FACULTAD DE INGENIERÍA - UNIVERSIDAD DE LA REPÚBLICA

PROYECTO DE GRADO

ENEAS

Respuesta automática a preguntas causales

Informe Final

19 de Julio de 2008

Sebastián Calvo
Ariel Guevara
Paula Imbriani

Tutores:
Diego Garat
Guillermo Moncecchi

Resumen

Los sistemas de búsqueda de respuestas intentan responder automáticamente a las preguntas realizadas por los usuarios en lenguaje natural. En el caso de este proyecto se aborda el problema de búsqueda de respuestas para el caso de preguntas causales en el idioma español. Éstas son las que interrogan acerca de la causa de un fenómeno, por ejemplo *¿Por qué se desató la primera guerra mundial?* o *¿ Cuáles son las causas de la inflación?*, etc.

Como corpus de documentos se utiliza Internet, y para obtener los mismos se usa el motor de búsqueda Google, por lo tanto, los documentos a analizar serán los retornados por éste. Para recuperarlos se definen diferentes reformulaciones de la pregunta recibida. Luego, a partir de los documentos devueltos por el buscador, se seleccionan segmentos candidatos a contener la respuesta.

Una vez seleccionados los segmentos de texto que podrían contener una respuesta, éstos se procesan con el objetivo de asignarles un puntaje. Éste procesamiento incluye encontrar las palabras clave de la pregunta e indicadores causales. El puntaje otorgado al segmento depende del tipo de indicador causal encontrado y del sentido de la causalidad del segmento.

Las respuestas obtenidas se presentan al usuario mediante una interfaz Web, con la característica destacable de que el sistema saca provecho de la comunicación asíncrona entre el navegador Web y el servidor, permitiendo que el usuario pueda obtener las respuestas a medida que se van generando sin tener que esperar a que finalice todo el procesamiento.

Finalmente, se realiza el ajuste y la evaluación del sistema. Con el ajuste se ve cuáles valores son los más adecuados para los parámetros cantidad de resultados de Google, que indica cuantos documentos retornados por Google se van a analizar, y cuáles son los mejores indicadores. Luego se realiza la evaluación general con preguntas causales que no se restringen a ningún domino específico, y se comparan los resultados con los de otro sistema. Se concluye que los resultados obtenidos son aceptables, ya que en general se responden las preguntas planteadas.

Palabras clave: Causalidad, Búsqueda de respuestas, Recuperación de información, Procesamiento de lenguaje natural.

Índice general

Capítulo 1

Introducción

Actualmente, la Web es el mayor repositorio de información existente, por lo que, a la hora de obtener información, es una de las principales fuentes a consultar. La forma más común de obtener documentos de la Web es utilizar un buscador, el objetivo de éstos es retornar documentos relacionados con las palabras clave ingresadas por el usuario a la hora de efectuar la búsqueda. Esto hace que el éxito obtenido por el usuario dependa de que ingrese las palabras clave correctas. Luego de obtenidos los documentos el usuario los debe consultar hasta encontrar lo que desee.

De forma de mejorar el acceso a la información se han desarrollado técnicas de recuperación de documentos [Gon03], cuyo objetivo es proveer a los usuarios de documentos que sean relevantes a la información deseada. Sin embargo hay casos en el que un usuario no desea obtener toda la información posible sobre un tema, sino que lo que quiere es obtener una respuesta concreta a una pregunta.

En la disciplina computacional conocida como Pregunta & Respuesta [Gon03], (abreviación en ingles: Q&A) se intenta responder de forma automática a preguntas formuladas en lenguaje natural. El objetivo es interpretar la pregunta correctamente, y así obtener la información necesaria, para luego procesarla y generar una respuesta que el usuario entienda y al mismo tiempo responda de forma concreta a la pregunta que ha realizado. De esta forma el usuario se evita tener que revisar los documentos uno por uno hasta encontrar las respuestas.

Como es sabido, se pueden hacer muchas preguntas, de diferentes tipos. Uno de los tipos de preguntas que interesa responder son las preguntas fácticas, de

éstas se espera una única respuesta concreta, y además tienen un pronombre interrogativo claro (cuándo, dónde, quién, etc.). Un ejemplo es "¿En qué ciudad está la Torre Eiffel?", donde se espera como respuesta "París". Este tipo de preguntas fueron objeto de análisis del proyecto de grado WebQA [CIM07].

Otro tipo de preguntas más complejas son las causales, en este caso no se cumple ninguna de las dos características mencionadas anteriormente. Puede haber más de una respuesta correcta, ya que la correctitud puede depender del contexto de quien informa la causa, incluso la información puede ser una opinión subjetiva. Tampoco tiene por que ser concreta, un evento puede tener varias causas, incluso un evento detectado como causal puede tener otras causas y la unión de todas éstas podría considerarse como una respuesta completa a la pregunta. Esto trae aparejado que es difícil saber donde comienza y termina una respuesta causal, dado que hay que averiguar cuando en un texto se hace referencia al tema planteado, y cuando se deja de "hablar" de este.

Incluso puede haber mención al tema del que se refriere la pregunta, pero sin que haya relación con la respuesta esperada. Además puede haber respuestas en que no haya una mención directa del tema, sino que éste se desprende del contexto. De esto se observa que se requiere cierta elaboración a la hora de encontrar automáticamente una respuesta a una cuestión causal.

Un ejemplo de éste tipo de preguntas es:

¿Por qué el cielo es azul?

Donde una respuesta posible podría ser:

El rayo violeta es el que se ha separado mas de la dirección del rayo blanco y ahí esta precisamente la explicación del color del cielo. La desviación es máxima para los rayos de longitud de onda corta (violeta y azul), y mínima para los de longitud de onda larga (amarillos y rojos), que casi no son desviados. Los rayos violetas y azules, una vez desviados, chocan con otras partículas de aire y nuevamente varían su trayectoria, y así sucesivamente: realizan, pues, una danza en zigzag en el seno del aire antes de alcanzar el suelo terrestre. Cuando, al fin, llegan a nuestros ojos, no parecen venir directamente del Sol, sino que nos llegan de todas las regiones del cielo, como en forma de fina lluvia. De ahí que el cielo nos parezca azul.

Las preguntas causales [Jac99] son de especial interés por el hecho de que la

causalidad constituye un medio privilegiado para acceder al contenido semántico de los textos. Las relaciones causales tienen en particular, que permiten organizar entre ellas los conocimientos volviéndolos inteligibles. Como estas relaciones son necesarias, tanto para la comprensión de los fenómenos como para la acción, su presencia en los textos es significativa.

A nivel práctico, la búsqueda de la información causal en textos puede ser justificada por necesidades varias. Por ejemplo:

- ayudar a los abogados a buscar en corpus de la jurisprudencia las explicaciones otorgadas por el tribunal sobre una sentencia dictada;
- buscar las causas de un problema en las actas de incidentes de una central nuclear;
- buscar causas posibles de una enfermedad emergente (por ejemplo, la enfermedad de la vaca loca);
- conocer las causas de un fenómeno económico (por ejemplo: la quiebra de un gran banco);
- conocer los efectos nocivos de un medicamento administrado accidentalmente a una mujer embarazada;
- etc.

Es necesario destacar, muy especialmente, la importancia del conocimiento causal en los ámbitos científico, técnico y estratégico, donde el conocimiento de relaciones de dependencia entre elementos del ámbito supervisado puede fundamentar tomas de decisiones estratégicas importantes.

Como paso previo a este trabajo se intentó ver ejemplos de otros sistemas que hayan abordado el tema anteriormente. Pese a que se encontraron algunos trabajos sobre causalidad para el idioma inglés, se tuvo mayor acceso a estudios desarrollados para el idioma francés. Dos sistemas resultantes de estos estudios y que vale la pena destacar son COATIS[Gar98] y SAFIR[Jac99], los mismos analizan la causalidad en textos del idioma francés, pero no intentan responder preguntas automáticamente. Ambos fueron desarrollados por el grupo LaLIC (Langages, Logiques, Informatique, Cognition - Université Paris Sorbonne), y utilizan el método de Exploración de Contexto, desarrollado por el mismo grupo.

También se vio que algunos sistemas analizaban textos de forma independiente del dominio, pero sin embargo elegían textos que desarrollaban explícitamente una problemática causal. Textos de este tipo pueden ser, estudios científicos, que intentan explicar la causa de algo, o textos periodísticos, que analizan las causas de los hechos que se informan. Otra característica destacable es que el análisis tomaba en cuenta todo el texto.

En nuestro caso no solo se analizan textos de cualquier dominio, sino que se analizan todo tipo de textos sin restringirse a textos que hablen explícitamente de causalidad. También se toma en cuenta todo el texto, pero luego de detectar segmentos candidatos se analiza en profundidad cada segmento. Esto se diferencia con lo realizado en el proyecto [CIM07], que analizaba solo el texto de previsualización que muestra Google en los resultados de su búsqueda, sin tener en cuenta el texto entero.

1.1. Objetivo

El principal objetivo de este trabajo es el desarrollo de un sistema de búsqueda de respuestas a preguntas causales en español. El mismo debe recibir como entrada una pregunta causal en lenguaje natural (*¿Por qué...?*, *¿ Cuáles son las causas de...?*, etc), y dar como salida una o varias respuestas para la misma. Estas respuestas se le deben presentar al usuario a través de una interfaz Web.

En Internet se pueden encontrar textos muy variados, formales o informales, con faltas de ortografía, con información falsa, artículos científicos, o literatura fantástica. Por lo que otro de los objetivos planteados consiste en evaluar si los textos encontrados en la misma, pueden ser un corpus útil a la hora de responder preguntas causales. Otra de las características de Internet, que se va a intentar explotar es la gran cantidad de información que contiene y la redundancia de información que hay ella, característica aprovechada con resultados positivos por el proyecto WebQA [CIM07] para el caso de las preguntas fácticas.

Otro tema importante es la velocidad de respuesta. Se intenta no solo desarrollar un sistema que encuentre correctamente respuestas a preguntas causales, sino que además se pretende que el análisis planteado se ejecute lo suficientemente rápido ante cada pregunta, como para que un usuario pueda interactuar con el mismo. El objetivo es que un usuario ingrese preguntas y obtenga respuestas, de la misma forma que un usuario interactúa actualmente con buscadores Web.

1.2. Contenido

El documento se divide de la siguiente manera:

El capítulo 2 presenta un marco teórico de los sistemas de búsqueda de respuestas, y se ve también una breve reseña de cómo se lidia con la causalidad en los textos.

En el capítulo 3 se describe el sistema desarrollado Eneas, detallando los algoritmos y técnicas utilizadas, explicando cada etapa que compone el proceso desde recibir una pregunta hasta retornar una respuesta al usuario.

En el capítulo 4 se describen puntos a destacar de la implementación del sistema. Se mencionan los aspectos técnicos más importantes de la solución.

El capítulo 5 trata de la evaluación del sistema. De aquí se pueden determinar los resultados obtenidos a partir de las técnicas utilizadas. Se presentan las medidas de evaluación tomadas, y cómo afectan los diferentes parámetros al rendimiento del sistema.

El documento finaliza con el capítulo 6, el cual contiene las conclusiones finales y los posibles trabajos a futuro que se podrían realizar, de modo de solucionar problemas detectados y mejorar el rendimiento general del sistema.

Capítulo 2

Marco Teórico

En este capítulo vamos a ver los conceptos en los cuales se basa nuestro trabajo. Podemos ubicar este trabajo en los sistemas Q&A, que se especializan en causalidad. Por lo que, por una parte vamos a presentar que ideas tomamos en cuenta a la hora de trabajar con la causalidad en los textos, cómo se presenta, qué características tiene, etc. Y también vamos a ver la arquitectura de los sistemas Q&A, en los cuales nos basamos para construir nuestro sistema.

2.1. Introducción

Existen recientes estudios lingüísticos, en particular en Francia, sobre el concepto de *causalidad*. Se puede observar que abordan el problema según perspectivas muy diferentes y complementarias [Par06]. Una contempla esencialmente las articulaciones (más bien argumentativas) de los conectores como *porque*, *puesto que*, *ya que*, etc. Todo indica que los conectores son relativamente poco frecuentes, y que las relaciones causales son frecuentemente indicadas por verbos como *conducir*, *causar*, *crear*, *implicar*, *conservar*, *favorecer*, *modificar*, etc. Otra perspectiva se concentra en la semántica verbal de algunos verbos que expresan una relación causal. Se puede citar la tesis de D. García [Gar98] y la de A. Jackiewicz [Jac99]que analizan algunos de los verbos anteriores en un enfoque de captación de un conocimiento técnico.

¿Qué es la causalidad?

A. Jackiewicz [Jac99] intenta responder esta pregunta analizando el concepto en distintas disciplinas. Por ejemplo, en Lingüística el concepto de causalidad puede abarcar varias categorías gramaticales y léxicas, su relación con la argumentación y la temporalidad vuelven este concepto indispensable en los análisis semánticos y pragmáticos de las producciones lingüísticas.

Pese a no tener una definición precisa, la causalidad se considera como una relación entre dos situaciones (hechos, entidades), o sea, describe los vínculos existentes entre las causas y los efectos. Esta manera de ver el concepto permite realizar el estudio de la causalidad no solo teniendo en cuenta los verbos causales sino también las distintas locuciones verbales.

La causa de un evento puede ser explicada de muchas formas y se caracteriza por la subjetividad y por depender estrechamente del contexto.

Veamos finalmente a varios investigadores que trabajan en el ámbito del tratamiento automático de las lenguas, y se interesan muy de cerca por el concepto de causalidad. La generación de textos: Laurence Danlos ([Dan85], [Dan88], [Dan95]). La generación de explicaciones causales: Farid Cerbah ([Cer92]). La adquisición de los conocimientos dirigida por un modelo conceptual causal: Jean Charlet ([Cha93]) y Chantal Reynaud ([Rey93]). En un sistema de preguntas y respuestas en lenguaje natural, responder automáticamente a cuestiones causales del tipo por qué...? : Adeline Nazarenko ([Naz94]). La descripción de los conectores en francés y en polaco en la perspectiva de traducción automática de textos: Renata Kozlowska-Heuchin ([KH96]).

2.2. Estudio de la Causalidad

Como se verá más adelante, para estudiar la causalidad en los textos nos basamos principalmente en la idea de los marcadores causales presentados por Agata Jackiewicz [Jac99]para el idioma francés.

Se presentarán los enfoques posibles que describe Jackiewicz [Jac99] para abordar la causalidad. Es de estos enfoques justamente de dónde surgen los diferentes indicadores causales. Los mismos se pueden agrupar de acuerdo al enfoque

al que pertenecen, y también a su "grado" de causalidad, es decir que tan bueno es el indicador para marcar una relación causal.

En el presente trabajo utilizamos los indicadores causales, pero no discriminamos por enfoque, porque no nos pareció necesario, sino que jerarquizamos cada indicador según su "grado" de causalidad. Cabe destacar que el valor de cada indicador en esta jerarquía fue definido por nosotros, independientemente de la tesis de Jackiewicz.

Los indicadores utilizados no fueron exactamente los que propone Jackiewicz para el francés, si no que se revisó la lista y se seleccionaron los que se adaptaron más al idioma español. Luego, a medida que se fue desarrollando el sistema se fueron ajustando las categorías de los indicadores, por ejemplo, en un principio los indicadores que se consideraron más causales fueron los más evidentes para el español como *"causa"* y *"porque"*, luego surgieron otros como *"provoca"*. Los textos utilizados para estudiar la causalidad en español fueron los documentos retornados por Google en las diferentes etapas del trabajo. Esto hace que los textos analizados sean tan variados como los que utiliza el sistema final.

2.2.1. Enfoques posibles al abordar la causalidad

Existen diferentes enfoques para abordar la causalidad. Entre los cuales se destacan el *enfoque cualitativo, el funcional, el analítico y el sintético*.

El *enfoque cualitativo* es el que corresponde a la acepción clásica de la causalidad. Intuitiva en términos de la lengua ordinaria, pone a continuación de una causa el efecto que la produce, y posee un elevado poder explicativo.

Además hay otras dos maneras diferentes de describir hechos causales. Una es presentando el proceso que conduce a la producción de un efecto (llamada *causalidad eficiente*) y la otra se centra en la existencia de una relación entre dos situaciones (llamada *relación causal*). Esta distinción no se refiere directamente a la relación causa-efecto, sino a la manera de comprenderla y de presentarla.

Existen dos tipos de verbos para expresar la acción causal eficiente. Los pertenecientes al primer conjunto son llamados *verbos de relación* y presentan la acción causal eficiente conforme a su orientación natural, es decir como yendo de la causa al efecto. Los verbos del segundo conjunto son llamados *verbos que precisan el efecto producido,* ya que se centran en la explicación del efecto. Estos

verbos son más ambiguos que los pertenecientes al primero, por lo tanto es indispensable el análisis del contexto en las frases en las que aparecen para determinar la causalidad .

En los cuadros 2.1y 2.2se muestran algunos ejemplos de los verbos descritos anteriormente.

Enfoque cualitativo
Verbos de relación
causar, provocar, desencadenar, implicar,
suscitar, generar, inducir, determinar, ocasionar,
nacer de, emerger de, surgir de, venir de,
llevar a, conducir a, conseguir que, desembocar en,
forzar a, obligar a, incitar a, invitar.

Cuadro 2.1: Indicadores de la acción causal eficiente.Verbos de relación

Enfoque cualitativo
Verbos que precisan el efecto producido
crear, producir, fabricar, mantener, sostener,
abastecer, alterar, afectar, obligar, perturbar,
favorecer, facilitar, apoyar, asistir, fomentar,
acentuar, afectar, consolidar, aclarar, calmar,
ayudar, estimular, controlar, reforzar.

Cuadro 2.2: Indicadores de la acción causal eficiente.Verbos que precisan el efecto producido.

A continuación presentaremos ejemplos de frases donde aparecen algunos indicadores de causalidad eficiente.

- *La modificación de las retenciones móviles para las exportaciones* ***suscitaron*** *un nuevo conflicto en el campo.*
- *Las autoridades ven con preocupación la cantidad de medios que* ***fomentan*** *la anorexia entre las jóvenes.*

El *enfoque funcional* expresa la causalidad tal como es definida por las herramientas de la ciencia. Su función es más predecir que explicar.

Enfoque Funcional
convertir, indicar, función, existe,
dependencia, interdependencia, correlación,
evolución, establecer, descubrir.

Cuadro 2.3: Indicadores de la causalidad funcional

El siguiente cuadro muestra algunos de los indicadores de la causalidad funcional:

El siguiente es un ejemplo de una frase que expresa un vínculo funcional:

*Los precios se establecen en **función** de la oferta y la demanda; cuando la demanda **aumenta**, o la oferta **disminuye** los precios se disparan.*

En el ejemplo anterior se establece que existe una relación (*función*) entre la oferta, la demanda y los precios, y luego se indican las características de ésta (*demanda aumenta, o la oferta disminuye los precios se disparan*).

El *enfoque analítico* se refiere a cuando un efecto no es producido directamente por una única causa. En el se distinguen dos casos, *la contribución causal y la influencia causal*.

Contribuir es tener parte en un resultado que viene determinado por varios factores. Así pues, si la situación X contribuyó a la situación Y, no es posible decir que X causó Y. La idea de contribución causal se basa en el reconocimiento de una multiplicidad de factores que participan simultáneamente en una situación causal. Por ejemplo:

*La alianza entre Hindenburg y Hitler **contribuyó** fuertemente a la toma de poder por parte de los nacional socialistas en 1933.*

En el ejemplo se dice que la alianza entre Hindenburg y Hitler ***contribuyó*** al ascenso de los nazis al poder, pero el uso del verbo **contribuir** nos indica que pudieron haber más causas para este hecho.

En el caso de la influencia causal, X puede influir en Y sin ser Y el objetivo de X, en este caso puede suponerse que no vale la pena considerar esta relación como

causal. Pero cuando se trata de explicar Y, todas las influencias sobre Y se vuelven importantes. La entidad que influye sobre un hecho no puede nunca tomarse como la causa, pero si como la causa del resultado de la influencia ejercida sobre este hecho. Por ejemplo:

La separación de los padres puede tener una ***influencia*** *negativa sobre las personas a la hora de formar una familia.*

En el ejemplo anterior se dice que la separación de los padres *influye* en las relaciones futuras de los niños, aunque es claro que éste no es el objetivo de la misma.

En los cuadros 2.4y 2.5se muestran algunos indicadores causales de este enfoque.

Enfoque analítico
Contribución causal
contribuir, participar, tomar parte,
intervenir, desempeñar,
indicar, servir para.

Cuadro 2.4: Indicadores de causalidad analítica. Contribución causal

Enfoque analítico
Influencia causal
influir, actuar, ejercer, pesar,
sensible a, incidir, impactar,
consecuencia.

Cuadro 2.5: Indicadores de causalidad analítica. Influencia causal.

A veces es la contradicción aparente entre dos hechos que coexisten lo que lleva a estudiar una relación que pueda unirlos. Este tipo de vinculo es llamado *sintético*.

El *enfoque sintético* pretende dar cuenta de las dependencias complejas que pueden existir entre los fenómenos. El concepto de la causalidad no puede reducirse al proceso efectivo de producción causal. Por ejemplo:

Ciertas Investigaciones demuestran que ***existe una relación entre*** *el envejecimiento prematuro y el estrés.*

En el ejemplo, no se dice cómo se pasa de un fenómeno A (*estrés*) a un fenómeno B (*envejecimiento prematuro*). Esta capacidad para dar cuenta de un vínculo aproximado, a veces indeterminado, lejos de ser un defecto del lenguaje natural, es una virtud, ya que está expresado de forma que se percibe.

El cuadro 2.6 muestra algunos indicadores de éste enfoque causal.

Enfoque sintético
existe vínculo, existe relación,
va junto a, se aproxima a,
esta conectado con, existe una correspondencia

Cuadro 2.6: Indicadores de causalidad sintética

2.2.2. Secuencias textuales causales

La pregunta ahora es: ¿cómo explotar los conjuntos de indicadores causales, así como los valores semánticos de estos elementos, para construir un procedimiento automático de localización de declaraciones causales en los textos?

Las secuencias textuales causales, son las formas en que se redactan los textos, cuándo éstos abordan un tema causal, Jackiewicz [Jac99] describe once tipos de estas secuencias: *secuencia argumentativa*, *reformulación directa*, *explicación causal citada*, *mecanismo causal*, *explicación por un modelo*, *respuesta a una cuestión causal*, *enumeración de efectos*, *secuencia que lleva un título causal*, *secuencia que resume*, *secuencia de opinión* y *secuencia del campo temático.*

En las *secuencias de opinión* se presenta la explicación causal como dependiente de un punto de vista particular. Se caracteriza por *"introductores"* como según X, de acuerdo con X, etc. Por ejemplo:

Según *organizaciones religiosas radicales, la verdadera causa del sida se debe a la perdida de los valores cristianos en el mundo.*

En el ejemplo se da un punto de vista particular (*organizaciones religiosas*

radicales), sobre un tema (*las causas del sida*), aclararlo, también es una forma de distanciarse de estas afirmaciones.

Por otro lado, en la *enumeración de efectos* se enumeran y aclaran los efectos de un fenómeno. La primer declaración de la secuencia debe contener un marcador que anuncia explícitamente la enumeración de los efectos (por ejemplo: tener (uno, dos, varios) efectos). La enumeración puede utilizar las marcas clásicas de ésta (guiones, etc). Puede también extenderse sobre varios apartados y usar marcadores de integración lineal (en primer lugar, en segundo lugar, etc). Por ejemplo:

La reforma de la seguridad social tiene ***dos efectos*** *importantes a corto plazo. En* ***primer lugar*** *se da atención a los niños que estaban fuera del sistema, y en* ***segundo lugar*** *los mayores de 60 años pueden ser atendidos sin pagar tickets de ningún tipo.*

En el ejemplo se dice que un hecho tuvo *dos efectos*, y luego se los enumera (*en primer lugar, en segundo lugar*).

Nosotros optamos por no trabajar directamente con estas secuencias, debido a que el objetivo de nuestro trabajo no es obtener todas las relaciones causales que aparecen en un texto, sino las que se refieren a un tema especifico (la pregunta del usuario). Por lo que preferimos ubicar el tema de la pregunta en el texto, y con los marcadores ver si se esta refiriendo a una cuestión causal.

Por otro lado hacemos uso en forma explícita de *secuencia que lleva un título causal* en la formulación especial de Wikipedia 3.5.3. Para más detalles sobre las secuencias textuales causales ver el trabajo de Jackiewicz [Jac99].

Secuencia que lleva un título causal

El texto que pertenece a una sección cuyo título incluye un término explícitamente causal (como causa, efecto, origen, consecuencia, etc.) tiene grandes chances de expresar causalidad. A continuación se da un ejemplo:

Efectos potenciales del calentamiento global. *Existen numerosos efectos del calentamiento global que afectarían el medio ambiente y a la vida humana. El principal es el incremento progresivo de la temperatura promedio. A partir de este, surgen una serie de diferentes efectos como el aumento del nivel del mar, cambios en los ecosistemas agrícolas, expansión de las enfermedades tropicales,*

aumento de la intensidad de los fenómenos naturales, etc.

2.2.3. Causalidad en el idioma ingles

Ahora vamos a ver otros enfoques utilizados para abordar la causalidad ([Gir03], [KCN]). Éstos utilizan patrones para indicar la causalidad, además también hacen uso de WordNet, la validación cruzada[Val07], y otras técnicas interesantes. Aunque los trabajos citados fueron realizados para el idioma ingles, nos parece interesante verlos debido a que las técnicas utilizadas podrían adaptarse al español.

Roxana Girju [Gir03] utiliza la relación semántica *CAUSE-TO* que es usada explícitamente en WordNet. El objetivo es encontrar un conjunto de patrones que indiquen causalidad, de forma que detectar relaciones causales.

Propone el siguiente método: buscar todos los patrones <FN1 verbo_causal FN2> (FN = Frase Nominal) que ocurren entre un sustantivo y otro sustantivo en la definición correspondiente. Un ejemplo de esto es la relación causal entre escuálido y hambre. La definición de *escuálido* es flaqueza extrema usualmente causada por el hambre o la enfermedad. WordNet 1.7 contiene 429 relaciones enlazando sustantivos de diferentes dominios, los más frecuentes de medicina (aproximadamente 58.28 %). Para cada par de sustantivos, buscar en Internet o cualquier otra colección de documentos y retener solo las sentencias conteniendo el par. De esas sentencias, determinar automáticamente todos los patrones <FN1 verbo_causal/expresión_causal FN2> El resultado es una lista de verbos/expresiones verbales que refieren a la causalidad. Por último se utiliza un método de validación cruzada de diez vueltas (*10-fold cross validation* [Val07]).

Para evaluar los resultados se definen las siguientes medidas precisión y recuperación (*recall*):

$$precisión = \frac{Numero\_de\_relaciones\_correctas\_recuperadas}{Numero\_de\_relaciones\_recuperadas}$$

$$recuperación = \frac{Numero\_de\_relaciones\_correctas\_recuperadas}{Numero\_de\_relaciones\_correctas}$$

El sistema retornó 138 relaciones, de las cuales 102 eran causales y 36 no causales, lo que implica una precisión de 73.91 % y recuperación de 88.69 %. Sin

embargo, se encontraron 38 relaciones causales expresadas por patrones no considerados, llevando a un cubrimiento (recuperación) del 66.6 %. Los errores se explican en gran parte por el hecho de que el patrón causal es muy ambiguo. El patrón léxico-sintáctico codifica muchas relaciones que son difíciles de desambiguar basándose solo en la lista de conectores. Los errores también fueron causados por el parseo incorrecto de frases nominales, el uso de reglas con poca precisión y la falta de reconocimiento de entidades con nombre en WordNet (ej: nombres de personas, lugares, etc). Para este experimento se usó un corpus bastante pequeño de solo 6523 ejemplos.

En el trabajo de Syin Chan, Christopher S.G. Khoo, and Yun Niu ([KCN]) se construye un conjunto de patrones gráficos que indican la presencia de una relación causal en una frase y qué parte de la frase representa la causa y cual el efecto. Los patrones se machean con los árboles sintácticos de parseo de las frases, y las partes del árbol de parseo que coinciden con los patrones son extraídas como la causa y el efecto. Los resultados de este estudio se consideran no satisfactorios, con una precisión que ronda el 0.51 al extraer la causa y el 0.58 al extraer el efecto. Los valores se refieren a la ”medida-F” que se define como una combinación de la precisión y la recuperación (*recall*) y se calcula con la siguiente fórmula:

*2*precisión*recall / (precisión + recall)*

En este caso la recuperación es el porcentaje de lugares llenados por los analistas humanos que son completados correctamente por el programa. Y la precisión es el porcentaje de lugares llenos por el programa que son correctos.

Si se incluyen tanto las relaciones implícitas como las explícitas los valores se reducen a 0.41 y 0.48 respectivamente.

2.3. Módulos clásicos de los sistemas Q&A

Ya vimos en que nos basamos para trabajar con la causalidad, ahora vamos a ver los criterios que usamos para construir nuestro sistema.

Para esto seguimos las etapas clásicas de Q&A, las cuales se componen de cuatro módulos básicos, donde los resultados de un módulo sirven de entrada a los siguientes:

- Análisis de la pregunta. Módulo en el que se procesa la pregunta expresada en lenguaje natural.
- Recuperación de documentos. Módulo en el que se realiza una primera selección de los documentos.
- Selección de párrafos. Módulo en el que se analizan los documentos con el fin de detectar aquellos fragmentos de texto en los que es más probable encontrar respuestas.
- Extracción de respuestas. Módulo que procesa los fragmentos de documento con la finalidad de localizar la respuesta correcta.

2.3.1. Análisis de la pregunta

La correcta clasificación de la pregunta es uno de los factores más importantes para que un sistema Q&A encuentre la respuesta correcta.

En general, el proceso se compone de dos tareas:

- Detectar el tipo de información que se espera como respuesta a la pregunta. Por ejemplo: una fecha, un nombre propio, una cantidad, etc.
- Seleccionar los términos de la pregunta que van a permitir la localización de los documentos que son susceptibles de contener la respuesta.

Una de las formas de identificar el tipo de respuesta esperada es determinar el tipo de pregunta, basándose en la estructura de la misma y en sus palabras clave, que representan la información sintáctica y semántica respectivamente.

Muchos de los sistemas Q&A que tratan de responder preguntas fácticas utilizan la técnica que describiremos a continuación para realizar el procesamiento de la pregunta. En primer lugar intentan identificar las entidades de la misma, como ser: persona, lugar, número, etc. Esto posibilita la reformulación de la pregunta sin perder información relevante para su clasificación. Por ejemplo, dada la pregunta *¿Quién es x?*, lo importante es saber que la respuesta debe contener un nombre propio y por tanto puede ser reformulada de la siguiente manera *¿Quién es <PERSONA>?*. En general esta tarea es realizada utilizando algún reconocedor

de entidades [Ent07] y reemplazando las mismas por el nombre de su clase. Luego se analizan las palabras que indican pregunta (como ser: Cómo, Cuándo, etc.), ciertas secuencias de palabras y algunos términos que son representativos para algún tipo de pregunta. Por ejemplo: una pregunta que comienza con la palabra Cuántos, probablemente su respuesta se refiera a una cantidad.

Este procedimiento es muy eficiente para preguntas del tipo Cuántos, pero existen algunos casos en los que se requiere un mayor análisis, como ser las preguntas que comienzan con la palabra Quién, que pueden referirse a una persona, una organización, etc. Por ejemplo, dada la pregunta:

¿Quién es el mayor productor de computadoras?

Sabemos que la respuesta se refiere a una organización, por tanto es necesario realizar un análisis semántico de la misma. Estas ambigüedades pueden ser resueltas analizando el contexto de toda la pregunta, mediante técnicas de PLN [PLN07] como ser: desambiguación del sentido de la palabra [FRA+06],Formas Lógicas [SnGC05], etc.

2.3.2. Recuperación de documentos

Los modelos de recuperación tienen como objetivo facilitar el proceso de comparación entre una consulta determinada y un conjunto de textos sobre los que se realiza la consulta [Gon03]. A continuación se describirán los modelos más utilizados.

Booleano

El modelo booleano es un modelo de Recuperación y Organización de la Información simple, basado en la teoría de conjuntos y el álgebra booleana. Las consultas se introducen como expresiones booleanas, lo que las dota de significado preciso. Debido a su simplicidad, es el modelo más utilizado para recuperación de información.

La idea principal del modelo es que una palabra clave puede estar ausente o presente en un documento y por tanto serán relevantes solo aquellos documentos

que contengan las palabras o combinaciones de palabras clave especificadas en la consulta. Al considerar presentes o ausentes a las palabras clave en los documentos, los pesos de las mismas siempre serán binarios (0, 1).

Las consultas estarán compuestas por palabras clave unidas por conectores booleanos (not, and, or). Este enfoque supone una gran desventaja frente a otros modelos, porque no se devolverán documentos que podrían ser relevantes a pesar de que no encajen a la perfección con la consulta. Estos problemas se pueden solucionar aumentando las posibilidades de pesos para las palabras clave, con lo que el modelo dejaría de ser booleano, evolucionando hacia el modelo vectorial.

Vectorial

El modelo de recuperación vectorial o de espacio vectorial propone un marco en el que es posible el emparejamiento parcial asignando pesos no binarios a las palabras clave de la preguntas. Estos pesos de los términos se usan para computar el grado de similitud entre cada documento considerado y la pregunta del usuario.

La idea básica de este modelo reside en la construcción de una matriz de términos o palabras clave (colocados en las columnas de la matriz) y documentos (colocados en las filas). De esta manera, un documento podría expresarse de la forma d_1= (1, 2, 0, 0, 0, ..., 1, 3) siendo cada uno de estos valores el número de veces que aparece cada término en el documento.

La segunda idea asociada a este modelo es calcular la similitud entre la pregunta y los m vectores de documentos construidos. Para calcular esta similitud se dispone de varias fórmulas, la más conocida es la Función del Coseno, cuya formula es la siguiente:

$$SIM(Q, D_{ij}) = \frac{\sum_{j=1}^{m} p_j d_{ij}}{\sqrt{\sum_{j=1}^{m} d_{ij}^2 \sum_{j=1}^{m} p_j^2}}$$

Probabilístico

La base principal de su funcionamiento es el cálculo de la probabilidad de un documento de ser relevante a una pregunta dada.

Dentro de la recuperación probabilística, se destaca el modelo de recuperación probabilístico de independencia de términos binarios dónde:

- La probabilidad de los términos es independiente (un término es independiente de los otros).
- Los pesos asignados a los términos son binarios

La equiparación probabilística se basa en que, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta.

Pese a estudiar las formas existentes para resolver la recuperación de documentos, y considerando que en nuestro caso el corpus a utilizar es la Internet, se decidió utilizar un buscador web que se encargue de ésta tarea. Dada nuestra experiencia y la conocida calidad técnica del buscador Google, éste es el que se decidió utilizar.

2.3.3. Selección de párrafos relevantes

La selección de párrafos relevantes se obtiene realizando un filtrado de los documentos obtenidos en el modulo anterior. Por lo general se recompensa a los fragmentos de texto que tengan mayor relación con la pregunta.

Este proceso, comúnmente se divide en los siguientes pasos:

1. Indexación, se genera un diccionario con la información relevante del documento.
2. Se seleccionan los documentos que contienen algún termino de la pregunta.
3. Para cada párrafo de cada documento, se realiza un cálculo que da un peso al párrafo que se esta analizando.
4. Se seleccionan los párrafos más relevantes para la extracción de la respuesta, y los mismos son ordenados según su similitud con la pregunta realizada.

2.3.4. Extracción de respuestas

El ultimo módulo de los sistemas Q&A es el de extracción de respuestas que se encarga de realizar un análisis más detallado del subconjunto de párrafos relevantes resultado del módulo anterior, con la finalidad de localizar y extraer la respuesta buscada.

El proceso de extracción de respuestas varía según cual sea el tipo de respuesta esperado, pero en general se divide en las siguientes etapas:

1. Cada párrafo relevante es analizado con la intención de seleccionar aquellas estructuras sintácticas que pueden ser respuesta a la pregunta.
2. Cada una de las repuestas posibles detectadas en los párrafos relevantes se puntúan con la intención de valorar en que medida, puede o no ser una respuesta correcta.
3. Las respuestas posibles se ordenan en función del valor obtenido en el punto anterior, y las de mayor puntaje son presentadas al usuario como respuesta a su pregunta.

Capítulo 3

El sistema Eneas

El principal objetivo de este proyecto es el desarrollo de un sistema de búsqueda de respuestas a preguntas causales. Se va a trabajar con el idioma español y como corpus la Web. Se decidió usar este corpus, debido a la gran cantidad de información que se puede encontrar, y así no restringir la variedad de preguntas que el usuario puede consultar. Otras de las razones, es para tratar de aprovechar la redundancia de información que hay en la Web, cosa bien aprovechada por el proyecto WebQA. Como contrapartida la información en la Web puede no ser fiable, por lo que aquí la redundancia de la información se vuelve clave a la hora de descartar datos erróneos. A continuación se presenta un ejemplo de como se pueden aprovechar las técnicas de redundancia para el caso de las preguntas fácticas:

¿Quién fue el primer hombre en pisar la Luna?

Respuestas obtenidas:

1. *Armstrong fue el primer hombre en pisar suelo lunar.*
2. *El primer hombre en pisar la Luna fue Neil Armstrong.*
3. *Neil Armstrong se convirtió en el primer hombre en llegar a la Luna.*
4. *Primero fue el Sputnik ruso en 1957.*

Luego de hacer un análisis básico se obtienen como posibles respuestas Armstrong, o Sputnik. Y de acuerdo a la frecuencia de las respuestas, se puede deducir que el primer hombre en pisar la Luna fue Neil Armstrong.

Estas técnicas se pueden aplicar porque se esperan respuestas concretas. En el caso de las preguntas causales, la respuesta es toda una redacción, es decir cada respuesta es única, y no se puede saber cuando hay un falso positivo (en el ejemplo, Sputnik). Por lo que la redundancia de información no puede ser aprovechada, pero de igual forma, el tener disponible una gran cantidad de información es de utilidad a la hora de buscar datos sobre los temas referidos en las preguntas, y así es más probable encontrar una respuesta adecuada.

A la hora de construir el sistema se decidió seguir las etapas clásicas de los sistemas Q&A, ya que se adaptaba a nuestras necesidades, pero se hicieron algunos cambios considerados pertinentes para resolver nuestro problema.

Las etapas seguidas en la resolución del sistema fueron las siguientes:

1. Análisis de la pregunta.
2. Recuperación de Documentos.
3. Selección de pasajes relevantes.
4. Ponderación de la Causalidad.
5. Presentación al usuario.

Con el *análisis de la pregunta*, a partir de una pregunta formulada por un usuario, se obtiene la información morfológica de las palabras contenidas en la misma, para luego poder reescribirla y enviarla al motor de búsqueda.

Con la *recuperación de documentos*, usando la información obtenida del análisis de la pregunta, y los indicadores causales primarios, obtenemos los documentos que pueden ser relevantes a la hora de encontrar una respuesta.

En la etapa *selección de pasajes relevantes*, se analizan los documentos retornados por el módulo de recuperación de documentos para encontrar los pasajes que contengan posibles respuestas a la pregunta.

Con la *ponderación de la causalidad*, se evalúa que tan probable es que un texto tenga carácter causal.

Al usuario se le *presentan las respuestas* marcándole las palabras u oraciones que el sistema consideró relevantes al momento de escogerlas, además de su valoración (puntaje de la respuesta), y de un link al documento original.

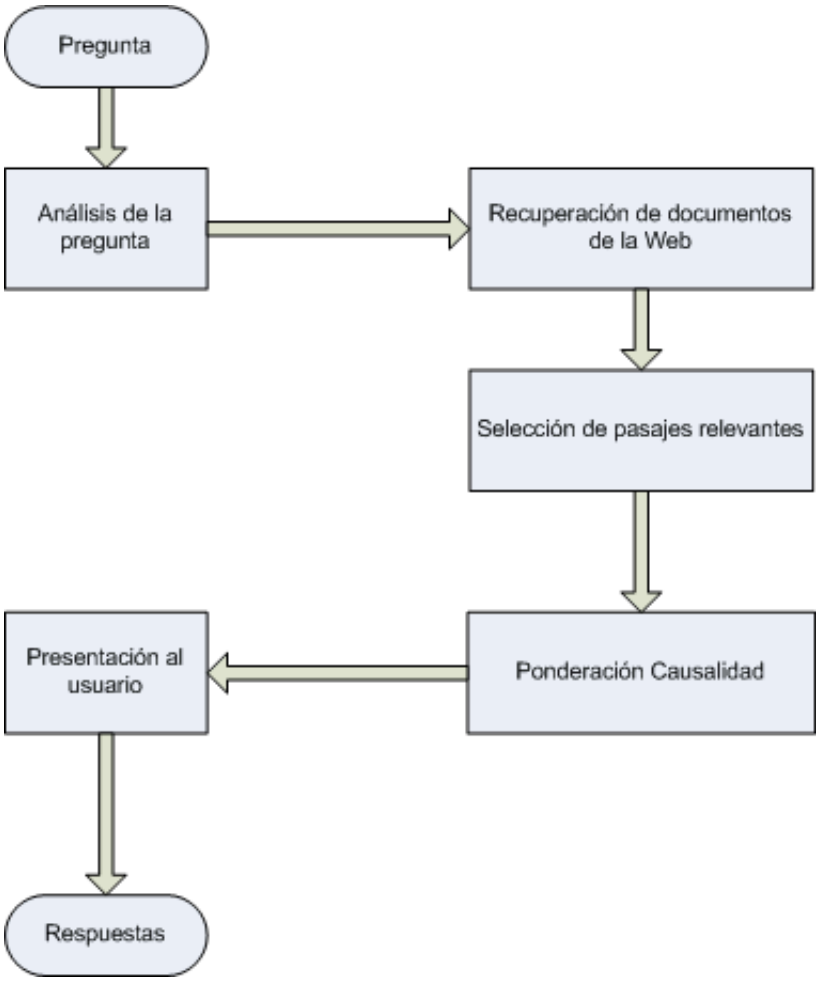


Figura 3.1: Etapas del sistema.

En las secciones siguientes, se describe como se resolvió cada etapa, además a modo de ejemplo, se ilustra como trabaja cada una de éstas con la pregunta *¿Por qué el cielo es azul?*

3.1. Indicadores causales

Luego de estudiar varios documentos del corpus utilizado se seleccionó una lista de frases, verbos, conectores y conjunciones que indican causalidad. En el cuadro 3.1 se muestran algunos ejemplos de estos indicadores.

Indicadores Causales
causa, provoca, desencadena, implica, suscita, genera, induce, ocasiona, lleva, porque, depende, deriva, produce, facilita, apoya, razón, debido a, causada por, etc.

Cuadro 3.1: Indicadores causales.

Una vez confeccionada la lista se sintió la necesidad de clasificarla, ya que no todos los indicadores indican causalidad con la misma intensidad, por ejemplo, se observó que el verbo *causar* indica causalidad en la mayoría de los textos en los que se encuentra, en cambio el verbo *llevar* puede tener o no un significado causal, dependiendo del contexto, por ejemplo en: *El gasto indiscriminado lleva a la recesión*, aquí *lleva* indica causalidad, pero en *Martín lleva una maleta roja*. *lleva* no tiene carácter causal.

La clasificación de indicadores realizada es la siguiente:

- Indicadores primarios: indicadores que expresan causalidad en la mayoría de los textos en los que se encuentran, como por ejemplo *causa*.
- Frases primarias: frases que indican causalidad en la mayoría de los textos en los que se encuentran, como por ejemplo *provocado por.*
- Indicadores secundarios: indicadores que expresan causalidad dependiendo del contexto en los que se encuentran, como por ejemplo *lleva*.
- Frases secundarias: frases que indican causalidad dependiendo del contexto en los que se encuentran, como por ejemplo *por lo tanto.*
- Indicadores terciarios: indicadores que dan dimensión a la causa, como por ejemplo *estimula*.

Dado que muchos de los indicadores encontrados son verbos, de forma de poder encontrar varias conjugaciones de los mismos al realizar la búsqueda, se decidió

colocar la raíz de éstos en la lista y no el verbo completo. Para identificar estos verbos se agregó a la lista información que indica si el verbo es completo o no.

Es importante destacar que en los únicos casos en los que aparecen las raíces de los verbos es en los indicadores clasificados como secundarios o terciarios. En el caso de los indicadores primarios, frases primarias y frases secundarias los indicadores aparecen completos. Esto es porque las frases no tienen conjugación y los indicadores primarios serán utilizados para las distintas reformulaciones como veremos más adelante. Como existen ciertos verbos que fueron clasificados como indicadores primarios, para no perder la posibilidad de encontrar alguna de las conjugaciones de los mismo, se los agregó como incompletos con su raíz en la lista de indicadores secundarios. Por ejemplo: podemos encontrar al verbo *causa,* en la lista de indicadores primarios y la raíz *caus* en la lista de indicadores secundarios.

Existe una propiedad que debe cumplir cada indicador (no importa su clasificación) para indicar causalidad, esta es el *sentido de la causalidad*, de la cuál se hablará en próximas secciones. A modo de introducción decimos que el sentido de la causalidad indica donde debe encontrarse el objeto de la pregunta para indicar causalidad, es decir, a la derecha o a la izquierda del indicador. Por ejemplo, para la pregunta *¿Cuáles son las causas de la gripe aviar?* dos oraciones causales serían:

- "*El virus H5N1 genera gripe aviar"*.
- *"La gripe aviar genera muerte"*.

Observamos que aunque ambas oraciones sean causales, solo la primera habla de las causas de la gripe aviar, ya que la segunda habla de sus consecuencias. Dado que el sentido de la causalidad depende del indicador, se decidió agregar a la lista de indicadores información que especifique el sentido de cada uno.

Para el caso de los verbos incompletos, de los que se quiere abarcar varias conjugaciones, notamos que el sentido de la causalidad depende de la conjugación. En el ejemplo anterior se vio que el verbo *generar* tiene sentido derecho cuando su conjugación es *genera*, pero en el caso de que su conjugación fuese *generada* el sentido sería izquierdo, como se observa en la siguiente oración:

- *"La gripe aviar es generada por el virus H5N1"*.

Además se vio que para algunas conjugaciones de verbos era necesario encontrar una palabra auxiliar (verbo, preposición, etc.) además del indicador, y que dicha palabra era relevante al momento de definir el sentido del verbo. En el ejemplo anterior se observa que la conjugación *generada* acompañada de la preposición *por* tiene sentido izquierdo, pero en el caso de la oración: *"El virus H5N1 ha generado la gripe aviar"* notamos que el sentido de la conjugación *generado* es derecho cuando está acompañada del verbo *ha*.

De forma de poder contemplar estos casos se decidió realizar una lista de palabras que concatenadas a cierto verbo hacen que se cambie el sentido de la causalidad, llamaremos a dichas palabras extensiones. La lista mantiene, para cada raíz de verbo, la siguiente información:

- extensión, auxiliar, sentido del auxiliar (esto es, lugar donde debe estar el auxiliar con respecto al indicador) y el sentido que toma el indicador en el caso que se encuentre extendido. Por ejemplo, para la raíz *gener* (del verbo *generar*), se tendrán las extensiones que se muestran en el cuadro 3.2.

Existen algunas frases, tanto primarias como secundarias, para las cuales también se debe buscar una palabra auxiliar, como en el caso de los verbos extendidos. Por ejemplo, si observamos la oración, *"La causa de la gripe aviar es el virus H5N1",* vemos que la frase primaria *causa de* debe estar acompañada del verbo *es,* además notamos que en este caso el objeto de la pregunta debe estar entre el verbo y el indicador. Para poder identificar este tipo de oraciones se agregó a la lista de frases primarias y secundarias información que indica el auxiliar que se debe buscar en caso que se encuentre una frase de este tipo.

Extensión	**Sentido Verbo**	**Sentido Auxiliar**	**Auxiliar**
ad	D	I	ha
ad	D	I	han
ad	D	I	he
ad	D	I	hemos
ad	I	D	por

Cuadro 3.2: Extensiones de verbos

En el anexo indicadores 8 se da una lista completa de todos los indicadores causales y su clasificación.

3.2. Análisis de la pregunta

En el caso de las preguntas fácticas esta etapa es una de las más importantes, ya que aquí se debe determinar de qué tipo es la pregunta y qué se quiere como respuesta (por ejemplo: fecha, lugar, nombre, etc.). En nuestro proyecto esta etapa es muy simple, debido a que trabajamos con un único tipo de preguntas (causales), y la respuesta no se puede reducir a una fecha, lugar, etc.

Como se explicó en el marco teórico 1, se asume que la causalidad es una relación binaria y que toda afirmación causal tiene una causa, un efecto, y un conector que los relaciona. De los tipos de palabras que se pueden encontrar en la pregunta, se consideró que algunos no aportan nada para encontrar la respuesta, como por ejemplo los artículos. Por otro lado se considera que los sustantivos, si se encuentran en un texto, indican que el mismo esta hablando sobre el tema del que trata la pregunta, y por lo tanto son de gran interés.

En esta fase solamente obtenemos las palabras trascendentes de la pregunta, como son los nombres propios, verbos, adverbios, adjetivos, cifras, etc. Y se descartan las conjunciones, pronombres, determinantes, interjecciones y preposiciones, dado que estas no son de utilidad a la hora de analizar los documentos.

En el cuadro 3.3 se presentan los tipos de palabra que no se descartan y la letra con la que se representan.

Tipo	**Letra**
Sustantivo	N
Adjetivo	A
Adverbio	R
Verbo	V
Cifra	Z
Moneda	m
Fecha y hora	W
Indicador causal primario	K

Cuadro 3.3: Tipos de palabras consideradas

A continuación se presenta un ejemplo de la entrada al módulo y la salida del mismo para la pregunta *¿Por qué el cielo es azul?*.

Entrada	Salida
¿Por qué el cielo es azul ?	*cielo:N es:V azul:A*

El *Por qué,* que es lo que da carácter causal a la pregunta, se descarta antes de realizar el análisis, dado que el sistema solo trabaja con preguntas causales, por lo que se supone que la pregunta es causal.

El *el* es un artículo, y no nos interesa a la hora de reformular la pregunta, por lo que también es descartado.

3.3. Recuperación de Documentos de la Web

Luego de que se hizo el análisis de la pregunta, es necesario obtener los documentos de la Web para analizar, para esto se construyen distintas reformulaciones de la pregunta, y cada una es enviada al motor de búsqueda. Estas reformulaciones se hacen utilizando los indicadores causales primarios, y la salida del módulo *análisis de la pregunta*.

Para el ejemplo 3.2, posibles reformulaciones serían:

- porque cielo es azul
- causas cielo es azul
- provoca cielo es azul
- razón cielo es azul
- debido a cielo es azul

Con cada reformulación extraemos un conjunto de documentos, de los cuales obtenemos el texto plano, y luego los analizamos para obtener los *pasajes relevantes*.

En un principio se pensó en analizar solo los *snippets*, y de esa forma evitar analizar un documento entero, lo que implicaría un ahorro importante en el tiempo de ejecución. Pero esta idea fue descartada rápidamente, dado que la información obtenida de los *snippets* es insuficiente en la mayoría de los casos al tratar de responder una pregunta causal. Por eso se decidió analizar los documentos en su totalidad.

3.4. Selección de pasajes relevantes

El objetivo de esta etapa es analizar los documentos recuperados de la Web, y obtener los segmentos donde podrían estar las posibles respuestas a la pregunta. A la hora de dividir el texto para hacer el análisis, evaluamos diferentes formas de hacerlo.

Primero se decidió dividir el texto en oraciones, y evaluar cada oración por separado. El problema de esto, además de la gran cantidad de tiempo que llevaría analizar los documentos (ya que se analizan todas las oraciones de todos los documentos), es que, como se analiza cada oración por separado, las respuestas que abarcan más de una oración se pueden perder, por lo que se decidió no seguir este camino.

Por ejemplo, para la pregunta *¿Por qué algunas aves migran?*, obtenemos el texto:

La respuesta descansa en el alimento. Mientras que las aves que viven en zonas cálidas cercanas al ecuador pueden conseguir alimento todo el año, los días en estas regiones son mucho más cortos – tan sólo 12 horas en el ecuador mismo. Debido a qué la mayoría de las especies se alimentan con la vista, esto limita el tiempo que pueden pasar alimentándose, un problema para las aves que tratan de reunir alimento suficiente para alimentar sus hambrientas crías.

Si analizamos cada oración por separado en el texto anterior, no encontraríamos la respuesta, dado que para que el sistema pueda responder, debe haber mención a las *aves*, a la *migración* y un indicador causal entre ellas. Cosa que ninguna oración por separado posee.

Luego se decidió separar el texto en párrafos, de esta manera la cantidad de texto a analizar es el mismo, lo que cambia es que se realiza un menor número de análisis, cada uno con un texto más amplio, además, de esta forma es más probable encontrar una respuesta. Dado que no es posible saber cuando comienza y termina un párrafo, debido a la heterogeneidad de los textos en la Web, se decidió definir párrafo como conjunto de oraciones, lo que llamaremos de ahora en más segmento. Si tomamos una cantidad fija de oraciones para cada segmento, podría pasar que las palabras de interés se encuentren en distintos segmentos, por lo que tenemos el mismo problema que teníamos antes con las oraciones.

Para evitar esto se construyeron los segmentos de forma solapada. Es decir, si

el primer segmento constaba de las oraciones uno al cinco, el segundo constaba de las oraciones cuatro al ocho, y así sucesivamente. Esto nos llevaba a aumentar el tiempo de análisis dado que había algunas oraciones que se analizaban más de una vez, además corríamos el riesgo de repetir respuestas ya que puede pasar que la respuesta a una pregunta se encuentre en una oración que cae en dos segmentos.

Para solucionar esto lo que se hizo fue buscar en el texto menciones al sujeto del que habla la pregunta, luego construimos los segmentos alrededor de esta palabra, y analizamos cada oración de cada segmento por separado. De esta forma evitamos analizar segmentos que no contienen ninguna palabra relevante.

Ejemplo: ***¿porqué el cielo es azul?*** *Para explicar el color azul del* ***cielo****, imaginemos que dejamos pasar un rayo de sol por un prisma de vidrio. La luz se abre en un abanico de colores (se dispersa) por refracción y como resultado de esta dispersión vemos una gama de colores: violeta, azul, verde, amarillo y rojo. El rayo violeta es el que se ha separado mas de la dirección del rayo blanco y ahí esta precisamente la explicación del color del* ***cielo****. La desviación es máxima para los rayos de longitud de onda corta (violeta y azul), y mínima para los de longitud de onda larga (amarillos y rojos), que casi no son desviados.Los rayos violetas y azules, una vez desviados, chocan con otras partículas de aire y nuevamente varían su trayectoria, y así sucesivamente: realizan, pues, una danza en zigzag en el seno del aire antes de alcanzar el suelo terrestre. Cuando, al fin, llegan a nuestros ojos, no parecen venir directamente del Sol, sino que nos llegan de todas las regiones del* ***cielo****, como en forma de fina lluvia. De ahí que el* ***cielo*** *nos parezca azul, mientras el Sol aparece de color amarillo, pues los rayos amarillos y rojos son poco desviados y van casi directamente en línea recta desde el Sol hasta nuestros ojos.*

En el ejemplo anterior se arman los segmentos alrededor de la palabra *cielo.*

El algoritmo de selección de pasajes relevantes implementado refleja la forma que encontramos para detectar respuestas causales a las preguntas realizadas por los usuarios. Como ya dijimos, éste se basa en encontrar palabras en el texto, diferenciando por un lado las palabras clave de la pregunta y por otro lado los indicadores causales. Para lograr esto se recibe un string conteniendo al principio todas las palabras de la pregunta y el indicador causal primario resultantes de la etapa de *análisis de la pregunta*, seguido por el texto plano resultante de la etapa de *recuperación de documentos*.

El análisis comienza buscando en el texto las ocurrencias de las palabras clave de la pregunta. Cuando se encuentra una ocurrencia se toma el texto que inicia

200 caracteres antes y finaliza 200 caracteres después de la misma. De este trozo de texto se quitan la primera y última oración, que por como fue construido el segmento seguramente estén incompletos. Este es el segmento candidato a contener la respuesta y el que se analiza en el resto del algoritmo. Se pueden encontrar cualquier cantidad de segmentos candidatos, dependiendo de cuántas palabras clave de la pregunta se detecten en él. Cuando se encuentran dos o más ocurrencias de una palabra clave, muy cerca, para evitar dos respuestas repetidas o con textos superpuestos, lo que se hace es tomar como segmento el texto que inicia 200 caracteres antes de la primer ocurrencia y finaliza 200 después de la última.

Luego de obtenido el segmento se debe analizar el mismo para ver si realmente responde a la pregunta. Para esto se evaluaron varios criterios.

En primer lugar se consideró que un segmento responde la pregunta si contiene las palabras clave de la misma y además un indicador causal. Utilizando sólo éste criterio se vio que había muchos casos en que las palabras clave estaban dispersas en el segmento y en realidad no hablaban del tema al que se refería la pregunta. Para evitar esto se decidió que se responde la pregunta cuando las palabras clave y el indicador se encuentran en la misma oración de un segmento. Este criterio, más restrictivo que el anterior, en general encontraba oraciones que hablaban de la pregunta y que eran causales, pero a su vez descartaba segmentos que no tenían todas las palabras clave en una oración pero que sin embargo si respondían. Por ésto se decidió dejar ambos criterios, pero asumiendo que los que cumplían el criterio más restrictivo respondían mejor que los otros.

Por cada segmento seleccionado se recorre cada una de las oraciones para detectar alguna que tenga todas las palabras clave. Las oraciones que tengan todas las palabras clave serán analizadas por el algoritmo de *Sentido de la Causalidad* que se explica a continuación. Si no se encuentra ninguna oración con todas las palabras clave, de todos modos se ve si las mismas están en el segmento.Con el resultado de éstos dos casos se define el puntaje final que se le asignará al segmento.

3.5. Ponderación de la Causalidad

Como las preguntas con las que se trabajan son las causales, se deben encontrar respuestas acordes a este tipo de preguntas. Para esto se analiza la causalidad de los segmentos considerados como candidatos a tener la respuesta. Estos segmen-

tos presentan un inconveniente, este es, que muchas veces contienen una oración causal pero ésta oración no tiene el sentido causal buscado. Por ejemplo, si se pregunta *¿Cuáles son las causas del calentamiento global?* se pueden encontrar las siguientes respuestas:

1. *El calentamiento global de la tierra es provocado por un aumento del efecto invernadero.*

2. *El calentamiento global provoca un dramático aumento del nivel del mar.*

En este ejemplo la primer oración habla de las causas del calentamiento global, y la segunda oración habla sus consecuencias. Ambas oraciones son causales y hablan del tema del que se esta preguntando, pero solo la primera es una respuesta válida a la pregunta.

Lo que se intenta descubrir en esta etapa es el sentido de la causalidad de un texto, es decir, se ve si se esta hablando de las causas de un hecho, o de las consecuencias del mismo.

3.5.1. Algoritmo de sentido de la causalidad

Al probar diferentes búsquedas se encontraron casos en los que la oración tenía todas las palabras clave de la pregunta, un indicador causal y sin embargo no respondía nada sobre las causas de lo que se estaba preguntando. Esto es porque hay casos en los que las oraciones son tan largas que hablan de varios temas y se puede dar que las palabras, al estar en diferentes partes de la oración no se relacionen entre sí; por ejemplo, en una parte se habla de un tema y en otra de las causas de otro.

Otro caso que nos llamó la atención fue en el que la oración era causal y hablaba del tema que nos interesaba pero la razón para no responder de forma adecuada, era que el sentido de la relación causal expresada en la oración, no se correspondía con el sentido esperado. Por ejemplo, al preguntar sobre las *causas de la lluvia ácida*, esperábamos una respuesta como ésta: *"La lluvia ácida es provocada por los humos y los gases emitidos por los automóviles y las industrias."* y de ahí se podía ver que las causas de la la lluvia ácida son *los humos y los gases emitidos por los automóviles y las industrias*. La misma oración pero con el sentido causal

opuesto, como: *"La lluvia ácida provoca alteraciones en el ciclo biológico de los hongos."* era considerada por nuestro sistema como una respuesta excelente, pese a que no dice nada sobre las causas de la lluvia ácida.

Al algoritmo que analiza oraciones y verifica que el sentido de la causalidad sea el adecuado fue denominado *Sentido de la Causalidad*. El mismo es aplicado solo a oraciones que contienen todas las palabras clave de la pregunta.

En las primeras versiones del algoritmo se pensó que simplemente recibiera una oración y retornara un valor booleano, que indicara si cumplía el sentido de la causalidad o no. Evaluando los resultados obtenidos se vio que para asegurarse de que una oración fuese causal y cumpliese el sentido de la causalidad había que utilizar varias condiciones muy restrictivas. Esto resultaba en que las que se consideraban válidas eran muy pocas, y entre todo el resto de oraciones descartadas había muchas que no cumplían los criterios pero sin embargo respondían. Por esta razón, en lugar de descartar éstas oraciones se decidió marcarlas como no tan buenas. De esta forma, en la segunda versión del algoritmo, se decidió que el mismo retornase una ponderación en lugar de devolver un valor booleano. Esta ponderación es utilizada por el algoritmo general para decidir el puntaje final del segmento.

Como se mencionó anteriormente una de las tareas de éste algoritmo es decidir si una oración es causal o no. Una oración se considera causal si contiene un indicador causal y cumple el sentido del indicador encontrado. A continuación se da el pseudocódigo del algoritmo:

Para cada oración del segmento.

1. Iterar sobre las frases primarias.
    - *a*) Se encuentra frase primaria en la oración. Se debe verificar si la oración cumple con el sentido indicado por la frase.
        - 1) Se cumple el sentido.
            - Si la frase debe tener una palabra auxiliar que la acompañe para ser causal, se busca dicha palabra.
                - Se encuentra la palabra auxiliar y ésta palabra cumple el sentido esperado:
                    - ponderación = frase primaria cumple sentido.
                    - Fin algoritmo.

- No se encuentra la palabra auxiliar o se encuentra la palabra pero la misma no cumple el sentido causal esperado:
  - ponderación = frase primaria no cumple sentido.
  - Vuelve al paso 1.

- Si la frase no debe tener una palabra auxiliar que la acompañe para ser causal:
  - ponderación = frase primaria cumple sentido.
  - Fin algoritmo.

2) No cumple el sentido:
   - ponderación = frase primaria no cumple sentido.
   - Vuelve al paso 1.

b) No se encuentra frase primaria, se va al paso 2 del algoritmo.

2. Iterar sobre los indicadores primarios.

a) Se encuentra indicador primario en la oración. Se debe verificar si la oración cumple con el sentido indicado por el indicador causal.

1) Se cumpla el sentido.
   - Si el indicador debe tener una palabra auxiliar que lo acompañe para ser causal, se busca dicha palabra.
     - Se encuentra la palabra auxiliar y ésta palabra cumple el sentido esperado:
       - ponderación = indicador primario cumple sentido.
       - Fin algoritmo.
     - No se encuentra la palabra auxiliar o se encuentra la palabra pero la misma no cumple el sentido causal esperado:
       - ponderación = indicador primario no cumple sentido.
       - Vuelve al paso 2.
   - Si el indicador no debe tener una palabra auxiliar que la acompañe para ser causal:
     - ponderación = indicador primario cumple sentido.
     - Fin algoritmo.

2) No cumple el sentido:
   - ponderación = indicador primario no cumple sentido.
   - Vuelve al paso 2.

b) No se encuentra indicador primario, se va al paso 3 del algoritmo.

3. Iterar sobre los indicadores secundarios.

   a) Se encuentra indicador secundario en la oración. Se debe verificar si la oración cumple con el sentido indicado por el indicador causal.

      1) Se cumpla el sentido.
         - Si el indicador debe tener una palabra auxiliar que lo acompañe para ser causal, se busca dicha palabra.
           - Se encuentra la palabra auxiliar y ésta palabra cumple el sentido esperado:
             - ponderación = indicador secundario cumple sentido.
             - Fin algoritmo.
           - No se encuentra la palabra auxiliar o se encuentra la palabra pero la misma no cumple el sentido causal esperado:
             - ponderación = indicador secundario no cumple sentido.
             - Vuelve al paso 3.
         - Si el indicador no debe tener una palabra auxiliar que la acompañe para ser causal:
           - ponderación = indicador secundario cumple sentido.
           - Fin algoritmo.
      2) No cumple el sentido:
         - ponderación = indicador secundario no cumple sentido.
         - Vuelve al paso 3.

   b) No se encuentra indicador secundario, se va al paso 4 del algoritmo.

4. Iterar sobre las frases secundarias.

   a) Se encuentra frase secundaria en la oración. Se debe verificar si la oración cumple con el sentido indicado por la frase.

      1) Se cumple el sentido.
         - Si la frase debe tener una palabra auxiliar que la acompañe para ser causal, se busca dicha palabra.
           - Se encuentra la palabra auxiliar y ésta palabra cumple el sentido esperado:
             - ponderación = frase secundaria cumple sentido.
             - Fin algoritmo.
           - No se encuentra la palabra auxiliar o se encuentra la palabra pero la misma no cumple el sentido causal esperado:

- ponderación de oración = frase secundaria no cumple sentido.
- Vuelve al paso 1.

- Si la frase no debe tener una palabra auxiliar que la acompañe para ser causal:
  - ponderación = frase secundaria cumple sentido.
  - Fin algoritmo.

2) No cumple el sentido:
   - ponderación = frase secundaria no cumple sentido.
   - Vuelve al paso 1.

b) No se encuentra frase secundaria, se va al paso 5 del algoritmo.

5. Iterar sobre los indicadores terciarios.

a) Se encuentra indicador terciario en la oración. Se debe verificar si la oración cumple con el sentido indicado por el indicador causal.

1) Se cumpla el sentido.
   - Si el indicador debe tener una palabra auxiliar que lo acompañe para ser causal, se busca dicha palabra.
     - Se encuentra la palabra auxiliar y ésta palabra cumple el sentido esperado:
       - ponderación = indicador terciario cumple sentido.
       - Fin algoritmo.
     - No se encuentra la palabra auxiliar o se encuentra la palabra pero la misma no cumple el sentido causal esperado:
       - ponderación = indicador terciario no cumple sentido.
       - Vuelve al paso 5.
   - Si el indicador no debe tener una palabra auxiliar que la acompañe para ser causal:
     - ponderación = indicador terciario cumple sentido.
     - Fin algoritmo.

2) No cumple el sentido:
   - ponderación = indicador terciario no cumple sentido.
   - Vuelve al paso 5.

b) No se encuentra indicador terciario:

- ponderación = no causal.
- Fin algoritmo.

NOTA: La verificación del sentido causal de una oración se realiza de la siguiente manera: si el indicador encontrado tiene sentido derecho entonces el objeto de la pregunta debe estar a la derecha del indicador, en caso contrario el objeto debe estar a la izquierda.

Por último, se puntúa el segmento dependiendo del tipo de indicador encontrado y de la proximidad del mismo con respecto al objeto de la pregunta.

3.5.2. Puntaje

Como se mencionó anteriormente, el puntaje final se obtiene de los algoritmos de análisis presentados. Los puntajes utilizados, y los casos en que se asignan a una frase son los siguientes:

- EXCELENTE: En una misma oración se encuentran todas las palabras de la pregunta y una frase primaria que cumple el sentido de la causalidad.
- MUY BUENO: En una misma oración se encuentran todas las palabras de la pregunta y un indicador primario que cumple el sentido de la causalidad. Además este indicador está próximo al objeto de la pregunta.
- BUENO: En una misma oración se encuentran todas las palabras de la pregunta y un indicador primario que cumple el sentido pero está lejos del objeto de la pregunta, o un indicador secundario que cumple el sentido y está próximo al objeto de la pregunta.
- ACEPTABLE: En una misma oración se encuentran todas las palabras de la pregunta y un indicador secundario que cumple el sentido pero está lejos del objeto de la pregunta, o se encuentra un indicador terciario o frase secundaria que cumple el sentido y está próximo al objeto de la pregunta.
- DEFICIENTE: No se encuentra nada de lo anterior.

3.5.3. La reformulación especial Wikipedia

Hasta ahora nos hemos basado principalmente en los indicadores causales, y el sentido de la causalidad para encontrar las respuestas. Ahora vamos a ver un caso especial, descrito en las *secuencias textuales* 2.2.2 como *Secuencia que lleva un título causal*.

Cuando se comenzó con la investigación de como implementar *Eneas*, se realizaron búsquedas en Google de las posibles preguntas que un usuario podría ingresar (como por ejemplo, *causas de la segunda guerra mundial*), de forma de poder analizar las páginas devueltas por el buscador y de encontrar una manera de devolverle al usuario las mejores respuestas, en caso que existiesen. Realizando este análisis se observó que en muchos casos las mejores respuestas estaban en páginas de Wikipedia en las cuales el artículo principal tenía como título *"Causas de X"* (siendo X el tema sobre el que se está preguntando) o el título del artículo mencionaba el tema del que se pregunta y además existía una sección especial del mismo llamada *Causas o Antecedentes.*

Para aprovechar éste hecho, luego de que se realiza la reformulación de la pregunta (3.3), se envía al buscador la consulta indicando el sitio en el que se quiere buscar. Para el ejemplo *3.2* la consulta sería: *site:es.wikipedia.org causa cielo es azul.*

Luego de obtenido el documento se analiza el título del artículo, si el mismo contiene la palabra causa seguida del objeto de la pregunta se da como posible respuesta el primer párrafo del artículo y la respuesta se rankea como EXCELENTE. Se puede observar que esto da muy buenos resultados, por ejemplo, para la pregunta *causas de la segunda guerra mundial* la respuesta que se devuelve es la siguiente:

Se considera comúnmente que las causas de la Segunda Guerra Mundial más directas son la Invasión de Polonia de 1939 y los ataques japoneses a China, los Estados Unidos de América y las colonias británicas y holandesas en Asia. En cada una de estas acciones, los ataques fueron el resultado de una decisión tomada por las élites gobernantes autoritarias en Alemania y Japón. La Segunda Guerra Mundial se inició después de que estas acciones agresivas recibieran como respuesta una declaración de guerra formal, una resistencia armada, o ambas cosas.

Si el título del artículo no contiene todas las palabras de la pregunta y el indi-

cador *causa,* se continúa analizando el artículo. Primero se busca la sección *causa* o *antecedente,* en caso que no exista ninguna de estas secciones el artículo es descartado. En caso contrario se da como posible respuesta las primeras frases de la sección encontrada; esta respuesta es rankeada según el siguiente criterio, en orden descendente:

- Si el título del artículo contiene todas las palabras de la pregunta, menos el indicador *causa* la respuesta es rankeada como MUY BUENA.
- Si se encuentran todas las palabras de la pregunta en la sección, la respuesta es rankeada como BUENA.
- En caso que se encuentre al menos una palabra de la pregunta en la sección, la respuesta es rankeada como ACEPTABLE.
- En caso que no se de ninguno de los casos de los puntos anteriores la respuesta es rankeada como DEFICIENTE.

3.6. Presentación de la respuesta al usuario

Esta etapa se encarga de mostrarle las respuestas al usuario. Por cada respuesta se mostrarán las palabras u oraciones que hicieron que el sistema la seleccionara como tal, el link a la página de donde se obtuvo la información y el puntaje otorgado por el sistema.

En la figura 3.2 se muestran algunas de las respuestas obtenidas a la pregunta, *¿Por qué el cielo es azul?*. Allí se puede observar que la respuesta a la pregunta se marca con negrita y además el puntaje se representa mediante estrellas.

Figura 3.2: Presentación de la respuesta.

Capítulo 4

Implementación

La construcción del sistema Eneas se caracterizó por el desarrollo de prototipos evolutivos desde el comienzo del proyecto. De esta forma se fueron superando muchos de los desafíos tecnológicos que fueron surgiendo al proponer diferentes soluciones a los problemas planteados. En el capítulo anterior se describieron, tanto la problemática planteada, como las dificultades que surgieron al avanzar en la construcción de Eneas, y las soluciones encontradas. En este capítulo se describe cómo éstas soluciones fueron plasmadas en un sistema real, funcional y utilizable.

Una de las características del sistema implementado es que puede ser fácilmente reutilizado. El hecho de que este proyecto, que se enfoca en responder preguntas causales, sea la continuación de WebQA, que responde preguntas fácticas, muestra claramente la conveniencia de poder integrarlos, y de esta forma contar con un sistema más general que responda ambos tipos de preguntas, primero que analice la pregunta, y luego invoque al sistema adecuado.

Para lograr esta característica se utilizó UIMA [UIM08], que también fue utilizado por WebQA. UIMA es una plataforma abierta, para uso industrial, que esta siendo estandarizada por OASIS [OAS], la cual propone una arquitectura basada en componentes con el objetivo de unificar el manejo de información no estructurada, de esta forma, un componente que analiza y procesa información no estructurada, puede ser reutilizado fácilmente por otro sistema.

El componente de nuestro sistema que recibe el documento en lenguaje natural y retorna las respuestas causales y los indicadores correspondientes, fue implementado como un componente UIMA. En nuestro caso la información no

estructurada es un texto en lenguaje natural, pero podría ser sonido o cualquier otro tipo de información. Aprovechando la reutilización de este componente y el hecho de que el proyecto WebQA[CIM07] también fue implementado como un componente UIMA, se podría implementar un tercer sistema que decida si la pregunta formulada es causal o fáctica e invoque el componente correspondiente en cada caso.

El paradigma utilizado para implementar el sistema fue el de orientación a objetos, en particular el lenguaje seleccionado fue Java [Jav08]. Todas las herramientas utilizadas fueron fácilmente utilizables desde Java, con la única excepción del POSTagger. El POSTagger utilizado fue Freeling [Fre08], ya que gracias al proyecto WebQA supimos que era mejor que OpenNLP [Ope08] y además lo comprobamos. Utilizamos el mismo método de WebQA para accederlo, escribir un archivo de texto con la pregunta, invocar directamente al ejecutable de Freeling para que lo procese y leer el archivo resultante.

4.1. Componentes de Eneas

En la estructura del sistema implementado se reflejan varias etapas que coinciden con etapas presentes en gran parte de los sistemas de respuesta automática a preguntas.

En la etapa de análisis de la pregunta se utiliza un POSTagger para obtener la morfología de cada palabra ingresada por el usuario y luego se descartan las palabras innecesarias. Esta etapa en particular se vio sumamente simplificada debido a que no se analizó el tipo de pregunta, ya que siempre son causales, y por lo tanto se asume que el usuario ingresa solo la entidad o suceso del que quiere saber la causa, no la pregunta entera.

En la etapa de reformulación de la pregunta surgió como problema el hecho de que *"calcular"* varias reformulaciones de una pregunta resulta en un sistema que demora mucho en responder. Para esto se propuso analizar cada reformulación de la pregunta en un hilo de ejecución separado. Esto implicó que la codificación se tornase más compleja, pero dado que en los primeros prototipos se vio que se podía implementar sin grandes inconvenientes y se vieron buenos resultados, se decidió continuar con la programación paralela.

La recuperación de documentos se hizo invocando a un buscador Web. Sim-

plemente se pasan las palabras que quedaron del análisis de la pregunta y el indicador causal primario correspondiente a la reformulación que se esté ejecutando. Es preciso hacer notar que las palabras no se pasan entre comillas, ni se invoca al buscador varias veces con las palabras en diferente orden. Simplemente se lo invoca una vez con las palabras separadas por espacio, esto permite que el buscador retorne los "mejores" documentos de Internet sin importar como las palabras se encuentren en el texto. Por lo tanto la capacidad de Eneas de recuperar buenos documentos esta totalmente delegada al buscador utilizado. Elegimos el buscador Google, por dos razones, primero porque se pudo reutilizar muy fácilmente el acceso implementado en el proyecto WebQA, y segundo porque es un buscador mundialmente conocido y nadie pone en duda su buen funcionamiento. Por razones obvias los resultados se limitan a que sean del idioma español.

Al momento de la selección de pasajes relevantes se vio que el buscador retornaba los documentos tal cual los retorna al explorador de Internet, o sea, en formato HTML con todas los tags necesarios para que el explorador los pueda interpretar de forma adecuada. Eneas asume que el texto es simplemente un conjunto de oraciones, y no tiene en cuenta el formato del mismo. En cambio, para implementar la reformulación especial que aprovecha la estructura de Wikipedia era necesario tener en cuenta la estructura del HTML. Esto resultó en que se utilizan dos detaggers diferentes, uno que retorna el texto plano correspondiente a una página web (HTML Parser [HP07]) y otro que además permite seleccionar partes del texto según su estructura de tags (Jericho [Jer07]). El análisis del texto plano para obtener las frases de interés, simplemente busca las palabras claves en todo el texto y analiza las oraciones circundantes, como se explicó en el capítulo anterior. Para poder analizar oraciones individuales se utiliza el Sentence Detector provisto por OpenNLP. Éste presentaba un pequeño problema pero gracias a que contábamos con el código fuente del mismo, pudo ser debuggeado y corregido fácilmente. La frase obtenida en esta etapa es la respuesta que retorna el sistema, por lo que no hay una etapa en la que se arme una respuesta.

Los componentes descritos anteriormente forman la capa lógica del sistema. Éste está conformado también, por la capa de presentación y por la capa de persistencia. Para la capa de presentación, además del acceso web que fue planteado desde un comienzo, se desarrollo un acceso batch de forma de poder ejecutar varias preguntas y obtener los resultados de cada una. Ambas presentaciones utilizan una misma interfaz, que publica las funcionalidades de la capa lógica.

Para la capa de persistencia se utilizó una base de datos para guardar la información obtenida durante el análisis, pero también se implementó la persistencia en archivos de texto, de forma de tener la posibilidad de no instalar un manejador

de base de datos para utilizar el sistema. La lógica utiliza una interfaz que publica las funcionalidades de la capa de persistencia, esta interfaz es implementada por clases diferentes para cada tipo de persistencia.

4.1.1. Arquitectura del sistema

Para ilustrar mejor los componentes descritos anteriormente vamos a ver un resumen de lo que es la arquitectura del sistema. Para más información ver [CGI08].

La arquitectura propuesta es en capas. En la figura 4.1 se muestra como interactúa cada una de las capas y a continuación se da una breve descripción de las mismas.

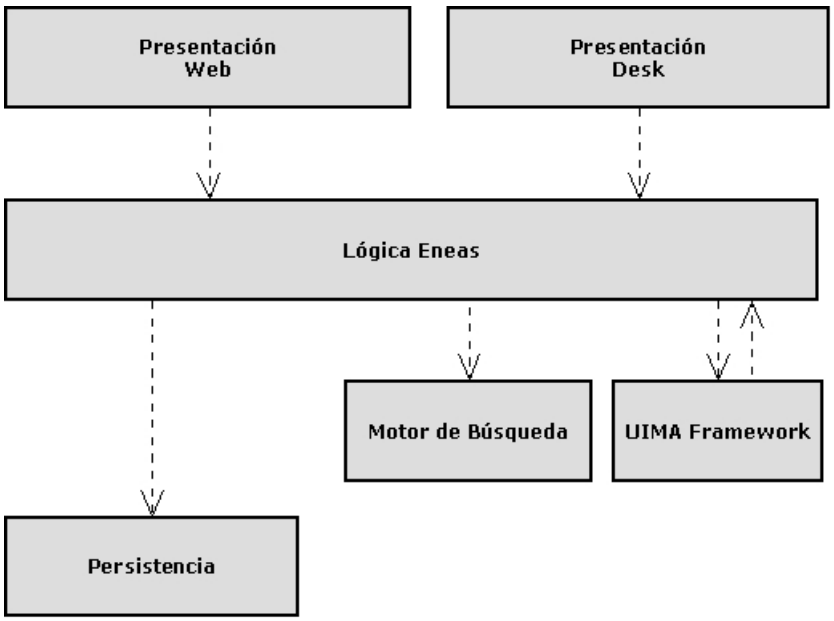


Figura 4.1: Arquitectura en capas de Eneas.

- **Presentación Web:** Esta capa es la encargada de permitir al usuario la interacción con el sistema. Para su diseño se utilizó el estilo de *Cliente Fino* de forma de que la aplicación pueda ser utilizada desde cualquier sistema operativo y cualquier explorador.

- **Presentación Desk:** El sistema provee una interfaz de escritorio que permite realizar la evaluación y ajuste de parámetros de forma más amigable. La misma brinda, entre otras cosas, gráficas que indican que tan buenos fueron los resultados obtenidos.

- **Lógica:** La capa Lógica es la encargada de manejar todos los pedidos de la presentación, la misma brinda las interfaces necesarias para resolver los pedidos de la interfaz de escritorio y los de la interfaz Web.

- **Persistencia:** Esta capa es la encargada de la interacción con el método de persistencia seleccionado en la configuración del sistema. Las posibilidades que Eneas brinda como persistencia son: Base de Datos o archivos de texto plano. Esta capa es fundamental para el sistema, ya que de allí se levantarán datos como por ejemplo los indicadores causales.

Los componentes *Motor de Búsqueda* (encargado de la interacción con el motor de búsqueda utilizado) y *UIMA Framework* forman parte de la capa lógica pero nos pareció importante destacarlos en la figura porque cumplen un rol muy importante en el diseño del sistema Eneas.

Los componentes de cada capa se muestran en la figura 4.2

Presentación

Este componente se encarga de obtener la pregunta realizada por el usuario, invocar las operaciones del sistema correspondientes a la búsqueda de respuestas y de presentar los resultados obtenidos, esto es realizado invocando a las operaciones brindadas por la interfaz . El sistema provee dos interfaces para ésto, una Web y otra batch.

En la interfaz Web, la pregunta es ingresada en una página Web, una vez que se obtienen las respuestas se muestran al usuario agregándolas a la página. Cada respuesta se muestra junto con un link al documento de donde ésta se obtuvo y la evaluación que el sistema dio a la misma.

En el caso de la interfaz batch, se lee un conjunto de preguntas de un archivo y luego se invocan las operaciones del sistema correspondientes para responder cada una de ellas.

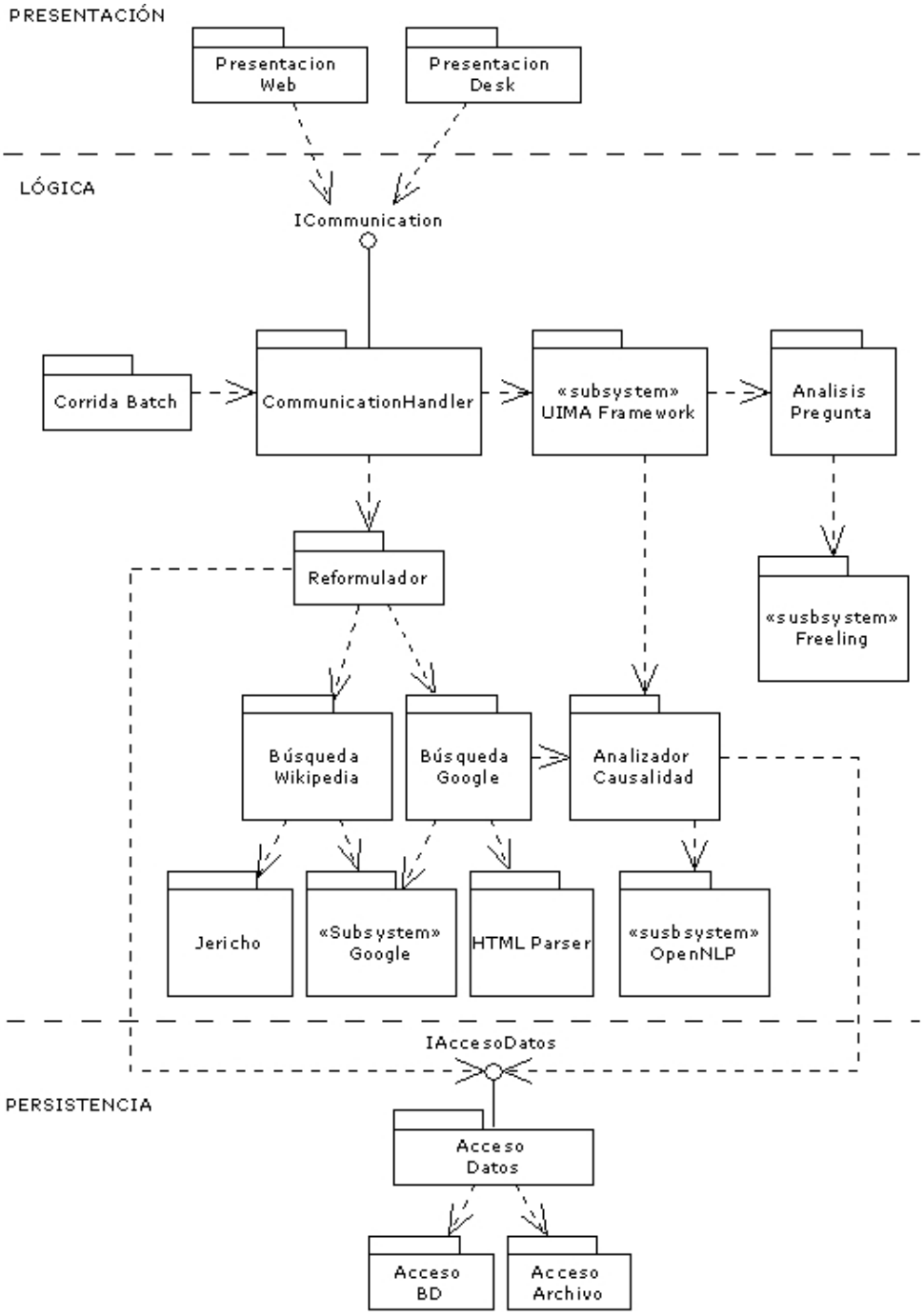


Figura 4.2: Diagrama de componentes.

Presentación de escritorio

El componente Presentación Desk se encarga de sacar estadísticas sobre los resultados obtenidos, es decir, en caso que se desee se podrán loguear todas las respuestas otorgadas por el sistema a cada una de las preguntas realizadas. Cuando se quiera realizar la evaluación, el usuario deberá leer las respuestas y marcarlas como correctas o incorrectas. Luego de esto el sistema mostrará ciertas gráficas que ayudarán al usuario en la evaluación del mismo.

Entre otras se podrán visualizar gráficas que indiquen:

- Cantidad de respuestas correctas e incorrectas por indicador causal.
- Cantidad de respuestas correctas e incorrectas por ranking otorgado.
- Cantidad de preguntas respondidas y no respondidas por el sistema.
- etc.

Comunicación

El componente *CommunicationHandler* es el encargado de manejar la comunicación entre las capas *Lógica y Presentación,* el mismo provee una interfaz que permite invocar operaciones útiles tanto para la presentación Web como para la presentación de escritorio.

Análisis Pregunta

Este componente es el encargado de realizar el análisis de la pregunta, es decir, dada una pregunta causal retorna las palabras relevantes de la misma, junto con el tipo de cada una.

Dado que al recibir la pregunta del usuario se asume que es causal y que solo incluye la entidad o la acción de la que se quiere averiguar la causa, éste componente se limita a descartar las palabras que no aportan información al análisis posterior. Esto se logra seleccionando los sustantivos, verbos, adverbios, cifras, fechas y descartando el resto de las palabras de la pregunta.

Reformulador

En este caso se reciben las palabras obtenidas por el modulo *Análisis Pregunta* y luego se invoca a un buscador Web para obtener los documentos que se analizarán posteriormente. El buscador es invocado una vez por cada reformulación, esto es, una vez por cada indicador primario y además se realiza una invocación especial en la que se indica que la página en la que se desea buscar es Wikipedia.

Esta etapa se corresponde con la etapa de recuperación de documentos, en nuestro caso, la recuperación de documentos fue delegada completamente al buscador Google. El comportamiento del buscador es crucial para que nuestro sistema obtenga los documentos más relevantes para cada reformulación enviada.

Luego de obtenidos los resultados del buscador, estos son procesados por un detagger para obtener el texto plano correspondiente, que es lo que retorna éste componente.

Analizador Causalidad

Este componente es el encargado de evaluar la causalidad de los textos obtenidos por el buscador. Recibe como entradas las palabras de la pregunta obtenidas por *Análisis Pregunta* y el documento a analizar.

En caso que el analizador encuentre respuestas las mismas son guardadas en memoria, para ser retornadas al usuario y además son enviadas al componente Persistencia, para que puedan ser analizadas posteriormente, en caso que se desee.

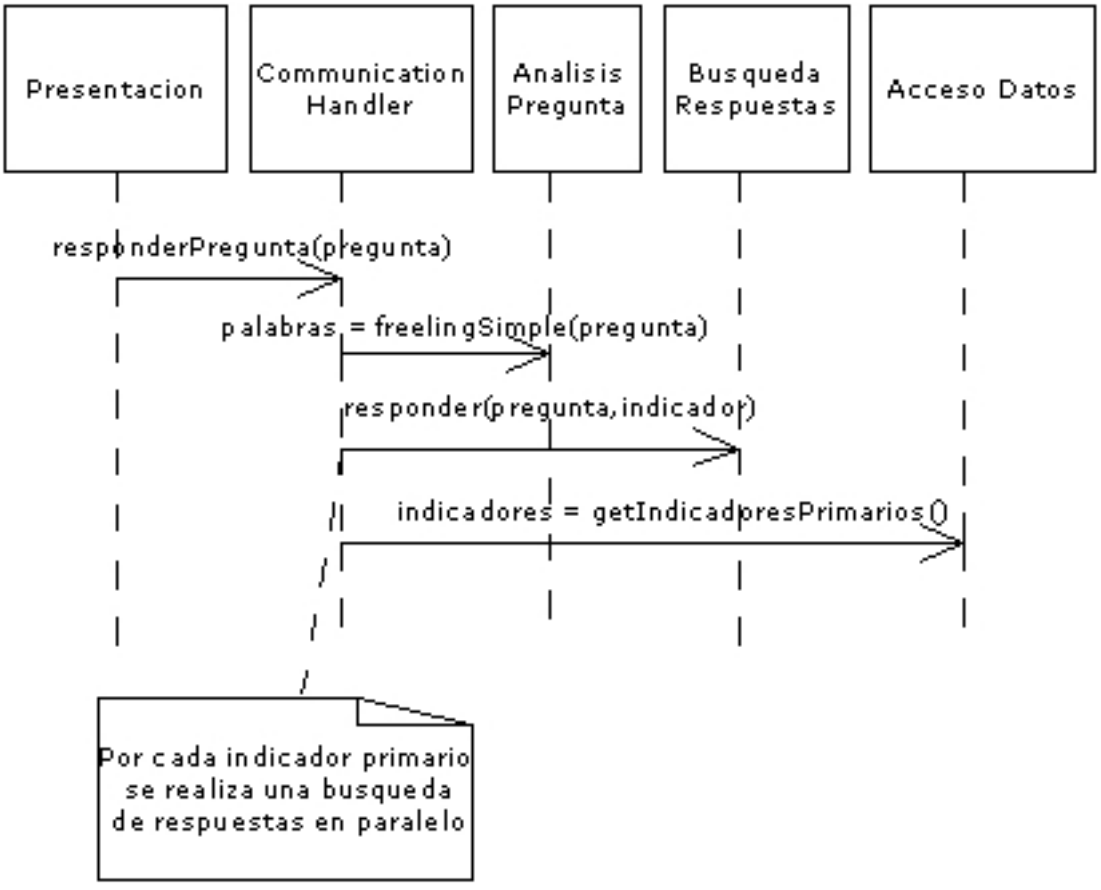
Acceso a Datos

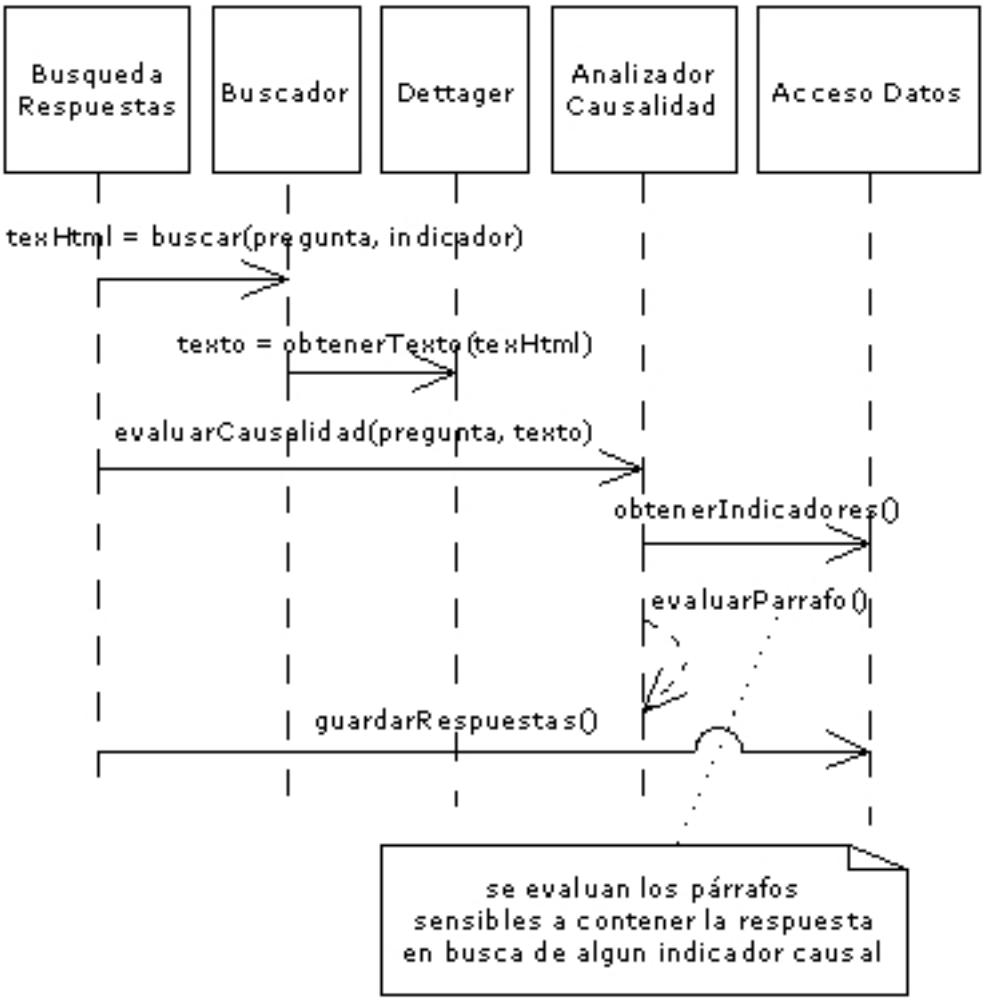
El componente *Acceso a Datos* es el encargado de manejar la comunicación entre las capas *Lógica* y *Persistencia.* El mismo provee una interfaz que permite invocar operaciones para obtener los indicadores causales y para guardar las respuestas otorgadas por el sistema. El sistema *Eneas* provee dos tipos de métodos de persistencia, base de datos y archivos.

En caso que se haya seleccionado como método de persistencia Base de Datos, el componente se encarga de hacer la conexión a la misma, de generar las

sentencias necesarias para guardar las respuestas en las tablas que corresponda y de obtener los indicadores causales. Como alternativa, este componente se puede configurar para persistir en archivo, en este caso se guardan las respuestas en archivos de texto plano creados en el servidor de aplicaciones, además el componente se encarga de parsear los archivos que contienen los indicadores causales.

El caso de uso que definió la arquitectura del sistema Eneas es *Responder Pregunta*. A continuación se muestra un diagrama que explica como interactúan cada uno de los componentes descritos anteriormente para su resolución.





4.2. Utilización de hilos en Java

Como se vio en la sección anterior, los hilos juegan un papel importante en el sistema Eneas. Por eso veremos con más detalle por qué y cómo los utilizamos.

Al principio del proyecto se planteó la posibilidad de enviar una consulta al buscador por cada indicador primario. En las primeras pruebas se vio que la ejecución de las consultas demoraba bastante, lo más notorio era que el programa se quedaba esperando la respuesta del buscador durante mucho tiempo. También se notó el mismo efecto en el uso del POSTagger. Considerando que el uso de éste implica una operación de E/S al acceder al disco duro, sin utilizar el recurso pro-

cesador, y por otro lado que la espera por el buscador no utilizaba ningún recurso de hardware, se decidió probar el uso de hilos para lograr que mientras un hilo espera resultados del buscador, otro pueda ejecutar instrucciones y otro pueda estar esperando E/S.

Ya en los primeros prototipos se vio que se reducía el tiempo de ejecución total. También se consideró el hecho de que la utilización de hilos implica resolver problemas de sincronización, sumando tiempo a la implementación, tanto por el hecho de utilizar hilos como por corregir problemas no conocidos que podrían surgir al utilizar un aspecto de la programación que no estábamos acostumbrados a manejar. Evaluando ventajas y desventajas, y el hecho de que actualmente los procesadores crecen en cantidad de núcleos, con lo que la utilización de hilos podría volverse más ventajosa, se decidió utilizarlos.

El funcionamiento básico consiste en iniciar cada uno de los hilos y esperar que terminen todos. Para lograr esto se mantiene una lista de hilos que se va cargando al iniciar cada uno. Luego de que todos los hilos se iniciaron, el software espera por cada uno de ellos. Pese a esto, por la forma en que esta hecha la implementación, cada hilo genera respuestas independientemente de los demás, y se pueden mostrar al usuario a medida que se van generando.

Los tres puntos de contacto de los diferentes hilos, y por lo tanto, las estructuras que hay que sincronizar son, los resultados que se van obteniendo del buscador, las respuestas que se van generando luego del análisis, y la persistencia de éstas respuestas, junto con la información relevante obtenida en el proceso de obtenerlas.

En el primero, cada hilo invoca al buscador de forma independiente, esto puede provocar que se obtengan dos resultados iguales en diferentes hilos, lo que implica descargar la página y pasarla por el detagger dos veces. Para evitarlo, cada hilo guarda los links y los documentos de los resultados obtenidos en una estructura compartida por todos los hilos. Cuando un hilo invoca al buscador, este le devuelve links, antes de descargar el documento correspondiente a cada link, el hilo chequea en la estructura si existe un documento para dicho link, en ese caso el documento es obtenido de la estructura; en caso contrario se descarga la página correspondiente, se la pasa por el detagger y luego se la almacena. Las operaciones para verificar si un link existe, agregarlo y obtenerlo de la estructura son accedidas concurrentemente por varios hilos, por lo tanto se declaran como mutuamente excluyentes.

Cada hilo puede generar varias respuestas, dependiendo de la cantidad de re-

sultados que se analicen y de cuántas respuestas se encuentren en cada texto analizado. Por eso, es deseable que el usuario no tenga que esperar a que un hilo termine para ver las respuestas. Por lo que, cuando se llega a obtener una respuesta, ésta es colocada en una lista. La interfaz Web esta implementada de forma que, periódicamente accede a esta lista y si hay nuevas respuestas las muestra al usuario. De ésta manera el usuario va viendo las primeras respuestas generadas mientras el sistema sigue procesando. Las operaciones que permiten agregar y obtener respuestas también son declaradas como mutuamente excluyentes.

Cuando se obtiene una nueva respuesta la misma es persistida en BD o en archivo según cual sea la opción de persistencia utilizada. En caso que la opción sea BD, la concurrencia a la misma es controlada por el motor de Base de Datos, en caso contrario el sistema debe controlar esta concurrencia ya que no pueden haber dos hilos escribiendo en el mismo archivo al mismo tiempo. Es por esto que la escritura en persistencia es declarada como mutuamente excluyente.

4.3. Acceso al buscador

La elección del buscador fue hecha en base a que se considera que Google es un buen producto en general, pero no se evaluaron diferentes buscadores y no hay razones objetivas para no haber utilizado otro.

Para acceder al buscador se reutilizó el código implementado en WebQA. En un principio se investigó la forma de acceder al servicio utilizando la API SOAP [API08] más nueva provista por Google, sin embargo, desde el 5 de Diciembre de 2006 no se dan más claves. La alternativa era utilizar Google AJAX Search, pero como es para utilizar con Java Script, nos resultó más fácil llamarla desde código Java utilizando lo ya probado por WebQA.

Luego de obtener las clases Java generadas a partir del WSDL propuesto por Google [WSD08] simplemente se invocan métodos Java para acceder al WebService, además se requiere disponer de la librería Apache Axis [AXI08].

Para hacer una búsqueda se utiliza la operación *'doGoogleSearch'* de la clase *'GoogleSearchBindingStub'*, la misma recibe como parámetros, la clave otorgada por Google, las palabras a consultar, la cantidad de resultados a retornar y el idioma al que se desea restringir los resultados, entre otros. El resultado de esta operación es un objeto del tipo *'GoogleSearchResult'*, que provee un método

'getResultElements' que retorna la cantidad de resultados solicitados.

Para cada uno de los resultados obtenidos anteriormente se puede obtener información, como por ejemplo la URL. Con la URL se pueden utilizar clases estándares de Java para obtener el documento Web correspondiente. Sin embargo, en nuestro caso se pasa la URL a los detaggers utilizados, para que ellos se encarguen de obtener el texto de Internet y procesarlo para retornar el texto plano.

4.4. Web

Como se mencionó anteriormente, el sistema permite que el usuario vaya obteniendo las respuestas a medida que son generadas por el sistema. Ésta funcionalidad se implementó utilizando la tecnología AJAX [Wik08] con JavaScript.

La metodología AJAX sirve para generar contenido dinámico en una página Web. Esto se realiza mediante eventos, scripts y rutinas que van al servidor a buscar datos, luego estos datos son usados para actualizar la página solo regenerando ciertas porciones de la misma, sin tener que volver a cargar todo el contenido del documento en el navegador.

Básicamente el sistema Eneas lo que se hace es preguntar, cada cierto tiempo, al servidor si tiene nuevas respuestas. En caso que las tenga, se actualiza la sección respuestas de la página y si es necesario también se actualiza el paginado de la misma, el usuario visualizará los cambios sin que se refresque toda la página. Cabe aclarar que la paginación permite que se visualicen hasta tres respuestas por página, por lo tanto en caso que la página ya esté mostrando dicha cantidad de respuestas el usuario solo verá cambios en el paginado.

Para cada respuesta obtenida el usuario verá los siguientes datos:

- *Puntaje otorgado por el sistema*. Se visualizarán cinco estrellas para las respuestas EXCELENTES, cuatro para las MUY BUENAS, etc.
- *Respuesta*. Además se remarcarán las palabras u oraciones que el sistema consideró relevantes para rankear la respuesta.
- *Link* a la página de donde se obtuvo la información.

Es importante aclarar que las respuestas mostradas al usuario serán las rankeadas como EXCELENTES, MUY BUENAS y BUENAS. Solo se mostrarán respuestas ACEPTABLES y DEFICIENTES en caso que no se haya encontrado ninguna respuesta que califique con los puntajes anteriores. Esto es porque se considera que es mejor mostrar pocas buenas respuestas, que mostrar muchas entre las cuales existan algunas incorrectas.

Capítulo 5

Evaluación y Ajuste de Parámetros

En este capítulo, se presenta un resumen de los resultados experimentales obtenidos por el sistema *Eneas*, descrito en los capítulos anteriores.

Primero se presenta el ajuste de los parámetros del sistema, que influyen en la calidad de las respuestas. Y luego la evaluación que fue realizada en base a un conjunto de preguntas seleccionadas en forma arbitraria. Las medidas de evaluación utilizadas son la precisión de las respuestas, y el MRR (*Mean Reciprocal Rank*).

Las preguntas utilizadas se presentan en el apéndice de evaluación 7. Las mismas abordan temas variados y no se restringen a ningún dominio particular.

5.1. Ajuste de Parámetros

El rendimiento del sistema *Eneas*, tanto en velocidad de respuesta, como en la calidad de las mismas, depende de varios factores. Por la forma en que esta estructurado, se considera que hay dos parámetros que influyen en su rendimiento.

El primero es la cantidad de resultados de Google analizados. Entre más resultados se utilicen, más documentos serán analizados y será más probable encontrar buenas respuestas. Este parámetro podría ser aumentado arbitrariamente, y en un caso extremo se analizarían todos los documentos retornados por Google que

contengan alguna palabra clave de la pregunta, aumentando la probabilidad de encontrar una respuesta correcta, sin embargo ésto aumentaría considerablemente el tiempo de respuesta. La cantidad de resultados de Google es un parámetro global a todo el sistema, por lo que utilizar un resultado más de Google implica analizar un documento más en cada hilo de ejecución.

El segundo parámetro es la cantidad de reformulaciones utilizadas. Entre más reformulaciones se utilicen, se obtendrán más resultados de Google. A diferencia del caso anterior, el aumento es tanto cuantitativo como cualitativo. Ya que al agregar una nueva reformulación, se agregan tantos documentos al análisis, como resultados de Google utilice el sistema. Por otro lado, estos nuevos documentos provienen de una búsqueda diferente, utilizando un indicador primario distinto, por lo que se agregan al análisis las expresiones causales que utilicen éste nuevo indicador.

La elección de éstos parámetros es fundamental para controlar el tiempo de respuesta. Por ejemplo, el utilizar un indicador primario para reformulaciones y un resultado de Google, implica analizar dos documentos, ya que también se debe sumar la reformulación de Wikipedia. Utilizar seis indicadores primarios y cinco resultados de Google implica analizar 31 documentos. En cada documento se pueden encontrar cero, una, o varias respuestas candidatas a analizar. Suponiendo se encuentre un promedio de tres candidatas por documento, se analizarían 93 frases por cada pregunta, lo que provoca un tiempo de respuesta bastante alto.

5.1.1. Cantidad de resultados de Google

Para evaluar cuál valor es el más adecuado para este parámetro, en cada respuesta se indica de que número de resultado de Google se obtuvo la misma. El sistema se ajusta con cinco resultados de Google, lo que provoca altos tiempos de procesamiento. Se seleccionó este valor porque por las pruebas efectuadas durante el desarrollo, se considera que es poco probable que una respuesta no se encuentre entre los primeros cinco resultados y sí se encuentre en alguno de los siguientes.

El criterio utilizado para decidir el valor óptimo de este parámetro es el de seleccionar el menor valor posible sin perder respuestas correctas. Lo esperado es que los resultados aporten respuestas correctas por igual, es decir si se eligen *n* resultados, se espera que cada uno aporte el $\frac{100}{n}$ % de las respuestas correctas.

5.1.2. Cantidad de reformulaciones

Los indicadores primarios fueron elegidos estudiando los textos y observando los resultados de los primeros prototipos. Para la selección se tuvo en cuenta que existen ciertos verbos y conectores en el idioma español que son comúnmente utilizados para formular textos causales, la inclusión de estas palabras en la lista hace más probable encontrar respuestas a preguntas causales. La lista completa de indicadores primarios seleccionados para hacer el ajuste se encuentra en el cuadro 8.1.

La forma de evaluación es igual a la mencionada en el punto anterior. Se analiza la cantidad de respuestas correctas provenientes de cada reformulación, y luego se hace un análisis de cuáles indicadores primarios son los que indican causalidad más frecuentemente y cuáles nunca la indican o lo hacen en muy escazas ocasiones. También se espera que cada formulación agregue respuestas correctas por igual.

5.2. Resultados del ajuste de parámetros

El ajuste de los parámetros se hizo con diez preguntas propuestas por nosotros en forma arbitraria, sin restringirse a ningún dominio en particular. Éstas se presentan en la sección 7.1.

5.2.1. Resultados de Google

La evaluación de parámetros se hizo con cinco resultados de Google para todo el conjunto de preguntas 7.1.

No se noto una variación en la cantidad de respuestas correctas e incorrectas según la posición del resultado. Pero como también se debe tener en cuenta los tiempos de respuesta, se decidió hacer la evaluación del sistema con tres resultados, para disminuir los tiempos de ejecución.

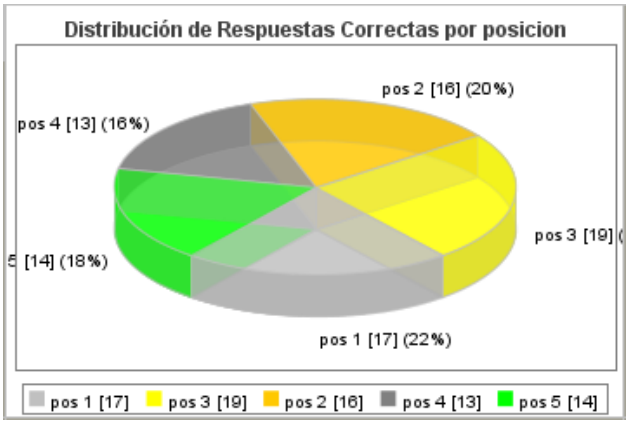


Figura 5.1: Resultados Google cinco posiciones (Correctas).

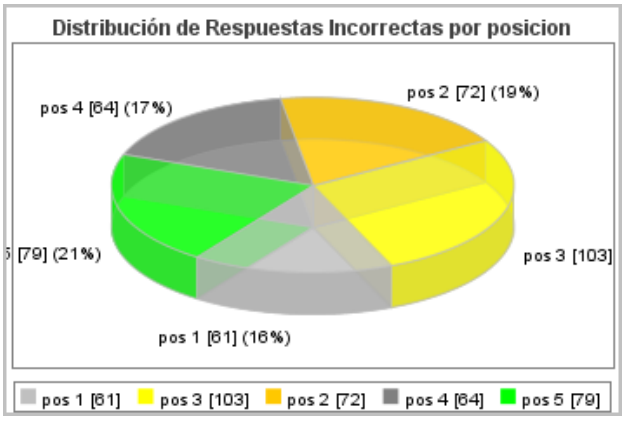


Figura 5.2: Resultados Google cinco posiciones (Incorrectas).

5.2.2. Reformulaciones

En las figuras 5.4 y 5.3 se muestran las gráficas correspondientes a las distintas reformulaciones que hace el sistema para realizar la búsqueda de respuestas.

Si observamos la gráfica que muestra las respuestas correctas por indicador y puntaje notamos que la cantidad de respuestas correctas surgidas de las reformulaciones *si y solo si*, *entonces*, y *wikipedia* son menores que las obtenidas por el resto de las reformulaciones. Por lo tanto, se decidió eliminar los indicadores *entonces* y *si y solo si*. Optamos por dejar la reformulación especial wikipedia porque ésta se ejecuta con un único resultado de Google, y por lo tanto no afecta significativamente el tiempo de ejecución del sistema.

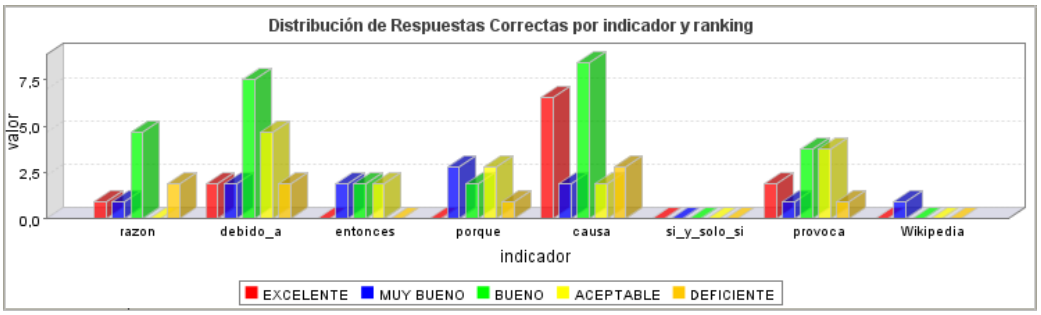


Figura 5.3: Respuestas correctas según indicadores y puntaje.

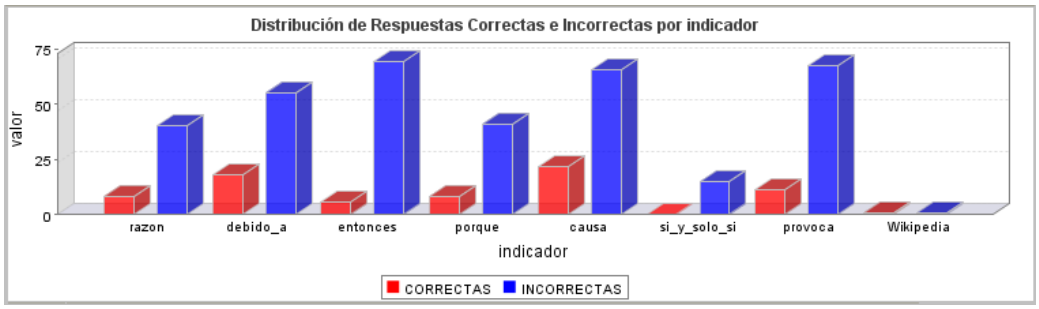


Figura 5.4: Respuestas correctas e incorrectas según indicadores.

En el cuadro 5.1 se detallan los valores presentados en la figura 5.4.

Indicador	**# respuestas correctas**	**# respuestas incorrectas**
wikipedia	1	1
por qué	9	44
debido a	19	59
provoca	12	72
razón	9	43
causa	24	71
entonces	6	74
si y solo si	0	16

Cuadro 5.1: Respuestas según indicadores

5.3. Evaluación

El objetivo de las medidas descritas anteriormente fue ajustar el sistema para obtener el mejor equilibrio posible entre calidad de las respuestas y tiempo de respuesta. Luego de hecho el ajuste, es interesante analizar los resultados teniendo en cuenta una mayor cantidad de preguntas. Estas noventa preguntas fueron propuestas por nosotros y por personas conocidas. La lista de preguntas utilizadas para la evaluación se muestra en 7.2.

5.3.1. Forma de evaluación

Se realizaron dos tipos de evaluaciones:

1. Midiendo la *Precisión* del sistema.
2. Observando que tan bien se asignan los puntajes, calculando el MRR.

Ambas se hicieron analizando en forma manual las respuestas obtenidas por el sistema *Eneas*. Las mismas eran leídas por un humano, que indicaba si la respuesta era correcta o no. Por lo tanto una respuesta correcta es aquella que satisface la necesidad de información según el criterio de un humano, independientemente de la forma en que el sistema obtuvo la misma o del puntaje que le asignó. Como ejemplo del criterio utilizado, se considera la pregunta *¿Cuáles son las causas de la lluvia ácida?*. A continuación se dan posibles respuestas y su clasificación:

- *"la lluvia se produce por la condensación del vapor de agua"* se considera incorrecta ya que se refiere a la lluvia y no a la lluvia ácida.
- *"en los últimos tiempos se habla mucho sobre las causas de la lluvia ácida"* se considera incorrecta porque no provee información sobre las causas de la lluvia ácida.
- *"la lluvia ácida se produce cuando los contaminantes presentes en la atmósfera se mezclan con las gotas de lluvia reduciendo su PH a valores menores a 5"* se considera correcta.

Las respuestas fueron recolectadas utilizando la modalidad batch, y no usando la interfaz Web, por razones de comodidad, y tiempo.

5.3.1.1. Precisión del sistema

El cálculo de la precisión, indica cuántas de las preguntas realizadas fueron contestadas correctamente. Consideramos que una pregunta ha sido contestada correctamente cuando, entre las respuestas a esa pregunta, hay al menos una que la responde.

Precisión = cantidad de preguntas contestadas correctamente/cantidad de preguntas realizadas.

5.3.1.2. Asignación del puntaje y MRR

En este caso se observa como el sistema puntúa las distintas respuestas, es decir evaluamos el comportamiento del algoritmo de asignación de puntaje. Para medir esto se calcula el MRR.

Éste se define de la siguiente forma:

$$MRR = \frac{\sum_{i=1}^{n} 1/pos(i)}{n}$$

Donde *n* corresponde a la cantidad de preguntas de prueba, y *pos(i)* indica la posición de la primera respuesta correcta para la pregunta *i*. El valor de $i/pos(i)$ será cero si no se ha encontrado ninguna respuesta. La importancia del MRR es que considera la posición en el puntaje de la respuesta correcta, a diferencia de la *Precisión*. O sea, no solo es importante que el sistema devuelva una respuesta correcta, también importa la posición en que se da la respuesta.

5.3.2. Resultados Generales

Aquí se van a presentar los resultados de la evaluación, enfocándonos en que tan bien el sistema clasifica las respuestas obtenidas.

En la gráfica 5.5 se puede ver que la mayoría de las preguntas marcadas como EXCELENTE son correctas. De un total de 78 respuestas clasificadas como EXCELENTE, 58 son correctas y 20 incorrectas.

Figura 5.5: Respuestas correctas e incorrectas según el puntaje.

Mientras se va bajando en el puntaje asignado, la proporción de correctas e incorrectas varía, es decir mientras más bajo sea el puntaje, es más probable que la respuesta sea incorrecta. Esto es razonable, y muestra que el comportamiento del algoritmo que asigna el puntaje es aceptable. En el cuadro 5.2 se detallan los valores para cada uno de los puntajes.

Puntaje	**% de respuestas correctas**	**# respuestas registradas**
EXCELENTE	74,34 %	78
MUY BUENO	44,83 %	116
BUENO	35,15 %	202
ACEPTABLE	13,91 %	496
DEFICIENTE	9,67 %	362

Cuadro 5.2: Respuestas correctas según puntaje

En la gráfica 5.6 se puede ver que un reducido porcentaje de las respuestas correctas fueron calificadas como DEFICIENTE (13 %), y que en el puntaje ACEPTABLE se encuentran el 24 % de las respuestas correctas. Esto indica que hay una cantidad importante de respuestas correctas (37 %) a las que el sistema les asigna un puntaje bajo. De todos modos, el porcentaje de respuestas correctas con puntaje medio (25 %) y alto (38 %) es mayor que el anterior.

Además si vemos la gráfica de las respuestas incorrectas 5.7, se puede ver que la clasificación se comporta mucho mejor que en el caso anterior. Los dos puntajes más bajos (ACEPTABLE Y DEFICIENTE) son asignados a casi el 80 % de las respuestas incorrectas. También es destacable el hecho de que la definición

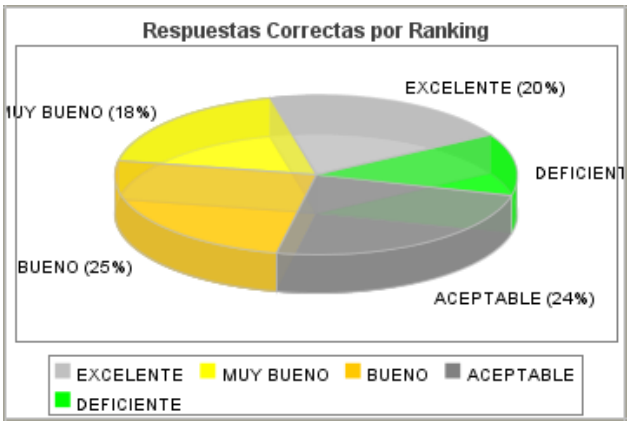


Figura 5.6: Distribución de respuestas correctas según el puntaje.

del puntaje EXCELENTE logro detectar el 20 % de las respuestas correctas, pero es lo suficientemente restrictivo como para solo aceptar el 2 % de las respuestas incorrectas.

Estos buenos resultados, sobre todo para el caso de respuestas incorrectas, evidencian que el algoritmo que clasifica las respuestas no necesita mayores modificaciones, aunque seguramente haya casos particulares que se puedan mejorar.

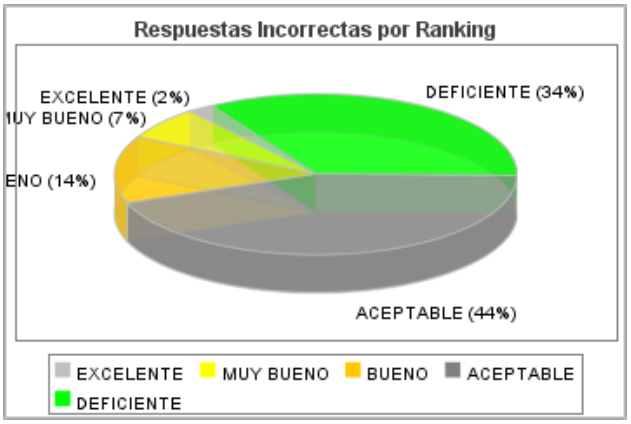


Figura 5.7: Distribución de respuestas incorrectas según el puntaje.

Se detectaron casos en los que el sistema no encuentra una respuesta pese a que el texto responde a la pregunta realizada. Un ejemplo es la pregunta *¿Por qué los ingleses manejan por la izquierda?*, la respuesta esperada sería algo como *los ingleses manejan por la izquierda porque...*, sin embargo las palabras *manejan* e

izquierda nunca se encuentran como respuesta, sino que las palabras encontradas son *conducen*, *lado izquierdo*, etc. Éstos casos, en los que las palabras de la pregunta no se encuentran en el segmento, pero igualmente se responde a la pregunta, se deben principalmente a que:

- Los verbos que se encuentran en la pregunta están conjugados distinto a los verbos encontrados en el segmento.
- En el segmento se encuentran sinónimos de las palabras de la pregunta.
- En la pregunta se encuentra alguna palabra en singular y en el segmento la misma palabra se encuentra en plural, o viceversa.
- En la pregunta se encuentra alguna palabra en femenino y en el segmento la misma palabra se encuentra en masculino, o viceversa.

Lo mencionado anteriormente se podría solucionar con análisis de correferencias, sinónimos, etc. Estas observaciones ponen de manifiesto que en un futuro se deberían utilizar más herramientas de procesamiento de lenguaje natural como clasificador de entidades nombradas, tener en cuenta sinónimos, hipónimos, hiperónimos y otros análisis morfológicos de cada palabra.

Dado que no se encontraron otros sistemas que aborden los temas que intenta atacar *Eneas*, no se tiene un punto de referencia con el cual comparar. Sin embargo, dada las características del sistema *WebQA*, que responde preguntas fácticas, que utiliza Internet como corpus y que es de dominio abierto. Y dado que una de las incógnitas planteadas consistía en evaluar si haciendo un sistema que no utilice técnicas avanzadas y complejas de procesamiento de lenguaje natural, se podían obtener buenos resultados aprovechando la gran cantidad de información que se puede encontrar en Internet, como lo hizo WebQA, pero en este caso con preguntas causales. Se decidió comparar los resultados con los de WebQA.

Sistema	**MRR**	**Precisión**
ENEAS	**0,4574**	**0,6000**
WebQA	0,4845	0,5442

Cuadro 5.3: Resultados comparativos con WebQA

Pese a que el problema que atacan ambos sistemas es diferente, y por lo tanto la forma de resolución no es comparable. Dado que se calcularon dos medidas

que se usan generalmente para sistemas de respuesta automática a preguntas, se considera que los resultados son aceptables, ya que el sistema, en general responde las preguntas planteadas.

Capítulo 6

Conclusiones y Trabajos Futuros

Con la realización del sistema *Eneas* se pudo comprobar la factibilidad de construir un sistema de respuesta automática a preguntas causales utilizando como corpus la Internet. Pese a que el corpus a utilizar no fue fijado de antemano, y se podía evaluar la posibilidad de utilizar un corpus de dominio cerrado, con documentos predefinidos, la idea siempre fue utilizar Internet, y como desde un principio se observó que se obtenían buenas respuestas, éste fue el corpus seleccionado. Dado que se contaron con pocas herramientas de procesamiento de lenguaje natural y que la respuesta automática a preguntas causales no es un tema que se haya estudiado mucho, fue necesario desarrollar una forma de encontrar las respuestas deseadas. Para lograr esto se debió, por un lado detectar la causalidad en los textos, y por otro discernir cuando un texto habla de un tema dado.

Para lograr los objetivos planteados en principio se analizó parte del corpus seleccionado. De este análisis se concluyó lo siguiente:

- Existen ciertos tipos de palabras, denominados por el equipo de proyecto como *indicadores causales*, cuya presencia en un texto expresa causalidad.

- Cuando las palabras que conforman una pregunta, se encuentran en un mismo segmento es probable que el segmento hable del mismo tema que la pregunta, y es más probable todavía cuando toda las palabras están en una misma oración del segmento. A medida que se fue desarrollando el sistema y se fue estudiando el corpus, se analizaron los resultados y se fueron mejorando los algoritmos utilizados.

De los resultados obtenidos, se puede concluir que en el idioma español se pueden encontrar indicadores que expresan causalidad con diferente intensidad, algunos expresan causalidad siempre que se encuentren en un texto, otros expresan causalidad la mayoría de las veces que se encuentran en un texto y existen otros que a veces expresan causalidad y a veces no. Por ejemplo cuando se encuentra la palabra *"causa"* en una respuesta, es más probable que sea causal que en el caso en el que se encuentra la palabra *"entonces"*.

También se pudo observar que la causalidad, además de relacionar una causa con su consecuencia, tiene un sentido, que depende del indicador. Según el indicador causal y la posición en la que se encuentren las entidades relacionadas, una puede ser causa o consecuencia.

Resultó claro que los resultados del sistema *Eneas* dependen muy fuertemente de los resultados de Google. Las preguntas que no tengan un buen resultado haciendo una búsqueda clásica en Google, tampoco van a tener un buen resultado en *Eneas*. Entre los documentos obtenidos por el buscador y analizados, puede haber documentos que contengan la respuesta y documentos que no la contengan. Por lo tanto *Eneas* debe, no solo detectar las respuestas en los documentos que las contienen, sino que también debe no detectar como respuestas textos que no la contienen.

La evaluación del sistema se realizó en base a segmentos que *Eneas* consideró candidatos a contener la respuesta. Si un segmento respondía la pregunta, pero el mismo no fue seleccionado como candidato a contener la respuesta, éste no fue considerado en la evaluación. En el futuro se podrían persistir todos los documentos encontrados por el buscador y evaluar cuántas respuestas se perdieron.

En un sistema completo de respuesta automática, se debería analizar el tipo de la pregunta, en éste caso se asumió que eran causales. Existen herramientas que podrían haberse usado para detectar el tipo de pregunta, como el POSTagger Freeling [Fre08], la desventaja de esta herramienta es que consume demasiado tiempo. Esto hace pensar que aunque se tengan disponibles herramientas avanzadas, tal vez su uso provoque una demora inaceptable en el tiempo de respuesta del software producido. Por ésto último tal vez deberían investigarse sistemas que hagan un uso completo de todas las herramientas de PLN [PLN07] disponibles sin tener en cuenta las demoras, para evaluar la calidad máxima de respuestas a la que se puede llegar y que habría que mejorar. Otra posibilidad sería procesar los textos y guardar información estructurada, para luego acceder a ésta estructura. Uno podría pensar que al utilizar la Web y hacer un análisis al vuelo, no se podría utilizar ésta técnica, pero todo depende de como se escale el sistema. Google es

un ejemplo de que se puede cachear e indizar una gran cantidad de documentos de la Internet.

Se notó que la relación causal es una relación transitiva. Si A causa B y B causa C, entonces A también se puede considerar una causa de C. En la actualidad *Eneas* no tiene en cuenta la transitividad de la relación causal, esto podría agregarse en un futuro.

También se vieron ejemplos de situaciones planteadas en [Jac99], como por ejemplo títulos o listas. Para esto sería necesario analizar no solo el texto, sino la estructura en que se presenta el mismo. Un ejemplo de aprovechamiento de la estructura fue la reformulación especial Wikipedia. Notamos que muchas respuestas se obtenían de Yahoo Answers, por lo que en un futuro se podría analizar y aprovechar la estructura de éstas páginas al igual que se hizo con las páginas de Wikipedia.

Finalmente se concluye que el sistema construido logra obtener respuestas adecuadas en muchos casos, pero todavía se puede mejorar en varios sentidos. Se considera que construir sistemas de respuestas automáticas que respondan siempre correctamente es posible pero todavía falta mucho. Las principales barreras son la calidad y accesibilidad de las herramientas auxiliares de procesamiento de lenguaje natural que se deberían usar y la capacidad de procesamiento de las computadoras actuales. Por un lado la calidad de las herramientas auxiliares impacta en el resultado del sistema, por ejemplo OpenNLP que no detecta correctamente la morfología de algunas palabras del idioma español. Por otro lado, si no es fácil utilizarlas e integrarlas al resto del software puede decidirse no usarlas, o usar otras de menor calidad pero que sean más accesibles. En nuestro caso un ejemplo sería la herramienta EuroWordNet que no fue utilizada, ya que es una herramienta con licencia restrictiva. La capacidad de procesamiento limita la cantidad de información que se puede procesar. Si en la computadora donde se ejecuta el sistema se van a poder procesar más documentos, entonces se aumentaría la probabilidad de encontrar respuestas correctas. En nuestro caso, si la capacidad de procesamiento hubiera sido mayor, podríamos haber hecho más consultas al buscador por cada pregunta, o analizado más documentos retornados por Google en cada consulta.

Capítulo 7

Apéndice Evaluación

7.1. Preguntas de Ajuste

1. ¿Por qué Plutón no es un planeta?
2. ¿Por qué se suicidan los japoneses?
3. ¿Cuáles son las causas de los huracanes?
4. ¿Cuáles son las causas de los eclipses de sol?
5. ¿Cuáles son las causas de la fiebre amarilla?
6. ¿Por qué se debe legalizar la marihuana?
7. ¿Cuáles son las causas del hipo?
8. ¿Por qué hay violencia en el fútbol?
9. ¿Cuáles son las causas de la sordera?
10. ¿Cuáles son las causas del agujero de la capa de ozono?

7.2. Preguntas de Evaluación

1. ¿Cuáles son las causas de la corrupción?

2. ¿Cuáles son las causas de la inflación?
3. ¿Cuáles son las causas del éxodo oriental?
4. ¿Cuáles son las causas del cáncer de útero?
5. ¿Cuáles son las causas del virus de la vaca loca?
6. ¿Cuáles son las causas de la enfermedad de la vaca loca?
7. ¿Cuáles son las causas de la gripe aviar?
8. ¿Cuáles son las causas del fundamentalismo islámico?
9. ¿Cuáles son las causas de la segunda guerra mundial?
10. ¿Cuáles son las causas de la primera guerra mundial?
11. ¿Cuáles son las causas de la guerra Vietnam?
12. ¿Cuáles son las causas de la guerra civil española?
13. ¿Cuáles son las causas de la revolución francesa?
14. ¿Cuáles son las causas de la revolución cubana?
15. ¿Cuáles son las causas de la revolución industrial?
16. ¿Cuáles son las causas de los terremotos?
17. ¿Cuáles son las causas de los tzunami?
18. ¿Cuáles son las causas de los sismos?
19. ¿Por qué erupcionan los volcanes?
20. ¿Cuáles son las causas de la lluvia ácida?
21. ¿Cuáles son las causas del calentamiento global?
22. ¿Por qué los árboles se deshojan en otoño?
23. ¿Por qué las plantas florecen en primavera?
24. ¿Por que el cielo es azul?
25. ¿Por qué a Microsoft le hacen juicios por monopolio?
26. ¿Por qué se instalan papeleras europeas en América Latina?

27. ¿Por qué tiraron una bomba atómica a Japón?
28. ¿Por qué la tierra es redonda?
29. ¿Por qué Marte es rojo?
30. ¿Por qué Saturno tiene anillos?
31. ¿Por qué la Tierra gira alrededor del sol?
32. ¿Por qué el día tiene 24 horas?
33. ¿Por qué Argentina está en contra papeleras?
34. ¿Por qué Gualeguaychú está en contra papeleras?
35. ¿Por qué Estados Unidos invadió Irak?
36. ¿Por qué Estados Unidos invadió Uruguay?
37. ¿Por qué se festeja Navidad?
38. ¿Por qué se construyeron pirámides?
39. ¿Por qué cristianismo se extendió en Europa?
40. ¿Cuáles son las causas del hambre?
41. ¿Por qué emigran las aves?
42. ¿Por qué las rosas son rojas?
43. ¿Por qué Júpiter es grande?
44. ¿Cuáles son las causas de la marea roja?
45. ¿Cuáles son las causas de la mareas?
46. ¿Cuáles son las causas de la malaria?
47. ¿Por qué comemos?
48. ¿Por qué Pablo se convirtió al cristianismo?
49. ¿Por qué se extinguieron los dinosaurios?
50. ¿Cuáles son las causas de la era glaciar?
51. ¿Por qué las mujeres musulmanas usan velo?

52. ¿Por qué mataron a Kennedy?
53. ¿Por qué los pingüinos no vuelan?
54. ¿Por qué los leones tiene melena?
55. ¿Cuáles son las causas del racismo?
56. ¿Por qué surgió el nazismo en Alemania?
57. ¿Cuáles son las causas de la inquisición?
58. ¿Por qué se hundió el Titanic?
59. ¿Cuáles son las causas del alcoholismo?
60. ¿Por qué los aviones vuelan?
61. ¿Por qué no se puede beber el agua de mar?
62. ¿Por qué rota la Tierra?
63. ¿Por qué los ingleses manejan por la izquierda?
64. ¿Por qué hay más personas derechas que izquierdas?
65. ¿Por qué los perros atacan a los gatos?
66. ¿Por qué el whisky es mejor añejo?
67. ¿Cuáles son las causas de la crisis del 29?
68. ¿Cuáles son las causas de la suba del petróleo?
69. ¿Por qué los animales se lamen las heridas?
70. ¿Por qué romper un espejo trae mala suerte?
71. ¿Por qué envejecemos?
72. ¿Por qué sudamos?
73. ¿Por qué se llora al cortar cebolla?
74. ¿Por qué los hombres son más altos que las mujeres?
75. ¿Por qué usamos el sistema decimal?
76. ¿Cuáles son las causas del arco iris?

77. ¿Por qué las hojas son verdes?
78. ¿Cuáles son las causas de la lluvia?
79. ¿Cuáles son las causas del cambio climático?
80. ¿Por qué cambian de pelo los animales?
81. ¿Cuáles son las causas de los calambres?
82. ¿Cuáles son las causas de la epilepsia?
83. ¿Por qué salen manchas en las uñas?
84. ¿Por qué parpadeamos?
85. ¿Cuáles son las causas de la muerte súbita?
86. ¿Por qué la fuerza de gravedad es más baja en la luna?
87. ¿Por qué ocurren los vientos?
88. ¿Por qué Artigas se retiró al Paraguay?
89. ¿Por qué los dinosaurios eran tan grandes?
90. ¿Por qué hacen erupción los volcanes?

Capítulo 8

Anexo Indicadores

Indicador	Sentido
causa	D
provoca	D
debido a	D
razón	D
porque	I
entonces	D
si y solo si	I

Cuadro 8.1: Indicadores primarios del ajuste

Indicador	Sentido
causa	D
provoca	D
debido a	D
razón	D
porque	I

Cuadro 8.2: Indicadores primarios de la evaluación

Cuadro 8.3: Indicadores terciarios

Indicador	Sentido	Completo
activ	D	0

actua	D	0
alenta	D	0
aport	D	0
apoy	D	0
asigna	I	0
asist	D	0
ayud	D	0
basad	I	0
beneficia	D	0
colabor	D	0
conduc	D	0
conduj	D	0
consec	I	0
constr	D	0
converg	D	0
cooper	D	0
crea	D	0
decid	I	0
demostr	I	0
demuestr	D	0
depend	I	0
deriv	I	0
descubr	D	0
desemboc	D	0
dispers	I	0
eman	I	0
empez	D	0
empiez	D	0
empuj	D	0
establec	D	0
estall	I	0
estimul	D	0
estremec	I	0
explo	D	0
fabric	D	0
facilit	D	0
favor	D	0
foment	D	0

forz	D	0
fruto	I	0
fuent	D	0
fuerz	D	0
funcion	D	0
hace	D	0
haci	D	0
impac	I	0
implic	D	0
imputa	I	0
incid	D	0
incit	D	0
iduc	D	0
infer	I	0
infier	I	0
influy	D	0
intensific	D	0
introd	D	0
llev	D	0
mostra	I	0
neces	I	0
oblig	D	0
ocasion	D	0
papel	I	0
part	I	0
particip	I	0
permit	D	0
peso	I	0
precipit	D	0
prob	I	0
proba	I	0
promov	D	0
prueb	D	0
relacion	I	0
reperc	I	0
requier	I	0
responsab	D	0
restaur	D	0

rol	I	0
serv	I	0
sirv	D	0
surg	I	0
suscit	D	0
viene	I	0
viniendo	I	0
vino	I	0

Cuadro 8.4: Extensiones de verbos

Verbo	**Extension**	**Sentido**	**Conector**	**Sentido Conector**
contrib	uid	D	he	I
provoc	ad	I	por	D
provoc	ad	D	hemos	I
provoc	ad	D	ha	I
desencaden	ad	I	por	D
desencaden	ad	D	han	I
desencaden	ad	D	hemos	I
desencaden	ad	D	ha	I
gener	ad	I	por	D
gener	ad	D	han	I
gener	ad	D	hemos	I
gener	ad	D	ha	I
desarroll	ad	I	por	D
desarroll	ad	D	han	I
desarroll	ad	D	hemos	I
desarroll	ad	D	ha	I
explic	ad	I	por	D
explic	ad	D	han	I
explic	ad	D	hemos	I
explic	ad	D	ha	I
contrib	uid	D	he	I

Cuadro 8.4: Extensiones de verbos

Verbo	Extension	Sentido	Conector	Sentido Conector
provoc	ad	D	han	I
contrib	uid	D	he	I
inic	ad	I	por	D
inic	ad	D	han	I
inic	ad	D	hemos	I
inic	ad	D	ha	I
impuls	ad	I	por	D
impuls	ad	D	han	I
impuls	ad	D	hemos	I
impuls	ad	D	ha	I
propici	ad	I	por	D
propici	ad	D	han	I
propici	ad	D	hemos	I
propici	ad	D	ha	I
produci	d	I	por	D
produci	d	D	han	I
produci	d	D	hemos	I
produci	d	D	ha	I
contrib	uid	D	he	I

Cuadro 8.4: Extensiones de verbos

Verbo	Extension	Sentido	Conector	Sentido Conector
provoc	ad	D	han	I
contrib	uid	D	he	I
orig	inad	I	por	D
orig	inad	D	han	I
orig	inad	D	hemos	I
orig	inad	D	ha	I
contrib	uid	I	por	D
contrib	uid	D	han	I
contrib	uid	D	hemos	I
contrib	uid	D	ha	I
provoc	ad	D	he	I
desencaden	ad	D	he	I
gener	ad	D	he	I
desarroll	ad	D	he	I
explic	ad	D	he	I
inici	ad	D	he	I
caus	ad	D	he	I
impuls	ad	D	he	I
propici	ad	D	he	I
produci	d	D	he	I
orig	inad	D	he	I
contrib	uid	D	he	I

Indicador	Sentido	Conector
debe a	I	
gracias a	I	
consecuencias de	I	
causa de	D	es
causas de	D	son
es la razón de	D	
razón de	D	es
razones de	D	son
es la causa de	D	
consecuencia	D	es
consecuencias	D	son
son las razones de	D	
son las causas de	D	
su causa es	I	
sus causas son	I	
deben a	I	
debidos a	I	
debió a	I	
es por	I	
son por	I	
consecuencias de	I	
causada por	I	
causadas por	I	
debieron a	I	
causados por	I	
causado por	I	
provocado por	I	
provocados por	I	
provocada por	I	
provocadas por	I	
es la razón de	D	
son las razones de	D	

Cuadro 8.5: Frases primarias

Indicador	Sentido	Completo
provoc	D	0
desencaden	D	0
gener	D	0
determin	D	0
produce	D	0
desarroll	D	0
resulta	I	0
implic	D	0
contrib	D	0
inici	D	0
entonces	D	1
cataliz	I	0
caus	D	0
produj	D	0
impuls	D	0
orig	D	0
propici	D	0
produci	D	0
producto	I	0
atribuid	I	0

Cuadro 8.6: Indicadores secundarios

Indicador	Sentido
Asi que	D
efecto de	I
parte de	I
parten de	I
por lo que	D
por lo tanto	D
por tanto	D
si y solo si	I
tener por	I
tenido por	I
ya que	I

Cuadro 8.7: Frases secundarias

Glosario

Adjetivo El adjetivo es la palabra que acompaña al sustantivo o nombre para determinarlo o calificarlo; expresa características o propiedades del sustantivo. Ejemplo: El Libro Verde, El Libro Grande.

Adverbio El adverbio es la clase de palabra que actúa como núcleo del Sintagma adverbial. En la morfología del español suelen ser invariables o con una variabilidad muy pequeña (algunos admiten sufijos: cerquita, lejísimos, lejitos).

API Conjunto de operaciones que proveen funcionalidades utilizables al desarrollar software.

Búsqueda de Respuestas (Question & Answering) Herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas, en base a la comprensión de preguntas y a partir del análisis de documentos escritos en lenguaje natural.

Corpus lingüístico Es un conjunto, normalmente muy amplio, de ejemplos reales de uso de una lengua. Estos ejemplos pueden ser textos (típicamente), o muestras orales (normalmente transcritas).

POSTagger Part Of Speech Tagger, marca las palabras de un texto, con el tipo correspondiente a cada una

Extracción de Información (EI) Consiste en extraer, de un texto o un conjunto de textos, entidades, eventos y relaciones entre ellos.

Lexicón Es el "diccionario" en el que se registran las palabras que conoce un hablante.

Palabra clave Palabra que es de relevancia dentro de una pregunta o fragmento.

Sujeto En la parte de la gramática denominada sintaxis, el sujeto es el sintagma de una oración cuyo núcleo concuerda en número con el núcleo del sintagma verbal o nominal predicativo.

Sustantivo En la Gramática del español, el nombre o sustantivo es la clase de palabra que puede funcionar, con articulo o sin él, como sujeto de la oración.

Verbo El verbo es la categoría gramatical que funciona como núcleo del predicado y suele indicar movimiento (llevar, correr, etc.), acción (pensar, creer, etc.) o estado (existir, vivir, permanecer, ser, estar, parecer etc.).

WordNet Lexicón semántico para el idioma inglés.

Bibliografía

[API08] Google SOAP Search API. http://code.google.com/apis/soapsearch/, 2008.

[AXI08] AXIS. http://ws.apache.org/axis/, 2008.

[Cer92] Farid Cerbah. Integrating qualitative reasonning and text planning to generate causal explanations. In *Proceedings of the fifteenth International Conference on Computational Linguistics*, volume II, pages 617–623. Integrating qualitative reasonning and text planning to generate causal explanations, 1992.

[CGI08] Sebastián Calvo, Ariel Guevara, and Paula Imbriani. Documentación técnica eneas, Julio 2008.

[Cha93] J. Charlet. « acte : A casual model-based knowledge acquisition tool », in second generation experts systems. berlin heidelberg : Springer-verlag. 1993.

[CIM07] Daniel Castelo, Jorge Isi, and Sebastián Martinez. Informe final web-qa, Junio 2007.

[Dan85] Laurence Danlos. Génération automatique de textes en langues naturelles. 1985.

[Dan88] Laurence Danlos. « connecteurs et relations causales », langue française. 1988.

[Dan95] Laurence Danlos. « un exemple d'étude linguistique du discours : les relations causales directes ».traitement des langues naturelles, 5ème école d'été organisée par le cnet. 1995.

[Ent07] Reconocimiento De Entidades. http://en.wikipedia.org/wiki/named-entity-recognition, Mayo 2007.

[FRA+06] Sergio Fernández, Sandra Roger, Antonia Aguilar, Antonio Ferrández, and Pilar López-Moreno. Nueva propuesta de desambiguación de sentidos de palabras para nombres en un sistema de búsqueda de respuestas. Technical report, Universidad de Alicante, 2006.

[Fre08] Freeling. http://garraf.epsevg.upc.es/freeling/, 2008.

[Gar98] Daniela Garcia. *Analyse automatique des textes pour l'organisation causale des actions Réalisation du système informatique COATIS*. PhD thesis, Université de Paris-Sorbonne (Paris IV) I.S.H.A. (Institut des Sciences Humaines Appliquées), Mayo 1998.

[Gir03] R. Girju. Automatic detection of causal relations for question answering. 2003.

[Gon03] José Luis Vicedo González. *Recuperación de información de alta precisión: Los sistemas de búsqueda de respuestas*. 2003.

[HP07] HTML-Parser. http://htmlparser.sourceforge.net/, Agosto 2007.

[Jac99] Agata Jackiewicz. *L'expression de la causalité dans les textes Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle*. PhD thesis, Université de Paris-Sorbonne (Paris IV) I.S.H.A. (Institut des Sciences Humaines Appliquées), Febrero 1999.

[Jav08] Java6. Java standar edition 6 http://java.sun.com/javase/6/, 2008.

[Jer07] Jericho. http://jerichohtml.sourceforge.net/doc/index.html, Agosto 2007.

[KCN] CHRISTOPHER S.G. KHOO, SYIN CHAN, and YUN NIU. Extracting causal knowledge from a medical database using graphical patterns.

[KH96] R Kozlowska-Heuchin. *Etude comparée des connecteurs en français et en polonais*. PhD thesis, Université Paris, Junio 1996.

[Naz94] A. Nazarenko. *Compréhension du Langage naturel : le problème de la causalité*. PhD thesis, Université Paris, 1994.

[OAS] OASIS. Organization for the advancement of structured information standards http://www.oasis-open.org/home/index.php, 2008.

[Ope08] OpenNLP. http://opennlp.sourceforge.net/, 2008.

[Par06] Wonsil Park. *Sémantique et représentation formelle de verbes qui expriment les relations causales : augmenter, conduire, créer, déclencher, diminuer, entraîner, entretenir, pousser, provoquer*. PhD thesis, Université de Paris-Sorbonne, 2006.

[PLN07] PLN. http://es.wikipedia.org/wiki/procesamiento-de-lenguajes-naturales, Mayo 2007.

[Rey93] Chantal Reynaud. Acquisition and validation of expert knowledge by using causal models. pages 517–539, 1993.

[SnGC05] Fernando Martínez Santiago and Miguel Ángel García Cumbreras. Identificación de formas lógicas en el caso del español: propuesta de un modelo basado en reglas y aprendizaje automático. Technical report, Universidad de Jaén, Abril 2005.

[UIM08] UIMA. http://incubator.apache.org/uima/, 2008.

[Val07] Cross Validation. http://en.wikipedia.org/wiki/cross-validation, Mayo 2007.

[Wik08] WikipediaESP. Ajax, http://es.wikipedia.org/wiki/ajax, 2008.

[WSD08] Google SOAP Search API WSDL. http://api.google.com/googlesearch.wsdl, 2008.