabras en la sentencias
- Métodos Supervisado
 - Datos enteramiento etiquetados manualmente
- Métodos no supervisados
 - Datos enteramiento sin etiquetar (clusters, textos paralelos)

Marco teórico - REST

- REST (Representational State Transfer) no es un protocolo, un formato de archivo, o un marco de desarrollo. Es un conjunto de restricciones de diseño, la hipermedia como el motor de estado de la aplicación.
- Utilizar los métodos del protocolo HTTP como son PUT, GET, POST y DELETE

Propuesta



Propuesta - Algoritmo de Lesk

```
for every word w[i] in the phrase
  let BEST_SCORE = 0
  let BEST_SENSE = null
  for every sense sense[j] of w[i]
    let SCORE = 0
    for every other word w[k] in the phrase, k != i
      for every sense sense[l] of w[k]
        SCORE = SCORE + number of words that occur in the gloss of
                           both sense[j] and sense[l]
      end for
    end for
    if SCORE > BEST_SCORE
      BEST_SCORE = SCORE
      BEST_SENSE = w[i]
    end if
  end for
  if BEST_SCORE > 0
    output BEST_SENSE
  else
    output "Could not disambiguate w[i]"
  end if
end for
```

Arquitectura

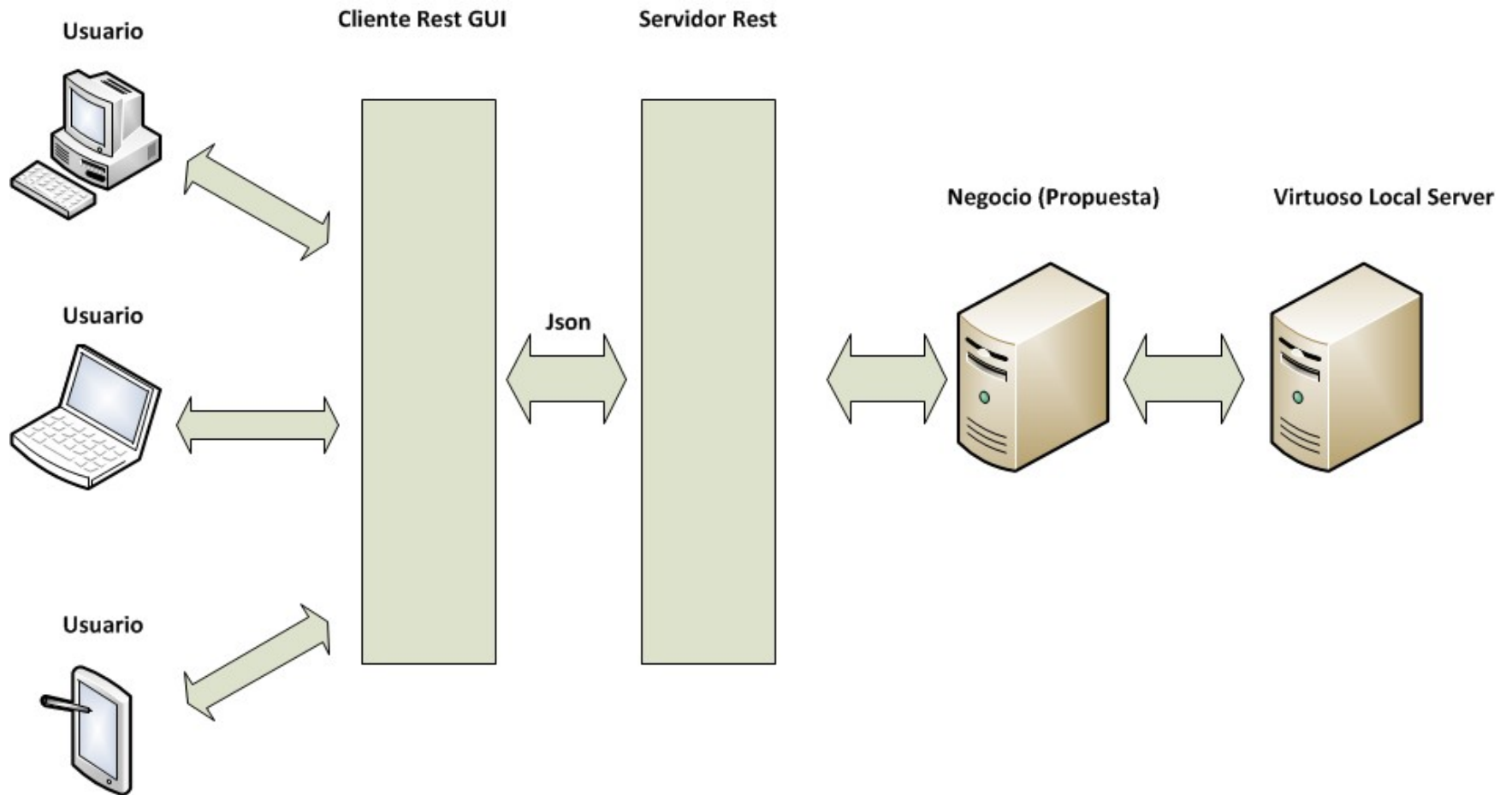


Diagrama de Secuencia - Etiquetado

ETIQUETAR

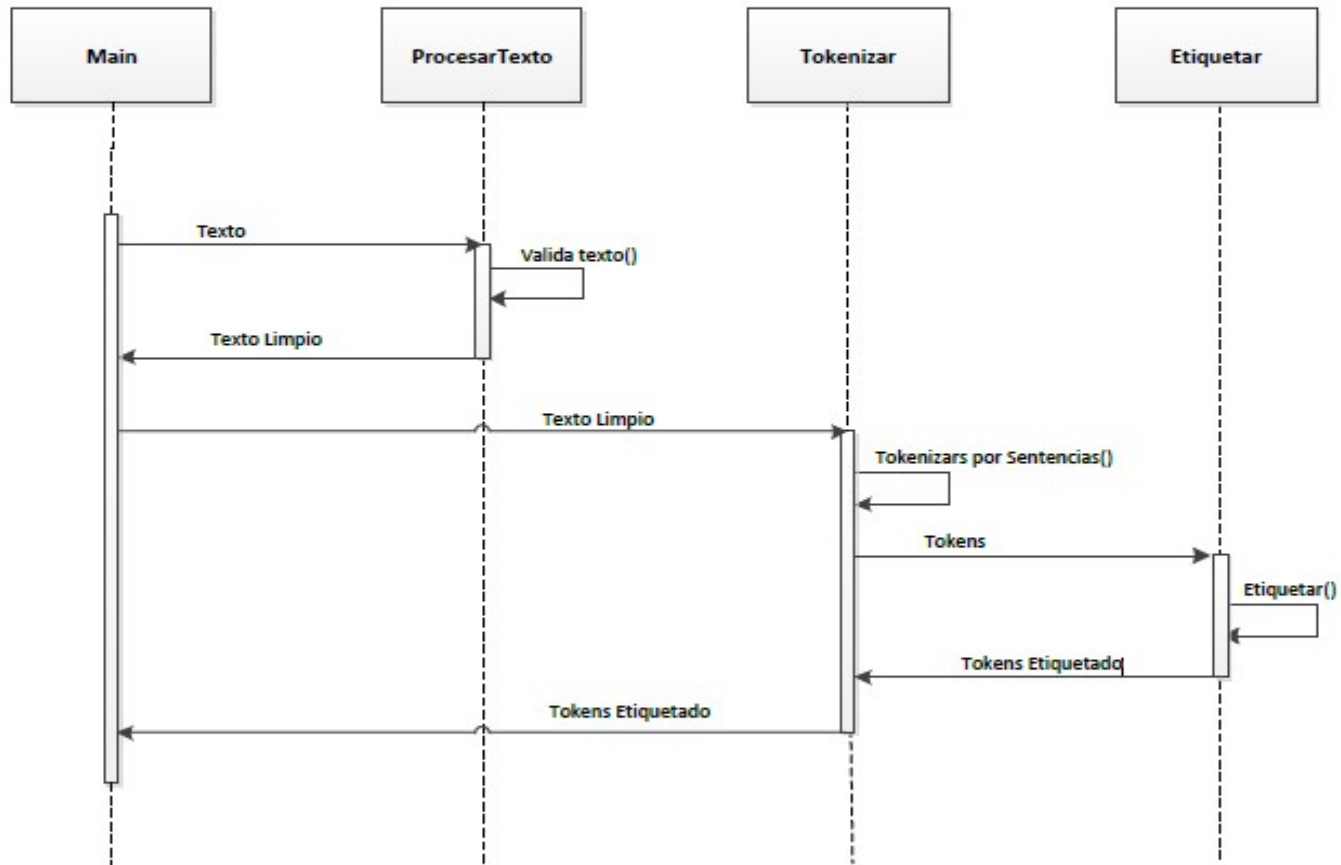


Diagrama de Secuencia - Etiquetado

EXTRACCIÓN ENTIDADES Y KEYWORDS

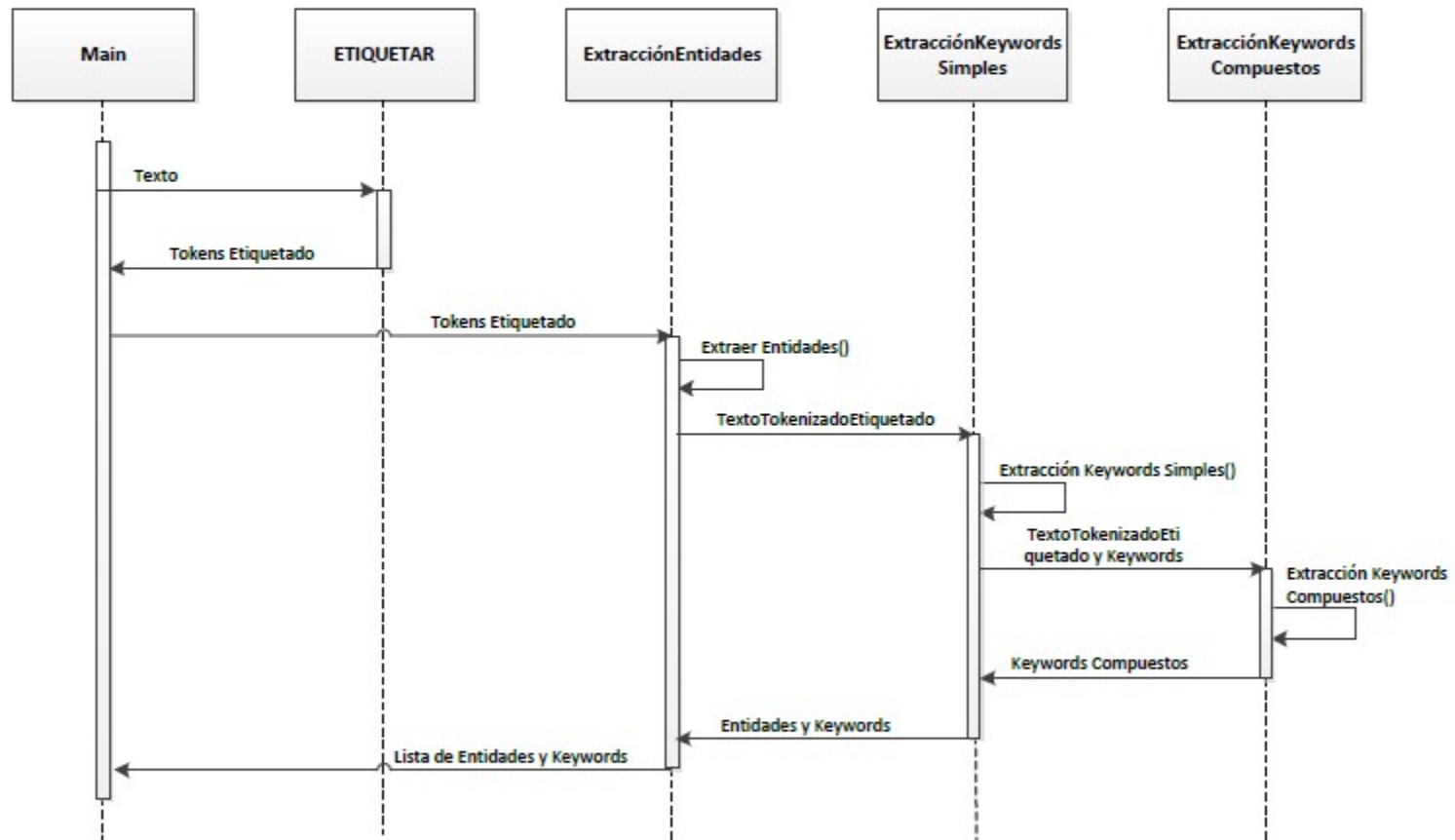
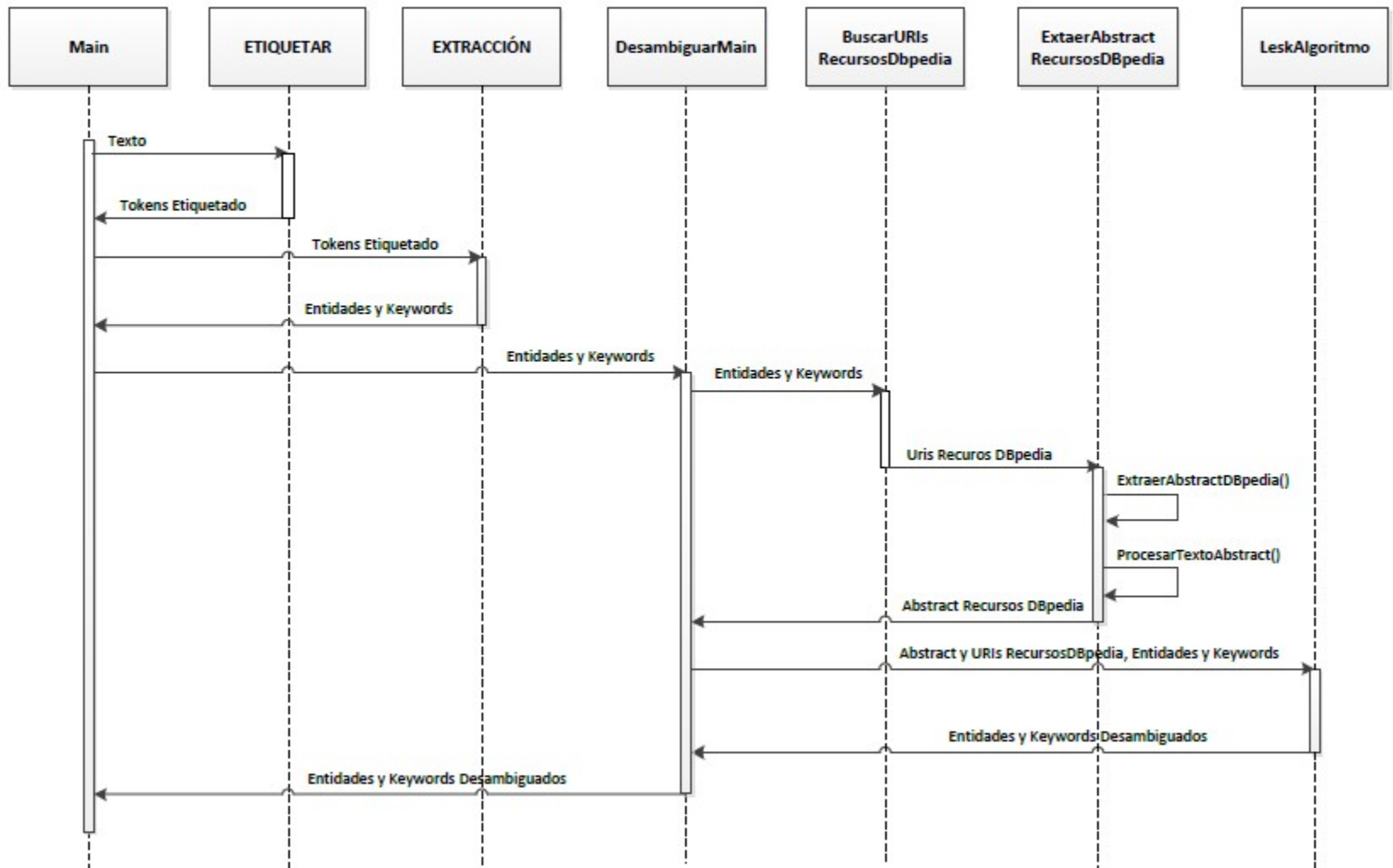


Diagrama de Secuencia - Etiquetado

DESAMBIGUACIÓN



Principales problemas encontrados

- Desconocimiento de las tecnologías necesarias para el desarrollo del sistema (Tecnologías)

Principales aprendizajes

- Adquisición de nuevos conocimientos
Linkend Data
- Procesamiento de lenguaje Natural
- Servicio Web – Rest

Preguntas ??

Gracias