Procesamiento del lenguaje natural

Fernando Mata[1] y Jennifer Ledezma[2]
CI-1441 Paradigmas computacionales
Escuela de Ciencias de la Computación e Informática
Facultad de Ingeniería
Universidad de Costa Rica
[1]*ffmm.14@gmail.com*, [2]*jennylv17@gmail.com*

Julio de 2015

Resumen

El procesamiento del lenguaje natural (PLN), es aplicado actualmente en diferentes actividades como traducción automática, sistemas de recuperación de la información, entre otros. Sin embargo, aunque se han logrado avances de gran impacto para el ámbito de la inteligencia artificial, todavía existen fundamentos teóricos del PLN que se encuentran en estado de desarrollo. Existen muchos obstáculos por superar, entre estos, capturar la diversidad cultural que existe en el lenguaje natural para lograr que ordenador tenga flexibilidad de comunicación con el usuario. Es indispensable dar una solución inicial que atienda esta limitante mediante la implementación de redes semánticas que sean capaces de responder a interrogantes a través de análisis de conocimiento adquirido, es decir se debe previamente dotar al ordenador de información gramatical.

Palabras clave: lenguaje natural, semántica, análisis.

1 Introducción

Actualmente existe una gran cantidad de información digital almacenada en el lenguaje natural y una gran necesidad administrar dicha información de manera automática. La inteligencia artificial ha permitido que dicha gestión sea posible a través de sistemas computacionales. Sin embargo, esta gestión se realiza de manera estructural mediante formalismos, lo que genera grandes dificultades, dado a que los medios computacionales no pueden interpretar el conocimiento suministrado tal como lo hace un ser humano.

Dicha insuficiencia ha influenciado a la búsqueda de alternativas que minimicen aquellos errores que se ocasionan en los diferentes procesos de gestión del lenguaje natural.

Es por esta razón que la presente investigación involucra redes semánticas para lograr transformar el lenguaje natural en un formalismo que logre que la máquina trate de dar respuestas a interrogantes de la manera más manera más natural posible basados en oraciones con gramáticas definidas.

2 Marco teórico

2.1 Lenguaje natural controlado

Limita la gramática que el usuario pueda usar para hacer más sencilla a la hora de analizar. Se imponen reglas como no usar voz pasiva o evitar el uso de pronombres [1]. Tiene la ventaja de permitir analizar y modelar el conocimiento contenido a un texto de una forma más correcta. Sin embargo, su interpretación es muy limitada.

2.2 Análisis de sentimiento

Determina el estado de ánimo o actitud en el que se encontraba el autor en el momento que escribió sobre un tema, esta puede ser el estado emocional o la intención comunicativa emocional que desea transmitir al lector Realiza análisis de textos para obtener el significado subjetivo de un texto en específico, haciendo uso de diferentes técnicas del procesamiento del lenguaje natural, análisis de texto y lingüística computacional [2]. Tiene la ventaja de que permite dar

una noción de lo que trata el contenido del texto y clasificarlo en la alguna categoría de conocimiento, pero debido a que funciona por medio de aproximaciones estadísticas no es muy acertado en cuanto al análisis de contexto

2.3 Búsqueda de respuestas

Consiste en la recuperación de información a partir de uno o varios textos para obtener una respuestas preguntas que se le puedan hacer [3]. Posee la ventaja de obtener información de un texto por medio de preguntas que se le hagan a la máquina, sin embargo se ve limitado a interactuar solo tipo con preguntas predeterminadas y algoritmos de análisis muy complejos.

3 El problema

Durante varios años muchos científicos en el área de IA han dedicado parte de su esfuerzo en el área de reconocimiento y análisis del lenguaje natural, con el objetivo de extraer información de textos que le puedan brindar conocimiento extra a la máquina acerca de lo que está procesando. Este problema trae consigo muchas otras interrogantes, como "¿Qué se entiende por significado?" o "¿Puede una máquina realmente entender lo que el texto quiere transmitir?". No es conocido si se le puede dotar a una máquina de conciencia de tal modo que pueda entender y ver mundo del mismo modo que lo hace un humano, pero sí sabemos que se le puede programar para poder analizar información que pueda reinterpretar para obtener nueva información acerca de lo que el texto transmite. El presente trabajo ambiciona con dar una respuesta parcial a este problema por medio del estudio del funcionamiento de estas máquinas y aportando sugerencias de árboles de parsing para frases en español, identificando la transitividad de las oraciones e implementando reglas en el lenguaje Prolog que generen redes semánticas basadas en los árboles, y respondan a interrogantes de usuarios ante ciertas frases ya limitadas.

4 Objetivos y cronograma

4.1 Objetivo general

- Desarrollar un programa en Prolog que realice análisis del contenido de un texto.

4.2 Objetivos específicos

- Estudiar cómo emplea Prolog el procesamiento de lenguaje natural.
- Representar árboles que analicen las categorías gramaticales aplicadas a la descomposición de sus partes.
- Implementar árboles sintácticos o árboles de parsing que representen la estructura sintagmática de oraciones de la lengua española.
- Identificar la transitividad de frases para responder a diferentes preguntas mediante la implementación de redes semánticas.

4.3 Cronograma

- Investigación bibliográfica: 16/04/2015
  Recolección de información para la definición del proyecto.

- Consulta con el profesor: 20/04/2015
  Aclaración de dudas sobre la delimitación del alcance y los objetivos del proyecto

- Limitación del problema: 21/04/2015
  Delimitación de la bibliografía y definición final del tema a desarrollar, así como la herramienta a utilizar "Prolog".

- Definición del producto y conceptualización: 8/06/2015
  Definición de la meta general del programa y cada una de las partes que llevan al alcance del prototipo esperado.

- Representación e inferencia: 9/06/2015
  Definir cómo se van a utilizar cada las partes definidas en la conceptualización que dará estructura a cada una de las partes del árbol de parsing hasta llegar a formar una red semántica.

- Diseño de la solución: 14/06/2015
  Analizar y diseñar la solución que se desea del producto

- Presentación del avance: 18/06/2015
  Exposición de los objetivos logrados hasta la fecha, ante el profesor y los compañeros de la clase

- Corrección de los errores detectados en la presentación: 19/06/2015
  Basados en los comentarios de los compañeros y el profesor implementar mejoras en el proyecto.

- Casos de uso del prototipo: 22/06/2015
  Dar una representación gráfica que modele la estructura lógica que deberá tener el árbol de parsing, así como una especificación de cómo será utilizado el prototipo por los usuarios finales.

- Creación del prototipo: 22/06/2015
  Creación del prototipo en Prolog, implementando únicamente reconocimiento de oraciones que se van a utilizar: árboles de parsing

- Implementación de redes semánticas: 28/06/2015
  Como parte final del proyecto implementar redes semánticas para el análisis de la información contenida en un texto.

- Informe final: 30/06/2015
  Elaboración de documento formal que contenga todos los resultados y planes futuros.

5 Propuesta de solución

Mediante el paradigma lógico con técnicas de producción interpretada "Prolog", se pretende diseñar e implementar un prototipo que permita realizar procesamiento de lenguaje natural, siendo en este caso el vocabulario español. La idea es que mediante el uso de gramáticas libres de contexto y árboles de análisis sintáctico, se pueda definir la estructura correcta del lenguaje ya mencionado, caracterizando principalmente las partes que se consideren más importantes de una oración como por ejemplo: artículo, sujeto, verbo y predicado. A su vez es importante considerar el uso de gramáticas con cláusulas definidas que permitan verificar el sentido de la frase, es decir, que pueda determinar, por ejemplo, de manera correcta la oración "Los niños comen helado" y de forma incorrecta "Los comen helado". Es importante además que las gramáticas con cláusulas definidas tomen en cuenta la concordancia de género y número, restringiendo de este modo oraciones del tipo "Los niño come helado". Una vez teniendo definidos en el prototipo dichos términos o gramáticas es de suma importancia incorporar razonamiento de nuestro lenguaje natural, esto con el fin de lograr la meta propuesta del prototipo, que es emplear transitividad entre frases definidas, logrando que el programa sea capaz de responder o procesar interrogantes del usuario sobre una o varias oraciones ya definidas.

Primero se hará una pasada por el texto para hacer el análisis léxico de las diferentes palabras que conforman el corpus e identificar su categoría morfológica para los posteriores análisis. Después se construirán los árboles de análisis sintáctico a partir del análisis anterior, dividiendo el texto y agrupando las palabras en los sintagmas para el posterior análisis del contenido del texto.

Por último se plantea hacer uso de redes semánticas para modelar el conocimiento que se pueda extraer del texto en específico, por medio de la red posteriormente obtener información acerca del texto que se está analizando.

6 Conceptualización

- oracion*:* Regla de inicialización, conformada por una frase que puede ser descompuesta o concatenada por dos listas distintas del tipo sintagma nominal y sintagma verbal.
- sintagmaNominal: Constituyente sintáctico dada por la concatenación de cláusulas de tipo artículo y sujeto, donde el sujeto puede ser un sustantivo o pronombre.
- sintagmaVerbal: Regla que procesa listas de tipo verbo y predicado donde el predicado puede corresponder también un sujeto o pronombre.
- articulo: Categoría morfológica que brinda el conocimiento del género y número del sustantivo o pronombre.
- sustantivo: Sintagma que será participante de la predicación verbal, puede designar un objeto, animal o cosa abstracta
- nombre_personal: Sintagma que será participante de la predicación verbal, denota un nombre de persona
- verbo: Sintagma léxico que representa la acción realizada por los constituyentes del sintagma nominal.
- predicado: Constituyente del sintagma verbal que da descripción de la acción realizada por el sujeto, puede describir el estado en el que se encuentran las cosas en de los constituyentes de la oración. Puede ser de dos tipos: predicado directo o indirecto.
- predicado_directo: Básicamente hace un nuevo llamado o uso del sintagma nominal, para permitir oraciones de la forma: "Valery come un helado".

- predicado_indirecto: Hay dos subtipos de predicados indirectos para lograr concordancia en entre el verbo y las preposiciones de referencia y espacio, por ejemplo: “kevin y valery corren en la cocina”, “kevin y valery comen helado de fresa”.
- preposicionS: Hace énfasis a las preposiciones de referencia (de, con, en) en concordancia con el verbo.
- preposicionE: Hace énfasis a las preposiciones de espacio (desde, hacia, en) en concordancia con el verbo.
- agregarArco: Permite crear la red semántica basado en aristas de los nodos, donde los nodos son los elementos que se capturan de la oración ingresada por el usuario.
- agregarArcoPredicados: Estos tipos de aristas se agregan con base a la cláusula anterior, sin embargo toma en cuenta los tipos de predicados mencionados para que la red semántica también tenga concordancia con las preposiciones y el verbo.

Tabla 1. Relaciones conceptuales

	Artículo	Sujeto	Verbo	Predicado
Oración	✓	✓	✓	✓
Sintagma Nominal	✓	✓		
Sintagma Verbal			✓	✓

7 Representación e inferencia

La definición de gramáticas se implementará por medio de árboles sintácticos, como se ha mencionado a lo largo del documento, el objetivo general de estos árboles es dar conocimiento de las reglas del lenguaje natural al paradigma “Prolog”. Para definir la gramática se necesita inicialmente definir la estructura sintáctica que debe tener una oración o frase, para esto se utilizarán reglas que deben determinar las cláusulas que contemplan la oración, cabe destacar, que Prolog al ser un paradigma lógico basado en reglas de producción, no permite crear estructuras de árboles como en otros lenguajes, sino que se limita al uso de listas, dando como resultado que el árbol sea generado mediante la recopilación de listas de listas, de esta manera el lenguaje interpretará la oración como la concatenación de dos listas: sintagma nominal y sintagma verbal, que dicho sea de paso la primer lista es el enlace entre artículo y sujeto o sustantivo (según sea el caso) respetando las reglas de género, por otro lado el sintagma verbal será una lista formado por la unión de las lista de verbos y predicados. Las listas que conforman el sintagma nominal y verbal, serán el resultado o clasificación de las palabras que pueden conformar un artículo, un sujeto, un verbo o predicado.

La red semántica se genera posteriormente a la creación del árbol de parsing de donde se extraen los diferentes elementos que se utilizan para modelar la red. La red semántica es modelada por medio de un grafo que contiene en los nodos los verbos y los sustantivos que conforman los objetos de las oraciones y en los vértices diferentes atributos con las preposiciones que generan conexiones entre los diferentes objetos. Por lo que el siguiente par de frases se modelaría con una estructura similar propuesta en la figura 2: Los niños comen helado. Ana come helado con cuchara.

8 Casos de uso

Como resultados finales de la aplicación se espera que usuario tenga la opción de interactuar con la aplicación de dos maneras:

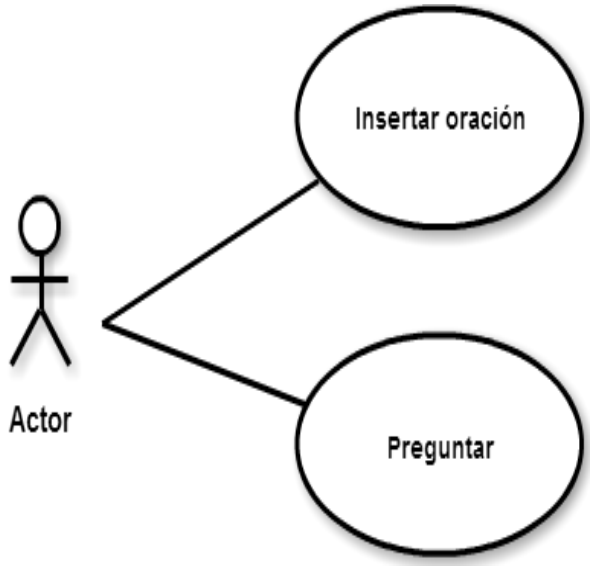

Figura 1. Casos de uso

- Insertar una oración: Prolog analizara si la oración es correcta según la gramática ya definida en el árbol de parsing creado, donde la frase insertada debe corresponder en su totalidad en género y número.
- Preguntar: el usuario podrá realizar preguntas limitadas con base a oraciones predefinidas, el tipo de preguntas serán en relación a la red semántica, tomando como "token" de la red al verbo de las gramáticas definidas.

9 Desarrollo, prueba y validación

9.1 Árboles de parsing

Es la representación de secuencia de derivaciones, por el cual, dada una frase, se debe descomponer en sus constituyentes sintácticos, es decir, los componentes de una frase se estructuran en diferentes categorías sintácticas. Posee nodos internos que representan los símbolos no terminales y hojas que constituyen los símbolos terminales. Basado al lenguaje natural estos símbolos terminales y no terminales, se encontrarán representados de acuerdo a la sintaxis composicional: sintagma nominal y sintagma verbal, que ya han sido descritos en la sección de la conceptualización. En la figura 2 se visualiza la estructura del árbol interpretando las clausulas del lenguaje Prolog, es por esta razón que no se hizo uso de tildes ni espacios entre palabras de un mismo nodo.

9.1.1 Validación de las gramáticas

La gramática del lenguaje natural será definida inicialmente por una notación de gramáticas que serán establecidas mediante un árbol de parsing con el paradigma Prolog, el objetivo primordial es establecer un conjunto de reglas de inferencia que ayuden a generar cláusulas o listas debidamente definidas por nombres y especificaciones.

Las gramáticas definidas como se ha mencionado a lo largo de esta investigación, deben obedecer a reglas de concordancia de número y género así como de verbo y preposición.

Según la gramática definida se deben permitir oraciones del tipo:

- Ana y Carlos comen.
- Ana y Kevin comen un helado rico.
- Kevin y Valery comen helado de fresa.
- Ana y Valery comen helado con cuchara.
- Carlos y Ana tienen una camisa linda.
- Kevin y Valery corren en la cocina.
- Kevin y Ana corren desde la cocina.
- Carlos y Valery caminan hacia el carro.

Mientras que no será permitido hacer uso de oraciones del siguiente tipo:

- Kevin y Valery come helado con cuchara: cómo podemos ver no hay concordancia de número entre los sujetos y el verbo.
- Carlos y Ana tienen un camisa linda: no existe concordancia en género entre el sustantivo y el artículo.
- Ana y Valery corren de la cocina: la preposición "de" ha sido definida del tipo referencia, y el verbo corren como un verbo que debe tener concordancia con preposiciones que hagan asociación con espacio o lugar.
- Kevin y Ana comen hacia la cocina: lo correcto es que la asociación entre el verbo y el predicado, y la familiaridad con el lenguaje natural sea: Kevin y Ana comen en la cocina.

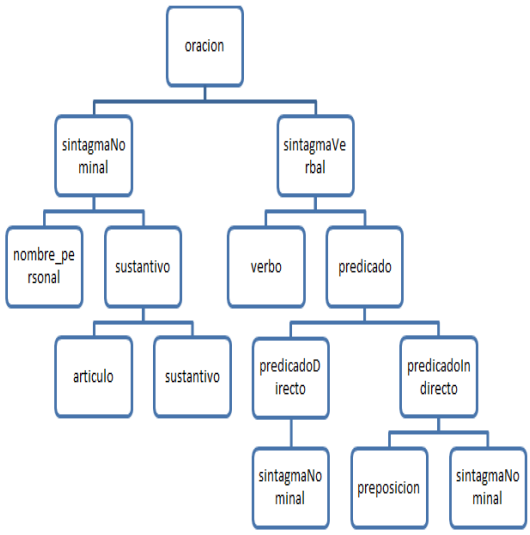

Figura 2. Árbol de parsing

Así mismo es importante mencionar la facilidad que da Prolog de definir gramáticas mediante el operador: "-->", que tiene una funcionalidad similar al popular ":-", su diferencia radica en la cantidad de elementos que puede tener una oración, es decir, se pueden crear oraciones recursivamente de diferentes tamaños sin necesidad de utilizar el operador "append", o crear

diferentes tipos de cláusulas basados en diferentes números de parámetros. Originalmente el operador: "-->", necesita de dos parámetros, donde para la definición común de gramáticas se incluye dentro del primer parámetro la oración, cada elemento separado por comas "," y el segundo y último parámetro quedara vacío como lo podemos ver a continuación:

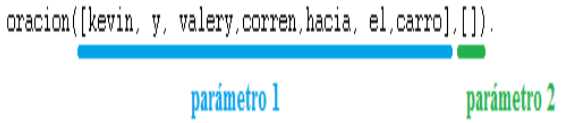

Figura 3. Detalles de ejecución

9.2 Red semántica y análisis de gramáticas

Mediante una red semántica se dará una muestra del razonamiento del lenguaje en la que los conceptos y sus interrelaciones tendrán como objetivo dar respuestas a interrogantes basados en una frase previamente definida con los árboles de parsing y mostrar la transitividad entre los elementos de una oración.

Figura 4. Redes semánticas

Para crear la red semántica se debe de hacer un análisis desde las hojas de árbol de parsing para identificar cada uno de los elementos que se van a agregar a la red. Después se debe de identificar la función de cada elemento en el texto para que sea mapeado en la respectiva red semántica o grafo. Por ejemplo:

"Kevin y Valery corren hacia la cocina linda" " se debe de mapear de la siguiente manera:

```
kevin-suj>corren
valery-suj>corren
corren-hacia>cocina
cocina-adj>linda
```

Mientras que la frase, "kevin come", debería verse en la red de la siguiente manera:

```
kevin-suj>come
come-obj>helado
helado-adj>rico
```

La base de preguntas debe ser creada desde la perspectiva de la red, donde sí se consulte quien come, pueda responder en base a los sujetos mapeados, o que hacen los sujetos, en este caso Kevin y Valery debería responder para la primera frase "corren" y por consiguiente para hacer preguntas espaciales, como, dónde corren Kevin y Valery o hacia dónde corren, debe de responder con la arista, hacia>cocina, desde el token corren, según lo mencionado en secciones anteriores.

Sin embargo, como parte de los resultados obtenidos, no se logró implementar que el usuario haga interrogantes basado en una oración previamente aceptada por el analizador de las gramáticas, ya que para esto se debe crear un tipo de base que guarde la oración recién insertada para que dé respuestas concisas, y no que por ejemplo al preguntar quién, devuelva todos los sujetos de las gramáticas definidas, esta forma perdería todo el sentido a la función de la transitividad de la red semántica.

10 Experimentación y análisis

Para la ejecución de las pruebas es necesario que el usuario haga uso de la cláusula principal: oracion, sin utilizar tildes ni mayúsculas.

Los resultados con gramáticas correctas mostraran la red semántica completa seguido de "true" como es definido propiamente en el lenguaje Prolog para los casos de unificaciones correctas, con casos de prueba validos (ver figura 5).

Pero por otro lado para los casos en que la oración insertada carezca de lógica o sintaxis gramatical, se mostraran resultados de la red hasta donde la definición dada fue correcta seguido de un "false" que cortara el mapeo de la oración en la red (ver figura 6).

```
17 ?- oracion([kevin, y, valery,corren,hacia, el,carro],[]).

kevin-suj>corren
valery-suj>corren
corren-hacia>carro
kevin-suj>corren
valery-suj>corren
true .

 18 ?- oracion([kevin, y, valery,tienen,una, camisa,linda],[]).

kevin-suj>tienen
valery-suj>tienen
tienen-obj>camisa
kevin-suj>tienen
valery-suj>tienen
camisa-adj>linda
tienen-obj>camisa
kevin-suj>tienen
valery-suj>tienen
true .

 19 ?- oracion([kevin, y, valery,corren,en, la,cocina],[]).

kevin-suj>corren
valery-suj>corren
corren-en>cocina
kevin-suj>corren
valery-suj>corren
true .

 20 ?- oracion([carlos, y, ana,caminan,hacia,el,carro],[]).

carlos-suj>caminan
ana-suj>caminan
caminan-hacia>carro
carlos-suj>caminan
ana-suj>caminan
true .

 21 ?- oracion([ana, camina,hacia,el,carro],[]).

ana-suj>camina
camina-hacia>carro
ana-suj>camina
true .
```

Figura 5. Ejecuciones correctas

```
 26 ?- oracion([kevin, y, valery,comen,helado,hacia, cocina],[]).

kevin-suj>comen
valery-suj>comen
comen-obj>helado
kevin-suj>comen
valery-suj>comen
false.

 27 ?- oracion([kevin, y, valery,corre,hacia, el,carro],[]).
false.

28 ?- oracion([kevin, y, valery,come,helado,en,la, cocina],[]).
false.

29 ?- oracion([kevin, y, valery,comen,helado,hacia,la, cocina],[]).

kevin-suj>comen
valery-suj>comen
comen-obj>helado
kevin-suj>comen
valery-suj>comen
false.

 30 ?- oracion([kevin, y, valery,comen,helado,en,la, cocina],[]).

kevin-suj>comen
valery-suj>comen
comen-obj>helado
kevin-suj>comen
valery-suj>comen
comen-obj>helado
comen-en>cocina
kevin-suj>comen
valery-suj>comen
true .

 31 ?- oracion([kevin, y, valery,come,helado,en,la, cocina],[]).
false.
```

Figura 6. Ejecuciones incorrectas

11 Problemas abiertos y problemas futuros

Una de las desventajas de esta aplicación en cuanto a la definición de gramáticas, es la limitación que existe actualmente en estas, como idea inicial tuvimos utilizar la herramienta "TreeTagger" adaptada en el lenguaje de programación Java, ya que brindaba la facilidad de hacer análisis sintáctico en diferentes idiomas, separando la oración en artículo, sujeto, verbo, predicados, etc. Sin embargo, se debía interconectar Java con Prolog y viceversa, y dado al poco tiempo restante entre el desarrollo y la entrega final de la investigación, se decidió hacer uso de nuestras propias gramáticas.

Otra de las desventajas es la ausencia de una interfaz gráfica más amigable, por ejemplo existen herramientas como XPCE/Prolog la cual permite crear una interfaz gráfica, u otra alternativa más sencilla como JPL, desde Java. Estas ideas se tuvieron presentes, sin embargo, la mala estimación en el tiempo y el desarrollo de las actividades del cronograma imposibilito llevar a cabo esta idea.

Por último como se mencionó en la sección de desarrollo, es necesario para lograr la transitividad de respuestas de oraciones, crear una base de conocimiento en tiempo de ejecución, que permita al usuario hacer preguntas en relación a la semántica generada desde la oración insertada.

12 Agradecimientos

Agradecemos la asesoría y colaboración del Dr. Alvaro de la Ossa Osegueda, profesor del curso, ya que sin su ayuda y sugerencias no hubiéramos alcanzado la mayoría de los objetivos de esta investigación, así como el aporte de bibliografía esencial para la definición de las gramáticas y delimitación de la investigación.

13 Bibliografía

[1] S. (2003). *Controlling Controlled English An Analysis of Several Controlled Language Rule Sets*. Consultado el 30 de Abril del 2015, de http://www.mt-archive.info/CLT-2003-Obrien.pdf

[2] Pang, B. & Lee, L. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques* Recuperado el 15 de Junio del 2015, de

http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

[3] L. Hirschman and R. Gaizauskas (2001). *Natural language question answering: the view from here. Natural Language Engineering*, 7, pp 275-300. doi:10.1017/S1351324901002807.

[4] Aho, A. (2008). *Compiladores: Principios, técnicas y herramientas.* México : Pearson.

[5] Bratko, I. Prolog Programming for Artificial Intelligence: Cap 21: "Language Processing with Grammar Rules" Pearson.