



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

09/07/2014

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)

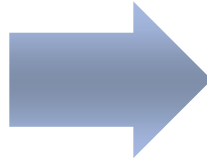
Resultados Esperados

- 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

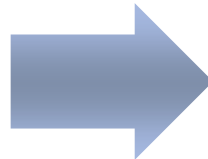


Investigación
Preliminar

- Marco teórico

Diseño

- Módulos
- Servicio - Cliente



Implementación y
Pruebas

- Código (SW)
- Doc Pruebas

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: %85
 % Avance Esperado: %90
 % Avance Desvío: **%15**

Observaciones:
 Retrasos en documentación
 Retrasos en implantación y pruebas

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	99%	95%	10%
Implementación de la solución	---	100%	90%	10%
Análisis de Resultados, Conclusiones y Recomendaciones	---	50%	20%	10%

Riesgos del Proyecto

Resp.	Riesgo	Fecha Creación	Fecha Cierre	Mitigación	Estado/ Impac.
Tesista	Falta de conociendo: conceptos, herramientas y tecnologías	03/12/2013	---	-Dedicación tiempo extra a la investigación durante la jornada	Abierto 
Tesista	No cumplir con el Cronograma establecido	03/12/2013		-Alargar tiempo de proyecto	Abierto 
Tesista	Tiempo	03/12/2013		-Alargar tiempo de proyecto	Abierto 

Marco Teórico

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

Marco Teórico (Linked 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, OWL, 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)

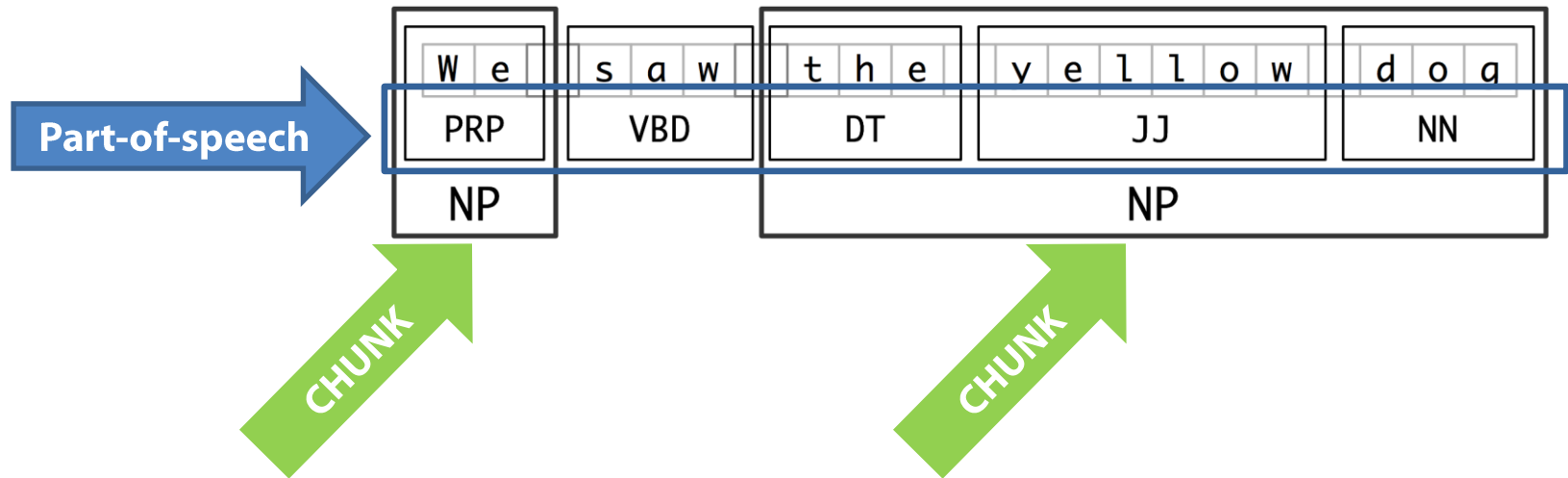
➤ Entender el lenguaje humano, una tarea que para las personas e inclusive animales es tan natural y cotidiana se vuelve un reto al tratar de interpretarlo mediante procesos computacionales a fin de comprenderlo y poder replicarlo.

PLN – Part of Speech Tagging

➤ Penn Treebank (Penn Treebank - Universidad de Pennsylvania)

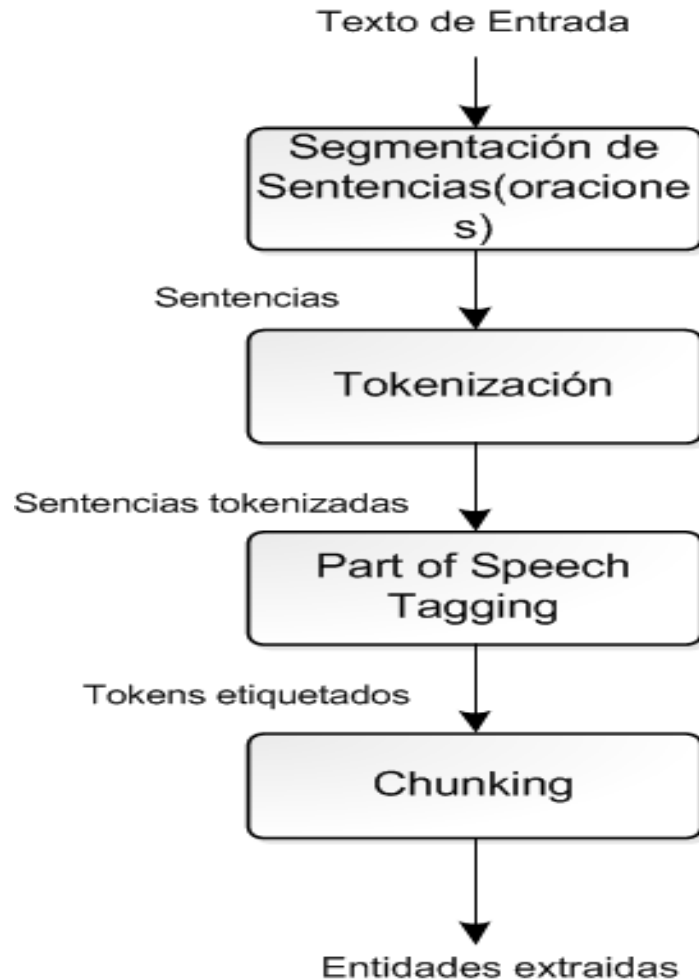
Tag	Meaning	Examples
ADJ	adjective	<i>new, good, high, special, big, local</i>
ADV	adverb	<i>really, already, still, early, now</i>
CNJ	conjunction	<i>and, or, but, if, while, although</i>
DET	determiner	<i>the, a, some, most, every, no</i>
EX	existential	<i>there, there's</i>
FW	foreign word	<i>dolce,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



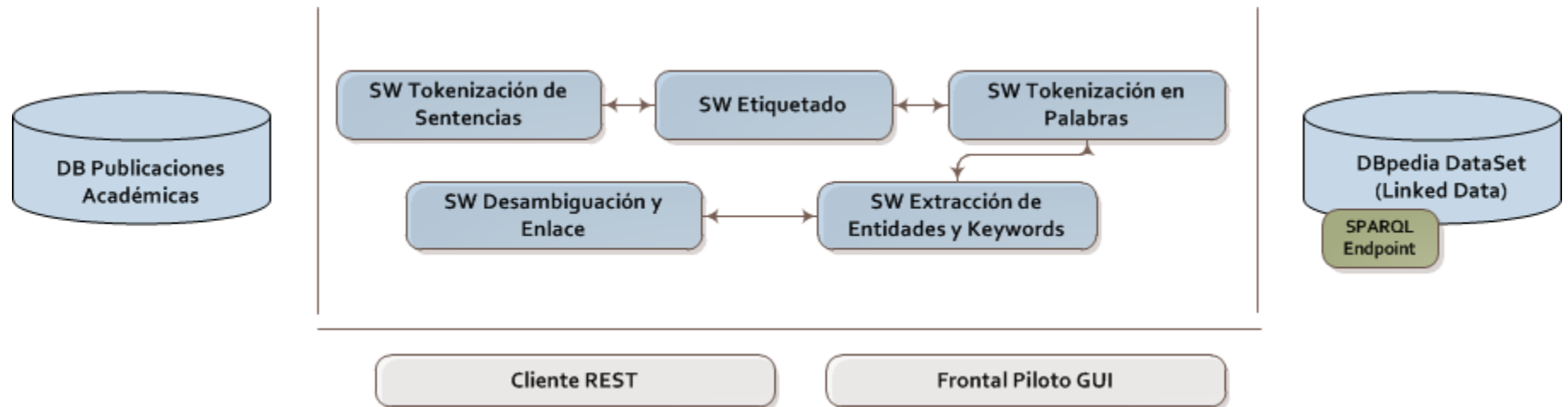
PLN – Desambiguación WSD

- 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

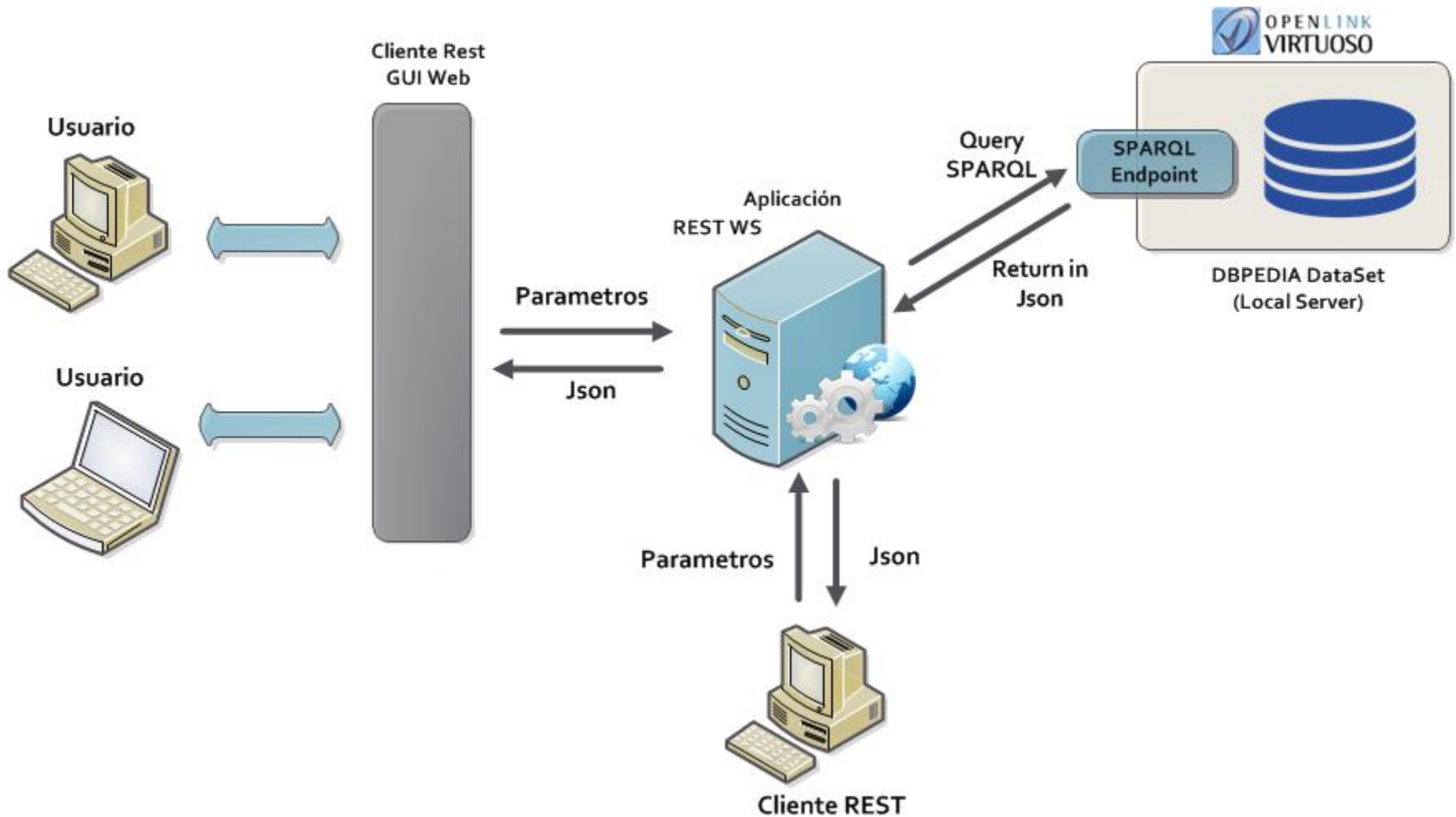
Propuesta



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



Arquitectura – Logica del negocio

- Flask (WS)
- SPARQLWrapper (Consulta DBpedia)
- NLTK (PLN)
- TreeTagger (PLN)

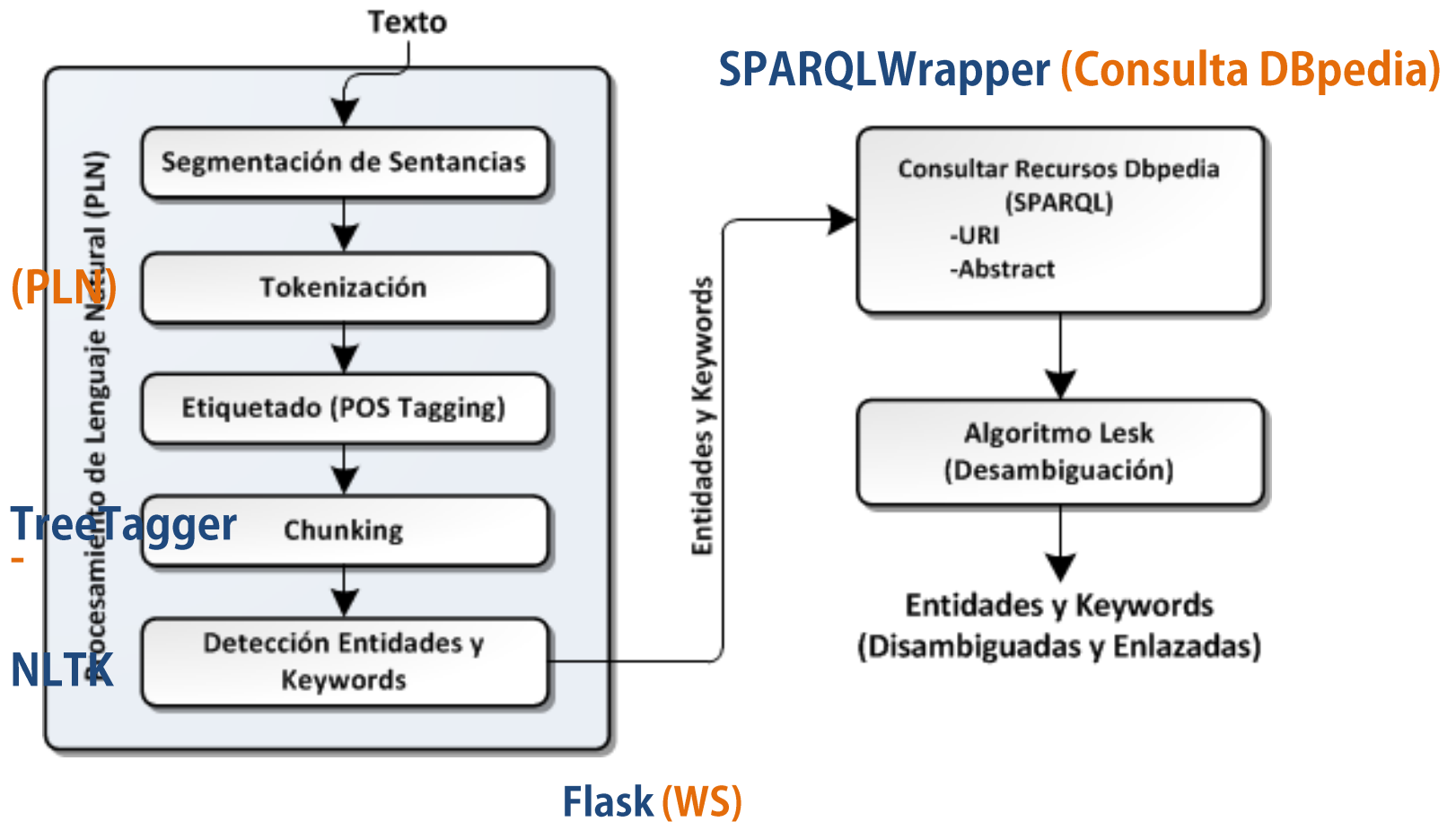


```

1 # -*- coding: utf-8 -*-
2 import urllib2
3 from SPARQLWrapper import SPARQLWrapper, JSON
4 import sys
5 import nltk
6 from nltk.corpus import stopwords
7 import os
8 import treeTaggerWrapper
9 import MySQLdb
10 import unicodedata
11 from urllib import quote_plus
12
13 def PrepararFiltroTexto(recursivo):
14     titlletyope = ""
15     for s in recursivo.split():
16         if s.lower() in stopwords.words('english') or s.lower() in stopwords.words('spanish'):
17             titlletyope = titlletyope + s.lower() + " "
18         else:
19             titlletyope = titlletyope + s.title() + " "
20     titlletyope = titlletyope[:-1]
21     return titlletyope
22
23 def ejecutarQuery(SparqlQuery, servidor):
24     sparql = SPARQLWrapper(servidor)
25     sparql.setQuery(SparqlQuery)
26     sparql.setReturnFormat(JSON)
27     results = sparql.query().convert()
28     return results
29
30 def extraerListaDesdeJson(results):
31     uris = []
32     for result in results["results"]["bindings"]:
33         if "ambi" in result.keys():
34             if result["ambi"]["value"] not in uris and "http://dbpedia.org/resource/" in result["ambi"]["value"]: uris.append(result["ambi"]["value"])
35         elif "redir" in result.keys():
36             if result["redir"]["value"] not in uris and "http://dbpedia.org/resource/" in result["redir"]["value"]: uris.append(result["redir"]["value"])
37         elif "amb" in result.keys():
38             if result["amb"]["value"] not in uris and "http://dbpedia.org/resource/" in result["amb"]["value"]: uris.append(result["amb"]["value"])
39         else:
40             if result["x"]["value"] not in uris and "http://dbpedia.org/resource/" in result["x"]["value"]: uris.append(result["x"]["value"])
41     return uris
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

```

Arquitectura – Logica del negocio



Arquitectura – Dataset Local DBpedia

Archivos importados de Dbpedia

<http://data.dws.informatik.uni-mannheim.de/dbpedia/3.9/>

- Labels de recursos:
labels_en.nt.bz2
- Datos personal de los recurso tipos Persona:
persondata_en.nt.bz2
- Resúmenes Corto de los recursos
short_abstracts_en.nt.bz2
- Links de Desambiguación de Wikipedia
disambiguations_en.nt.bz2
- Redirecciones entre Recursos
redirects_en.nt.bz2

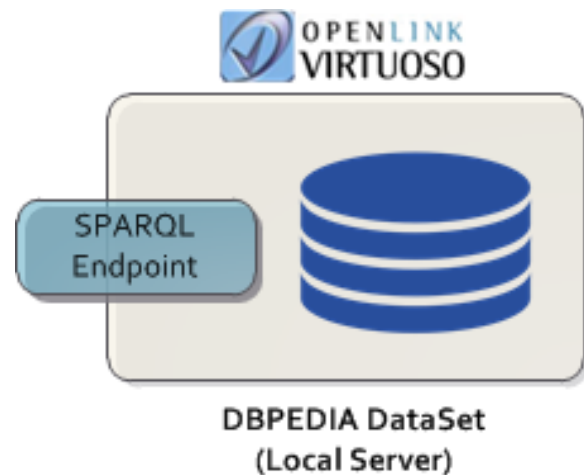


Diagrama de Secuencia – Etiquetado/Tokenización

ETIQUETAR

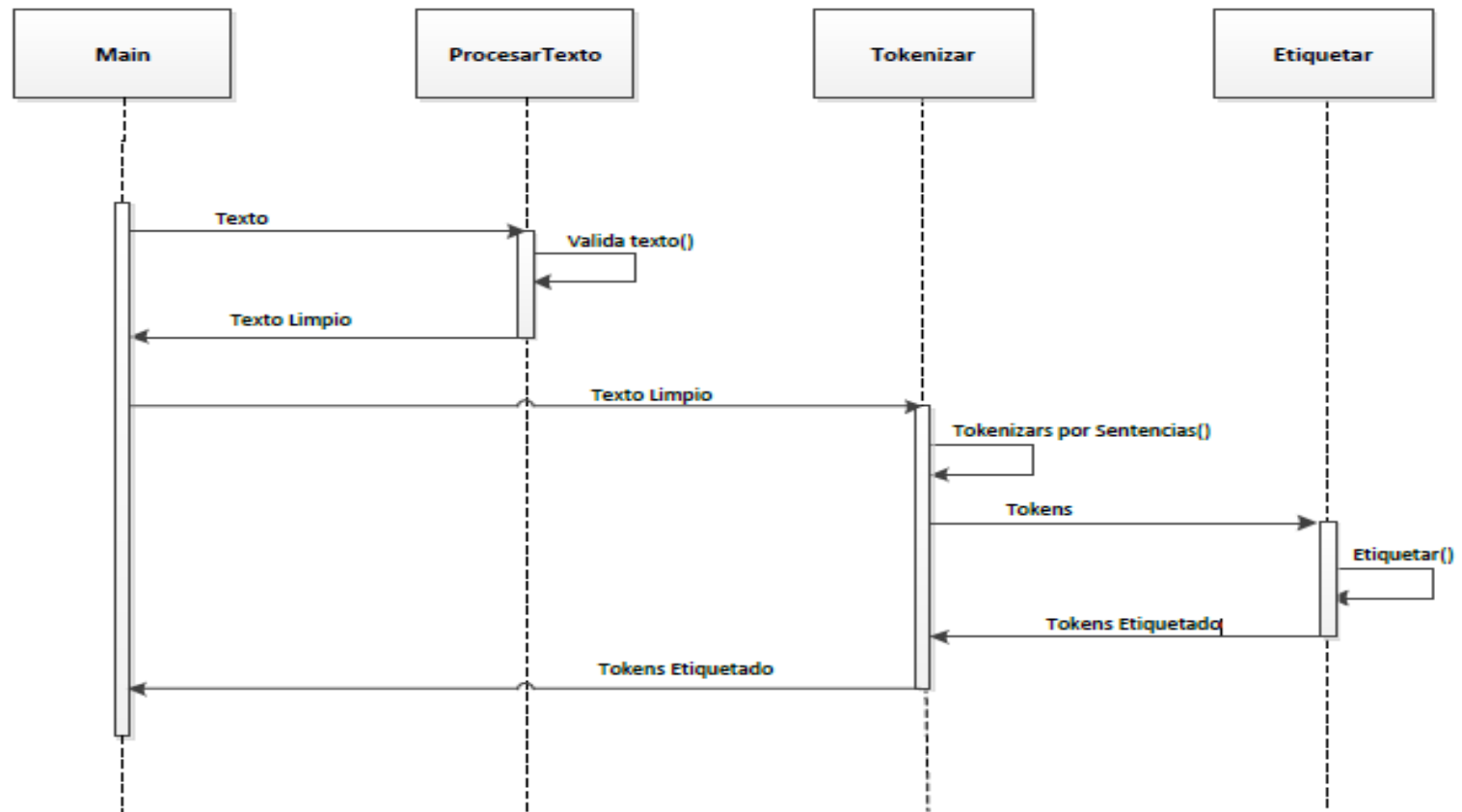


Diagrama de Secuencia – Extracción de Entidades y Keywords

EXTRACCIÓN ENTIDADES Y KEYWORDS

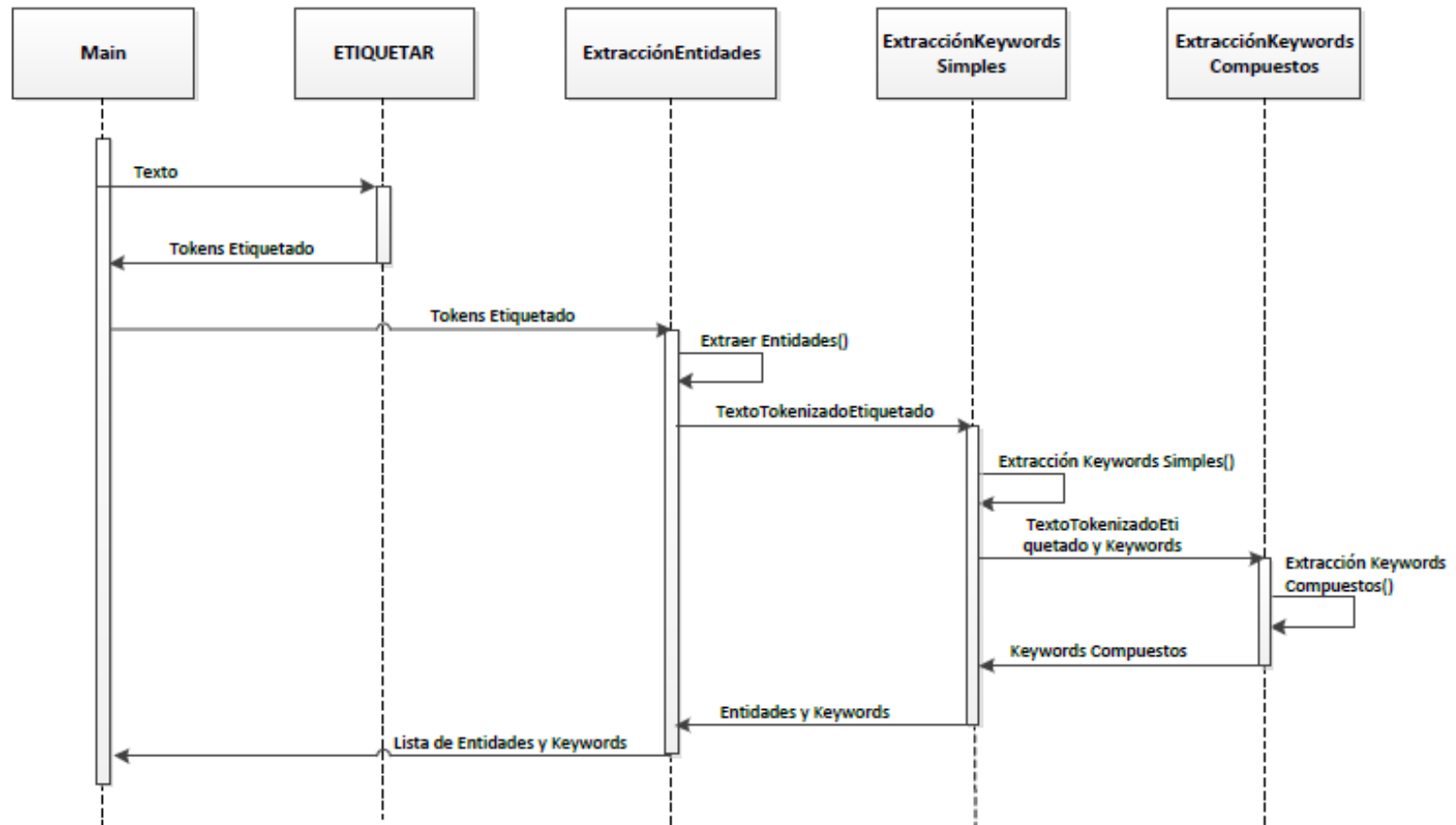
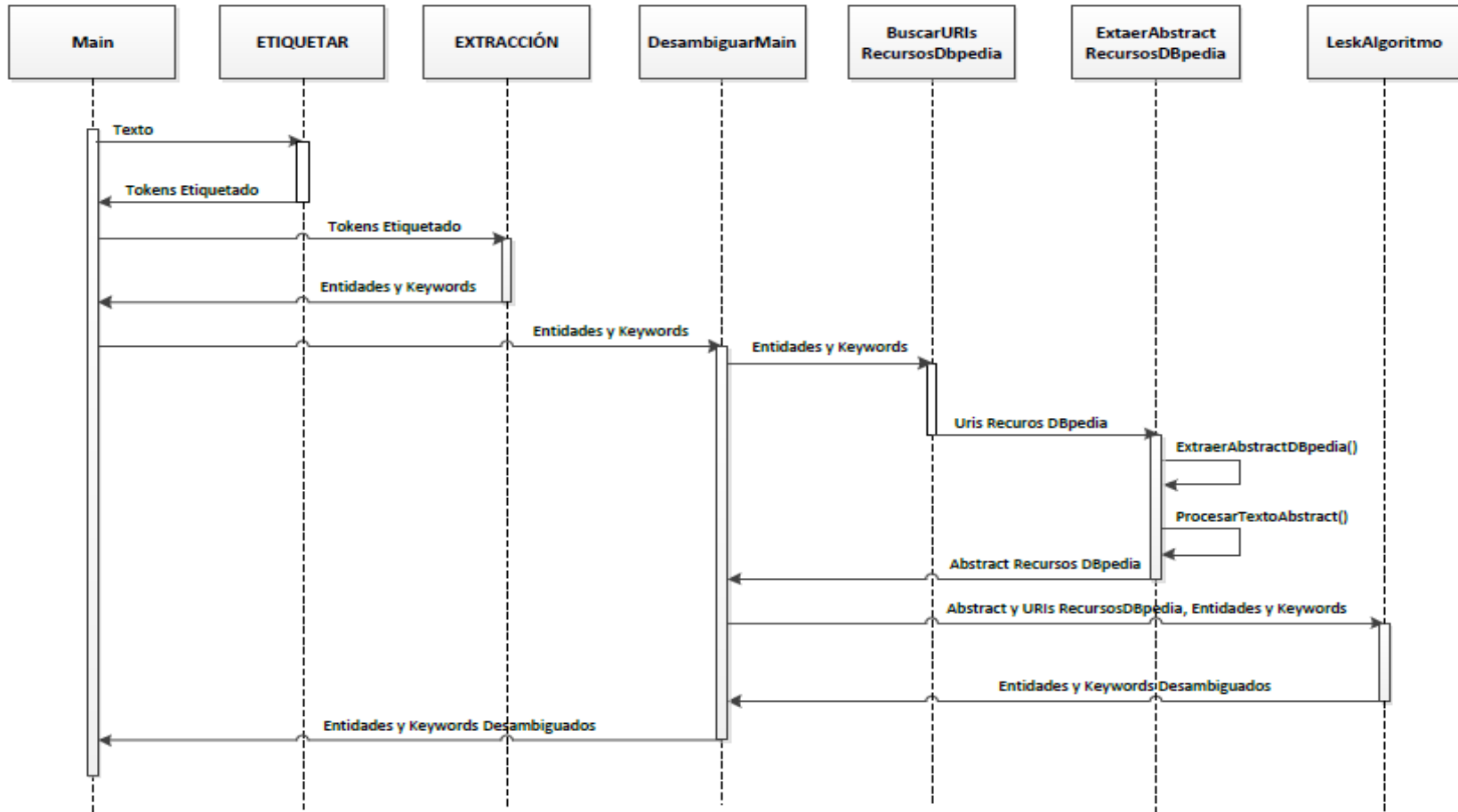


Diagrama de Secuencia - Desambiguación

DESAMBIGUACIÓN



Demostración de Servicios Web

- **WS - Tokenización de Sentencias**
- **WS - Tokenización de Palabras**
- **WS - Etiquetado**
- **WS - Extracción de Entidades y Keywords**
- **WS - Desambiguación**

Principales problemas encontrados

- Tiempo
- No cumplir con el Cronograma establecido
- Falta de herramientas o 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
- Métodos de Desambiguación en base a conocimiento

Preguntas ??

Gracias