FACULTAD DE CIENCIAS E INGENIERÍA

MÉTODO DE EXTRACCIÓN E INDENTIFICACIÓN DE LUGARES DEL MUNDO REAL EN TEXTOS EN ESPAÑOL DEL GÉNERO LITERARIO

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

Melissa Zevallos Franco

ASESOR: Fernando Alva Manchego

Lima, abril del 2015

RESUMEN

En este proyecto de fin de carrera se ha escogido abarcar el tema de Turismo Literario, resaltando principalmente la integración de una herramienta que ha existido durante años, el libro, con las nuevas tecnologías de Reconocimiento de Entidades Mencionadas (REM). Se propone implementar un método de extracción de lugares que se encargue de procesar las obras literarias con la finalidad de identificar los nombres de los lugares mencionados en dichos textos; para que éstos, finalmente, sean validados en el mundo real con el apoyo de una librería de información geográfica. Con el método implementado se va a obtener información, la cual puede ser utilizada para la construcción de herramientas que permitan difundir y aprovechar el Turismo Literario. Esta clase de turismo busca difundir los lugares reales que son mencionados en las obras literarias Estos escenarios pueden ser parques, restaurantes, pasajes, avenidas, entre otros lugares.

FACULTAD DE
CIENCIAS E INGENIERÍA
ESPECIALIDAD DE
INGENIERÍA INFORMÁTICA

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO

TÍTULO: **MÉTODO DE EXTRACCIÓN E IDENTIFICACIÓN DE LUGARES DEL MUNDO REAL EN TEXTOS EN ESPAÑOL DEL GÉNERO LITERARIO**

ÁREA: Ciencias de la computación
PROPONENTE: Mag. Fernando Emilio Alva Manchego
ASESOR: Mag. Fernando Emilio Alva Manchego
ALUMNO: Melissa Zevallos Franco
CÓDIGO: 20084015
TEMA N°: S69
FECHA: San Miguel, 14 de diciembre de 2014

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

DESCRIPCIÓN

En este proyecto de fin de carrera se ha escogido abarcar el tema de Turismo Literario, resaltando principalmente la integración de una herramienta que ha existido durante años - el libro - con las nuevas tecnologías de Reconocimiento de Entidades Mencionadas (REM). Se propone implementar un método de extracción de lugares que se encargue de procesar las obras literarias con la finalidad de identificar los nombres de los lugares mencionados en dichos textos; para que éstos finalmente sean validados en el mundo real con el apoyo de una librería de información geográfica. Con el método implementado se va a obtener información, la cual puede ser utilizada para la construcción de herramientas que permitan difundir y aprovechar el Turismo Literario. Esta clase de turismo busca difundir los lugares reales que son mencionados en las obras literarias Estos escenarios pueden ser parques, restaurantes, pasajes, avenidas, entre otros lugares.

OBJETIVO

Este proyecto tiene como objetivo general implementar un método de extracción e identificación de lugares del mundo real en textos en español del género literario.

OBJETIVOS ESPECÍFICOS

Los objetivos específicos del presente proyecto son:

- Recolectar un conjunto de textos literarios en español.
- Especializar un Reconocedor de Entidades Mencionadas (REM) en español para que identifique únicamente lugares en textos literarios.
- Implementar una aplicación que valide si los lugares extraídos en los textos literarios existen en el mundo real.

Av. Universitaria 1801
San Miguel, Lima – Perú
Apartado Postal 1761
Lima 100 – Perú
Teléfono:
(511) 626 2000 Anexo 4801

- Desarrollar un prototipo que sirva para mostrar los resultados obtenidos luego de extracción e identificación de lugares del mundo real en cada texto literario procesado.

ALCANCE

Este proyecto utilizará un REM existente orientado al idioma español y que se adaptará a las necesidades del proyecto.

Por otro lado, los textos a ser analizados son únicamente del género literario y en español. Asimismo, la historia de estas obras se debe centrar en Lima Metropolitana, ya que se ha limitado el radio de búsqueda de lugares reales a sólo la ciudad de "Lima, Perú". Además, es necesario mencionar que integridad de esta validación está sujeta a la herramienta de información geográfica utilizada.

Por último, el prototipo permitirá analizar una obra literaria digitalizada en formato .TXT y se presentará el listado de lugares reales que son mencionados en dicha obra.

Máximo: 100 páginas

Av. Universitaria 1801
San Miguel, Lima – Perú

Apartado Postal 1761
Lima 100 – Perú

Teléfono:
(511) 626 2000 Anexo 4801

DEDICATORIA

El día de hoy la etapa universitaria que inicié hace unos años ha culminado satisfactoriamente y a lo largo de este recorrido he tenido y compartido momentos de éxito, felicidad, tristeza, amargura, entre otras emociones.

Dedico este trabajo a mi familia que siempre ha deseado lo mejor para mí y me ha enseñado que con perseverancia y muchas ganas todo es posible.

AGRADECIMIENTO

Al Mag. Fernando Alva Manchego que aceptó apoyar desde un inicio este proyecto con absoluta confianza. Gracias a su apoyo, paciencia, palabras de aliento, sabiduría y, especialmente, por ser parte de esta investigación y ayudarme a construir el camino para lograr este objetivo.

A mi papá que siempre ha dado todo de él para que yo logre mis objetivos y me ha enseñado a enfrentar mis miedos y desafíos con inteligencia y buen humor.

A mi hermana Andrea por ayudarme a leer las obras literarias y por brindar ideas que fueron determinantes para incluir la literatura en este proyecto.

Al Doc. Andrés Melgar, por presentarme un lado diferente de mi carrera y ayudarme a defender y consolidar mi trabajo.

ÍNDICE GENERAL

LISTADO DE ILUSTRACIONES

LISTADO DE TABLAS

LISTADO DE EJEMPLOS

1. DEFINICIÓN DEL PROYECTO

Este capítulo tiene como finalidad presentar un panorama general de lo que trata el trabajo de fin de carrera a desarrollar. Este trabajo tiene como objetivo general implementar un método de extracción e identificación de lugares del mundo real en textos en español del género literario. Por esta razón, se va a presentar a continuación una serie de apartados que van a contextualizar, concretar, delimitar y justificar este proyecto.

1.1 Problemática

El desarrollo del turismo a través del tiempo ha demostrado que esta actividad puede convertirse en un medio de promoción y difusión de la cultura nacional, así como en una efectiva herramienta generadora de ingresos y de ayuda en la construcción de un desarrollo sustentable. Prueba de ello, es que el turismo ha crecido un 3.8% a nivel mundial en el año 2012 (Tourism, 2012) y se cuenta que hubo más de 1 billón de turistas en el mundo (UNWTO, 2013).

El turista es el principal actor en esta actividad, por ello se le debe brindar el mejor servicio. Para esto se han desarrollado de la mano con la tecnología, diferentes herramientas que han consolidado esta actividad. Entre éstas encontramos la compra de tickets en línea (LAN, 2013), museos virtuales, generadores de paquetes turísticos (PromPeru, 2013), entre otras.

Asimismo, existen áreas poco convencionales, como lo es la literatura, que pueden ser mejor aprovechas para generar un turismo diferente y ampliar el público interesado en la lectura. Por ejemplo, en Europa encontramos las rutas de El Quijote (Quijote.es, 2013) o las del Código Da Vinci (Fodor, 2013), con las cuales el turista descubre las ciudades de la Mancha e Italia desde una perspectiva más cultural, casi como si fueran parte de la historia de dichos países. Otro ejemplo lo tenemos en Londres con el programa "*Get London Reading*", cuyo objetivo es mostrar en un mapa virtual una serie de libros ubicados en lugares que han influenciado a los autores a escribir sus obras a lo largo de Londres (Limited, 2013).

Por otro lado, en el ámbito nacional, esta alternativa de generar turismo usando literatura es poco explotada, a pesar de que el tipo de turismo más popular es el cultural. La popularidad del turismo cultural se debe principalmente a la gran cantidad de visitas a sitios arqueológicos, museos e iglesias (PromPeru, 2012). Sin embargo, nuestro país está repleto de lugares tangibles que son escenario de grandes historias

literarias, las cuales pueden ser aprovechadas por su valor literario. Estos escenarios pueden ser parques, restaurantes, pasajes, avenidas, entre otras, que en conjunto representan el contexto de la historia que se quiere transmitir en la obra literaria. Esto lo encontramos especialmente en las novelas del género conocido como realismo mágico (Achitenei, María, 2006), en las que se referencia una realidad fundida con lo fantástico.

Un ejemplo nos lo brinda PromPeru con sus guías literarias, una de ellas se titula “La Lima de Mario Vargas Llosa”. En ella se resaltan los lugares más ilustres de nuestra ciudad de Lima que son mencionados en las obras más representativas de Mario Vargas Llosa. Esta iniciativa tiene la misión de difundir y revalorizar la literatura peruana reforzando la identidad nacional entre los peruanos (PromPeru, 2011).

Para la elaboración de una guía similar a la explicada anteriormente, es necesario dejar de lado el contenido literario de las obras y leer de manera meticulosa con el objetivo de reconocer lugares de interés. Asimismo, se requiere conocimiento previo de los lugares para validar que estos existan en la realidad, porque de lo contrario no serán tomados en cuenta. Sin embargo, este procedimiento puede ser automatizado utilizando sistemas basados en Reconocedores de Entidades Mencionadas (REM) y librerías de información geográfica.

Un REM tiene como objetivo identificar y clasificar las entidades nombradas presentes en un texto, como nombres propios de personas, lugares, compañías, entre otros (LIU, 2013). Estos sistemas prometen mejorar la calidad de las búsquedas, especialmente en grandes colecciones de documentos (Althobaiti, 2012) ya que reducen significativamente la cantidad de tiempo invertido en analizarlos (Cunningham, 1999). Sin embargo, estos sistemas dependen mucho del tipo de textos que procesan (noticias, textos académicos, obituarios, comentarios en foros, etc), ya que cada tipo está redactado y estructurado de una manera distinta. A pesar de eso, estos sistemas permiten ser especializados a un determinado tipo de texto. Esto resulta beneficioso para el proyecto porque se puede adaptar esta tecnología para los textos literarios que se desea analizar.

Por tanto, en este proyecto de fin de carrera se propone aprovechar esta tecnología para implementar un método de extracción de lugares que se encargue de procesar las obras literarias con la finalidad de identificar los nombres de los lugares mencionados; para que éstos finalmente sean validados, con el apoyo de una librería de información geográfica, en el mundo real. Con el método implementado en el

presente trabajo se va a poder obtener información, la cual puede ser utilizada para la construcción de herramientas que permitan difundir y aprovechar el Turismo Literario.

1.2 Objetivo General

Este proyecto tiene como objetivo general implementar un método de extracción e identificación de lugares del mundo real en textos en español del género literario.

1.3 Objetivos Específicos

Para que este proyecto pueda alcanzar el objetivo general presentado en 1.2, se ha divido el trabajo en 4 objetivos específicos para un mejor desarrollo, estos son presentados a continuación:

a) **Objetivo Específico 1:** Recolectar un conjunto de textos literarios en español.
b) **Objetivo Específico 2:** Especializar un Reconocedor de Entidades Mencionadas (REM) en español para que identifique únicamente lugares en textos literarios.
c) **Objetivo Específico 3:** Implementar una aplicación que valide si los lugares extraídos en los textos literarios existen en el mundo real.
d) **Objetivo Específico 4:** Desarrollar un prototipo que sirva para mostrar los resultados obtenidos luego de extracción e identificación de lugares del mundo real en cada texto literario procesado.

1.4 Resultados Esperados

Para comprobar el logro de cada uno de los objetivos específicos presentados se han planteado resultados esperados para cada uno de ellos, los cuales se presentan a continuación:

a) **Resultado 1 para el Objetivo Específico 1:** Conjunto de textos literarios en español de manera digitalizada.
b) **Resultado para el Objetivo Específico 2:** Un REM en español especializado en identificar únicamente lugares de obras literarias con un indicador de calidad mayor a 50%.
c) **Resultado para el Objetivo Específico 3:** Sistema de validación de los lugares extraídos en los textos literarios existen en el mundo real.
d) **Resultado para el Objetivo Específico 4:** Prototipo que sirva para mostrar los resultados obtenidos luego de extracción e identificación de lugares del mundo real en cada texto literario procesado.

1.5 Alcance

El presente proyecto se enfoca en proponer un método de extracción e identificación de lugares del mundo real en textos literarios. Se pretende brindar conocimiento pertinente para armar rutas turísticas y, así, difundir el Turismo Literario.

Este trabajo se divide en dos fases: (i) En la primera parte se van a identificar los lugares extraídos de los cuentos literarios en español un Reconocedor de Entidades Mencionadas (REM). (ii) En la parte complementaria, se procede a validar si los lugares extraídos anteriormente existen en el mundo real a través de una herramienta de información geográfica.

1.5.1 Limitaciones de alcance

Se han establecido las siguientes limitaciones de alcance para este proyecto:

a) Los textos son únicamente literarios peruanos y en español. La historia de estas obras se debe centrar en Lima Metropolitana.

b) Se va a validar la autenticidad de los lugares con el mundo real solamente de Lima, Perú. En el Ejemplo 1-A se muestra caso que explica esta limitación.

> En la siguiente oración no se puede diferenciar si se están refiriéndose a la avenida que queda en Miraflores o a la que queda en San Miguel:
> *"…estaban en la **avenida** **Arequipa**"*

Ejemplo 1-A: Limitaciones por geografía.

c) Solo se va a trabajar con lugares reales que existen en la actualidad. En el Ejemplo 1-B se muestra caso que explica esta limitación.

1. *"…me citó en el **Davory**, a las dos."* el lugar identificado era una pastelería en Miraflores que ya no existe.
2. *"vamos al **colegio** **San Peregrino**"* el lugar identificado fue inventado por el autor.

Ejemplo 1-B: Limitaciones por existencia.

d) No se va a trabajar con lugares reales cuyo contexto no me permita identificarlos automáticamente. En el Ejemplo 1-C se muestra caso que explica esta limitación. Puede que se identifique un lugar real, pero el contexto no brinda suficiente

información sobre éste para validar su autenticidad en el mundo real. En el Ejemplo 1-2 se muestra caso que explica esta limitación.

1. *"…sucede exacto que en la **Virgen del Pilar**: todo **San Isidro**, cada familia en su banquita."* el lugar **Virgen del Pilar** puede que no sea identificada como la iglesia existente en San Isidro por falta de información.

2. *"Mangongo llega a la **iglesia** **San Felipe**"* en cambio en esta oración

Ejemplo 1-C: Limitaciones por contexto

1.6 Justificativa

El Internet es un repositorio cada vez más grande de datos que van desde textos e imágenes hasta audios y vídeos. Con este rápido incremento de datos en la web, también ha aumentado la necesidad de obtener información más exacta y pertinente. Esta necesidad impulsa a las personas a perfeccionar las tecnologías de administración de grandes cantidades de datos con la finalidad de transformarlos en información para generar conocimiento.

Se pueden aprovechar estos avances para procesar diferentes tipos de información en distintos escenarios y explotarlas al máximo para generar beneficios, como accesibilidad y rapidez. En este proyecto de fin de carrera se ha escogido abarcar el tema de Turismo Literario, resaltando principalmente la integración de una herramienta que ha existido durante años - el libro - con las nuevas tecnologías de Reconocimiento de Entidades Mencionadas.

Los libros pueden despertar en el lector la necesidad o curiosidad de querer comprobar el parecido existente entre la realidad y la descripción plasmada en las obras como cafés, parques, ciudades, playas y demás lugares. Empujado por este interés, el lector puede verse motivado a visitar los lugares mencionados. Por esta razón, el libro actuaría como una vía indirecta de promoción, favoreciendo así una mejor segmentación del mercado determinada por el perfil de quien disfrute leer.

Por otro lado, al existir un método como el que se va a desarrollar en este proyecto, se tendrá disponible la información de los lugares del mundo real que han sido mencionados en las distintas obras literarias. Con esto se podrán construir diferentes herramientas partiendo de ese conocimiento sin necesidad de ser un experto. También, se puede utilizar para construir desde un tour en un buses turísticos, como es el caso Literatour que ofrece un recorrido por los sitios más representativos de

Miraflores de las obras de Mario Vargas Llosa (Municipalidad de Miraflores, 2013), hasta aplicaciones móviles que utilizan realidad aumentada para mostrar distintos eventos de las obras en cada determinado lugar.

Por esta razón, la difusión de este tipo de turismo, pretenderá también expandir el interés de las personas por enriquecerse de cultura y de la literatura de cada país. Asimismo, el Turismo Literario, no sólo difunde reconocidas ciudades y capitales, también pequeños rincones del mundo que también son escenarios de muchas exitosas obras.

1.7 Riesgos

En la Tabla 1-A se muestran los riesgos relacionados a este proyecto; asimismo, se muestra el impacto y las medidas correctivas a realizar para mitigar cada uno de ellos.

Tabla 1-A: Riesgos del Proyecto

Riesgo identificado	Impacto en el proyecto	Medidas correctivas para mitigar
a) Los resultados del proyecto pueden verse afectados por la falta de información de los lugares que no pertenecen al repositorio de Google o son muy desactualizados.	No se encuentran los lugares en su totalidad.	Agregar los lugares al repositorio geográfico.
c) Retraso en las entregas de los avances	Realizar seguimiento del avance del proyecto y verificar que se está cumpliendo el plan de proyecto.	Priorizar el proyecto de fin de carrera reduciendo tiempos en otras actividades para nivelar el proyecto según el cronograma.
d) Falta de soporte de la versión de la API de Google utilizada.	Revisar constantemente las versiones de la API de Google utilizada y verificar su funcionalidad.	Asumir los cambios realizados pero mantener la versión elegida para la realización del proyecto, pues ya contemplan las funcionalidades que se requiere.

1.8 Viabilidad

Para este proyecto de fin de carrera se ha estimado un periodo de 4 meses, en el que se usarán herramientas gratuitas lo que evitará que se tengan limitaciones financieras y de accesibilidad en el proyecto.

Asimismo, se cuenta con apoyo para conseguir recursos adecuados sobre tema escogido, gracias al grupo GRPIAA-PUCP (Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada de la Pontificia Universidad Católica del Perú); y, al constante desarrollo de la tecnología de Extracción de Información en comunidades de internet.

Para la propuesta de solución mencionada en este proyecto de fin de carrera se ha planteado implementar un reconocedor de lugares existentes en textos literarios que existan en el mundo real. Para lograrlo, se deben conocer los conceptos y procedimientos necesarios que permiten realizar el proyecto de la mejor manera. Este capítulo tiene como finalidad desarrollar los conceptos necesarios para entender tanto el problema como la solución que se pretende plantear.

2.1 Procesamiento de lenguaje natural (PLN)

El procesamiento del lenguaje natural (PLN) es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje de máquina o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por computadora expresada en lenguaje humano para determinadas tareas, como la traducción automática (BIR, 2009).

2.2 Entidad mencionada (EM)

El término "entidad mencionada" es ampliamente utilizado en el Procesamiento del Lenguaje Natural, fue acuñado por la Sexta Conferencia de Entendimiento Mensaje (MUC-6). En MUC-7 se clasifica las entidades mencionadas en las categorías y subcategorías (NER LITERATURE SURVEY 2012) siguientes:

a) Entidad: persona, organización, lugar
b) Expresión de tiempo: fecha, tiempo
c) Expresión numérica: moneda, porcentaje

2.3 Extracción de la información (EI)

La extracción de la información es el proceso en el cual se analizan diversos textos con el fin de identificar y caracterizar el significado del lenguaje humano (Maynard, 2003) representado en un formato específico (Cunningham, 1999), e información relevante acerca de estos. Esta salida (output del proceso) puede ser integrada en los sistemas basados en el conocimiento para obtener resultados más precisos (TURMO, 2006) sobre un dominio en específico.

2.3.1 Tareas de EI

Cunningham distingue cinco tareas en EI, las cuales cada una representa niveles distintos de extracción (Cunningham, 1999). En la Tabla 2-A se muestra una breve

descripción de cada una de estas tareas, junto con el desarrollo de un ejemplo para mejorar el entendimiento de cada una.

2.4 Reconocedor de Entidades Mencionadas (REM)

El Reconocedor de Entidades Nombradas (REM) tiene como objetivo identificar y clasificar las entidades presentes en un texto, tales como nombres propios, compañías, instituciones, expresiones relacionadas con fechas, porcentajes, números, etc. (LIU, 2013) Esta actividad es reconocida como una de las más importantes sub-tareas de EI (Cunningham, 1999), debido a que está ganando popularidad en las aplicaciones cotidianas y promete mejorar la calidad de las búsquedas, especialmente en grandes colecciones de documentes tales como intranets, páginas web, entre otras (Althobaiti, 2012).

Tabla 2-A: Tareas de la Extracción de Información

Ejemplo: ***La hermosa y saludable bebe nació el domingo pasado. Sus padres, Emily y Pedro le pusieron el nombre de Mariana.***

Tarea	*Ejemplo*
1. Reconocimiento de Entidades mencionadas (REM) Tarea de identificar y clasificar las entidades (fechas, lugares, organizaciones, personas).	Entidades: *Bebe, domingo, Mariana, Emily, Pedro, padres.*
2. Resolución de correferencias (RC) Tarea de identificar expresiones en el texto que hacen referencia a la misma entidad.	Se descubre que **Mariana** se refiere al ***bebé***. También que ***Emily*** y ***Pedro*** se refieren a los ***padres***.
3. Construcción de plantillas de elementos (PE) Tarea de añadir información descriptiva al resultado de REM.	Se descubre que ***Mariana*** es ***hermosa*** y ***saludable***.
4. Construcción de plantillas de relación (PR) Tarea de identificar las relaciones entre las diferentes REMs.	Se descubre que ***Emily*** y ***Pedro son los padres*** de ***Mariana***.
5. Extracción de plantillas de escenario (PS) Tarea de reunir los resultados de PE en el escenario específico (Hechos).	Se descubre que el evento fue el ***nacimiento de un bebe en el cual todos las entidades están relacionadas.***

2.4.1 Métricas de evaluación de un REM

Para evaluar los resultados de los sistemas REM se utiliza la *F-measure*, la cual es un promedio ponderado de las dos métricas: *Precisión* y *Recall* que son dos indicadores para la medición del rendimiento más utilizado. La *Precisión* muestra los números de elementos identificados correctamente como proporción del número total de elementos total identificados (Zechner, 1997), con el objetivo de analizar cuántos de los resultados son relevantes. Mientras que el *Recall* muestra el número de elementos identificados correctamente como proporción del número total de elementos correctos a identificar (Zechner, 1997), con el objetivo de analizar cuántos de los resultados son correctos. En la Fórmula 2-A se puede apreciar expresión matemática utilizada para hallar la *F-measure* (LIU, 2013).

$$F - measure = 2.\frac{precision.recall}{precision + recall}$$

Fórmula 2-A: F-Measure

2.4.2 Definiciones relacionadas al entrenamiento de un REM

a) **Data anotada:** La anotación de data es agregar información adicional a las palabras o frases existentes en un texto.
b) **Aprendizaje de Máquina (AM):** El aprendizaje de máquina se centra en la predicción, basada en las propiedades conocidas extraídas de un corpus de data anotada. En otras palabras, son sistemas capaces de optimizar un criterio de desempeño usando datos de ejemplo o experiencia pasada (Althobaiti, 2012). Asimismo son capaces de generalizar comportamientos a partir de información no estructurada en forma de ejemplos, por lo tanto, es un proceso
c) **AM Supervisado:** El objetivo de aprendizaje supervisado es entrenar un modelo con una gran colección de datos anotados y estudiar las características de ejemplos positivos y negativos. El modelo entrenado será capaz de reconocer las instancias que tienen las mismas características que los datos anotados en cualquier ejemplo (Althobaiti, 2012).
   Métodos: Hidden Markov Models, Arboles de Decisión, Modelos de Entropía Máxima, Support Vector Machines (SVM) and Conditional Random Fields (CRF)
d) **AM Semi-Supervisado:** El Aprendizaje Semi-Supervisado explota los datos tanto anotados y no-anotados. Resulta útil manejar estos dos tipos de datos, porque los datos anotados son escasos, mientras que los no-anotados son abundantes y de fácil acceso (Althobaiti, 2012).
   Métodos: Bootstrapping

e) **Corpus para entrenamiento de un modelo AM:** Un corpus (plural, "corpus") es un conjunto de documentos (o, a veces, las frases individuales) que se han anotado a mano con los valores correctos que deben ser aprendidos según un interés en particular (PUSREJOVSKY, 2012). Una vez que se tiene el corpus anotado, este puede ser usado en un modelo de aprendizaje de máquina. La forma más común de hacerlo es dividirlo en dos partes: (1) corpus de desarrollo y (2) corpus de evaluación. Asimismo este último tipo de corpus se divide, también, en dos corpus: (2.1) corpus de entrenamiento y (2.2) corpus de prueba. En la Ilustración 2-A se puede apreciar mejor estas divisiones de corpus. (BIRD, 2009)

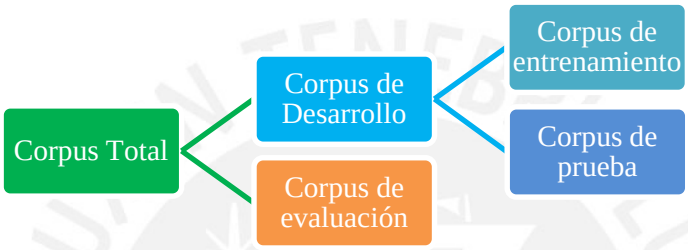

Ilustración 2-A: División de corpus para entrenamiento.

El corpus de entrenamiento es usado para entrenar el modelo y el de prueba es usado para realizar el análisis de error. El corpus de prueba brinda la evaluación final del modelo. Por otro lado, el corpus de evaluación sirve para comprobar, manualmente, que el modelo brinda buenos resultados de acuerdo a la métrica de evaluación resultante. Cabe mencionar que la distribución de corpus se realiza de manera aleatoria (PUSREJOVSKY, 2012), tratando de seguir la división presentada en la Ilustración 2-B.

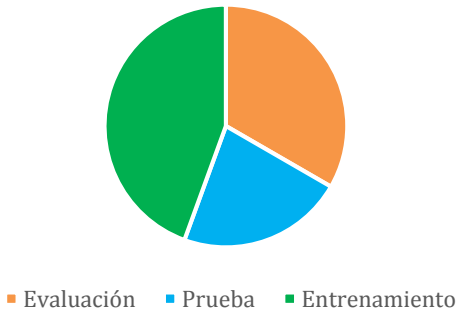

Ilustración 2-B: Distribución de corpus para entrenamiento, adaptación de (PUSREJOVSKY, 2012).

A continuación se presentarán una serie de aplicaciones de los sistemas de extracción de información que han sido desarrolladas para resolver diferentes necesidades en áreas profesionales donde se maneja información. Se decidió utilizar el Método de Revisión Sistemática. Es importante mencionar que en el mercado, actualmente, no existe una herramienta que realice lo planteado por este proyecto.

3.1 Objetivos de la revisión del estado del arte

Se han establecido los siguientes objetivos para esta Revisión Sistemática:

1. Identificar ejemplos exitosos de sistemas que utilicen el método de extracción de información.
2. Identificar ejemplos de reconocedores de entidades mencionadas en español aplicables al problema planteado.
3. Identificar ejemplos de sistemas de información geográfica aplicables al problema planteado.

3.2 Características de la Revisión Sistemática

En la Tabla 3-A se muestran las características del método utilizado.

Tabla 3-A: Características de la Revisión Sistemática

Método utilizado	Revisión Sistemática.	
Buscadores utilizados	IEEE	ACM
Pregunta Planteada	(("Abstract":Named Entity Recognition) OR "Abstract":Information Extraction) Publication Year: 2012 - 2014	(Abstract:"Named Entity Recognition") or (Abstract:"Information Extraction") Publication Year: 2012 - 2014
Criterios de búsqueda	Todos los estudios primarios con las palabras siguientes claves: - *Named Entity Recognition (NER)* - *Information Extraction (IE)* Año de publicación 2012-2014	
Criterios de exclusión	Se van a considerar los resultados que satisfagan lo siguiente: - Trabajen con los siguientes dominios: Social, Turismo y Académico. - Trabaje principalmente con la tarea de Reconocimiento de Entidades Mencionadas	

3.3 Resultados de la revisión sistemática

Se escogieron 4 resultados más representativos después de una selección previa, éstos se muestran detallados a continuación en las tablas Tabla 3-B, Tabla 3-C, Tabla 3-D, Tabla 3-E. Se estableció una plantilla para ordenar la información recolectada. Únicamente los trabajos que cumplían con todos campos requeridos a continuación, fueron elegidos.

a) Dominio de aplicación
b) Idioma
c) Año de publicación
d) Descripción del sistema
e) Métodos, procedimientos y herramientas usadas

Tabla 3-B: Resultado escogido 1ero de la revisión sistemática

[RSI 2] SEED: Un Marco para la extracción de Eventos Sociales en Noticias (Salvatore, 2013)		
Dominio de la aplicación	**Idioma**	**Año de publicación**
Social	Ingles	2013
Sistema		
Este trabajo propone una solución novedosa para descubrir los eventos sociales de las noticias de prensa real editada por seres humanos. En este trabajo se propone SEED, un marco para automáticamente descubrimiento eventos sociales desde una colección de noticias de prensa no estructurada proporcionada por la oficina editorial de una empresa de la Web en el mundo real.		
Métodos, procedimientos, herramientas usadas		
Concretamente, nuestro método se divide en dos etapas, cada una de ellas frente a una extracción de información (IE) tarea específica: primero, utilizamos una técnica para reconocer automáticamente cuatro clases de nombre-entidades de noticias de prensa: fecha, lugar, lugar, y el artista. Además, detectamos eventos sociales mediante la extracción de las relaciones ternarias entre dichas entidades, explotando también evidencia de fuentes externas (es decir, la Web). Finalmente, evaluamos ambas etapas de nuestra propuesta de solución en un conjunto de datos del mundo real.		
Modelo, arquitectura de la propuesta		
El módulo que reconoce las entidades mencionadas analiza las noticias e identifica todas las entidades que extrae. Se distinguen los siguientes sub-módulos: Date tagger, N-Gram tokenizer, Location tagger, Place tagger, Artist tagger.		

Tabla 3-C: Resultado escogido 2do de la revisión sistemática

[RSI 1] Tendencias ocultas en 90 años de la revista de negocios de Harvard (Tsai, 2012)		
Dominio de la aplicación	**Idioma**	**Año de publicación**
Académico	Inglés	2012
Sistema		
En este trabajo se demuestra y se discuten los resultados encontrados en el recopilado de publicaciones de la revista de negocios de Harvard entre 1922 y 2012. Para descubrir tendencias ocultas en los resúmenes de las publicaciones se han empleados técnicas como analizar n-gramas, el análisis básico de confianza y reconocimiento de entidades mencionadas.		
Métodos, procedimientos, herramientas usadas		
Empleamos las herramientas de Stanford PNL para extraer entidades nombradas en los resúmenes y para analizar las oraciones en los resúmenes. El reconocimiento de entidades mencionadas ayuda a encontrar los nombres de los países y las empresas.		

Tabla 3-D: Resultado escogido 3ro de la revisión sistemática

[RSI 3] El reconocimiento de entidades en Tweets (LIU, 2013)		
Dominio de la aplicación	**Idioma**	**Año de publicación**
Social	Inglés	2013
Sistema		
Este trabajo propone un sistema NER novela de tweets original bajo un marco de aprendizaje semi-supervisado.		
Métodos, procedimientos, herramientas usadas		
Para construir un REM para tweets se propone utilizar : 1. Dominio de adaptación, que pretende reutilizar el conocimiento del dominio de origen en un dominio de destino. 2. Aprendizaje de maquina semi-supervisado, cuyo objetivo es utilizar los abundantes datos no etiquetados para compensar la falta de datos anotados.		
Modelo, arquitectura de la propuesta		
Se propone un nuevo método que consiste en tres elementos básicos: 1. Normalización de tweets. 2. Combinación de un clasificador K-Nearest Neighbors (KNN) con un modelo condicional lineal Random Fields (CRF). 3. El uso de un framework de aprendizaje semi-supervisado.		

Tabla 3-E: Resultado escogido 4to de la revisión sistemática

[RSI 4] La identificación de entidades mencionadas en intranet universitaria (Althobaiti, 2012)		
Dominio de la aplicación	**Idioma**	**Año de publicación**
Académico	Inglés	2012
Sistema		
El objetivo de este trabajo es construir un mecanismo exclusivamente para reconocer las tres tipos de entidades mencionadas, que están en constante referencia en el dominio de la Universidad de Essex: nombres, códigos de curso, y los números de las habitaciones. Mientras que el nombre persona es considerada una entidad con nombre común, los códigos de los cursos y el número de habitaciones son específicos del dominio universitario.		
Métodos, procedimientos, herramientas usadas		
Método escogido: Aprendizaje de Máquina Herramientas mencionadas: OpenNLP, ANNIE		
Modelo, arquitectura de la propuesta		
Su sistema NER consta de tres componentes: 2.2 Pre-procesamiento la página web (Incluye la eliminación de datos que no son esenciales en el archivo web, extraer el contenido y tokenizar el texto. 2.3 Reconocimiento de ME-basado. 2.4 Post-procesamiento de la página web incluye elementos de red, destacando referencias cruzadas EM a documentos relacionados en el ámbito de la Universidad, y volver a mostrar la página web.		

En este capítulo se presentarán las herramientas a utilizar en este proyecto. Se ha dado principal énfasis en explicar la razón por la que fueron escogidas.

4.1 Reconocedor de Entidades Mencionadas (REM)

Para la propuesta de solución mencionada en este proyecto de fin de carrera, se planteó utilizar un REM que identificar los lugares mencionados en textos literarios el cual satisfaría la primera parte de este trabajo. Por esta razón, se procedió a investigar sobre REM ya existentes e implementados que estén orientados al idioma español. Se dejó de investigar una vez encontrada una herramienta que cumpliera con todas las características presentadas en la Tabla 4-A. Éstas se explican a continuación:

a) **Español:** La herramienta debe soportar analizar textos en el idioma español. Como dato curioso, se probó una herramienta orientada al idioma ingles para ver cómo esta se comportaba con un texto en español.
b) **Demo online:** Esta característica no es elemental para el análisis, sin embargo ayudó a agilizar el proceso de elección de herramientas, porque se pudo obtener rápidamente resultados mediante la consulta de un demo online.
c) **Disponible para desarrollo:** Si la herramienta es de libre distribución esta puede ser adaptada a este proyecto en particular, en caso sea necesario.
d) **Permite ser especializada:** Si la herramienta permite ser especializada, es decir entrenada en un tipo de texto en particular, ésta puede brindar mejores resultados en comparación a su configuración inicial.

Tabla 4-A: Resumen de herramientas investigadas para el REM

Herramienta	Español	Demo Online	Disponible para desarrollo	Permite ser especializada
NLTK	✔	✔	✔	✘
ALCHEMY	✔	✔	✘	-
NLP STANFORD	✘	✔	-	-
FREELING	✔	✔	✔	✔
LINGPIPE	✔	✘	✔	✘

Adicionalmente a las características solicitadas, se evaluó cada herramienta con el Ejemplo 4-A, ya que, si bien es cierto, no todas permiten ser entrenadas, existía aún

la posibilidad de que no sea necesario hacer ninguna modificación porque ya se obtenían los resultados deseados.

> *Pero se hizo más famoso todavía un poco después, apostando una carrera al amanecer, desde la Plaza San Martín hasta el Parque Salazar, con Quique Ganoza, éste por la buena pista, Pichulita contra el tráfico. Los patrulleros lo persiguieron desde Javier Prado, sólo lo alcanzaron en Dos de*

Ejemplo 4-A: Ejemplo para la evaluación de las herramientas para el REM

En la tabla 4-B se muestra el resumen de los resultados obtenidos después de evaluar el Ejemplo 4-A - en el Anexo 4 A se puede ver el detalle completo de esta evaluación. Sin embargo, se puede apreciar que no obtuvieron buenos resultados con lo que se descartó completamente utilizar una herramienta distinta a FreeLing. Se puede asegurar que la herramienta escogida servirá para este proyecto ya que nos brinda los nombres propios de los lugares deseados, además que permite ser entrenada y con esto se pueden obtener aún mejores resultados.

Tabla 4-B: Resumen de resultados de la evaluación del Ejemplo 2-A

Lugar	NLTK	ALCHEMY	NLP STANFORD	FREELING
Plaza San Martin	-	-	✔	✔
Parque Salazar	-	-	-	✔
Javier Prado	-	-	-	-
Dos de Mayo	✔	-	-	✔

4.1.1 FreeLing

(FreeLing, 2012) es una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas y estos servicios pueden ser utilizados en Java y son de libre distribución. Esta herramienta divide el REM en los siguientes dos módulos:

a) **Módulo NER:** El primero de denomina NER, el cual detecta entidades mencionadas independientemente de su tipo.
b) **Módulo NEC:** La misión de este segundo módulo es asignar una clase a las entidades mencionadas obtenidas por el módulo anterior. Está herramienta distingue las siguientes cuatro clases: (i) Persona → NP00SP0, (ii) Ubicación geográfica → NP00G00 (iii) Organización → NP00O00 (iv) Otros →NP00V00.

Para que esta herramienta nos brinde los resultados del NEC, se procesa una oración por diferentes etapas previas. Éstas son explicadas a continuación y se presenta en el Ejemplo 4-B una oración que va a pasar por cada una de las etapas mencionadas y se explica el resultado de cada una en las imágenes indicadas.

Los patrulleros lo persiguieron desde Javier Prado.

Ejemplo 4-B: Ejemplo etapas de FreeLing.

1. **Se extraen los tokens de la oración es decir las palabras que conforman la oración y se incluyen los signos de puntuación.** Se puede apreciar en la Ilustración 4-A los resultados de esta etapa para el Ejemplo 4-B.
2. **Se extraen los lemas.** En esta etapa se procede a obtener por cada palabra la forma única con la que se nombra un lexema (raíz). Muchas palabras pueden variar su forma de escritura pero todas responden a un mismo lexema. Por ejemplo vivió, vive, vivirá se reduce a vivir (el verbo en infinitivo). Se puede apreciar en la Ilustración 4-B los resultados de esta etapa para el Ejemplo 4-B.
3. **Se realiza un análisis morfológico y etiquetado de PoS.** En esta etapa se procede coloca la etiqueta PoS correspondiente a la función gramatical de cada palabra, éstas pueden ser sustantivo masculino singular, verbo, nombre propio de persona, etc. Freeling ya tiene establecidas sus etiquetas y la que corresponde a las entidades mencionadas es propios NP00000. Se puede apreciar en la Ilustración 4-C los resultados de esta etapa para el Ejemplo 4-B.
4. **Se realiza la clasificación de entidades mencionadas y desambiguación de etiquetas**. En esta etapa se reemplazan las etiquetas PoS de las entidades mencionadas identificadas (NP00000) por las etiquetas correspondientes a su tipo de entidad. Se puede apreciar en la Ilustración 4-D que la etiqueta de la entidad "javier_prado" fue reemplazada. Y finalmente se eligen las etiquetas más adecuadas para cada palabra de acuerdo a las probabilidades definidas y al contexto de la oración. Se puede apreciar la elección de etiquetas para el Ejemplo 4-B en la Ilustración 4-D, asimismo se visualiza que no necesariamente coge la probabilidad mayor ya que la etiqueta para la palabra "lo" es una con 0.27%.

Ilustración 4-A: Etapa 1 de Ejemplo 4-B.

Los	patrulleros	lo	persiguieron	desde	Javier_Prado	.
el	*patrullero*	*el*	*perseguir*	*desde*	*javier_prado*	.

Ilustración 4-B: Etapa 2 de Ejemplo 4-B.

Los	patrulleros	lo	persiguieron	desde	Javier_Prado	.
el DA0MP0 0.976481	*patrullero* NCMP000 0.64273	*el* DA0NS0 0.457533	*perseguir* VMIS3P0 1	*desde* SPS00 1	*javier_prado* NP00000 1	. Fp 1
lo PP3MPA00 0.0234632	*patrullero* AQ0MP0 0.35727	*lo* PP3CNA00 0.271177				
lo NCMP000 5.5732e-05		*lo* PP3MSA00 0.271177				
		lo NCMS000 0.000112943				

Ilustración 4-C: Etapa 3 de Ejemplo 4-B.

Los	patrulleros	lo	persiguieron	desde	Javier_Prado	.
el	*patrullero*	*lo*	*perseguir*	*desde*	*javier_prado*	.
DA0MP0	NCMP000	PP3CNA00	VMIS3P0	SPS00	NP00SP0	Fp

Ilustración 4-D: Etapa 4 de Ejemplo 4-B.

4.2 Herramienta de información geográfica

Para la identificación de lugares del mundo real se ha decidido utilizar la herramienta de Google Maps, la cual se explica a continuación:

4.2.1 Google Maps

Se ha decidido utilizar la herramienta de Google Maps debido a que proporciona una licencia gratuita para consultar sus servicios en todo momento y asegura que la información que brinda es íntegra y actualizada. Los servicios a utilizar se presentan a continuación (GOOGLE, 2013):

a) **Servicios web:** Son servicios web los cuales utilizan solicitudes de URL para acceder a información de lugares, de elevación, de direcciones y de codificación geográfica.
b) **API de JavaScript de Google Maps:** Este servicio permite insertar un mapa de Google interactivo en una página web utilizando JavaScript.

En el presente capítulo se van a desarrollar los primero dos objetivos específicos los cuales fueron denominados “Recolectar un conjunto de textos literarios en español” y “Especializar un Reconocedor de Entidades Mencionadas (REM) en español para que identifique únicamente lugares en textos literarios.”.

Se tiene como finalidad obtener un REM el cuál pueda satisfacer los objetivos mencionados anteriormente y sea utilizado en este proyecto. Para poder alcanzar este resultado se han utilizado diversas herramientas, desarrollado métodos y establecido definiciones los cuales van a ser explicados a continuación.

5.1 Herramientas utilizadas

La herramienta que ha sido utilizada en este capítulo ha sido explicada en el capítulo de Herramientas, y se menciona a continuación la importancia del uso de esta en este capítulo en particular:

a) **Freeling:** Esta herramienta nos brinda un REM ya implementado (nativo) el cual permite el reconocimiento y clasificación de entidades mencionadas. Entre los tipos de entidades que clasifica está LUGAR, y es precisamente ese tipo con el que se desea trabajar en este proyecto. Asimismo, esta herramienta permite ser adaptada a diferentes tipos de textos, y se puede utilizar esta opción para especializarse en textos literarios y obtener mejores resultados.

5.2 Desarrollo

A continuación se va a presentar las diferentes actividades y decisiones que se han realizado para alcanzar el objetivo definido en este capítulo. Asimismo en el ANEXO 5 H se muestran los resultados de todas las evaluaciones realizadas en este capítulo para un ejemplo en particular.

5.2.1 Definición de Tipos de Entidad

Se describen los tipos de entidades que se han considerado desde un inicio en este proyecto, en base a la experiencia de haber leído los libros.

a) Entidad Completa: Es aquella entidad que tiene un atributo que le brinda información adicional. Se ha asumido que el atributo es el sustantivo inmediatamente antes de la entidad. Este tipo se puede apreciar en el Ejemplo 5-

A. Esta característica es importante para poder contextualizar a la entidad y poder distinguirla.

Sin embargo, al asumir esto perdemos los casos presentados en el Ejemplo 5-B, pero es un riesgo que se ha aceptado para poder obtener esta característica necesaria. Se trabajó el caso contrario, queriendo incluir el Ejemplo 5-B, pero aumento la complejidad y era necesario que las entidades extraídas sean validadas personalmente con lo que ya no iba a ser automático.

1. *"…Descendió en la avenida **Pardo**, a una cuadra de su casa…".*
2. *"… hacia el muro blanco vio por la calle **Dos de Mayo**…"*

Ejemplo 5-A: Entidad Completa

1. *"…se fueron por el centro de **Lima**…"*
2. *"..en la esquina de **Solari**"*

Ejemplo 5-B: Entidades Completas no consideras

b) Entidad Real: Es aquella entidad que existe en el mundo real. Asimismo, con la finalidad de reducir la ambigüedad entre las entidades, explicada en el Ejemplo 1-A, se ha decidido limitar este mundo real al territorio de Lima Perú. Este tipo se puede apreciar en el Ejemplo 5-C.

1. *"…Ludo anduvo por las calles animadas de La Parada…".*
2. *"… les dio una propina y pronto nos fuimos, hacia ~~Paris~~…"*

La Parada es reconocida como entidad Real, en cambio Paris no lo es.

Ejemplo 5-C: Entidad Real

5.2.2 Recolección y digitalización de textos literarios

El objetivo de recolectar textos literarios es para contextualizar las pruebas realizadas en este proyecto y, con los resultados obtenidos, evaluar la mejor toma de decisiones para alcanzar el Objetivo General. Adicionalmente, esta actividad sirvió para comprobar que, efectivamente, para realizar lo que se plantea en este proyecto se debe invertir gran cantidad de tiempo y hacer una lectura meticulosa de cada obra literaria.

Para lograr esta recolección, fue necesario investigar qué libros cumplían con las características de contar con entidades mencionadas del tipo lugar y que un gran porcentaje de estas sean entidad real. Afortunadamente, se contaba con experiencia previa de lectura de literatura peruana y se logró seleccionar 6 libros - la información sobre éstos se encuentra en el ANEXO 5 A.

Estos libros fueron leídos en su estado físico y se marcaron las hojas en donde se presentaban lugares. Con estas hojas marcadas se escogieron 23 intervalos de páginas que aseguraban tener, entre líneas, lugares mencionados. Estos intervalos fueron digitalizados obteniendo, así, 23 archivos de texto que serán utilizados para diferentes experimentos explicados a lo largo de este trabajo.

5.2.3 Identificación de lugares en textos recolectados

Una vez obtenidos los 23 archivos digitalizados, se identificaron nuevamente los lugares que se mencionan en estos libros en una bitácora de pruebas, para poder tener ordenado los resultados, ya que esta información es determinante para lo que continúa del proyecto.

Asimismo, se identificaron características por cada entidad de manera manual para cada archivo recolectado – un ejemplo de esta información se puede apreciar en el ANEXO 5 B - éstas se detallan a continuación:

a) **Nombre:** en este campo se colocó el nombre propio identificado del lugar en particular, tomando en cuenta la definición de una entidad mencionada.
b) **Atributo:** en este campo se colocó el sustantivo inmediatamente antes del nombre del lugar, tomando en cuenta la definición de una entidad completa.
c) **Si es Completa:** en este campo se colocó SI en el caso que el lugar en particular tuviera un atributo.
d) **Si es Real:** en este campo se colocó SI en el caso que el lugar en particular existiera en el mundo real, tomando en cuenta la definición de una entidad real.

5.2.4 Evaluación del módulo NER nativo

Como se mencionó la herramienta elegida divide el reconocedor de entidades mencionadas en dos módulos. El primero denominado NER el cual identifica entidades independientemente del tipo a la que estas pertenecen fue probado con los textos literarios recolectados en 5.2.2 y se lograron identificar un 95% de las entidades de los lugares anotados en la bitácora de pruebas - en el ANEXO 5 C se puede ver el detalle de esta evaluación. Con este resultado se decidió que no era necesario entrenar este

módulo y se ha utilizado tal como ha venido configurado por defecto en la herramienta para lo que resta de este proyecto.

5.2.5 Evaluación del módulo NEC nativo

Si bien es cierto que el módulo de NER nos brinda altos resultados, aún faltaba evaluar el segundo módulo de FreeLing denominado NEC, el cual clasifica las entidades identificadas. Entonces se probó este módulo con cada uno de los textos literarios recolectados en 5.2.2 y se lograron identificar solamente un 53% de los lugares anotados en la bitácora de pruebas - en el ANEXO 5 D se puede ver el detalle de esta evaluación. Con este resultado se decidió que era necesario especializar este módulo con la finalidad de obtener un porcentaje mayor al obtenido con la configuración por defecto de la herramienta escogida.

En el Ejemplo 5-D se puede apreciar que a pesar de haber reconocido la entidad correctamente, esta no necesariamente está bien clasificada y esto genera una pérdida de información importante para este proyecto.

1. *"...en la **avenida** Javier Prado"*, la entidad es reconocida como *Persona*.
2. *"...en la **avenida** Dos de Mayo", la entidad es reconocida como Fecha.*

Ejemplo 5-D: Problemas de clasificación de entidades.

5.2.6 Recolección de textos del CONLL2002

El objetivo de recolectar textos adicionales a los literarios, es el de identificar cuántas entidades son necesarias para un entrenamiento ya antes elaborado. Para lograr esta recolección se investigó qué corpus en español ya existía y estaba accesible en internet. Entonces se escogieron los archivos pertenecientes al CONLL2002 por estar disponibles en su página web (CNTS, 2005) y por estar orientadas al idioma español. Además, se investigó y resultó que este conjunto de textos fueron utilizados para el módulo de NEC nativo de FreeLing (PADRÓ, 2012) pero solamente se consiguieron los textos. Después de revisarlos y analizarlos, se contaron 103 lugares y se tomó este valor como mínima cantidad de lugares anotados para cualquier especialización de dicho módulo de NEC.

5.2.7 Especialización del módulo NEC

En este apartado se van a detallar los tipos de corpus que se han planteado en este proyecto con el objetivo de mejorar los resultados obtenidos con el NEC nativo.

Asimismo, es importante resaltar, que para poder llevar a cabo cada especialización de los corpus elegidos primero éstos fueron estructurados según el formato establecido por FreeLing - más información en el apartado 5.2.7.1 - y después de realizar la especialización respectiva - más información en el apartado 5.2.7.2 - estos fueron evaluados con los textos recolectados en 5.2.2 - ver ANEXO 5 G para ver el detalle de cada evaluación. Finalmente en la Tabla 5-A se presenta un resumen de los resultados obtenidos de cada evaluación.

a) CORPUS-A: Este corpus contiene los textos literarios recolectados en 5.2.2. La finalidad de contar con este corpus es que se va a contar con una especialización orientada sólo a textos literarios.

Para la elaboración de este corpus, se eligieron 8 archivos aleatoriamente, dando un total de 122 lugares (según lo investigado en 5.2.6 este valor debe ser mayor a 103) y se generaron los siguientes dos archivos que fueron utilizados para el entrenamiento: (i) Archivo de entrenamiento → que contiene los archivos “input 4, 8, 10, 14, y 15” y (ii) Archivo de prueba → que contiene los archivos “input 6, 18 y 21”.

Una vez evaluado este corpus, se pudo confirmar que efectivamente la calidad aumento casi en 20%, ver Tabla 5-A. Sin embargo, aún se deseaba obtener mejores resultados y es por eso que se crearon los CORPUS B y C.

b) CORPUS-B: Este corpus contiene el CORPUS-A y textos del CONLL2002 recolectados en 5.2.3. Se decidió adjuntar estos últimos archivos ya que estos habían servido satisfactoriamente para la especialización de otra herramienta; con esta premisa se esperaba que aumente los resultados obtenidos por el CORPUS A.

Para la elaboración de este corpus, se conservaron los archivos generados para el CORPUS A y se añadió a cada uno los textos del CONLL2002 divididos en una proporción de 70% - 30% para el archivo de entrenamiento y prueba respectivamente.

Una vez evaluado este corpus, se aprecia que la precisión aumenta, es decir se obtienen más lugares deseados porque se aumentó más ejemplos al corpus. Sin embargo, estos ejemplos eran generales y no estaban orientados a la literatura

con lo que el recall disminuyó, se obtenían más resultados innecesarios, ver Tabla 5-A.

c) **CORPUS-C:** Este corpus contiene el CORPUS-A con la diferencia de que solo se van a tomar en cuenta las oraciones que tengan como mínimo un lugar mencionado.

El objetivo de elaborar este corpus es brindar ejemplos más específicos, porque se tuvo la idea de que las demás oraciones no daban ningún valor agregado a la especialización. Para este corpus, se modificaron los archivos generados para el CORPUS A y se conservaron las líneas con las oraciones indicadas.

Una vez evaluado este corpus, se aprecia que la recall aumenta notablemente, es decir ya no se obtenían resultados innecesarios, justamente por tener ejemplos más específicos. Sin embargo, las oraciones que no contienen lugares mencionados también son importantes para que el modelo aprenda lo que no debe reconocer como lugar y se logré calibrar el reconocedor y así, obtener más lugares deseados, porque la precisión disminuyo demasiado usando este corpus, ver Tabla 5-A.

Tabla 5-A: Información y resultados de cada corpus utilizado para el NEC

	Tipo de textos	Oraciones	Entidades (Lugares)			Precisión	Recall	F-measure
			Totales	Completas	Reales			
Corpus A	Literarios	TODO	259	77	221	0,680	0,771	72%
Corpus B	Literarios + CONLL2002	TODO	362	85	324	0,754	0,643	69%
Corpus C	Literarios	Solo las que contienen lugares	259	77	221	0,448	0,847	58%

5.2.7.1 Elaboración de los corpus de entrenamiento y prueba

Para elaborar los corpus de entrenamiento y prueba se siguió el formato establecido en el manual de H-FREE (Freeling, 2014). Este formato exigía que cada palabra del texto y sus elementos estén en una sola línea y se deje una línea vacía entre oraciones. En la Tabla 5-B se presenta un ejemplo de los elementos esenciales para la palabra Lima y a continuación se presenta la explicación y cómo se obtuvo cada uno de ellos:

1. **Token:** en este campo se coloca la palabra que se lee del archivo de texto.
2. **Etiqueta BIO:** en este campo se coloca B cuando la palabra es el lugar previamente identificado, en caso contrario se coloca O.
3. **Lemma:** en este campo se coloca la forma única con la que se nombra un lexema. Muchas palabras pueden variar su forma de escritura pero todas responden a un mismo lexema. Por ejemplo vivió, vive, vivirá se reduce a vivir (el verbo en infinitivo). Para obtener esta información se reutilizaron funciones definidas por la herramienta y no se realizó modificación alguna.
4. **Función gramatical:** en este campo se coloca la etiqueta PoS correspondiente a la función gramatical de cada palabra, éstas pueden ser sustantivo masculino singular, verbo, nombre propio de persona, etc. Freeling ya tiene establecidas sus etiquetas y la que corresponde a los nombres propios de lugares es NP00G. Para obtener esta información se reutilizaron funciones definidas por la herramienta. Sin embargo, como se tenía el conocimiento que el NEC no clasificaba bien todos los lugares, se decidió colocar manualmente NP00G a las palabras que correspondían y retirar se retiró esa etiqueta de los lugares mal identificados.
5. **Separador:** en este campo siempre se coloca el símbolo # para indicar el inicio de las formas de lectura.
6. **Formas de lectura**: en este campo se deben colocar las distintas formas de lectura que puede tener una palabra. Por ejemplo en el caso de la palabra “como” que puede ser utilizada como verbo, preposición, etc. En el caso particular de este trabajo se decidió colocar la misma función gramatical de la parte d), porque los lugares al ser nombres propios y al estar en mayúsculas no iba a existir ninguna confusión, además que el tiempo invertido en esta parte no iba a justificar ninguna mejora para el proyecto.
7. **Fin de forma de lectura:** en este campo siempre se coloca -1 por cada forma de lectura definida en f). En este trabajo en particular, al solo tener una forma de lectura siempre la línea terminará en “-1”.

Tabla 5-B: Ejemplo del formato requerido por FreeLing para el entrenamiento.

Formato por palabra establecido por FREELING							
Ejemplo	**Lima**	**B**	**Lima**	**NP000G**	**#**	**lima NCFG000**	**-1**
Elementos	Token	Etiqueta BIO	Lemma	Función gramatical		Formas de lectura	

5.2.7.2 Especialización del módulo de NEC

Para la especialización del módulo de NEC, se siguió lo indicado en el manual de Freeling. En él se especifica que si se tiene un corpus anotado, los modelos pueden

ser especializados, es decir entrenados, utilizando los scripts en *src/utilidades/NERC* y se deben seguir las indicaciones explicadas en el archivo README.txt. En resumen se realizan las siguientes actividades:

1. **Instalar Freeling 3.0.**
2. **Preparar los archivos de entrenamiento y prueba** – ver apartado 5.2.7.1 para ver cómo se le da el formato correspondiente.
3. **Crear las listas gazeetters.** Para elaborar este paso se ejecuta el script prepare-corpus.sh para el idioma español.
4. **Definir un conjunto de reglas de extracción.** Estas reglas definen como orientar la clasificación de las entidades para que se distinga entre los tipos persona, lugar, organización y otros. Cabe resaltar que se escogió el archivo F34 correspondiente a las reglas de extracción usadas en el módulo de NEC nativo y no se realizó modificación alguna porque el objetivo de la especialización era orientar la herramienta a textos literarios. Para elaborar este paso se ejecuta el script encode-corpus.sh para cada uno de los archivos obtenidos en el paso 2.
5. **Entrenar el modelo AdaBoost:** Aquí es donde el entrenamiento se realiza. Este es el proceso en donde se aplica el algoritmo de aprendizaje de máquina Adaboost y este convertirá un conjunto de clasificadores débiles cuya performance sea apenas por encima de la elección al azar, en un clasificador fuerte arbitrariamente preciso. Para elaborar este paso se ejecuta el script nec-adaboost.sh.
6. **Trasladar los archivos generados,** por todos los pasos anteriores, a producción y modificar el archivo de configuración correspondiente al idioma español.

5.3 Consideraciones finales

Finalmente, se completaron los primeros dos objetivos específicos cumpliendo con sus resultados esperados, los cuales se presentan a continuación:

a) **Resultado para el Objetivo Especifico 1:** Conjunto de textos literarios en español de manera digitalizada.
b) **Resultado para el Objetivo Especifico 2:** Un REM en español especializado en identificar únicamente lugares de obras literarias con un indicador de calidad mayor a 50%.

Cabe resaltar que utilizando la herramienta escogida con su configuración inicial ya se había cumplido con el objetivo 2. Sin embargo se realizaron 3 evaluaciones adicionales para poder mejorar los resultados por defecto, obteniendo una configuración con un F-Measure de 72% la cual será utilizada en este proyecto.

Una vez extraídos los nombres de los lugares en las obras literarias, se va a identificar cuáles de estos existen en el mundo real. Para ello, en el presente capítulo se va a desarrollar el tercer objetivo específico denominado "Implementar una aplicación que valide si los lugares extraídos en los textos literarios existen en el mundo real".

Se tiene como finalidad obtener un método el cuál pueda satisfacer el objetivo planteado y lo integre con el REM obtenido en el capítulo anterior. Para alcanzar este resultado se han realizado diversas actividades las cuales se explican a continuación.

6.1 Herramientas utilizadas

Las herramientas que fueron utilizadas en este capítulo han sido explicadas en el capítulo de Herramientas, y se menciona a continuación la importancia de cada una en este capítulo en particular:

a) **Freeling:** Es importante utilizar esta herramienta para rescatar la información del etiquetado PoS durante el proceso del REM. Con la finalidad de identificar los atributos diferenciadores de las entidades – ver "Entidad Completa" en 5.2.1- los cuales se han definido como el sustantivo inmediatamente antes de la entidad mencionada del tipo lugar. En el desarrollo de este capítulo se va a apreciar la importancia de esta información complementaria.
b) **Google Maps:** Esta herramienta se utiliza para la validación de lugares extraídos en el mundo real. Ésta nos proporciona un servicio de predicción de consultas para búsquedas geográficas basadas en un texto de consulta. En este capítulo es importante ya la predicción brindada se aproxima aún más al lugar que se quiere encontrar y nos complementa la información se tiene.

6.2 Desarrollo

Los textos literarios recolectados en 5.2.2 cuentan con un 80% de entidades reales – ver Tabla 6 A, por lo que se puede concluir que hay un gran número de nombres de lugares en estos textos que existen en Lima Metropolitana.

Tabla 6-A: Porcentaje de entidades totales en textos recolectados

Entidades Totales		
Lugares	Completas	Reales
279	57	221
	80% incompletas	80%

Sin embargo, los nombres de los lugares pueden no ser suficientes para determinar su existencia en el mundo real; asimismo, hay gran probabilidad que suceda esto, ya que un 80% de lugares carece de un atributo diferenciador – ver Tabla 6 A. Se ha llegado a esta conclusión después de releer y analizar los textos seleccionados. Es por ello que se han identificado las siguientes 3 situaciones que refuerzan esta afirmación:

a) **CASO 1 - Ambigüedad geográfica:** El nombre de un lugar puede existir en más de una oportunidad, ya que se refieren a distintos lugares geográficos. En el Ejemplo 6-A se puede apreciar un ejemplo de este caso.

En la siguiente oración no se puede diferenciar si se están refiriéndose a la avenida que queda en Miraflores o a la que queda en San Miguel:

*"…estaban en la avenida **Arequipa**"*

Ejemplo 6-A: Caso 1 – Ambigüedad geográfica.

a) **CASO 2 - Falta de información de entidades:** La entidad no es completa y carece de información adicional para determinar cuál es lugar del que se está haciendo referencia. En el Ejemplo 6-B se puede apreciar un ejemplo de este caso.

En la siguiente oración no se puede diferenciar si se están refiriéndose a la avenida o a la universidad,

*"…estaban en **Ricardo Palma**"*

Ejemplo 6-B: Caso 2 – Falta de información de entidades

b) **CASO 3 – Falta de relación entre entidades:** Se pueden identificar dos entidades que se refieran al mismo lugar pero sean identificadas de manera independiente debido a su diferente capitalización de palabras. En el Ejemplo 6-C se puede apreciar un ejemplo este caso.

Las siguientes dos oraciones hacen referencia al mismo lugar en particular

"…estaban en la Plaza San Martin" y "…estaban en la plaza San Martin"

De considerarse los lugares por separado, se estaría identificando San Martín como un lugar distinto y podría generar problemas.

Ejemplo 6-C: Caso 3 – Falta de relación entre entidades

6.2.1 Definición de criterios estandarización de entidades

Se ha establecido que los nombres de los lugares extraídos por el REM deben pasar por una estandarización previa pasando por los 3 criterios explicados a continuación, con la finalidad de reducir el número de entidades duplicadas y aumentar el número de entidades completas.

a) **Criterio 1 - Entidades idénticas:** Se asume que dos entidades son idénticas si los dos atributos son iguales y si los nombres de las entidades también son iguales. Se va a conservar la última entidad leída. En el Ejemplo 6-D se puede apreciar mejor este criterio.

Se identifica la entidad **Paracas**, asimismo se identifica otra entidad **Paracas**. Entonces se concluye que se las dos entidades son idénticas entre sí.

Ejemplo 6-D: Entidades idénticas

b) **Criterio 2 - Entidades equivalentes:** Se asume que dos entidades son equivalentes si la combinación de atributo + nombre es igual al atributo + nombre de la segunda entidad. Se va a conservar la última entidad completa leída. En el Ejemplo 6-E se puede apreciar mejor este criterio.

Se identifica la entidad **Salazar** con el atributo **parque**, asimismo se identifica otra entidad **Parque Salazar**. Entonces se concluye que se las dos entidades son equivalentes entre sí.

Ejemplo 6-E: Entidades equivalentes

c) **Criterio 3 – Entidades similares en el mismo nivel:** Se asume que dos entidades son similares en el mismo nivel si se encuentra después de una entidad

completa, una incompleta con el mismo nombre. En el Ejemplo 6-F se puede apreciar mejor este caso.

> *Se identifica la entidad Arequipa con el atributo avenida, asimismo se identifica otra entidad* ***Arequipa*** ***sin atributo*** *en el mismo nivel. Entonces se concluye que se las dos entidades son similares en el mismo nivel entre sí.*

Ejemplo 6-F: Entidades similares en el mismo nivel

Asimismo, se han establecido dos factores que determinan si dos entidades están en el mismo nivel, los cuales se describen a continuación:

1. **ALPHAPLACES:** Este factor valida el rango de mención entre dos entidades. Por ejemplo si la ENTIDAD 1 fue extraída en el orden 2 y la ENTIDAD 2 fue extraída en el orden 20 entonces se valida si (20-2 <ALPHAPLACES). Este factor es importante porque pueden haber muchas entidades identificadas en un rango de líneas muy corto y no necesariamente se va a cumplir lo visto en el Ejemplo 6-F.

2. **ALPHALINE:** Este factor valida el rango de líneas entre dos entidades. Por ejemplo si la ENTIDAD-1 fue extraída en la línea 5 y la ENTIDAD-2 fue extraída en la línea 15 entonces se valida si (15-5 < ALPHALINE). Este factor es importante porque pueden haber muy pocas entidades identificadas en un rango de líneas muy largo y no necesariamente se va a cumplir lo visto en el Ejemplo 6-F.

   Ambos factores se deben cumplir para poder concluir que dos entidades están en el mismo nivel. En este proyecto los valores de ALPHAPLACES y ALPHALINE eran de

6.2.2 Implementación del método

Para el desarrollo de la aplicación, primero se obtuvieron los atributos, de existir alguno, de cada lugar extraído en el capítulo anterior. Para realizar esto, durante el proceso de clasificación de entidades mencionadas, se consideraron los sustantivos inmediatamente de una entidad mencionada del tipo lugar, gracias a su etiqueta PoS, y éstos fueron almacenados junto a los nombres de los lugares. Obteniendo así, una dupla *(ATRIBUTO Y ENTIDAD)* con la que se necesitaba trabajar.

Cada vez que se procesa un texto en particular por el REM se obtiene un listado de todas las duplas *(ATRIBUTO Y ENTIDAD)*. Este listado pasa por la estandarización de entidades – explicada en 6.2.1 – como se puede apreciar en la Ilustración 6-A.

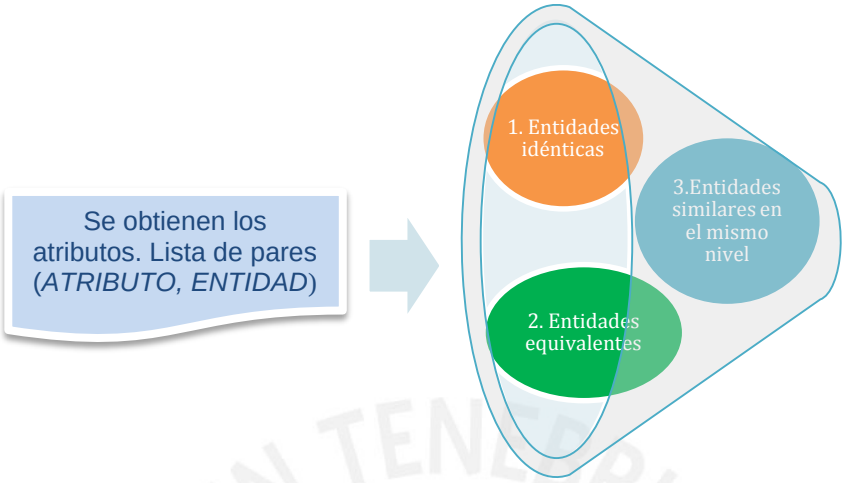

Ilustración 6-A: Recopilación de atributos y criterios de estandarización

Una vez estandarizadas las entidades se procede a la validación con el mundo real. Para esta etapa se ha utilizado las funciones establecidas Google Maps. En resumen, se recibe un texto consulta que va a ser procesado y se obtiene la predicción más aproximada al lugar solicitado. Con la información recibida es posible obtener la ubicación correspondiente al lugar consultado.

Asimismo, con la finalidad de reducir la ambigüedad entre las entidades, se ha decidido reducir el rango de búsqueda al territorio de Lima Perú para las entidades. Entonces el texto de consulta sería la concatenación de la siguientes información: ATRIBUTO + NOMBRE + “Lima, Perú”.

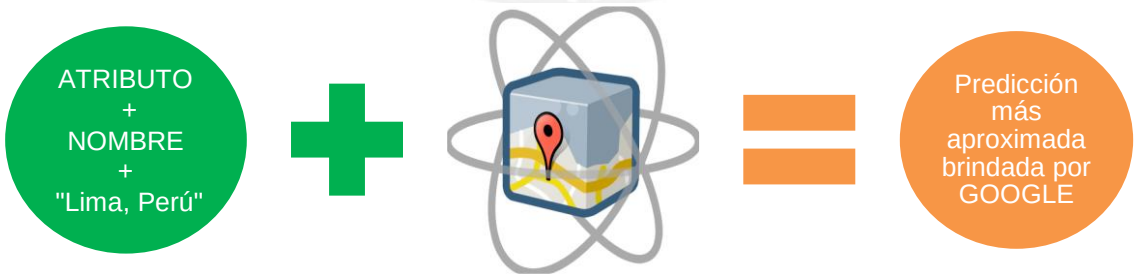

Ilustración 6-B: Esquema de validación de lugares con Google

6.2.3 Desempeño del método

Se realizaron evaluaciones con los libros recolectados en 5.2.2, es decir se juntaron todos los archivos relacionados con un libro en un solo archivo para poder simular un

escenario real y evaluar cuántos de los lugares reales que tenían que ser identificados por la herramienta de información geográfica se obtenían. En resumen, el valor de desempeño de la validación con el mundo real es 56% y éste valor resultó ser bastante alto porque los lugares que se dejaron de encontrar correspondían, en su mayoría, a entidades sin información adicional (incompletas).

6.3 Consideraciones finales

Finalmente, se completó el tercer objetivo específico cumpliendo con su resultado esperado, el cual se presentan a continuación:

a) **Resultado para el Objetivo Específico 3:** Sistema de validación de los lugares extraídos en los textos literarios existen en el mundo real.

Cabe resaltar que utilizando la herramienta geográfica escogida inicialmente sin hacer ninguna modificación a los resultados del REM, ésta arrojaba resultados muy lejanos a los que se deseaban obtener, casi ninguno correcto. Por esta razón, se decidió que era necesario obtener información complementaria a los nombres de los lugares. Entonces se extrajeron los atributos diferenciadores respectivos y, también, se definieron criterios de estandarización para poder mejorar los resultados. Con estas mejoras se obtuvo un desempeño total de 80% de este proyecto, ya que el método obtenido en este capítulo está integrado con el REM obtenido en capítulo anterior.

7. DESARROLLO DEL PROTOTIPO

El objetivo que se va a desarrollar en este capítulo fue denominado “Desarrollar un prototipo que sirva para mostrar los resultados obtenidos luego de extracción e identificación de lugares del mundo real en cada texto literario procesado”.

Se tiene como finalidad desarrollar las funciones necesarias para integrar lo trabajado en los capítulos anteriores en un prototipo y lograr así el objetivo general de este proyecto, el cual es “Implementar un método de identificación y extracción de lugares del mundo real en textos en español del género literario”.

Para poder alcanzar este resultado se han utilizado diversas herramientas, desarrollo de métodos y tomado decisiones. La totalidad de estos elementos se va a explicar a continuación.

7.1 Herramientas

La herramienta que ha sido utilizada en este capítulo ha sido explicada en el capítulo de Herramientas, y se menciona a continuación la importancia del uso de ésta en este capítulo en particular:

a) **Google Maps:** Esta herramienta se utiliza para mostrar los lugares extraídos de las obras literarias en un mapa. Al visualizar los lugares se puede corroborar visualmente que efectivamente los lugares existen o no en el mundo real.

7.2 Desarrollo

A continuación se va a presentar las diferentes actividades que se han realizado para alcanzar el objetivo definido en este capítulo.

7.2.1 Definición de los Pre Requisitos

Se ha establecido que los textos literarios deben estar digitalizados en archivos formato TXT.

7.2.2 Definición del objetivo

En la Ilustración 7-A se enumeran las actividades que este prototipo debe realizar para para poder cumplir, en su totalidad, el Objetivo General de este proyecto.

Ilustración 7-A: Actividades del Prototipo

7.2.3 Definición de Actores

El actor principal es la persona responsable de este trabajo de fin de carrera, porque la finalidad de este prototipo es validar el funcionamiento de su investigación.

7.2.4 Definición de la arquitectura

La arquitectura que se va a utilizar en este prototipo se muestra en la Ilustración 7-B. Se puede apreciar que se ha decidido utilizar 3 capas, las cuales corresponden al patrón Modelo Vista Controlador.

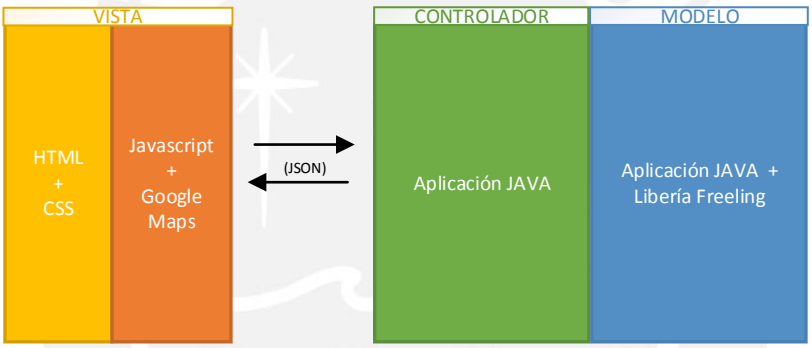

Ilustración 7-B: Arquitectura del Prototipo

7.2.5 Definición del Flujo Principal

En la Ilustración 7-C se puede apreciar una secuencia de actividades que representan completamente la funcionalidad del prototipo, se puede identificar que se ha integrado lo trabajado en los dos capítulos anteriores. Asimismo, cada actividad está representada de un color representativo a la capa que corresponden.

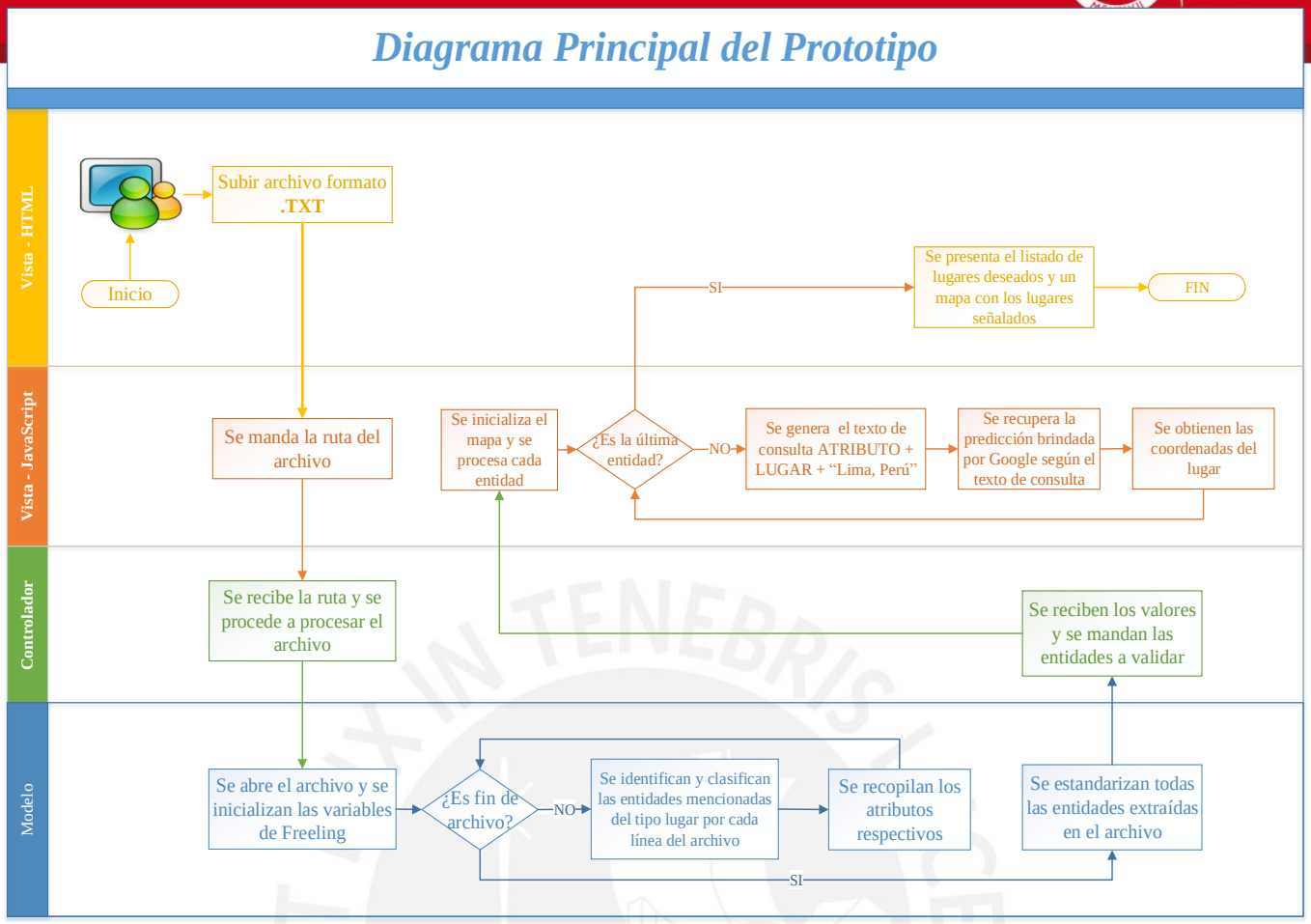

Ilustración 7-C: Diagrama Principal del Prototipo

7.2.6 Definición de las pantallas principales

A continuación se van a describir las dos pantallas principales de este prototipo:

a) **Ingreso de datos:** Esta pantalla inicial, el único botón que se muestra permite al usuario a seleccionar el texto que se desea procesar por este prototipo.

b) **Presentación de resultados:** Los nombres de los lugares extraídos de las obras literarias e identificadas en el mundo real son presentados en una lista. Se puede apreciar más detalle en la Ilustración 7-E. En esta pantalla el usuario podrá dar clic al nombre del lugar mostrará el lugar seleccionado en el mapa, esto se implementó ya que no todos los lugares iban a estar reunidos en un mismo cuadrante y de esta manera todos iban a poderse visualizar en el mapa.

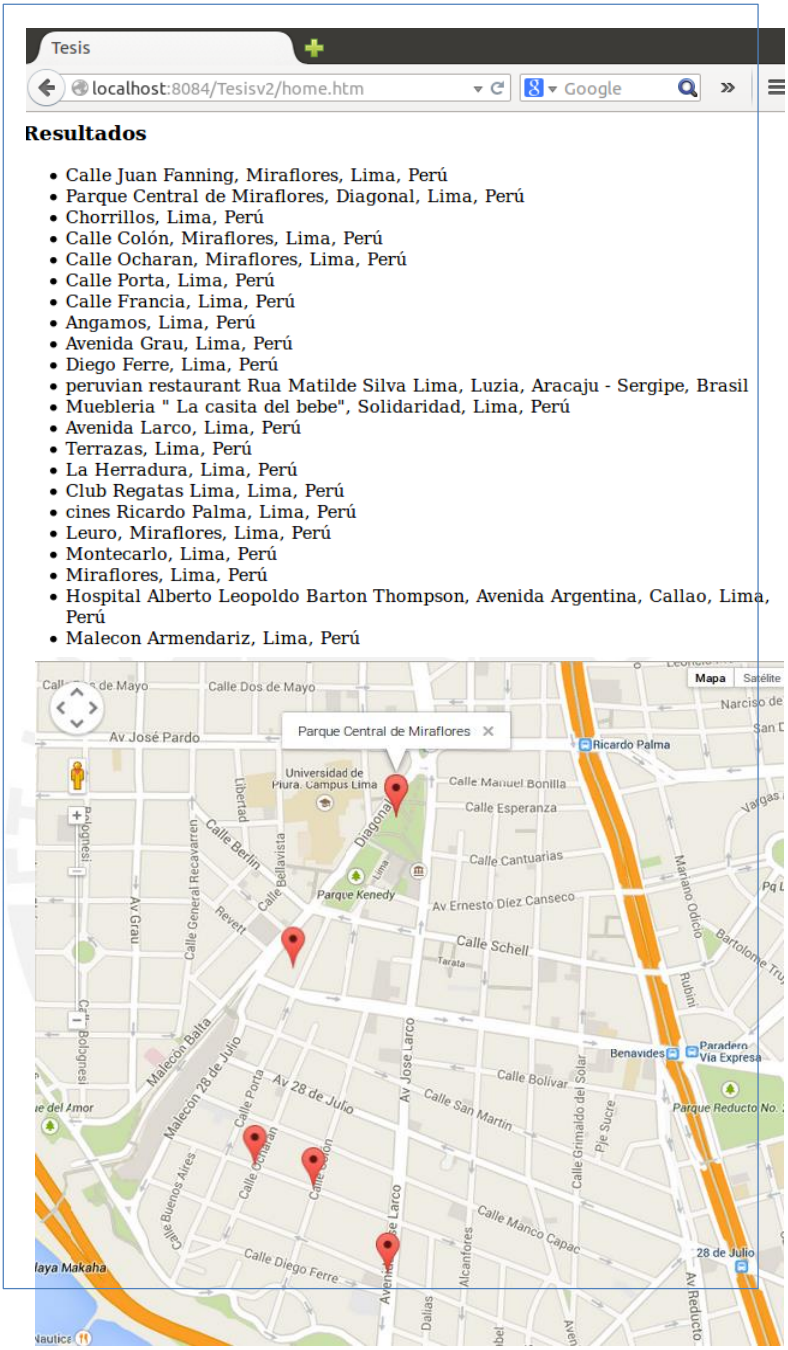

Ilustración 7-D: Pantalla de presentación de resultados

8. CONCLUSIONES FINALES

Este proyecto de fin de carrera tiene como objetivo general “Implementar un método de extracción e identificación de lugares del mundo real en textos en español del género literario” con la finalidad que la información brindada por este método se pueda utilizar para difundir el Turismo Literario y, de esta forma, rescatar el valor literario de los lugares mencionados en las obras.

Asimismo, se buscó explotar las tecnologías de procesamiento de información, porque de otra forma, para realizar lo deseado se tendría que invertir una gran cantidad de tiempo en analizar los libros, además que se necesita tener conocimiento previo de los lugares existentes en el mundo para poder determinar la autenticidad de estos en la realidad.

Entonces, se plantó dividir este trabajo en dos fases: (i) En la primera parte se van a identificar los lugares extraídos de los cuentos literarios en español un Reconocedor de Entidades Mencionadas (REM). (ii) En la parte complementaria, se procede a validar si los lugares extraídos anteriormente existen en el mundo real a través de una herramienta de información geográfica.

Para alcanzar objetivo de este proyecto, primero se procedió recolectar textos literarios con la finalidad de tener material base para corroborar que resultados de las diversas evaluaciones realizadas brindaban la información deseada. Además, existía la predisposición que iba a ser necesario que el REM sea orientado a la literatura – por lo investigado en la parte de herramientas. Estos textos recolectados fueron analizados meticulosamente y en el transcurso de esta actividad se pudo comprobar las siguientes dos afirmaciones (i) que para realizar lo que plantea este proyecto se debe invertir gran cantidad de tiempo y, también, (ii) que la forma de redacción literaria de generaba diversos problemas de ambigüedad y limitaciones de contexto lo cual iba afectar directamente a la identificación de lugares en el mundo real.

Después de analizar los textos y ordenar su información respectiva, ya todo estaba listo para proceder a evaluar la herramienta elegida para el REM. Cabe resaltar que utilizando la herramienta con su configuración inicial ya se obtenían más de la mitad de los de lugares deseados, sin embargo, los lugares no identificados generaban una pérdida de información importante para este proyecto. Es por eso que se decidió realizar 3 evaluaciones adicionales para poder mejorar dichos resultados y se comprobó que al especializar el REM a textos literarios este daba mejores resultados

(aumentó 20% en comparación de sus resultados por defecto). En esta etapa es importante mencionar que el REM elegido se puede utilizar para cualquier contexto en particular, pero el gran problema está en la disponibilidad de conjuntos de ejemplos suficientemente grandes y correctamente identificados (como los textos literarios usados en este proyecto) que permitan especializar a la herramienta.

Por otro lado, para la parte segunda parte, lo único que se iba a realizar –en un primer momento- era identificar los lugares extraídos en el mundo real, sin embargo si se utilizaba la herramienta geográfica escogida sin hacer ninguna modificación a los resultados del REM, ésta arrojaba valores muy lejanos a los que se deseaban obtener. Esto se debía, principalmente, a que el REM solo reconoce nombres propios de lugares y se concluyó que al tener sólo los nombres no era suficiente para determinar la existencia de los lugares en el mundo real. Por esta razón, se decidió que era necesario realizar las siguientes 3 consideraciones: (i) obtener información complementaria mediante la recopilación de atributos, (ii) estandarizar las entidades y (iii) limitar el territorio de búsqueda a Lima Metropolitana. Con estas mejoras la herramienta geográfica pudo identificar un 56% de los lugares reales deseados, éste valor resultó ser bastante alto porque los lugares que se dejaron de encontrar correspondían, en su mayoría, a entidades sin información adicional (incompletas).

En conclusión, se logró alcanzar el objetivo general de este proyecto planteado inicialmente y con el método resultante se va a poder procesar cualquier archivo de texto literario en menos de 10 minutos (dependiendo del tamaño) sin necesidad de tener conocimiento previo de los lugares existentes en el mundo real disminuyendo, así, los problemas planteados en este proyecto. Asimismo, se logró extraer 72% de lugares mencionados en los libros e identificar unos 56% de los lugares reales, unos valores bastante altos en comparación a los resultados por defecto, antes explicados.

8.1 Trabajo futuro

Como trabajo futuro se recomienda realizar las siguientes actividades para mejorar los resultados obtenidos en este trabajo:

a) Recolectar más del doble de textos literarios y dividir los ejemplos en categorías con la finalidad de cubrir la mayoría de casos posibles. Se ha comprobado en este trabajo que si se aumenta la cantidad de corpus, no necesariamente se mejora el indicador F-Measure del REM; pero al aumentar calidad al corpus, con ejemplos distintos, puede que este indicador si aumente.

b) Agregar una validación de lugares como un paso manual al final del procesamiento. En él se presenta la línea que donde se menciona el lugar en particular y las 4 mejores predicciones de lugares (brindada por GOOGLE) para que el usuario pueda elegir la que corresponde al lugar de acuerdo a lo que entiende de la oración brindada.

Zechner, Klaus.

1997. "A literature Survey on Information Extraction and Text Summarization." 1997, pág. 30.

Achitenei, María.

2006. "El realismo mágico. Conceptos, rasgos, principios y métodos". 2006, Babab , pág. 12.

Turmo, Ageno, Catala.

2006. "Adaptive Information Extraction". 2006, pág. 47.

Marrero, Mónica, y otros.

2009. "Evaluation of Named Entity Extraction Systems". 2009, A. Gelbukh (Ed.), pág. 12.

Fodor.

2013. "Fodor's Travel Guides". [En línea] 2013. [Citado el: 15/11/2013.] http://www.fodors.com/news/story_905.html.

Google.

2013. "API Google Places." [En línea] 2013. [Citado el: 15/11/2013.]

Tsai, Chia-Chi, y otros.

2012. "Hidden Trends in 90 Years of Harvard Business Review." 2012, Conference on Technologies and Applications of Artificial Intelligence, pág. 6.

Althobaiti, Maha, Kruschwitz, Udo y Poesio, Massimo.

2012.

"Identifying Named Entities on a University Intranet." 2012, Computer Science and Electronic Engineering Conference (CEEC), pág. 6.

Cunningham, Hamish.

1999.

"Information Extraction - A User Guide." 1999, pág. 18.

Abolhassani, Mohammad.

2003. "Information Extraction and Automatic Markup for XML documents." 2003, pág. 18.

LAN.

2013. LAN PERU. [En línea] 2013. [Citado el: 15/11/2013.] http://www.lan.com/es_pe/sitio_personas/index.html.

Limited, Evening Standard.

2013. Get London Reading. [En línea] 2013. [Citado el: 15/11/2013.] http://www.standard.co.uk/news/get-london-reading/.

LINGPIPE.

2013. LingPipe. [En línea] 2013. [Citado el: 10/11/2013.]

Municipalidad de Miraflores.

2013. LiteraTour. [En línea] 2013. http://www.miraflores.gob.pe/MVLL/index2.html.

LIU, XIAOHUA, y otros.

2013. "Named Entity Recognition for Tweets." 2013, ACM Transactions on Intelligent Systems and Technology, pág. 15.

PromPeru.

2013. ¿Y tú que planes? [En línea] 2013. [Citado el: 14/11/2013.] http://www.ytuqueplanes.com/.

PromPeru

2011. La Lima de Mario Vargas LLosa. [En línea] 2011. [Citado el: 15/11/2013.] http://media.peru.info/catalogo/Attach/vargas_10549.pdf.

PromPeru

2012. Perfil del Vacacionista Nacional 2012. [En línea] 2012. [Citado el: 12/11/2013.] http://intranet.promperu.gob.pe/IMPP/2012/TurismoInterno/Demanda%20Actual/Perfil%20del%20Vacacionista%20Nacional%202012/Tips%20P

VN%202012/PVN%202012%20Total%20y%20Ciudad%20de%20Residencia/TipsPVN2012porCiudades.pdf.

Quijote.es.

2013. Quijote. [En línea] 2013. [Citado el: 15/11/2013.] http://www.quijote.es/IVCentenario_RutaDonQuijote.php.

Orlando, Salvatore, Pizzolon, Francessco y Tolomei, Gabriel.

2013. "SEED: A Framework for Extracting Social Events from." 2013, World Wide Web Conference, pág. 9.

Hobbs, Jerry R.

1993. "The Generic Information Extraction System" 1993, pág. 6.

Tourism, World Travel.

2012. World Travel & Tourism. [En línea] 2012. [Citado el: 05/11/2013.] http://www.wttc.org/site_media/uploads/downloads/world2012.pdf.

Maynard, Bontcheva, Cunningham.

2003. Towardds a semantic extraction of named entities. 2003, pág. 7.

UNWTO.

2013. Tourism Highlights . [En línea] 2013. [Citado el: 05/11/2013.] http://dtxtq4w60xqpw.cloudfront.net/sites/all/files/pdf/unwto_highlights13_en_lr_0.pdf.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward.

2009. Natural Language Processing with Python.

PUSREJOVSKY James, Stubbs Amber

2012. Natural Language Annotation for Machine Learning

PADRÓ Lluis

2012. FreeLing 3.0: Towards Wider Multilinguality

FREELING

2013. FreeLing User Manual

GOOGLE

2013. Ayuda de Google Maps. [En línea] 2013 [Citado el: 15/11/2014.] http://support.google.com/maps/?hl=es