UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA
La Universidad Católica de Loja

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)

Fabricio Montaño

Ing. Nelson Piedra

05/06/2014

DATOS DEL PROYECTO

Propósito del Proyecto

- Propósito:
  - Descubrimiento de entidades y conceptos.
    - Proceso de **Desambiguación**.
    - Enlace de Datos con fuentes externas.
    - Levantar Servicio Web - Rest

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 delas 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 par que esas cosas puedan ser referenciadas
  - Representar los datos en RDF y SPARQL como lenguaje de consulta
  - Incluir enlaces hacia otra cosas, para descubrir más cosas

Marco Teórico (Lingüística Computacional o PLN)

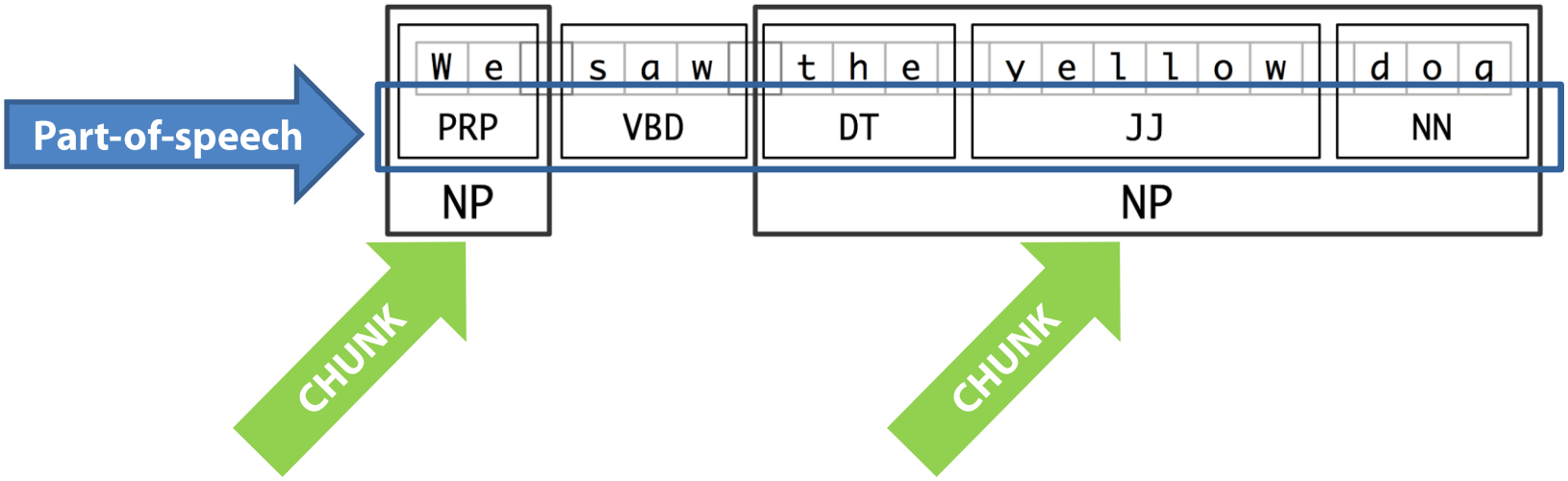
- Entender el lenguaje humano, una tarea que para las personas e inclusive animales es tan natural y cotidiana se vuelve un reto al tratar de interpretarlo mediante procesos computacionales a fin de comprenderlo y poder replicarlo.

PLN – Part of Speech Tagging

- **Penn Treebank (Penn Treebank - Universidad de Pennsilvania)**

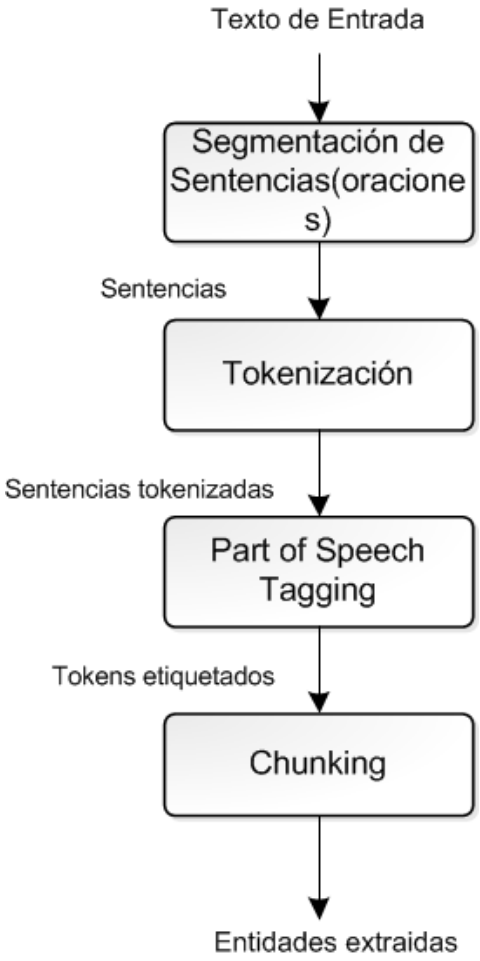
Tag	Meaning	Examples
ADJ	adjective	*new, good, high, special, big, local*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
EX	existential	*there, there's*
FW	foreign word	*dolce, ersatz, esprit, quo, maitre*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

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 sentecias
- 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

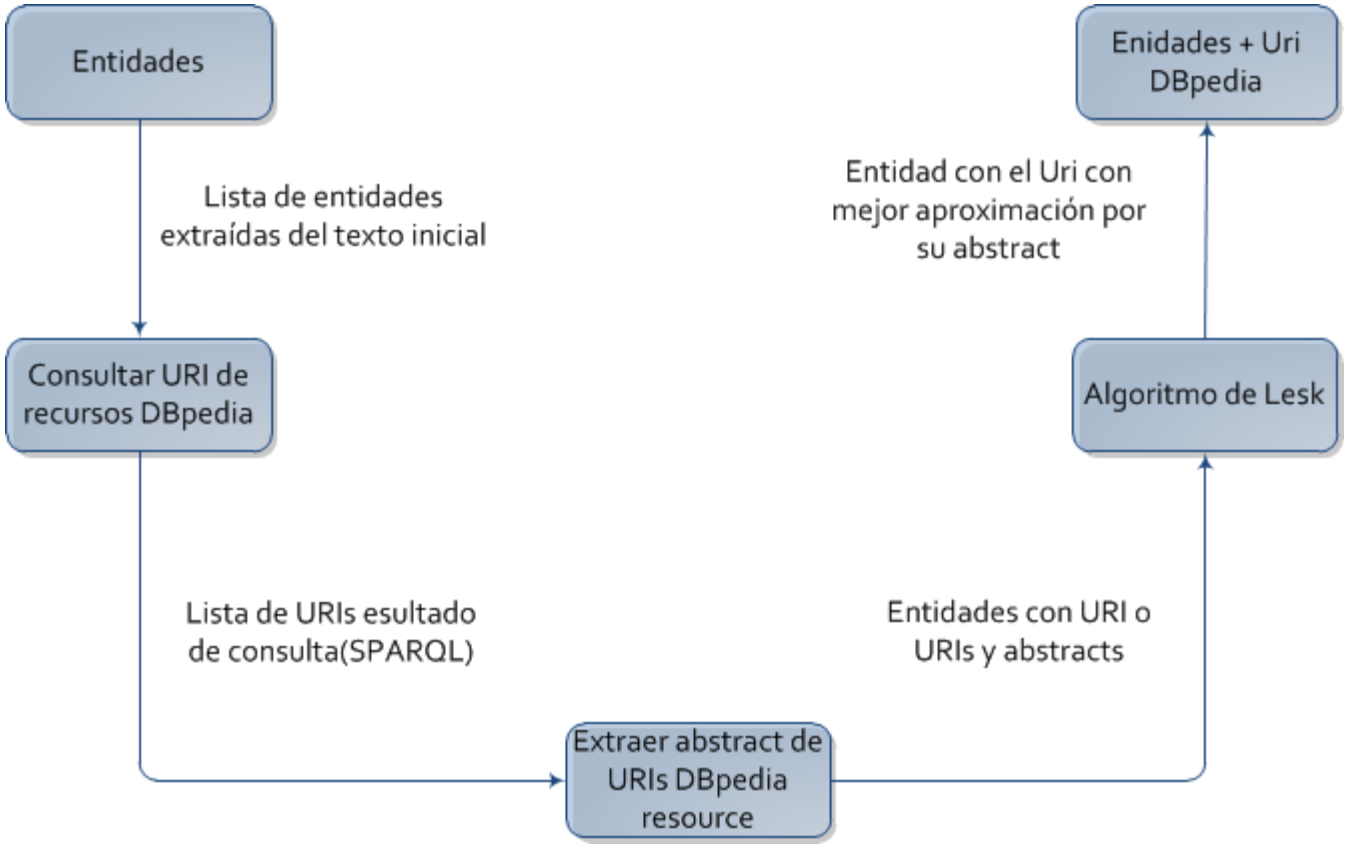
Propuesta:

- Desambiguación WSD y Enlce con Recursos de BDpedia mediante una adaptación al Algoritmo de Lesk

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
```

Propuesta


Entidades
Lista de entidades extraídas del texto inicial
Consultar URI de recursos DBpedia
Lista de URIs esultado de consulta(SPARQL)
Extraer abstract de URIs DBpedia resource
Entidades con URI o URIs y abstracts
Algoritmo de Lesk
Entidad con el Uri con mejor aproximación por su abstract
Enidades + Uri DBpedia

Arquitectura

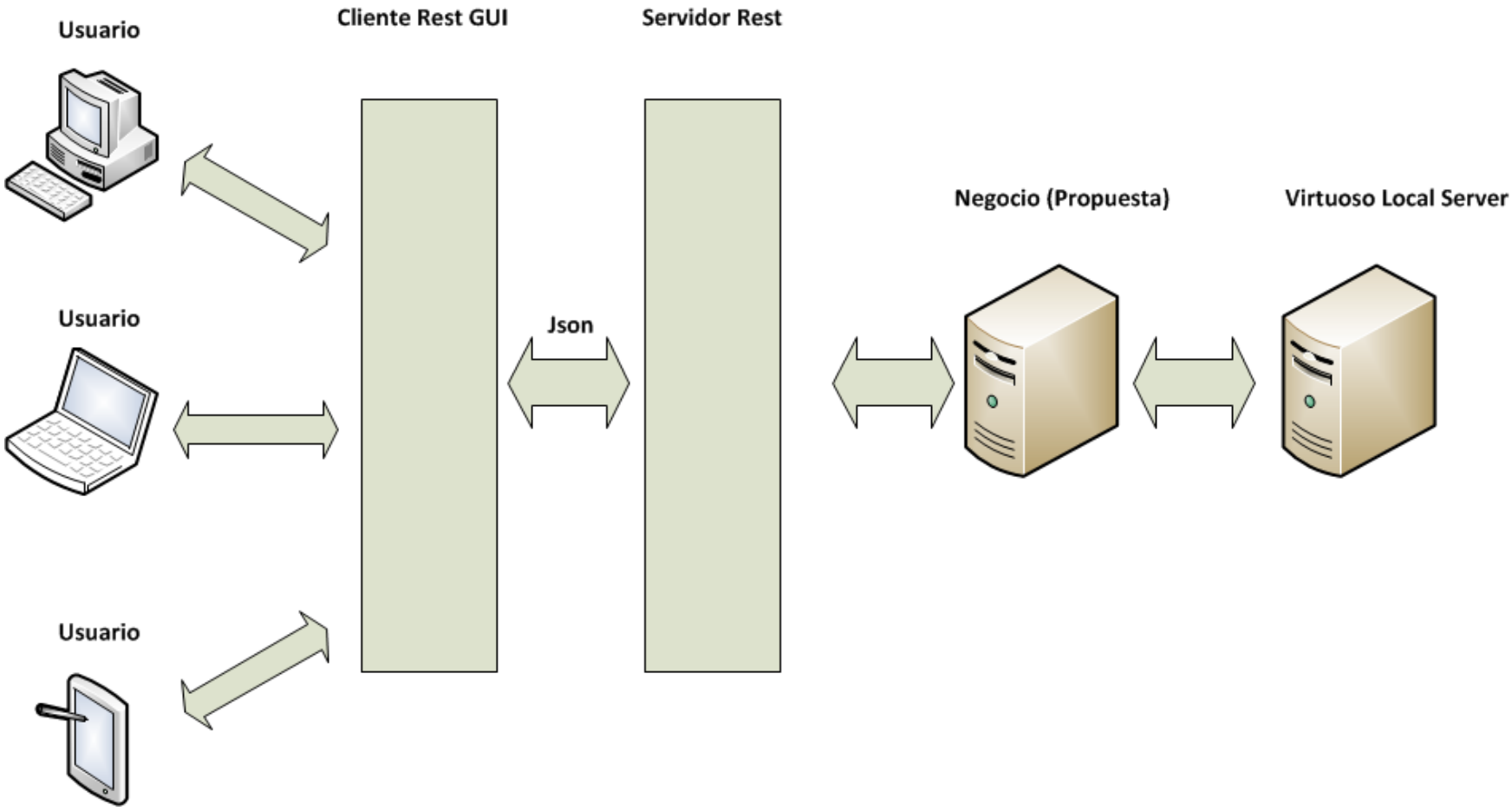


Diagrama de Secuencia - Etiquetado

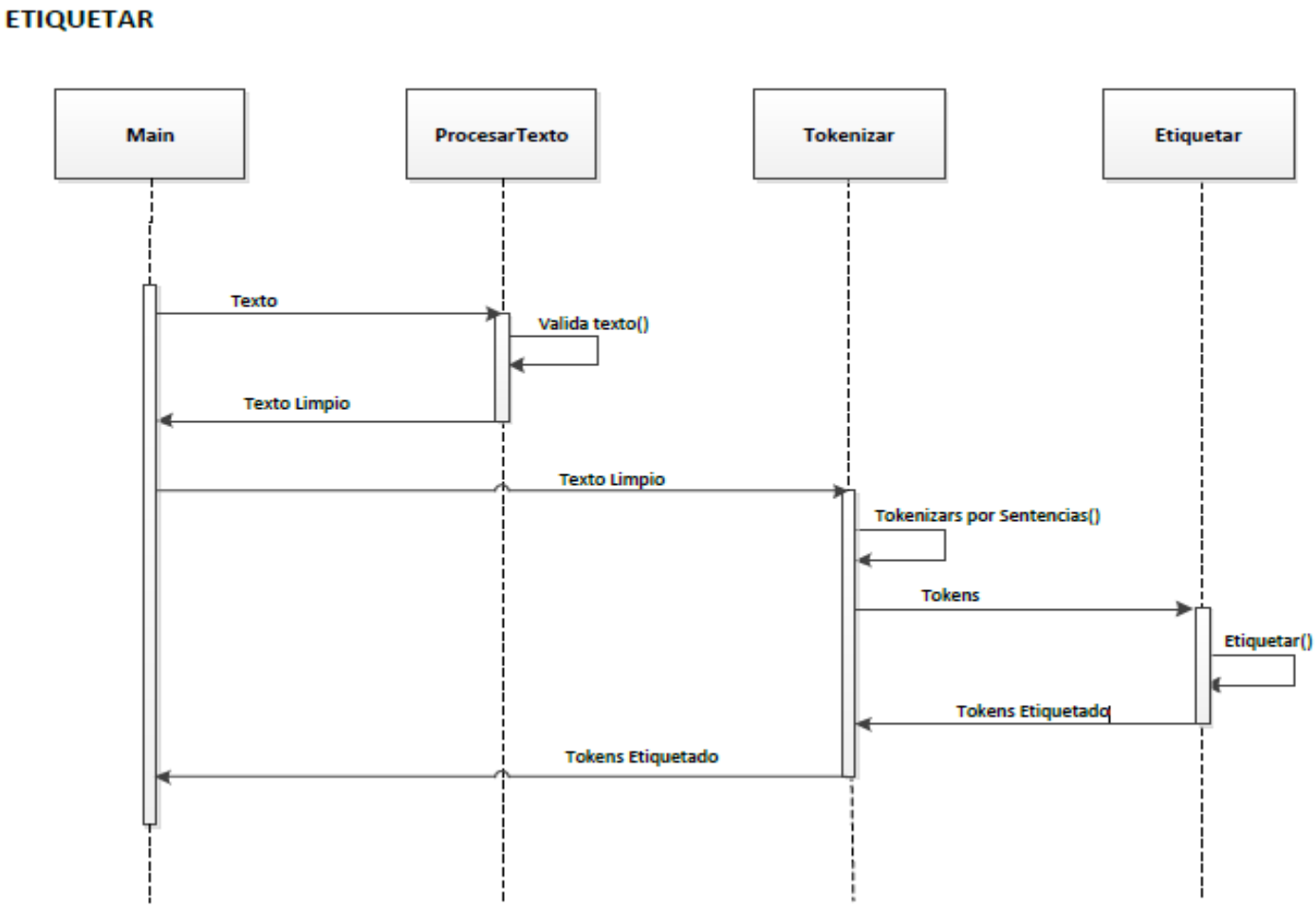


Diagrama de Secuencia - Etiquetado

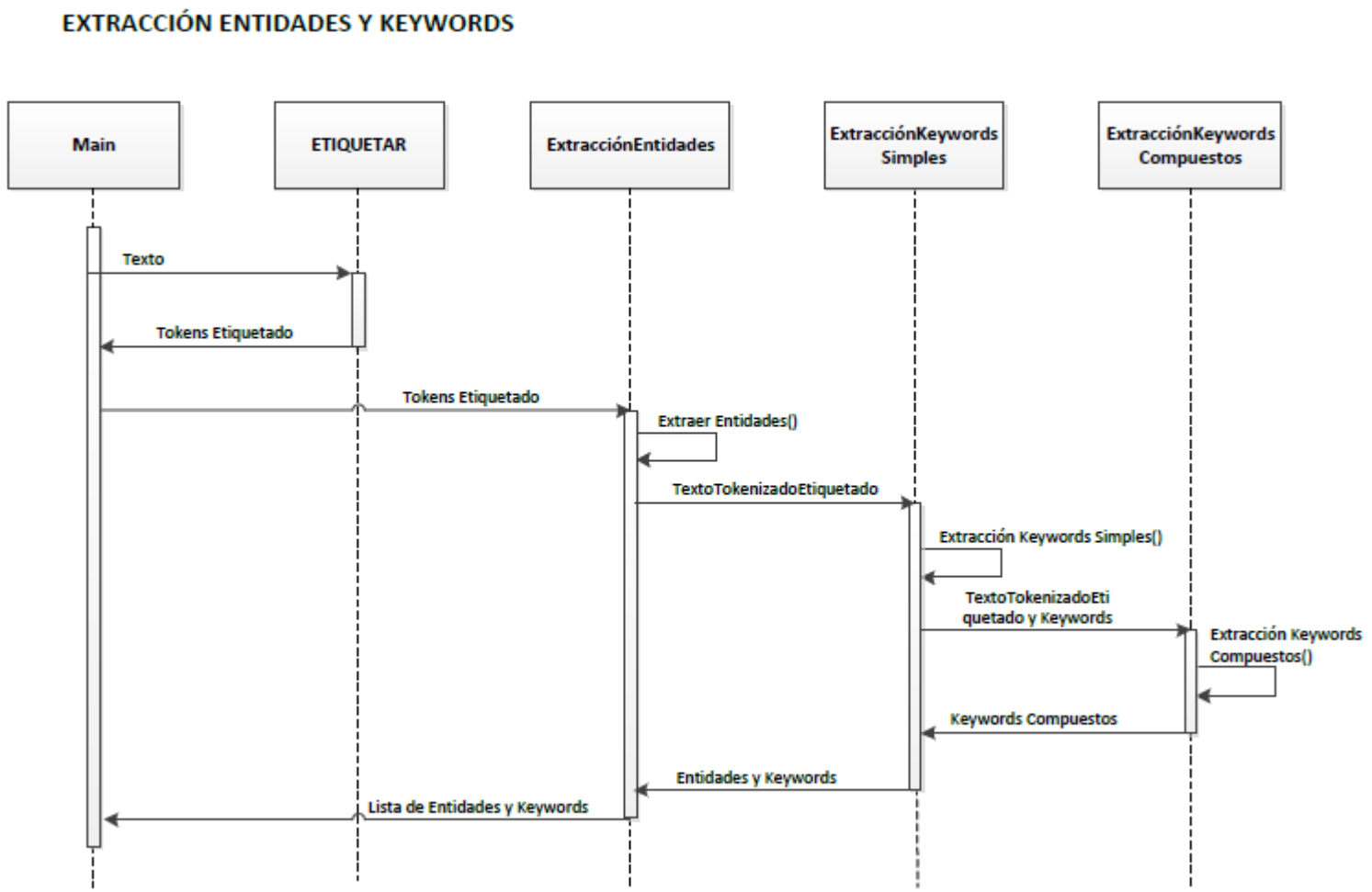
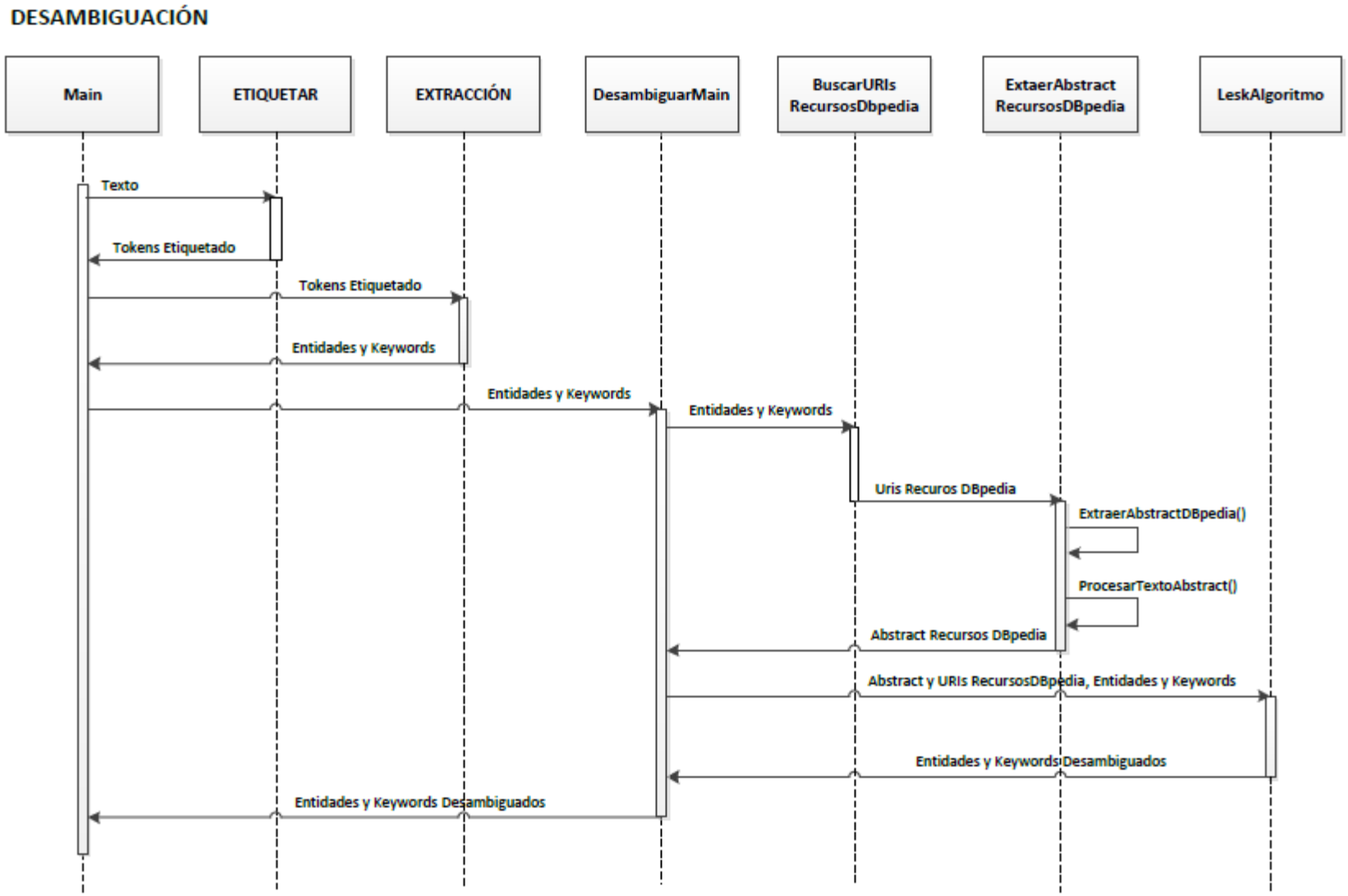


Diagrama de Secuencia - Etiquetado



Preguntas ??

Gracias