Validación de Respuestas Reconociendo la Implicación Textual

M.C.C. Alberto Téllez Valero

Coordinación de Ciencias Computacionales

Instituto Nacional de Astrofísica, Óptica y Electrónica

Tesis sometida como requisito parcial para obtener el grado de

Doctor en Ciencias Computacionales

Septiembre 2009

Resumen

Un sistema de búsqueda de respuestas es un tipo de motor de búsqueda que permite recuperar información concreta a partir de grandes colecciones de documentos de texto. La característica de este tipo de sistemas es que la petición del usuario es expresada como una pregunta para la cual piezas específicas de información (*i.e.,* fragmentos de texto en lugar de documentos completos) son retornadas como una respuesta. Desafortunadamente el desempeño actual de estos sistemas en muchos casos no ha resultado ser el esperado. Tal como ocurre en el español, donde hasta la fecha el mejor sistema de esta clase sólo ha contestado correctamente a un 53 % de las preguntas de un conjunto de prueba en este idioma. Con el propósito de mejorar dicho desempeño en esta tesis se presenta un método de validación de respuestas. Este método permite crear un sistema que etiqueta como válida o errónea a cada una de las respuestas de los sistemas de búsqueda de respuestas. En particular, el sistema de validación de respuestas utiliza un clasificador basado en aprendizaje supervisado para etiquetar cada respuesta. La característica principal del sistema es que emplea atributos novedosos para evaluar la implicación textual junto con atributos que verifican la compatibilidad entre pregunta–respuesta. Esta combinación de atributos le permite al sistema seleccionar respuestas válidas para las preguntas mientras descarta las erróneas. Los experimentos en preguntas y respuestas en español muestran la efectividad del sistema. Los resultados obtenidos son motivadores, éstos superan a los alcanzados por otros sistemas similares. Pero sobre todo, estos resultados permiten incrementar el mejor desempeño alcanzado en la búsqueda de respuestas en español. Esto último principalmente por utilizar el sistema de validación de respuestas para combinar las respuestas de múltiples sistemas de búsqueda de respuestas.

Abstract

A question answering system is a kind of search engine that allows retrieving concrete information from large text document collections. The characteristic of this type of systems is that requests from users are expressed as questions for which specific pieces of information (*i.e.,* text fragments instead of complete documents) must be returned as answer. Unfortunately, in many cases, the current performance of these systems has not been as expected. Such is the case of spanish, where, to date, the best system of this kind has only correctly answered to 53 % of the questions from a given test set in this language. In order to improve this performance in this thesis is presented an answer validation method. This method allows creating a system that labels as valid or erroneous each one of the answers from the question answering systems. In particular, the answer validation system uses a classifier based on supervised learning to label the answers. The principal characteristic of the system is that it uses novel attributes to evaluate the textual entailment along with attributes that verify the compatibility between question–answer. This combination of attributes allows the system to select valid answers for the questions while it discards the erroneous ones. The experiments in a set of questions and answers in spanish show the effectiveness of the system. The obtained results are encouraging since they outperform the results achieved by other similar systems; but mainly, because they allow increasing the best performance reached in spanish question answering. This last result mainly produced by the application of the answer validation system to combine the answers from multiple question answering systems.

A Noemi, Judith y Porfirio

Agradecimientos

Quisiera expresar mi más profundo y sincero agradecimiento a los directores de esta tesis, el Dr. Manuel Montes y Gómez y el Dr. Luis Villaseñor Pineda por su ayuda, soporte, amistad y buen humor a lo largo del presente trabajo.

A la Dra. Angélica Muñoz Meléndez, el Dr. Aurelio López López, el Dr. Gustavo Rodríguez Gómez, el Dr. Saúl Eduardo Pomares Hernández y el Dr. Anselmo Peñas Padilla, el comité revisor de esta tesis, cuyos oportunos comentarios y críticas sirvieron para llevar por buen camino la investigación. También quisiera hacer una mención muy especial a todos mis compañeros, sin excepción alguna, del Laboratorio de Tecnologías del Lenguaje de INAOE sin cuyos ánimos y apoyo habría resultado muy difícil la consecución del presente trabajo. Y a todos mis compañeros del Departamento de Ciencias Computacionales que me han acompañado y alentado durante mi trayectoria investigadora en este instituto, INAOE, al cual agradezco todas las facilidades y apoyos recibidos durante mi estancia. En ese sentido, también agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de estudios que me otorgo (no. 171610) y sin la cual no hubiera sido posible hacer este trabajo.

Sobre todo, quiero agradecer a Noemi, mi mujer, su apoyo incondicional y la confianza que siempre ha depositado en mí. Ella es la que más ha sufrido las consecuencias de la intensa dedicación que ha requerido este trabajo. Espero poder compensarla del tiempo robado durante la realización de esta tesis.

Y a mis padres, Judith y Porfirio, gracias por contagiarme su dedicación y gusto al trabajo. Estoy en deuda con ustedes.

Índice general

Índice de figuras

ÍNDICE DE FIGURAS

Índice de tablas

ÍNDICE DE TABLAS

Capítulo 1

Introducción

En este capítulo se describe la finalidad de la presente investigación. Los detalles del problema abordado así como la motivación para afrontarlo son presentados en la sección 1.1. Posteriormente, la sección 1.2 muestra el propósito de la tesis incluidos la hipótesis y los objetivos planteados. Finalmente la estructura del resto del documento es presentada en la sección 1.3

1.1. Problema y motivación

En las últimas décadas se ha presentado un crecimiento exponencial en la cantidad de textos electrónicos disponibles. Un claro ejemplo de este gran aumento de información es la Web, cuyo contenido día a día sigue incrementando[1]. Este auge de información digital ha contribuido a que actualmente podamos acceder desde una computadora a casi cualquier tema en el mundo. Desafortunadamente, el exceso de contenidos ha dado origen al problema de *cómo localizar la información deseada* entre todo ese inmenso volumen de documentos digitales.

[1]La Web o WWW (del inglés *World Wide Web*) es el sistema de documentos, llamados páginas web, interconectados por enlaces de hipertexto que se ejecutan en Internet (fuente Wikipedia, `http://es.wikipedia.org/wiki/Web`). Debido al diario crecimiento de la Web actualmente se desconoce su tamaño. Un reporte del año 2008 señaló que se conocían un trillón de enlaces a diferentes páginas web (más detalles en el sitio `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`).

1. Introducción

Las técnicas de *recuperación de información* representan hoy en día un avance importante para tratar con el problema de ubicar la información requerida. La aplicación más exitosa de estas técnicas es la *recuperación de documentos*, la cual a partir de una gran colección de textos digitales permite localizar documentos relevantes a una petición[1]. Los motores de búsqueda en la Web son casos sobresalientes de esta aplicación. Por ejemplo Google que ya desde el año 2006 con acceso a una colección de más de 25 billones de páginas Web respondía a más de 400 millones de consultas al día[2].

Evidentemente los sistemas de recuperación de documentos han hecho posible procesar los grandes volúmenes de información textual digital. Aunque si la petición del usuario es una pregunta específica, estos sistemas no pueden responder de manera concisa a dicha petición. Por ejemplo, un sistema de este tipo no puede retornar como respuesta el fragmento de texto exacto `Jacques Chirac` para contestar a la pregunta `¿Quién era el presidente de Francia durante las pruebas de armas nucleares en el Pacífico Sur?` De hecho, los sistemas de recuperación de documentos no fueron propiamente desarrollados para afrontar este tipo de peticiones. Puesto que una vez que el sistema retorna al usuario una lista de documentos relevantes a su consulta, ella o él aún tienen que inspeccionar los textos para encontrar la información deseada.

Un estudio reportado por Radev *et al.* (2001) muestra que de las peticiones planteadas a un motor de búsqueda en la Web aproximadamente un 8 % son preguntas específicas. Mientras que una parte importante del resto de las consultas (planteadas como un conjunto de palabras) son hechas por los usuarios teniendo en mente también una pregunta. Este estudio refleja la importancia de proveer recursos que permitan acceder de una forma más sofisticada a información concreta, en especial para evitarle al usuario tener que leer todo un documento cuando su petición de información no lo requiere.

Acertadamente desde hace ya cinco décadas se desarrollan sistemas de *búsqueda de respuestas* (BR)[3]. Estos sistemas se caracterizan porque la petición del usuario es expresada como una pregunta para la cual piezas específicas de información (*i.e.,* fragmentos de texto en lugar de documentos completos) son retornadas como una respuesta. En la

[1] Una petición o consulta es el conjunto de palabras que expresan la necesidad del usuario.

[2] Fuente Wikipedia, `http://en.wikipedia.org/wiki/Google_search`

[3] Existen registros de sistemas de BR desde finales de los años 60's, aunque el desarrollo de estos sistemas empieza a tener auge hasta finales de los años 90's. Esto principalmente motivado por el crecimiento y popularidad de la Web así como por el surgimiento de foros donde se evalúan a los sistemas de BR.

actualidad para el desarrollo de estos sistemas ya se tiene identificada una arquitectura típica, además de una amplia variedad de técnicas que han sido propuestas para su implementación. Un ejemplo de un sistema de BR es DFKI, el cual permite buscar respuestas en la Web[1].

Sin embargo, a pesar de los esfuerzos el desempeño hasta ahora alcanzado por la mayoría de los sistemas de BR no ha resultado ser el esperado. Por ejemplo lo que sucede en el español, donde de acuerdo a las evaluaciones hechas a estos sistemas ninguno ha logrado contestar a más de un 53 % de las preguntas de prueba en este idioma[2]. Sin duda, para satisfacer adecuadamente a peticiones sobre información concreta es necesario mejorar el desempeño actualmente alcanzado por muchos de los sistemas de BR, en particular de aquellos sistemas aplicados a idiomas populares como el español[3]. El problema es *cómo mejorar dicho desempeño.*

De las diversas propuestas que existen para mejorar el desempeño de los sistemas de BR, una de las más recientes consiste en utilizar más de un sistema para responder a las preguntas. Esto se debe a que se ha mostrado que las respuestas de diferentes sistemas de BR se complementan. Es decir, ciertos sistemas contestan correctamente a preguntas que otros sistemas no pueden responder, pero a su vez estos últimos responden correctamente a preguntas donde los primeros fallan en sus respuestas. Por lo tanto, la combinación de las respuestas correctas de distintos sistemas de BR complementarios permite incrementar el máximo número de preguntas contestadas correctamente por las respuestas de un único sistema. Tal es el caso del español, donde el máximo número de preguntas contestadas correctamente por un sistemas de BR (un 53 %) puede aumentar en un 24 % (alcanzando un 77 %) si se combinan sus respuestas con las respuestas de otros sistemas para este idioma (Magnini *et al.*, 2007). Ahora el problema es *cómo combinar de forma automática tales respuestas.*

Entre los métodos que ya se tienen para combinar las respuestas de diversos sistemas de BR complementarios, el más popular se basa en preferir para cada pregunta

[1]El sistema DFKI está libremente accesible en el sitio `http://experimental-quetal.dfki.de`

[2]Desde el 2003 se realiza una evaluación anual de sistemas de BR aplicados al español. Los resultados de tales evaluaciones se encuentran en (Magnini *et al.*, 2003), (Magnini *et al.*, 2004), (Vallin *et al.*, 2006), (Magnini *et al.*, 2007), (Giampiccolo *et al.*, 2007a) y (Forner *et al.*, 2008).

[3]El español es el tercer lenguaje con más presencia en la Web y el segundo más utilizado para hacer peticiones sobre ésta con Google (fuentes Estadísticas Mundiales de Internet `http://www.internetworldstats.com/stats7.htm` y CDT `http://www.cdtinternet.net/modules/news/article.php?storyid=3201`).

la respuesta más frecuente entre las respuestas de todos los sistemas. Esta técnica proporciona buenos resultados cuando los sistemas a complementar en general obtienen un alto desempeño y por lo tanto hay una alta frecuencia en las respuestas correctas; por ejemplo en el inglés donde se reportan sistemas que responden correctamente a más de un 60 % de las preguntas de prueba (*p. ej.,* los resultados presentados en (Voorhees, 2004)). Sin embargo, en casos como el español donde en promedio las respuestas de los sistemas son erróneas en más de un 70 % de las preguntas (Peñas *et al.*, 2008), esta técnica no produce los resultados deseados debido a la poca frecuencia que tienen las respuestas correctas.

Recientemente un nuevo enfoque conocido como la *validación de respuestas* (VR) ha sido propuesto para tratar con el problema de combinar las respuestas de diversos sistemas de BR. El propósito de un sistema de VR es clasificar automáticamente como válida o errónea a cada una de las respuestas producidas por los sistemas de BR (Peñas *et al.*, 2007). De manera que después de clasificar las respuestas sea posible seleccionar de aquellas etiquetadas como válidas, si existen, una para contestar a la pregunta. Entre las ventajas de este nuevo enfoque están que no depende de la frecuencia de las respuestas ni de algún tipo de información acerca de los sistemas a combinar. Es decir, un sistema de VR en cada ejecución sólo toma como entrada la pregunta y la respuesta de un sistema de BR sin importar como fue obtenida.

Cabe hacer notar que un sistema de VR además de permitir combinar sistemas de BR complementarios, también puede ser fácilmente integrado a la arquitectura de un único sistema de BR. En este caso después de que el sistema localiza la respuesta a una pregunta dada, dicha respuesta es validada y si se detecta que es errónea se le pide al sistema que encuentre una nueva respuesta. Este proceso puede ser repetido hasta encontrar una respuesta que sea clasificada como válida o bien hasta decidir que el sistema fue incapaz de contestar a la pregunta.

En (Gómez-Soriano *et al.*, 2005) se da evidencia de la importancia de integrar una validación de respuestas en la arquitectura de los sistemas de BR. El trabajo muestra que el desempeño del sistema descrito en (Pérez-Coutiño *et al.*, 2006) puede aumentar en un 18 % (respondiendo correctamente a un 60 % de las preguntas de un conjunto de prueba) si en lugar de proporcionar una respuesta por pregunta, que es lo que comúnmente se espera de estos sistemas, mejor proporciona veinte. En este caso

integrar la validación de respuestas permitiría al sistema de BR filtrar una respuesta correcta y evitar retornar al usuario las veinte respuestas por pregunta.

Los trabajos reportados por Tatu *et al.* (2007) y Glöckner *et al.* (2007) han demostrado que la validación de respuestas permite complementar exitosamente sistemas de BR para los idiomas inglés y alemán, respectivamente. Así como los trabajos de Harabagiu & Hickl (2006) y Ferrández *et al.* (2008a) han mostrado el éxito de incorporar una validación de respuestas en la arquitectura de un sistema de BR; en ambos casos aplicado al idioma inglés. En todos estos trabajos los sistemas de VR empleados se caracterizan por aplicar un *reconocimiento de la implicación textual* (RIT) para validar las respuestas. Bajo esta perspectiva, el problema de validar una respuesta se convierte en el problema de determinar cuándo, dados dos fragmentos de texto, el significado de un fragmento puede razonablemente ser inferido o textualmente implicado desde el significado del otro fragmento (Dagan *et al.*, 2005). Desafortunadamente, en otros idiomas donde también existen las condiciones para mejorar el desempeño de los sistemas de BR con la validación de respuestas, el enfoque basado en el RIT aún no ha tenido éxito.

Motivado por mejorar el desempeño actual de los sistemas de BR, en esta tesis se planteó desarrollar un sistema de VR inspirado como otros en el enfoque de reconocer la implicación textual. En particular, un sistema de VR aplicado al español uno de los idiomas donde se ha mostrado que una correcta validación de respuestas permite mejorar de manera importante el máximo desempeño hasta ahora alcanzado por un sistema de BR en este lenguaje. Esto a pesar de que las respuestas de los sistemas de BR en español en más de un 70 % son erróneas, haciendo que la tarea de identificar las respuestas válidas sea aún más complicada que en otros idiomas donde existe una mayor proporción entre las respuestas válidas y erróneas.

1.2. Propuesta de la tesis

Tomando en cuenta que el desempeño de los sistemas de BR puede ser mejorado con una validación de respuestas aplicada tanto para *i*) combinar las respuestas de diferentes sistemas de BR complementarios, así como para *ii*) filtrar las respuestas de un único sistema de BR; el propósito de esta tesis fue mejorar el desempeño de los

Tabla 1.1: Respuestas obtenidas por sistemas de BR para una pregunta

Sistema	*Respuesta*	*Pasaje de soporte*
1	ONU	sanciones económicas impuestas por la ONU contra Irak desde que invadió Kuwait en 1990
2	Kuwait	Kuwait fue un cercano aliado de Irak durante la guerra Irak-Iran
3	Kuwait	Kuwait fue invadido por Irak en 1990

sistemas de BR en español por medio de la validación de sus respuestas. Por lo tanto, la pregunta de la que partió esta investigación fue *cómo desarrollar un sistema de VR que permita mejorar el desempeño de los sistemas de BR en español.*

Inspirado principalmente por los trabajos de Tatu *et al.* (2007) y Glöckner (2007) se propusó validar las respuestas reconociendo la implicación textual. Esto significa que, una respuesta es clasificada como válida si se distingue que el significado de una afirmación compuesta por la pregunta junto con dicha respuesta se infiere del significado de la respuesta y su contexto de donde fue extraída. Al fragmento de texto formado por la respuesta con su contexto se le conoce como *pasaje de soporte*, el cual también es proporcionado por los sistemas de BR para justificar su respuesta.

Por ejemplo, para la pregunta `¿A qué país invadió Irak en 1990?` la tabla 1.1 muestra la respuesta (junto con su pasaje de soporte) de tres distintos sistemas de BR. Para cada una de estas respuestas en la tabla 1.2 se presenta una afirmación compuesta con la pregunta. Estas afirmaciones permiten ver que sólo el significado de la última se infiere por el significado del correspondiente pasaje de soporte (en la primera afirmación no es posible inferir del pasaje de soporte que invadieron a ONU, mientras que en la segunda afirmación el pasaje de soporte no menciona nada sobre una invasión realizada por Irak). Por lo tanto después de validar las respuestas, aquella obtenida por el sistema tres es la que debe ser seleccionada para responder a la pregunta.

Inconvenientemente los sistemas de VR que han alcanzado su propósito (*i.e.,* mejorar el desempeño de los sistemas de BR) se caracterizan por recurrir a recursos sofisticados para el procesamiento del lenguaje natural. Es decir, para distinguir si reconocen o no la implicación textual estos sistemas se valen de un análisis profundo del lenguaje (*p. ej.,* resolución de anáfora), así como de diversas fuentes de conocimiento extra (*p. ej.,* ontologías con relaciones de hiponimia o hiperonimia como WordNet). Muchos de

Tabla 1.2: Afirmaciones para evaluar la implicación textual en la validación de respuestas

Afirmación de la pregunta junto con la respuesta	*Pasaje de soporte*
Irak invadió a ONU en 1990	sanciones económicas impuestas por la ONU contra Irak desde que invadió Kuwait en 1990
Irak invadió a Kuwait en 1990	Kuwait fue un cercano aliado de Irak durante la guerra Irak-Iran
Irak invadió a Kuwait en 1990	Kuwait fue invadido por Irak en 1990

estos recursos no están disponibles o no alcanzan los mismos resultados para idiomas como el español.

Al parecer el problema de reconocer la implicación textual requiere de tratar de entender el lenguaje para su solución. Esto se puede ver en los métodos propuestos para su estudio, los cuales recurren cada vez más a herramientas de un análisis profundo del lenguaje como se muestra en los reportes presentados en (Dagan *et al.*, 2005), (Bar-Haim *et al.*, 2006) y (Giampiccolo *et al.*, 2007b). Sin embargo, cabe hacer notar que a pesar de que en el español hay una escasez de recursos para el procesamiento profundo del lenguaje, los sistemas de BR aplicados a este idioma son capaces de localizar sus respuestas. Por lo tanto, la hipótesis que se planteó en esta tesis fue que:

> **Las respuestas de los sistemas de BR en español se pueden validar reconociendo la implicación textual con un análisis superficial del lenguaje.**

En otras palabras, se supuso que un análisis léxico-sintáctico del texto, sin llegar a lo semántico, permite la validación de respuestas en español. Además de que dicha validación de respuestas mejora el desempeño de los sistemas de BR en este idioma.

Entonces, el objetivo general de la investigación fue el siguiente:

> **Mejorar de manera significativa el máximo desempeño alcanzado en la búsqueda de respuestas en español mediante la validación de respuestas.**

Donde por *mejorar de manera significativa* se entiende que una prueba de significancia estadística confirma que la mejora lograda no es producto de la casualidad.

Respecto a los objetivos particulares, éstos fueron los siguientes:

- **Desarrollar un sistema de VR para el español basado en el reconocimiento de la implicación textual.**
- **Aplicar el sistema de VR desarrollado para combinar las respuestas de sistemas de BR complementarios y mejorar así el máximo desempeño alcanzado individualmente por cada sistema.**
- **Incorporar el sistema de VR desarrollado a la arquitectura típica de un sistema de BR para mejorar su desempeño.**

En general, la investigación aquí presentada pretende formar parte de esa visión recientemente propuesta de complementar dos áreas de estudio como son la búsqueda de respuestas y el reconocimiento de la implicación textual, mediante la validación de respuestas. En particular, el trabajo intenta mostrar *cuál es el alcance de un análisis superficial del lenguaje en la validación de respuestas en español.*

1.3. Estructura del documento

El resto del documento está organizado de la siguiente manera:

- *Capítulo 2.* En este capítulo se presentan los conceptos básicos para abordar el contenido de la tesis, los cuales incluyen conocimientos de la búsqueda de respuestas y del reconocimiento de la implicación textual.

- *Capítulo 3.* Este capítulo muestra el trabajo relacionado a la investigación. El cual consiste en describir los métodos que están siendo utilizados por los sistemas de VR basados en reconocer la implicación textual. En el capítulo también se incluye un análisis del tipo de implicación textual que intentan reconocer los sistemas de VR aplicados al español y se discute sobre lo oportuno de utilizar ciertos niveles de análisis lingüístico para resolver el problema.

- *Capítulo 4.* La descripción del método propuesto se presenta en este capítulo. En el cual se detalla la arquitectura propuesta para un sistema de validación y selección de respuestas, esto último útil para incorporar el sistema de VR a los sistemas de BR. Durante el capítulo también se hace énfasis de las diferencias del método con el trabajo relacionado.

- *Capítulos 5 y 6.* La evaluación del método propuesto se resume en estos capítulos. En el capítulo 5 se muestra la evaluación intrínseca del método que consiste en medir su capacidad para distinguir las respuestas válidas de las erróneas. Mientras que en el capítulo 6 se presenta la evaluación extrínseca del método la cual muestra su utilidad para mejorar el desempeño de los sistemas de BR.

- *Capítulo 7.* Finalmente, en este capítulo se resumen las aportaciones de la presente investigación las cuales se ubican en la validación de respuestas. También al final del capítulo se describe el trabajo futuro y las líneas de investigación abiertas.

Capítulo 2

Conceptos básicos

Este capítulo presenta una descripción general de las dos áreas de estudio involucradas en la tesis, la búsqueda de respuestas y la implicación textual. En particular, para cada una de estas áreas en las secciones 2.1 y 2.2 se describe su entorno así como la arquitectura general de los sistemas aplicados a cada una de estas tareas.

2.1. La búsqueda de respuestas

La *búsqueda de respuestas* (BR) se puede definir como la tarea que, dada una colección de documentos, tiene la finalidad de encontrar respuestas concretas a necesidades precisas de información. Con la característica de que dichas necesidades se expresan como preguntas en lenguaje natural (Maybury, 2004).

Aunque históricamente la idea de crear computadoras capaces de contestar a nuestras preguntas ha formado parte de los objetivos de la inteligencia artificial, es hasta la última etapa de los 90's cuando los sistemas de BR comienzan a multiplicarse. Esto impulsado en gran medida por el surgimiento de foros de evaluación para la tarea.

El primer foro en incluir su evaluación fue el TREC (La conferencia de recuperación de textos, del inglés *Text REtrieval Conference*) en el año 1999, un foro especializado en tareas de recuperación de información para el idioma inglés. Posteriormente, en el año 2001, el foro NTCIR (La colección de prueba NII para sistemas de recuperación

de información, del inglés *NII Test Collection for IR Systems*) incluyó la evaluación de la tarea para idiomas asiáticos como el japonés y el chino. Finalmente, otro foro que decidió incluir esta tarea en sus evaluaciones fue el CLEF (el foro de evaluación multilingüe, del inglés *Cross-Language Evaluation Forum*), que a partir del año 2003 busca impulsar el desarrollo de sistemas de BR para los idiomas de Europa, entre estos el idioma español. Los foros de evaluación actualmente siguen vigentes y son realizados una vez por año.

2.1.1. Condición de la búsqueda de respuestas

Antes de presentar la condición actual de la búsqueda de respuestas es necesario mostrar su complejidad. A dicha complejidad se le suele dividir en cuatro niveles tomando en cuenta las necesidades del usuario (Carbonell *et al.*, 2000). En la tabla 2.1 se describe cada uno de estos niveles.

De acuerdo a los resultados de los sistemas de BR en los foros de evaluación se puede decir que la situación actual de esta tecnología aún se encuentra en su nivel uno. Es decir, las preguntas que hasta hoy tratan de responder estos sistemas corresponden a las de un usuario casual. A este tipo de preguntas se les ha clasificado como factuales y de definición (la tabla 2.2 muestra algunos ejemplos de estas preguntas). Las *preguntas factuales* son aquellas que esperan como respuesta un dato concreto como una fecha, una cantidad o el nombre de algo o alguien. Mientras que las *preguntas de definición* son aquellas que esperan como respuesta un párrafo con la descripción de una persona, organización o concepto.

Cabe mencionar que si bien el problema de buscar respuestas para preguntas de nivel uno está aún vigente, algunos sistemas de BR ya intentan responder a preguntas de un nivel superior como son las de lista y las enlazadas. Las *preguntas de lista* son aquellas que esperan como respuesta una lista de datos generalmente del tipo factual, las cuales corresponden a preguntas de un recopilador de información (nivel dos). Por ejemplo, las preguntas `¿Qué actores son los protagonistas de ‘‘El Bueno, el Feo y el Malo’’?`, `¿Cómo se llaman las bolas que se utilizan en Quidditch?` y `¿Cuáles son los ingredientes de la sangría?`

Respecto a las *preguntas enlazadas*, dada una pregunta llamada la cabecera, existen otras preguntas conocidas como dependientes que se mantienen en el contex-

Tabla 2.1: Clasificación de la complejidad de la búsqueda de respuestas

Nivel	*Descripción*
1	*El usuario casual.* Un usuario con necesidades de información puntual acerca de hechos concretos y cuya contestación puede encontrarse en un único documento expresada, generalmente, de forma simple (*p. ej.*, `¿Cuál es la capital de China?`).
2	*El recopilador de información.* Un usuario con preguntas cuya contestación requiere del proceso de recopilar varias piezas de información, posiblemente en múltiples documentos, para su posterior combinación como respuesta final (*p. ej.*, `¿Qué países tienen frontera con Brasil?`).
3	*El periodista.* Un usuario que tiene el objetivo de redactar un artículo relacionado a un evento determinado (*p. ej.*, un terremoto en la ciudad de Shanghai), entonces sus preguntas son tanto de datos concretos del suceso (*p. ej.*, `¿Cuál fue su intensidad?`) así como de información histórica que le permita enmarcar el evento (*p. ej.*, `¿Cuántos terremotos ha habido anteriormente en la zona?`). En este nivel es necesario mantener en contexto las series de preguntas por parte del usuario, además de que la información puede aparecer en diferentes idiomas.
4	*El analista profesional.* Un usuario experto en temas concretos que requiere obtener conclusiones y tomar decisiones. Por ejemplo, las preguntas de un analista de la policía que intuye cierta conexión entre dos grupos terroristas: `¿Qué evidencia hay de conexión, comunicación o contacto entre estos dos grupos terroristas o sus miembros conocidos?`, `¿Cuándo y dónde planean realizar alguna acción conjunta?`

Tabla 2.2: Preguntas que tratan de responder los sistemas de BR

Preguntas factuales	*Acerca de*
¿Quién era el protagonista de la película ‘‘Siete años en el Tíbet’’?	*Persona*
¿En qué país nació el Papa Juan Pablo II?	*Localidad*
¿Qué altura tiene la Torre Eiffel?	*Medida*
¿Cuántos habitantes tiene Longyearbyen?	*Cantidad*
¿Qué multinacional francesa cambió su nombre por el de Grupo Danone?	*Organización*
¿Cuándo fue la coronación oficial de Isabel II?	*Fecha*
¿Cómo se le llama también al Síndrome de Down?	*Otro*
¿Qué organismo presidió Simón Peres *después de morir Isaac Rabin*?	*Organización (con restricción)*
¿Cuántos habitantes tenía Hong Kong *en 1993*?	*Cantidad (con restricción)*
¿Cómo se llama la colección de pinturas que hizo Goya *entre 1819 y 1823*?	*Otro (con restricción)*
Preguntas de definición	*Acerca de*
¿Quién es Iosif Kobzon?	*Persona*
¿Qué es la quinua?	*Objeto*
¿Qué es la Asociación por la Paz?	*Organización*
¿Qué es el Big Bang?	*Otro*

to de la pregunta cabecera. Tal como ocurre con las preguntas de un usuario periodista (nivel tres); por ejemplo, la pregunta cabecera `¿En qué colegio estudia Harry Potter?` y sus preguntas dependientes `¿Cuál es el lema del colegio?`, `¿En qué casas está dividido?` y `¿Quién es el director del colegio?`

Los resultados muestran que el bajo desempeño de los sistemas decrece aún más con preguntas de un nivel superior al primero. Por ejemplo en el español, donde los resultados de evaluación reportados en Giampiccolo *et al.* (2007a) y Forner *et al.* (2008) muestran una considerable disminución del desempeño de los mejores sistemas en las preguntas enlazadas (como máximo un sistema de BR sólo a contestado correctamente a un 18 % de estas preguntas).

2.1.2. Arquitectura típica de un sistema de búsqueda de respuestas

Típicamente, en la arquitectura de un sistema de BR se consideran tres procesos básicos (ver figura 2.1); éstos son los siguientes: *i*) el *análisis de la pregunta*, *ii*) la *recuperación de documentos/pasajes* y *iii*) la *extracción de la respuesta* (Vicedo, 2003). A continuación se dan detalles de cada proceso.

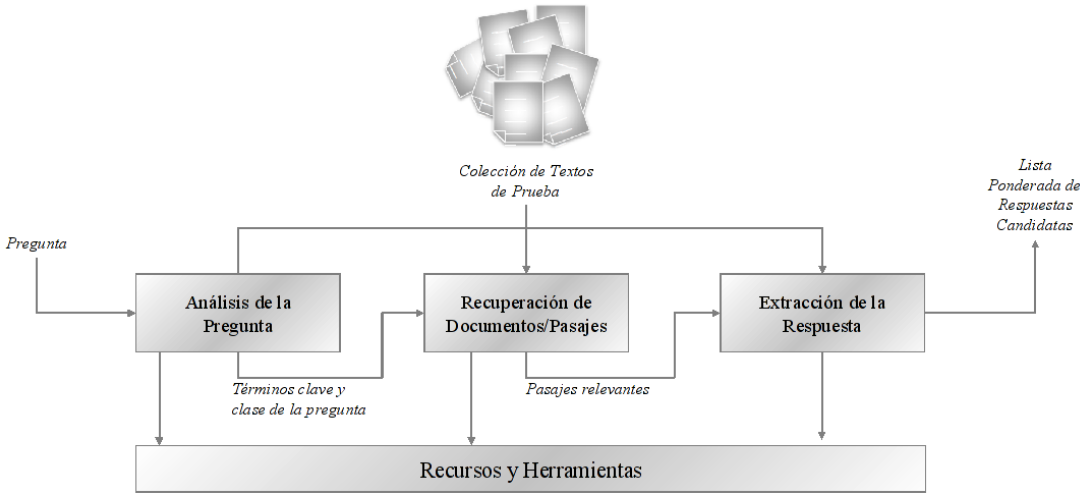

Figura 2.1: Arquitectura típica de los sistemas de búsqueda de respuestas.

Análisis de la pregunta

El objetivo principal de esta primera etapa es obtener de la pregunta sus términos clave y su clase.

Los *términos clave* son palabras o frases consideradas importantes para encontrar la respuesta. Comúnmente estos términos son todas aquellas palabras que resultan de eliminar en la pregunta los símbolos de interrogación, la partícula interrogativa (`Qué`, `Quién`, `Cuándo`, `Dónde`, `Cuánto`, etc.) y las palabras vacías[1]. Por ejemplo, en la pregunta `¿A qué país invadió Irak en 1990?`, los términos clave son `país, invadió, Irak` y `1990`. En algunos sistemas de BR el valor de importancia de cada término clave es ponderado, por ejemplo dar más valor a los verbos y nombres propios (`invadió, Irak`) o a restricciones en la pregunta (`1990`) que a sustantivos comunes (`país`).

La *clase de la pregunta* es una etiqueta que se le asigna a la pregunta y que refleja el tipo de respuesta que se espera obtener. Los métodos actuales para clasificar las preguntas varían desde los basados en emparejamiento de patrones que generalmente son construidos manualmente, hasta aquellos basados en técnicas de aprendizaje automático comúnmente supervisado. La granularidad de las clases asignadas por el clasificador también varía. Por ejemplo, para la pregunta antes citada (`¿A qué país invadió Irak en 1990?`) algunos sistemas de BR simplemente le asocian la categoría general *FACTUAL* para indicar que es un hecho concreto, en contraste con los sistemas que asocian una categoría más específica como por ejemplo *FACTUAL-NOMBRE_PROPIO-LOCACIÓN*.

Recuperación de documentos/pasajes

En esta segunda etapa, dada una colección de documentos, el objetivo es recuperar fragmentos de texto relevantes a los términos clave de la pregunta. El resultado es una lista de pasajes ordenados en base a su relevancia (generalmente entre más términos clave contiene un pasaje, éste es más relevante). En la lista los pasajes con mayor relevancia son los que aparecen primero. Los pasajes con un valor de relevancia menor a un umbral predeterminado en los sistemas son descartados de la lista.

[1] Palabras funcionales en el texto sin un significado del contenido del mismo, como sucede con preposiciones y conjunciones. En contraste con los verbos o sustantivos considerados como parte de los términos de contenido.

Tabla 2.3: Pasajes relevantes para una pregunta

Número	*Pasaje de soporte*
1	...El ministro ruso de Asuntos Exteriores tiene previsto llegar hoy a Bagdad para estudiar las ‘‘medidas constitucionales’’ que se deben tomar para reconocer oficialmente la soberanía y la independencia de Kuwait, país invadido por Irak en 1990 ...
2	...El general Hasan es la personalidad más importante del régimen iraquí que deserta del país, desde que Irak invadió Kuwait en agosto de 1990 ...
3	...Carlos Cardoen, quien dejó de fabricar armas en 1990 tras la invasión de Kuwait por parte de Irak ...
4	...Irak sigue en posesión de unos 9.000 equipos militares incluidos misiles y cohetes que robó a Kuwait durante su invasión y ocupación de el emirato en 1990 ...
5	...El régimen de Bagdad invadió Kuwait en agosto de 1990 y proclamó al emirato decimonovena provincia de Irak ...

Para lidiar con la variabilidad del lenguaje, en la recuperación de los pasajes se prefiere incluir una expansión de los términos clave (p. ej., sinónimos) así como trabajar con las raíces morfológicas de las palabras[1]. Esto con el propósito de darle mayor cobertura a la recuperación.

Para ejemplificar el resultado de esta etapa, la tabla 2.3 presenta una pequeña lista de pasajes relevantes para la pregunta `¿A qué país invadió Irak en 1990?`

Adicionalmente a los métodos que dada la pregunta recuperan los pasajes, existen sistemas de BR con enfoques predictivos que *a priori* a la pregunta tratan de encontrar la información relevante. Este tipo de sistemas utilizan patrones construidos manual o semi-automáticamente que, principalmente, capturan en la colección de documentos de prueba las relaciones entre *ENTIDAD–ENTIDAD* y *ENTIDAD–DEFINICIÓN*. Entonces, el emparejamiento de los patrones permite reconocer tanto pasajes relevantes como posibles respuestas a futuras preguntas. Por ejemplo, para preguntas del tipo `¿Quién invadió a` *ENTIDAD_2*`?` y `¿A quién invadió` *ENTIDAD_1*`?`, se tienen patrones como:

[1]Las palabras sin inflexiones; por ejemplo la raíz morfológica de `invadió` e `invadido` es `invadir`.

⟨*ENTIDAD_1*⟩ `invadió` ⟨*ENTIDAD_2*⟩
⟨*ENTIDAD_2*⟩ `fue invadido por` ⟨*ENTIDAD_1*⟩
`invasión de` ⟨*ENTIDAD_2*⟩ `por parte de` ⟨*ENTIDAD_1*⟩

El resultado de empatar estos patrones —y de tomar el texto empatado— es una tabla con registros compuestos por los campos *ENTIDAD_1*, *ENTIDAD_2* y *pasaje de soporte* (generalmente la oración donde se logró emparejar el patrón). Por lo tanto para la pregunta `¿A quién invadió Irak en 1990?` las respuestas candidatas son los fragmentos de texto localizados en el campo *ENTIDAD_1* de los registros en la tabla que contienen en su campo *ENTIDAD_2* el texto `Irak`.

Los resultados con el enfoque predictivo generalmente obtienen mayor *precisión*[1] pero menor *cobertura*[2] que los obtenidos con el método no predictivo. Este problema de cobertura se debe a su pobre generalización. Es decir, para contestar a la infinidad de preguntas posibles, en el enfoque predictivo es necesario construir patrones por cada pregunta y para todas las variantes del lenguaje que introducen la respuesta.

Extracción de la respuesta

El objetivo de esta última etapa es localizar y extraer la respuesta a la pregunta a partir de los pasajes relevantes (o de los registros relevantes en el enfoque predictivo). Si la lista de pasajes (o registros) está vacía, la respuesta otorgada por el sistema es un *nil* para indicar que no se consiguió la respuesta. Lo cual puede ocurrir en casos donde la pregunta no tiene respuesta en la colección de documentos de prueba o en ninguna otra colección; por ejemplo, la pregunta `¿Cuál es la capital del país de Nunca Jamás?`

Probablemente el recurso más utilizado para la extracción de la respuesta es el *reconocimiento y clasificación de entidades*. Mediante este recurso en una primera fase se identifica y clasifica a las entidades de los pasajes. Posteriormente, de todas las entidades se toman como respuestas candidatas aquellas que coinciden en clase con la pregunta. Por ejemplo, de los pasajes relevantes presentados en la tabla 2.3, para la pregunta `¿A qué país invadió Irak en 1990?` tenemos las respuestas candidatas `Bagdad` (en dos pasajes) y `Kuwait` (en cinco pasajes), la entidad `Irak` se descarta por ser parte de la pregunta.

[1] Porcentaje de respuestas correctas del total de sus respuestas.
[2] Porcentaje de preguntas contestadas del total de las preguntas.

Finalmente, de las respuestas candidatas se selecciona la respuesta final. La estrategia clásica de selección es tomar la respuesta más frecuente. En nuestro ejemplo `Kuwait` es la respuesta seleccionada. Algunas características extra a la frecuencia son el valor de relevancia de los pasajes y la distancia de la respuesta candidata a cada uno de los términos clave que aparecen en el pasaje de soporte. Es decir, además de la respuesta más frecuente, los sistemas de BR prefieren seleccionar aquella respuesta que proviene del pasaje más relevante y que en un mínimo contexto contiene a los términos clave junto con la respuesta. En caso de que ninguna de las respuestas candidatas satisfaga los criterios de selección, los sistemas prefieren no entregar ninguna respuesta (*i.e.,* responden un *nil*).

En la actualidad es común que los sistemas empleen técnicas de aprendizaje automático supervisado para seleccionar una respuesta entre sus candidatas, esta técnica permite combinar fácilmente diversas características para la selección de la respuesta. También cabe mencionar a los sistemas de BR que tratan al problema de selección de la respuesta como un problema de demostración lógica, donde se prefiere como respuesta final aquella que sea más probable de deducir de la información en la pregunta y el pasaje que la soporta.

2.1.3. Validación de respuestas en la búsqueda de respuestas

Entre los diversos esfuerzos para incrementar el desempeño de los sistemas de BR se encuentran aquellos que buscan extender la arquitectura típica de estos sistemas. Dos de esas extensiones son los llamados sistemas de BR multi-flujo y los sistemas de BR con validación de respuestas. A continuación se dan detalles de cada una de estas extensiones, las cuales están relacionadas a los objetivos de esta tesis.

Sistemas de BR multi-flujo

Un *sistema de BR multi-flujo* es el resultado de combinar superficialmente diferentes sistemas de BR complementarios. En otras palabras, para una pregunta dada, un sistema de BR multi-flujo selecciona de las respuestas de diversos sistemas de BR (los flujos) una que considera es la correcta, o bien retorna un *nil* si considera que ninguna de las respuestas otorgadas por los flujos es válida.

Estos sistemas son el contraste de los conocidos como *meta sistemas de BR*, los cuales en un único sistema de BR (un flujo) internamente combinan diferentes técnicas dentro de los procesos de la arquitectura típica; por ejemplo, los meta sistemas de BR reportados por Pizzato & Molla-Aliod (2005) y Chu-carroll *et al.* (2003) que, entre otras cosas, combinan diversas técnicas de recuperación de información dentro del proceso de recuperar los documentos/pasajes.

Actualmente la mayoría de sistemas de BR multi-flujo son adaptaciones de técnicas multi-flujo utilizadas en la recuperación de documentos (Belkin *et al.*, 1995; Lee & Ho, 1997). En estos sistemas dado un conjunto de respuestas proporcionadas por los sistemas de BR, la respuesta final es la más redundante en el conjunto o bien la otorgada por el sistema en que más confianza se tiene. La tabla 2.4 describe los métodos que siguen estas ideas.

Recientemente un nuevo método está siendo utilizado en los sistemas de BR multi-flujo, este método se basa en utilizar la validación de respuestas para combinar superficialmente los diferentes sistemas de BR complementarios. Las ventajas de este nuevo enfoque es que no depende de la redundancia de las respuestas o de la confianza en los sistemas como se describe en (Rodrigo *et al.*, 2008). La figura 2.2 muestra la arquitectura de un sistema de BR multi-flujo basado en la validación de respuestas. En la figura se muestra como n sistemas de BR proporcionan una respuesta para la pregunta de entrada, entonces con la validación de respuestas se detectan las respuestas válidas del conjunto de respuestas y se proporciona una como salida. En este caso n puede ser cualquier número de sistemas de BR disponibles, por ejemplo el sistema descrito en (Glöckner *et al.*, 2007) combina tres sistemas de BR para el idioma alemán.

Sistemas de BR con validación de respuestas

Analizando los resultados de sistemas de BR, en especial los aplicados al idioma inglés, se observa que aquellos que aplican algún tipo de validación de respuestas mejoran sus resultados (*p. ej.*, los sistemas de BR con los mejores resultados en la evaluación reportada en (Voorhees & Dang, 2005)). Comúnmente la validación empleada por estos sistemas se basa en medir, en grandes colecciones de textos como la Web, la redundancia que existe de los términos de la pregunta junto con los de la respuesta candidata. Entonces, a mayor redundancia mayor posibilidad hay de que la respuesta candidata

Tabla 2.4: Métodos comúnmente utilizados por los sistemas de BR multi-flujo

Ordenamiento ligero. Éste selecciona la respuesta de acuerdo a la confianza que tiene en cada uno de los flujos (los sistemas de BR). En otras palabras, éste prefiere seleccionar la respuesta del flujo en que confía más y que su respuesta no es *nil* (en el caso de que sea *nil* elige la respuesta de aquel flujo del resto en que confía más, esto se repite hasta tener una respuesta). La confianza en cada uno de los flujos es estimada midiendo su exactitud en una colección de preguntas de entrenamiento. Algunos sistemas de BR multi-flujo basados en este enfoque son descritos por Clarke *et al.* (2002) y Jijkoun & de Rijke (2004).
Caballo negro. Éste, al igual que el ordenamiento ligero, selecciona la respuesta de acuerdo a la confianza que tiene en cada uno de los flujos. Pero en este caso, la confianza es asignada por tipos de preguntas. Es decir, para las preguntas factuales prefiere las respuestas de un sistema distinto al que prefiere para las preguntas de definición, esto tomando en cuenta que el mismo sistema no obtuvo los mejores resultados en ambos tipos de preguntas. Jijkoun & de Rijke (2004) describen un sistema de BR multi-flujo basado en este enfoque.
Coro. Éste selecciona la respuesta basado en su repetición a través de los diferentes flujos. Esto es, para cada pregunta este método prefiere la respuesta más frecuente entre todas las respuestas candidatas. En el caso de que dos o más respuestas tengan la misma máxima frecuencia, éste elige aleatoriamente una de estas respuestas con máxima frecuencia. Sistemas de BR multi-flujo que implementan este enfoque son descritos en (Burger *et al.*, 2002; de Chalendar *et al.*, 2002; Jijkoun & de Rijke, 2004; Rotaru & Litman, 2005; Roussinov *et al.*, 2005).
Coro Web. Éste selecciona la respuesta basado en el número de páginas Web, recuperadas por Google, que contienen los términos de la pregunta (sin la partícula interrogativa) junto con los términos de la respuesta. Similar al método previo (el coro), en el caso que dos o más respuestas obtengan la misma máxima calificación, la respuesta final es elegida aleatoriamente de esas con los mejores valores. Este enfoque fue propuesto por Magnini *et al.* (2001) para los sistemas de BR y fue subsecuentemente evaluado en un sistema de BR multi-flujo en (Jijkoun & de Rijke, 2004).
Coro-Caballo negro. Éste considera una combinación de criterios tomados de los enfoques coro y caballo negro. En este enfoque híbrido se seleccionan las respuestas basadas en su repetición a través de los diferentes flujos. Cuando varias respuestas obtienen la misma máxima frecuencia, éste aplica el método de caballo negro —en las respuestas con la máxima frecuencia— para seleccionar la respuesta final. El sistema de BR multi-flujo descrito en (Jijkoun & de Rijke, 2004) es un ejemplo de este tipo de enfoque.

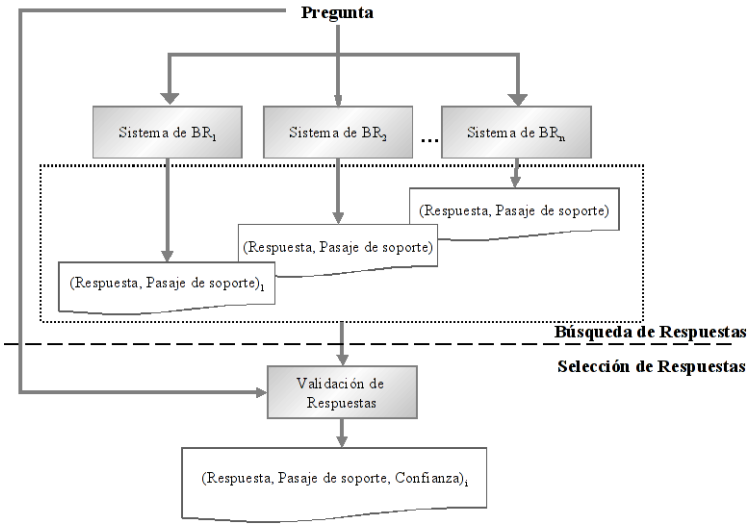

Figura 2.2: La validación de respuestas en la búsqueda de respuestas multi-flujo.

sea válida; esto bajo la hipótesis de que es más común que los términos de la pregunta co-ocurran con las respuestas válidas que con las erróneas. El sistema descrito por Magnini *et al.* (2001) es un ejemplo de este tipo de validación de respuestas incorporada en un sistema de BR.

Sin embargo, utilizando fuentes externas de documentos no es posible determinar si el pasaje de soporte justifica a la respuesta. Por ejemplo, en la pregunta `¿Quién es Diego Armando Maradona?` y su respuesta candidata `futbolista argentino`, si consultamos la Web con Google encontramos que aproximadamente 58,300 páginas contienen los términos `Diego Armando Maradona` y `futbolista argentino`, ésta es suficiente redundancia para confirmar que la respuesta es válida. Pero, si observamos su pasaje de soporte `el futbolista argentino Claudio Caniggia, amigo de Diego Armando Maradona, será el próximo delantero del Roma italiano`, encontramos que es una validación incorrecta (*i.e.*, no todos los amigos de Claudio Caniggia son futbolistas y además argentinos).

Más recientemente comenzaron a aparecer algunos sistemas que aplican una validación de respuestas más robusta que la basada en redundancia. Esta nueva clase de validación de respuestas está inspirada en el reconocimiento de la implicación textual. Ejemplos de estos sistemas son los descritos por Harabagiu & Hickl (2006) y Ferrández *et al.* (2008a) para el idioma inglés, los cuales extienden la arquitectura tradicional

de los sistemas de BR por incorporar un proceso extra de validación y selección de respuestas, como se muestra en la figura 2.3. La ventaja de este nuevo enfoque es que además de verificar que las respuestas son correctas, también revisa si están justificadas por su pasaje de soporte.

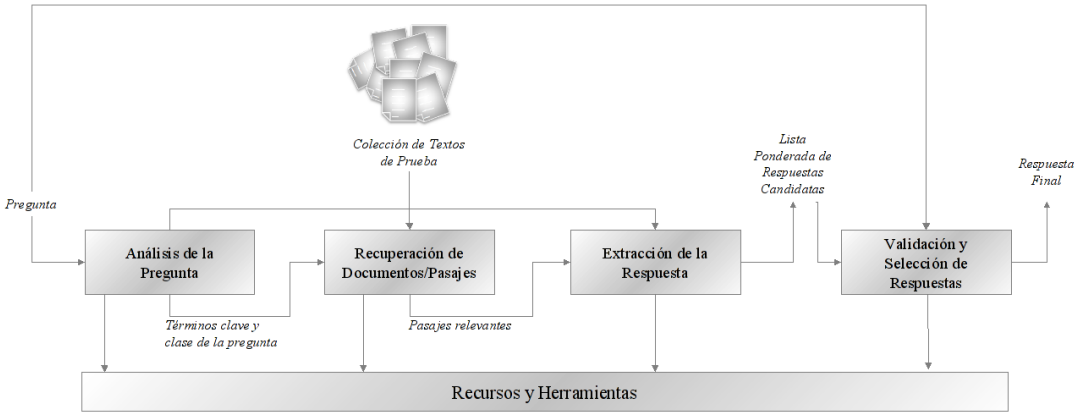

Figura 2.3: La validación de respuestas en un sistema de búsqueda de respuestas - Además de los tres procesos básicos (análisis de la pregunta, recuperación de documentos/pasajes y extracción de la respuesta) también se le agrega un cuarto proceso para la validación y selección de respuestas.

2.2. La implicación textual

El término *implicación textual* se utiliza para indicar la situación en la que la semántica[1] de un texto se puede inferir de la semántica de otro texto, ambos escritos en lenguaje natural (Dagan *et al.*, 2005). En otras palabras, si la verdad de un enunciado implica la verdad de otro se dice que existe una implicación textual. Por ejemplo, en las siguientes oraciones se puede ver que la semántica de la segunda se infiere de la semántica de la primera:

[1]La semántica consiste en interpretar el significado de los enunciados generados por la sintaxis y el léxico.

1) `Yahoo adquirió Overture`
2) `Yahoo posee Overture`

Entonces, la tarea del *reconocimiento de la implicación textual* (RIT) consiste en determinar si hay o no una implicación textual entre dos textos. Es decir, el RIT es un problema de clasificación binaria. A los textos involucrados en esta tarea se les conoce como el *texto* (T) y la *hipótesis* (H), donde T es el implicador y H el implicado.

La automatización del RIT puede ubicarse hoy en día como un reto para el *Procesamiento del Lenguaje Natural* (PLN). Aunque existen trabajos previos relacionados principalmente a la detección de paráfrasis (Bosma & Callison-Burch, 2007), los verdaderos desafíos comenzaron apenas a descubrirse. Principalmente con el surgimiento de foros relacionados a la tarea, como fue el caso del primer foro de evaluación RTE-PASCAL (del inglés *The Recognizing Textual Entailment Challenge-Pattern Analysis, Statistical Modeling and Computational Learning*) o el Taller EMSEE-ACL (del inglés *Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Association of Computational Linguistic*), ambos ocurridos en el año 2005 y enfocados al idioma inglés.

La importancia del RIT es su amplia variedad de aplicaciones, las cuales van desde la recuperación de documentos hasta la generación de resúmenes multi-documento (Dagan *et al.*, 2005). Actualmente entre los métodos aplicados al RIT se puede detectar que son dos los procesos básicos que intervienen, *i*) el *análisis de similitud* y *ii*) la *decisión del reconocimiento*. En las siguientes subsecciones se describe brevemente cada uno de estos procesos.

2.2.1. Análisis de similitud entre T y H

El propósito de este proceso es medir la semejanza entre el texto T y la hipótesis H. De manera que el valor de semejanza permita establecer si la semántica de H se infiere o no de la semántica de T. Dicho valor de semejanza generalmente se obtiene por contar las coincidencias de H con T.

Por ejemplo, para el caso de `T:Yahoo adquirió Overture` y `H:Yahoo posee Overture` se tiene que dos de los tres términos coinciden (`Yahoo` y `Overture`), pero falta reconocer que "adquirió" y "posee" coinciden. En esta situación se necesita establecer

una relación semántica entre las palabras (*i.e.,* establecer que se refieren al mismo concepto "compró"). Algo que hasta el momento es un problema abierto en esta tarea.

Otros métodos buscan hacer más preciso el análisis de similitud al incluir un análisis lingüístico más sofisticado. Un caso es el análisis de dependencias, que se usa para detectar las relaciones sujeto-acción-objeto en las oraciones. Ahora las coincidencias no sólo se miden por los términos, también por sus relaciones. Por ejemplo, en los textos T y H mencionados en el párrafo anterior los verbos "adquirió" y "posee" tienen como sujeto y objeto los mismos términos (`Yahoo` y `Overture`); por lo tanto además de coincidir en dos términos T y H también coinciden en sus relaciones de dependencias.

2.2.2. Decisión del reconocimiento

La forma más simple de decidir si se reconoce o no la implicación textual es utilizar *restricciones y valores por defecto*, los cuales consisten en determinar a partir de cuántas coincidencias entre T y H se puede concluir que existe la implicación textual.

Por otro lado hay propuestas que aplican una *clasificación automática sobre características de similitud*. Comúnmente estas propuestas utilizan métodos de aprendizaje supervisado para construir un clasificador, el cual para decidir si hay o no una implicación textual entre T y H toma como entrada los valores que representan la similitud entre este par de textos.

También existen métodos que aplican un *demostrador lógico* para revelar si ocurre o no el reconocimiento de la implicación. El problema esencial de este enfoque es que la baja tasa de coincidencias le afecta más que a otros métodos. Esto se debe a que la implicación se reconoce sólo si es posible inferir lógicamente la hipótesis a partir del texto. Sin embargo, incluir el conocimiento extra necesario en este caso puede ser más natural que en los demás métodos, además de ser capaz de deducir nuevo conocimiento a partir del que se tiene (*i.e.,* conocimiento implícito).

Capítulo 3

Validación de respuestas analizando la implicación textual

En esta sección se resume el trabajo relacionado a la tesis, el cual tiene que ver con los métodos de validación de respuestas (VR) que siguen el enfoque de reconocer la implicación textual (RIT). La sección 3.1 brevemente describe en qué consiste dicho enfoque. Posteriormente, en la sección 3.2 se muestran los procesos típicos de los métodos de VR que siguen este enfoque. Finalmente, en la sección 3.3 se discute sobre lo oportuno de los métodos hasta ahora propuestos para validar las respuestas de los sistemas de BR en español.

3.1. Función de un sistema de VR

Un sistema de *validación de respuestas* (VR) es un sistema capaz de emular la evaluación humana de los sistemas de BR (Peñas *et al.*, 2008). En otras palabras, un sistema de VR debe decidir cuándo la respuesta de un sistema de BR es válida o errónea para contestar a una pregunta dada. Por ejemplo, de las tres respuestas en la tabla 3.1 obtenidas por diferente sistemas de BR para contestar a la pregunta `¿A qué país invadió Irak en 1990?`, un sistema de VR debe ser capaz de identificar

que sólo la segunda respuesta es válida, tal que además de ser correcta está justificada por su pasaje de soporte.

Para lograr su propósito, actualmente la mayor parte de los sistemas de VR se basan en métodos que tratan de reconocer si existe o no la implicación textual entre el pasaje de soporte (el texto T) y una afirmación de la pregunta junto con la respuesta (la hipótesis H). De manera que si estos sistemas reconocen que T implica textualmente a H, entonces concluyen que la respuesta es válida para contestar a la pregunta. En la descripción de tales sistemas es posible detectar una arquitectura básica, así como un conjunto de métodos generales para su implementación. Los cuales se detallan en la siguiente sección.

3.2. Arquitectura de un sistema de VR basado en el RIT

Bajo el enfoque de reconocer la implicación textual, la arquitectura de un sistema de VR tiene tres procesos fundamentales: la *generación de la hipótesis*, el *análisis de similitud entre T y H* y la *clasificación de la respuesta*. A continuación se dan detalles de cada uno de estos procesos.

3.2.1. Generación de la hipótesis

La hipótesis H comúnmente se define como una afirmación compuesta por la pregunta junto con la respuesta. En los trabajos reportados la manera más simple de generar H es tomar la pregunta (sin los símbolos de interrogación) y concatenarle, al principio o al final, la respuesta candidata. Por ejemplo, para la pregunta y primera respuesta en la tabla 3.1 una hipótesis puede ser H `:A qué país invadió Irak en 1990 Kuwait`.

Evidentemente la forma antes descrita de construir H produce una sentencia que no refleja la afirmación deseada en el análisis de una implicación textual. Sin embargo, esta hipótesis ya permite medir la similitud entre T y H necesaria en el RIT. El sistema descrito en Cumbreras *et al.* (2007b) es un ejemplo de un sistema de VR que genera H de esta forma.

En contraste al método anterior, en (Peñas *et al.*, 2006) se muestra una forma de construir una H afirmativa que conserva la semántica que se espera de una respuesta

Tabla 3.1: Entrada dada a un sistema de validación de respuestas

Pregunta:	¿A qué país invadió Irak en 1990?
Respuesta:	Kuwait
Pasaje de soporte:	Irak - Kuwait - aniversario prensa irak critica eeuu y kuwait en cuarto aniversario crisis El Cairo, 2 ago (efe). - La prensa iraquí critica duramente hoy, martes, la política ‘‘conspirativa’’ de Estados Unidos de Kuwait respecto a Irak, su fuerza militar y sus recursos económicos, y dice que el ejército iraquí ‘‘se vio a atacar obligado (Kuwait) para acabar con el complot estadounidense contra nosotros’’. En el cuarto aniversario de la invasión iraquí de Kuwait, la prensa, co
Respuesta:	Kuwait
Pasaje de soporte:	Irak invadió Kuwait en agosto de 1990, y menos de dos semanas después se lo anexionó convirtiéndolo de la ONU en Kuwait. ‘‘La zona de Al Abdali está libre de agricultores iraquíes desde el pasado 28 de febrero’’, dijo Kubaj en una declaración transmitida hoy a la prensa kuwaití. De acuerdo con la nueva demarcación de la línea fronteriza entre Irak y Kuwait, de 210 kilómetros de largo, la frontera kuwaití ha sido desplazada en algunos puntos unos 600 metros en el interior
Respuesta:	ONU
Pasaje de soporte:	El primer ministro subrayó también los esfuerzos hechos para restaurar la solidaridad y conseguir la reconciliación entre los países árabes, así como para lograr el levantamiento de las sanciones económicas impuestas por la ONU contra Irak desde que invadió Kuwait en 1990

válida. Este método consiste en crear manualmente, para cada pregunta, un patrón que al ser instanciado con la respuesta candidata forma la hipótesis. Por ejemplo, para la pregunta de la tabla 3.1 se puede tener el siguiente patrón `Irak invadió el país de <RESPUESTA> en 1990`. En el patrón la etiqueta <*RESPUESTA*> se debe sustituir por la respuesta a ser evaluada. Continuando con el ejemplo, para la respuesta `Kuwait` H quedaría como `Irak invadió el país de Kuwait en 1990`. Un ejemplo de sistema que aplica esta técnica para generar H es el descrito en (Rodrigo *et al.*, 2007a).

3.2.2. Análisis de similitud entre T y H

Analizar la similitud entre T y H es el recurso utilizado por los sistemas de VR para evaluar el RIT. Donde, a mayor valor de similitud mayor es la probabilidad de concluir que la respuesta es válida. A continuación describimos los métodos generales que se utilizan en el análisis de la similitud, los cuales los dividimos en niveles de acuerdo al tipo de procesamiento aplicado al texto. Estos niveles son el *léxico-morfológico*, el *sintáctico* y el *semántico*.

Nivel léxico-morfológico

A este nivel la similitud entre el par (T, H) es calculada principalmente por: comparar palabras, comparar raíces morfológicas de las palabras, comparar partes de la oración de las palabras y comparar entidades nombradas junto con expresiones numéricas y temporales.

En el caso de comparar palabras simplemente se cuentan las coincidencias de éstas tal como aparecen entre los textos T y H. El sistema descrito en (Kozareva *et al.*, 2006) es un ejemplo de este tipo de análisis de similitud.

Con el propósito de incrementar la tasa de coincidencias entre T y H, a estos textos se les aplica un análisis morfológico que permita conocer los lemas de las palabras. Por ejemplo, las palabras `ocurre` y `ocurrió` tienen el mismo lema `ocurrir`; por lo tanto al comparar sus raíces morfológicas pueden coincidir en contraste a comparar las palabras originales. Sistemas que comparan las raíces morfológicas de las palabras son descritos en (Tatu *et al.*, 2007), (Herrera *et al.*, 2006), y (Cumbreras *et al.*, 2007b).

Además, con el objetivo de hacer más preciso el análisis de similitud, algunos sistemas prefieren no comparar todas las palabras de los textos. En lugar de eso estos sistemas utilizan el análisis morfológico para conocer las partes de la oración de las palabras, entonces limitan el conteo de coincidencias a aquellas palabras que tienen asignadas ciertas categorías morfológicas. Por ejemplo, el sistema presentado por (Ferrández *et al.*, 2008b) sólo compara las palabras consideradas como de contenido (sustantivos, verbos, adjetivos y adverbios) y omite las funcionales (preposiciones, conjunciones y determinativos).

Algunos sistemas más sofisticados emplean herramientas para el reconocimiento de entidades nombradas, expresiones numéricas y expresiones temporales. Esto para lograr que ciertos términos compuestos coincidan en su totalidad entre T y H, por ejemplo el nombre de una persona con sus apellidos. El sistema descrito en Rodrigo *et al.* (2007b) evalúa la similitud entre T y H comparando sólo términos de este tipo.

Adicionalmente para el análisis de la similitud a este nivel, diversos sistemas han evaluado métodos de *distancia de edición* y de *secuencias de términos*. Esto con el propósito de comparar términos similares de los que no se tiene su lema o bien en aquellos casos donde hay cambios en la escritura (*p. ej.,* `Irak` e `Iraq`), así como para comparar adecuadamente secuencias de palabras que tienen un significado propio. Por ejemplo, `aprendizaje automático` tiene un sentido distinto a los términos por separado `aprendizaje` y `automático`. Además, las secuencias de términos permiten reflejar cierta estructura del texto sin llegar a un nivel superior en el análisis del lenguaje.

Respecto a la distancia de edición, ésta mide el costo de insertar, eliminar o sustituir letras de una palabra A para que sea igual a otra palabra B. Por ejemplo, dadas las palabras `ocurrió` y `ocurre` y un costo de uno para cada una de las tres operaciones permitidas, la distancia de edición entre estas dos palabras es dos debido a que en la primera palabra se necesita 1) sustituir `i` por `e` y 2) eliminar la `ó` (o bien en la segunda palabra sustituir `e` por `i` e insertar al final la `ó`). Ejemplos de sistemas que utilizan esta técnica son reportados en (Herrera *et al.*, 2006), (Bosma & Callison-Burch, 2007) y (Ferrández *et al.*, 2008b). Entonces, entre mayor es la distancia de edición de los términos comparados menor es su similitud.

Comúnmente, los sistemas que utilizan la distancia de edición hacen una normalización de la misma y establecen un umbral para decidir a partir de qué valor los

términos comparados siguen coincidiendo. Por ejemplo, en el caso antes descrito la distancia de edición normalizada de los términos es calculada de la siguiente manera:

$$\frac{distancia\_edicion(\texttt{ocurrió}, \texttt{ocurre})}{max(longitud(\texttt{ocurrió}), longitud(\texttt{ocurre}))} = 2/7 = 0{,}29 \tag{3.1}$$

De manera que si el umbral de coincidencia es igual a *distancia_edición_normalizada* $<$ 0.3, podemos concluir que los dos términos coinciden con un valor de similitud de $1 -$ *distancia_edición_normalizada* $= 0.71$.

En el caso del análisis de secuencias de términos, los métodos más utilizados son obtener *n-gramas* y obtener la *sub-secuencia común más larga*. Los n-gramas son secuencias de términos contiguos de tamaño n, donde n es definido por el usuario. Mientras que la *sub-secuencia común más larga* es la secuencia de términos no necesariamente contiguos del tamaño máximo posible.

Por ejemplo, entre T :`el ejército de Irak invadió Kuwait en 1990` y H :`Irak invadió el país de Kuwait en Agosto de 1990` tenemos como máximo dos n-gramas con $n = 2$, éstos son `Irak invadió` y `Kuwait en`. Y para el mismo par de textos tenemos la sub-secuencia común más larga `Irak invadió Kuwait en 1990`. Ejemplos de sistemas que utilizan estas técnicas son descritos en (Kozareva *et al.*, 2006) y (Herrera *et al.*, 2006).

Nivel sintáctico

A este nivel no es sólo los términos lo que se compara, sino además la manera en que éstos se relacionan en el texto. Por ejemplo, en T: `Yahoo adquirió Overture` y H: `Yahoo compró Overture` además del traslape de los términos `Yahoo` y `Overture`, estás oraciones comparten la estructura sujeto-acción-objeto (llamadas *dependencias*) donde el sujeto y el objeto de ambas son el mismo (`Yahoo` y `Overture`). De manera que la similitud de ambas no sólo es $2/3 = 0{,}66$ (el número de términos que coinciden entre el total de términos), sino que además se pueden tomar en cuenta las relaciones de dependencia sujeto y objeto dando como resultado una similitud de $4/5 = 0{,}8$ (términos y relaciones que coinciden entre el total posible).

Un sistemas que analiza la similitud a este nivel es descrito en (Tatu *et al.*, 2007), el cual incluye como formulas lógicas las relaciones sujeto y objeto.

Nivel semántico

A este nivel lo que se pretende principalmente es establecer conexiones entre conceptos equivalentes pero que están escritos de diferentes maneras y donde los lemas o la distancia de edición no son suficientes para encontrar la conexión (*p. ej.,* el caso de los sinónimos compró y adquirió). Actualmente el sistema de VR que parece estar más a este nivel es el descrito en (Tatu *et al.*, 2007), este sistema además de establecer relaciones entre palabras por medio de diccionarios u ontologías, también utiliza recursos para desambiguar el sentido de las palabras y resolver correferencias con el propósito de alcanzar un mayor entendimiento del texto.

También a este nivel es común que los sistemas añadan diversas fuentes de conocimiento externo para intentar incrementar la similitud de H con T. Tal es el caso del sistema descrito en (Bosma & Callison-Burch, 2007), el cual agrega conocimiento extra por medio de sustituir algunos de los conceptos en H por otros considerados similares; el resultado de cada sustitución representa una nueva hipótesis. Para realizar dicha sustitución, los nuevos conceptos son tomados de una base de datos construida a partir de una colección de documentos multilingüe. En esta colección se alinean fragmentos de textos en diferentes idiomas y se seleccionan aquellas palabras distintas pero que comparten los mismos contextos. Un ejemplo de tales conceptos es la palabra `cadáveres`, para la cual tienen conceptos como `muertos` y `cuerpos` para sustituirlo.

Para concluir esta subsección cabe mencionar que para varios de los análisis aquí descritos existen diferentes variantes; por ejemplo, las diversas aproximaciones utilizadas en el sistema descrito en (Ferrández *et al.*, 2006) para evaluar la distancia de edición. Sin embargo, sin ser exhaustivos, en el estudio se presentó una idea general de lo que se está utilizando para el análisis de la similitud en la validación automática de respuestas.

3.2.3. Clasificación de la respuesta

Finalmente para decidir si la respuesta es válida o errónea los sistemas de VR evalúan la similitud previamente analizada. Actualmente para evaluar la similitud los sistemas recurren a alguno de los tres métodos hasta ahora propuestos: *i*) aplicación de restricciones y valores por defecto, *ii*) clasificación automática sobre características de similitud y *iii*) aplicación de demostradores lógicos.

En el caso de la *aplicación de restricciones y valores por defecto*, una respuesta es válida sólo si su similitud satisface una restricción previamente definida. Por ejemplo, en el sistema reportado por Rodrigo *et al.* (2007a) se establece que una respuesta es válida sólo si cada una de las expresiones numéricas, temporales y entidades nombradas en H coinciden con alguno de los términos presentes en T.

Respecto a la *clasificación automática sobre características de similitud*, actualmente la mayoría de los sistemas de VR utilizan este método. En el cual un algoritmo de aprendizaje automático generalmente supervisado se utiliza para construir un clasificador. Dicho clasificador es el encargado de decidir si la respuesta es o no válida. Entre los atributos que se utilizan para representar cada una de las instancias a clasificar los más comunes son la tasa de términos y relaciones similares entre T y H, o bien la tasa de secuencias de términos similares entre T y H. Ejemplos de sistemas de VR que siguen este método son reportados en (Herrera *et al.*, 2006), (Bosma & Callison-Burch, 2007) y (Ferrández *et al.*, 2008b), cabe mencionar que en la mayoría de estos sistemas el algoritmo de aprendizaje utilizado es las Máquinas de Vectores de Soporte (SVM, por las siglas en Inglés de *Support Vector Machines*). Este algoritmo ha demostrado obtener buenos resultados para tareas relacionadas al procesamiento automático de textos (Joachims, 1998).

Finalmente, en la *aplicación de demostradores lógicos* los sistemas que utilizan este método en primer lugar representan el contenido de los textos como formulas lógicas. Posteriormente, si por medio de un demostrador lógico el sistema infiere lógicamente todos los predicados de H a partir de los de T, entonces se puede concluir que la respuesta es válida. El sistema reportado en (Tatu *et al.*, 2007) es un ejemplo de un sistema de VR que utiliza este método. Este sistema le aplica al texto un análisis léxico, sintáctico y semántico para construir las formulas lógicas; además, para facilitar la demostración éste incluye al proceso de inferencia predicados formados manual y automáticamente a partir de diversas fuentes de conocimiento externo. El resultado es un conjunto de predicados que representan principalmente lo siguiente: sustantivos, verbos, preposiciones, conjunciones, negación, dependencias, temporalidad de un evento, relaciones de los términos en diccionarios de sinónimos y WordNet así como de fenómenos del lenguaje como la aposición (para más detalles consultar (Harabagiu & Moldovan, 1998; Moldovan & Novischi, 2002; Moldovan & Rus, 2001; Moldovan *et al.*, 2005)). Además, si el sistema es incapaz de inferir lógicamente todos los predicados de H en lugar de

clasificar inmediatamente como errónea la respuesta, este sistema aplica un proceso llamado *relajación* (Moldovan *et al.*, 2007). Este proceso consiste en calcular la tasa de predicados inferidos lógicamente en la hipótesis y compararla contra un umbral α previamente definido, de modo que si el valor de la tasa de predicados inferidos es mayor o igual que α entonces la respuesta es válida, en caso contrario la respuesta es etiquetada como errónea. Esto último se puede ver como un caso de aplicación de restricciones y valores por defecto.

3.3. Validación de respuestas en español

En esta sección se muestra un estudio útil para distinguir el tipo de implicación textual que ocurre en la validación de respuestas en español. Además, con base en dicho estudio también se discute lo conveniente de los métodos hasta ahora propuestos para tratar el problema en este idioma.

3.3.1. Tipo de implicación textual a resolver

Para presentar una clasificación del tipo de implicación textual a analizar en la validación de respuestas en español, en este estudio al igual que Zaenen *et al.* (2005) se prefiere emplear el término *inferencia textual* más que implicación textual para referirnos a aspectos más generales del problema.

Antes de presentar el estudio es importante aclarar que cuando se habla de inferencia textual estamos interesados sólo en la verificabilidad no en la verdad. Es decir, podemos verificar que H :`Apple posee Microsoft` se infiere textualmente de T :`Apple adquirió Microsoft` aunque esta información no sea verdadera. Esta característica de la inferencia textual se adapta bien a la validación de respuestas, donde si la respuesta se justifica entonces se evalúa como correcta a pesar de que en la realidad esa información sea obsoleta o incorrecta.

Para iniciar el estudio se tiene que una inferencia textual puede ser de tres tipos: implicación, inferencia convencional e inferencia conversacional (Zaenen *et al.*, 2005). A continuación se dan detalles.

- *Implicación.* Este tipo de inferencias suceden sólo cuando se cumple que si T ocurre entonces H también ocurre. Por ejemplo, H :`Tony estuvo en Bagdad el domingo en la noche` se infiere textualmente de T :`Tony arribó a Bagdad el domingo en la noche`, aseverando que hablamos del mismo Tony, el mismo Bagdad y el mismo domingo. Es importante hacer notar que este tipo de inferencia es lo que evalúan los sistemas de VR cuando analizan la similitud por medio de contar coincidencias, donde la coincidencia de términos simplemente son las aseveraciones.

- *Inferencia convencional.* En este tipo de inferencias aparte de hacer aserción, también es necesario suponer ciertas cosas. Un ejemplo es lo que sucede en los fenómenos de aposición, por ejemplo: H :`Ames fue un exitoso espía` se infiere textualmente de T_1 :`Ames, el exitoso espía, resolvió el robo`. Además, este tipo de inferencias también ocurren frecuentemente en apreciaciones; por ejemplo, la hipótesis anterior se infiere textualmente de T_2 :`Ames fue un exitoso espía`. Sin embargo, la inferencia textual de H puede cambiar con T_3 :`Según la prensa, Ames fue un exitoso espía`, en este caso depende de considerar una opinión.

- *Inferencia conversacional.* En este tipo de inferencia además de las aserciones y suposiciones de los dos casos anteriores, también es necesario suponer que "a falta de evidencia de lo contrario, se puede decir tanto como sea posible". Por ejemplo: H :`No todos los soldados murieron` es inferida textualmente por T :`Algunos soldados fueron asesinados`.

Con el propósito de obtener una idea del tipo de inferencia textual a resolver en la validación de respuestas en español, de manera aleatoria se analizó la mitad de las respuestas válidas presentes en un conjunto de prueba de preguntas y respuestas en este idioma. En total los casos analizados fueron 335 de 671, esto representa una confianza en los resultados del análisis de un 94.7 % de acuerdo a una prueba estadística para seleccionar el tamaño de la muestra (donde $n = \frac{4\sigma^2}{28}$ para un $\sigma \approx 14$ tomando en cuenta que en promedio en el conjunto de datos un 29 % (55) de las respuestas de cada sistema de BR son válidas).

Del análisis tenemos que aproximadamente el 68 % de los casos son inferencias del tipo implicación, mientras que el 32 % restante son inferencias convencionales donde

ocurre un fenómeno de aposición. Estos valores son un reflejo de la complejidad actual de la búsqueda de respuestas; es decir, preguntas de nivel uno con respuestas expresadas de forma simple. En otras palabras, actualmente los sistemas de BR en español para contestar a las preguntas no realizan suposiciones distintas a las de fenómenos de aposición o de otros fenómenos del lenguaje similares (*p. ej.*, H :`El acrónimo ONU significa Organización de las Naciones Unidas` se infiere textualmente de T :`ONU (Organización de las Naciones Unidas)`); sin duda, esto se debe a que los sistemas en este idioma carecen de recursos sofisticados para el procesamiento profundo del lenguaje.

Para complementar el análisis de los tipos de inferencias, también se realizó un estudio de las fuentes de conocimiento que se requieren para reconocer la inferencia textual en los casos donde las aserciones no son inmediatas debido a las variaciones del lenguaje. Las fuentes que se estudiaron son dos:

- *Conocimiento de las convenciones del uso del lenguaje.* Donde el conocimiento de cualquier hablante competente del lenguaje es suficiente para reconocer la inferencia textual. Por ejemplo, en el caso de H :`La catástrofe de Chernobyl ocurrió en el año 1986` y T :`décimo aniversario de la catástrofe de Chernobyl, ocurrida en la noche del 25 al 26 de abril de 1986` no se necesita conocimiento extra al presente en los textos para reconocer que H es inferida textualmente por T.

- *Conocimiento de la semántica del lenguaje.* Cuando cierto conocimiento de fondo es necesario para establecer la inferencia textual. Por ejemplo, para reconocer la inferencia textual de H :`La catástrofe de Chernobyl ocurrió en el año 1986` por T :`se inició un año después del accidente nuclear de Chernobyl, que se produjo en 1986` es necesario conocer que un accidente nuclear es o produce una catástrofe.

Analizando nuevamente la muestra de respuestas válidas, tenemos que aproximadamente en un 34 % de los casos es necesario conocimiento de la semántica del lenguaje, mientras que el 66 % restante requiere únicamente de conocimiento de las convenciones del uso del lenguaje. Esto refleja como los sistemas de BR en español prescinden de fuentes de conocimiento tales como diccionarios u ontologías para reforzar su búsqueda

de respuestas. Sobre todo si tomamos en cuenta que en la mayoría de los casos donde se requiere conocimiento de la semántica del lenguaje, este conocimiento no es simple de obtener ni siquiera manualmente de los recursos que actualmente hay disponibles no sólo para el español (*p. ej.*, determinar la relación de accidente nuclear con catástrofe).

Como resultado de este estudio tenemos que, en la validación de respuestas en español, la inferencia textual que actualmente ocurre en las respuestas válidas son principalmente del tipo *i*) implicación donde sólo hay aserciones y *ii*) algunas del tipo inferencia convencional donde las suposiciones son referentes a fenómenos del lenguaje como la aposición. Esto es un reflejo de como los sistemas de BR en español pueden encontrar respuestas válidas a las preguntas sin llegar a un entendimiento del lenguaje.

3.3.2. Relevancia de un análisis a nivel semántico

Después del estudio del tipo de implicación textual a resolver y del tipo de conocimiento de fondo que se requiere en la validación de respuestas en español (Sección 3.3.1); parece ser que para la mayoría de las respuestas válidas en este idioma, un sistema de VR no requiere llegar a un costoso análisis semántico del texto para encontrar la similitud entre la mayor parte de los términos de T y H.

Por ejemplo, en uno de los experimentos previos de esta tesis se realizó la evaluación de un sistema de VR que clasifica las respuestas por medio de aplicar un demostrador lógico. Como se describe en (Téllez-Valero *et al.*, 2006), este sistema transforma el texto en formulas lógicas que representan los términos de contenido y las dependencias sujeto-acción-objeto. Adicionalmente, como parte del proceso de clasificación de la respuesta, el sistema utiliza la información del análisis pregunta-respuesta (*p. ej.*, el tipo de la pregunta y de la respuesta así como la compatibilidad entre éstos) y agrega una relajación a la demostración similar a la utilizada en el sistema presentado en (Tatu *et al.*, 2007). Pero en contraste con los sistemas que aplican demostradores lógicos, el sistema evaluado no consulta ninguna fuente de conocimiento extra y sólo utiliza una representación léxico-sintáctica de los textos bastante alejada de la semántica del lenguaje.

El resultado de este sistema en una colección de prueba formada por preguntas y respuestas en español fue una precisión de 0.4862 (la tasa de respuestas clasificadas correctamente como válidas del total de respuestas que el sistema clasifico como válidas)

y una cobertura de 0.7862 (la tasa de respuestas clasificadas como válidas del total de respuestas válidas en la colección de prueba) alcanzando una medida-F de 0.59 (una combinación lineal de la precisión y el recuerdo).

Si tomamos en cuenta que el sistema de VR reportado en (Tatu *et al.*, 2007) obtuvo en la misma colección de prueba una medida-F de 0.61 (el máximo resultado hasta ahora alcanzado en dicha colección según lo reportado en (Peñas *et al.*, 2007)), al parecer la sofisticada representación de los textos así como las diversas fuentes de conocimiento extra que incluye este sistema no se está reflejando en sus resultados finales; superando por sólo un 0.2 de medida-F a un sistema similar que no incluye todos estos recursos de análisis a nivel semántico.

3.3.3. Problemas con un análisis a nivel léxico-sintáctico

En la validación de respuestas aplicando restricciones o una clasificación automática para clasificar la respuesta es posible prescindir de un análisis a nivel semántico del texto así como de consultar fuentes de conocimiento extra para concluir casos como H_1 :`Lucy fue con los padres de Luis` es implicada textualmente por T_1 :`Lucy visita a los padres de Luis`.

Sin embargo, estos enfoques a nivel léxico-sintáctico presentan problemas cuando existe una alta tasa de coincidencias de términos pero no existe la implicación textual (*p. ej.*, H_1 con T_2 :`Lucy odia a los padres de Luis`), una situación que es común en la validación de las respuestas erróneas. En ocasiones la alta tasa de coincidencias es el resultado de calcular la similitud tomando en cuenta todos los términos en los textos, entre éstos los términos que son frecuentes a cualquier tópico pero que no proporcionan información de la existencia de una implicación textual (*p. ej.*, preposiciones y conjunciones). El resultado es que aquellos sistemas de VR que siguen estos métodos pero que omiten los términos funcionales en el análisis de similitud generalmente obtienen mejores resultados que los que no los omiten.

Además, el problema de la alta tasa de coincidencias en respuestas erróneas también puede existir cuando el texto T contiene a todos, o casi a todos, los términos de contenido presentes en H; pero, la información planteada en ambos textos está en diferentes contextos (*p. ej.*, H_1 con T_3 :`Lucy fue con el médico después de que sus padres y Luis se lo sugirieron`). Este problema empeora aún más en el caso

de que H si está en el contexto de T pero ésta fue formada con una respuesta extraída incorrectamente durante la búsqueda de respuestas; por ejemplo, la respuesta tres en la tabla 3.1.

La idea actual de analizar la similitud de sub-secuencias de términos puede servir para evitar algunos casos donde existe una alta tasa de coincidencias pero la relación entre términos no es la misma (*p. ej.*, H_1 con T_4 :`Luis fue con los padres de Lucy`). Sin embargo, a las estrategias actuales para calcular las secuencias les afectan tanto los textos grandes que tienen mayor probabilidad de contener sub-secuencias más largas (*p. ej.*, H_1 con T_3), así como casos de paráfrasis donde puede suceder que los agentes de la acción principal en H se inviertan en T y en consecuencia las secuencias sean más cortas (*p. ej.*, H_1 con T_5 :`Con los padres de Luis fue Lucy`).

En el siguiente capítulo se describe el método propuesto en esta tesis, el cual se distingue de los presentes en el estado del arte principalmente por reducir el número de errores causados por un alto grado de coincidencia entre T y H a pesar de que la respuesta correspondiente sea errónea.

Capítulo 4

El método propuesto

En este capítulo se describe el método propuesto, el cual consiste de dos etapas principales: la *validación de respuestas* y la *selección de la respuesta* (ver figura 4.1). En el caso de la validación de respuestas (primera etapa) dadas como entrada una pregunta y una respuesta con su pasaje de soporte; en esta etapa se clasifica a la respuesta como *válida* o *errónea* y se le asigna un valor de confianza β (entre 0 y 1) para indicar la certeza que se tiene en su categoría. Posteriormente, en la selección de respuestas (segunda etapa) tomando como entrada un conjunto de respuestas previamente clasificadas como válidas (junto con su valor de confianza); la salida de esta etapa es una de las respuestas de dicho conjunto de entrada. Esta respuesta de salida, llamada la *respuesta final*, es la considerada más pertinente para contestar a la pregunta. En el caso de que ninguna respuesta sea clasificada como válida en la primera etapa, el método en su segunda etapa retorna un *nil* como la respuesta final. En las siguientes secciones se describen los procesos realizados en cada una de estas dos etapas.

4.1. Validación de respuestas

Con el objetivo de reconocer si una respuesta es válida o errónea para contestar a una pregunta, en el método en su etapa de validación de respuestas se propuso aplicar cinco procesos principales: el *pre-proceso del texto*, la *generación de la hipótesis*, el

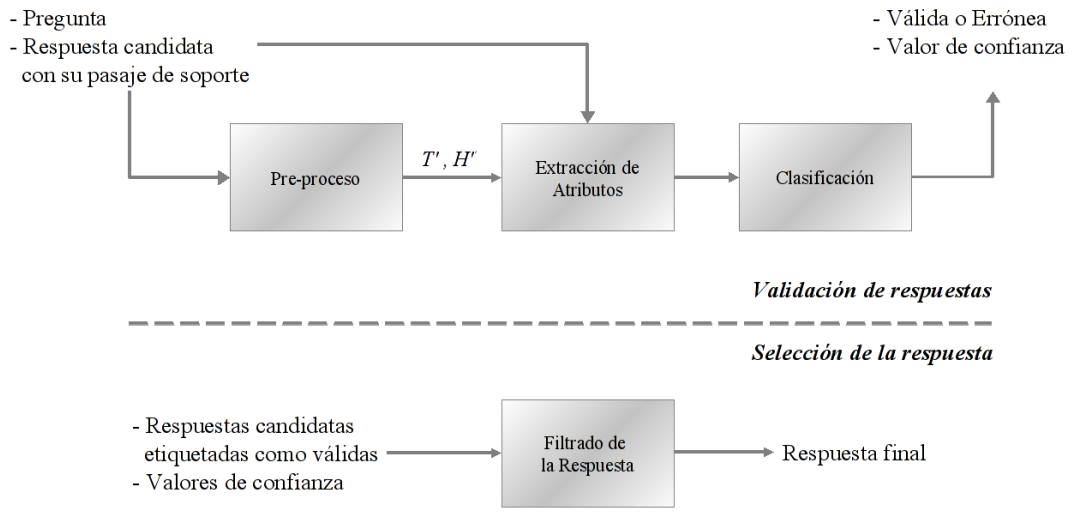

Figura 4.1: Procesos generales del método propuesto

análisis de similitud entre T y H, el *análisis de la relación pregunta-respuesta* y la *clasificación de la respuesta*. En las siguientes subsecciones se describe a detalle cada uno de estos procesos.

4.1.1. Pre-proceso del texto

El propósito de este proceso es extraer los principales elementos de contenido para la pregunta y la respuesta así como de su pasaje de soporte. Estos elementos son posteriormente utilizados para decidir la validez de la respuesta evaluada. En el proceso se consideran dos tareas básicas. Por un lado, la *identificación de los constituyentes* en la pregunta y la respuesta (los que definen a H en la implicación textual). Por otro lado, la *detección del núcleo* del pasaje de soporte así como la consecuente eliminación de la información innecesaria (el texto resultante es el T en nuestro análisis de la implicación textual).

Identificación de los constituyentes

Los constituyentes detectados en la pregunta son tres: la *acción principal*, los *actores de la acción* y, si existe, la *restricción de la acción*. A manera de ejemplo, considere la

pregunta de la tabla 3.1, en este caso la acción principal está representada por el verbo `invadir`, sus actores son los sintagmas `A qué país` e `Irak` y la restricción de la acción es descrita por el sintagma preposicional `en 1990`.

Para detectar dichos constituyentes, en primer lugar se le aplica a la pregunta un análisis sintáctico superficial. Entonces, desde el árbol sintáctico resultante (P_{sint}) se construye una nueva representación de la pregunta (llamada P') por medio de detectar y etiquetar los siguientes elementos:

1. *El constituyente de la acción.* Éste corresponde al sintagma en P_{sint} que incluye el verbo principal.

2. *El constituyente de la restricción.* Éste es representado por el sintagma en P_{sint} que tiene al menos una expresión temporal explícita (*p. ej.,* `en 1990`) o que incluye una preposición tal como `después` o `antes`.

3. *Los constituyentes de los actores.* Estos constituyentes están formados por el resto de los elementos en P_{sint}. Éstos están comúnmente divididos en dos partes. El primero, de aquí en adelante llamado *constituyente del actor oculto*, corresponde al sintagma que incluye la partícula interrogativa (*p. ej.,* `qué`, `cuándo`, etc.) y está generalmente localizado en el lado izquierdo del constituyente de la acción. El segundo, al que se llamo el *constituyente del actor visible*, está formado por el resto de los sintagmas, los cuales generalmente están localizados a la derecha del constituyente de la acción.

Finalmente, en este proceso también consideramos un *constituyente de la respuesta.* Éste simplemente es el análisis morfológico de la respuesta evaluada (denotado por R').

La figura 4.2 muestra los constituyentes identificados para la pregunta y primera respuesta candidata presentadas en la tabla 3.1.

Detección del núcleo del pasaje de soporte

Comúnmente, el pasaje de soporte es un fragmento de texto corto que proporciona el contexto necesario para justificar su correspondiente respuesta. Sin embargo, en muchos casos, este pequeño fragmento de texto contiene información que va más allá del

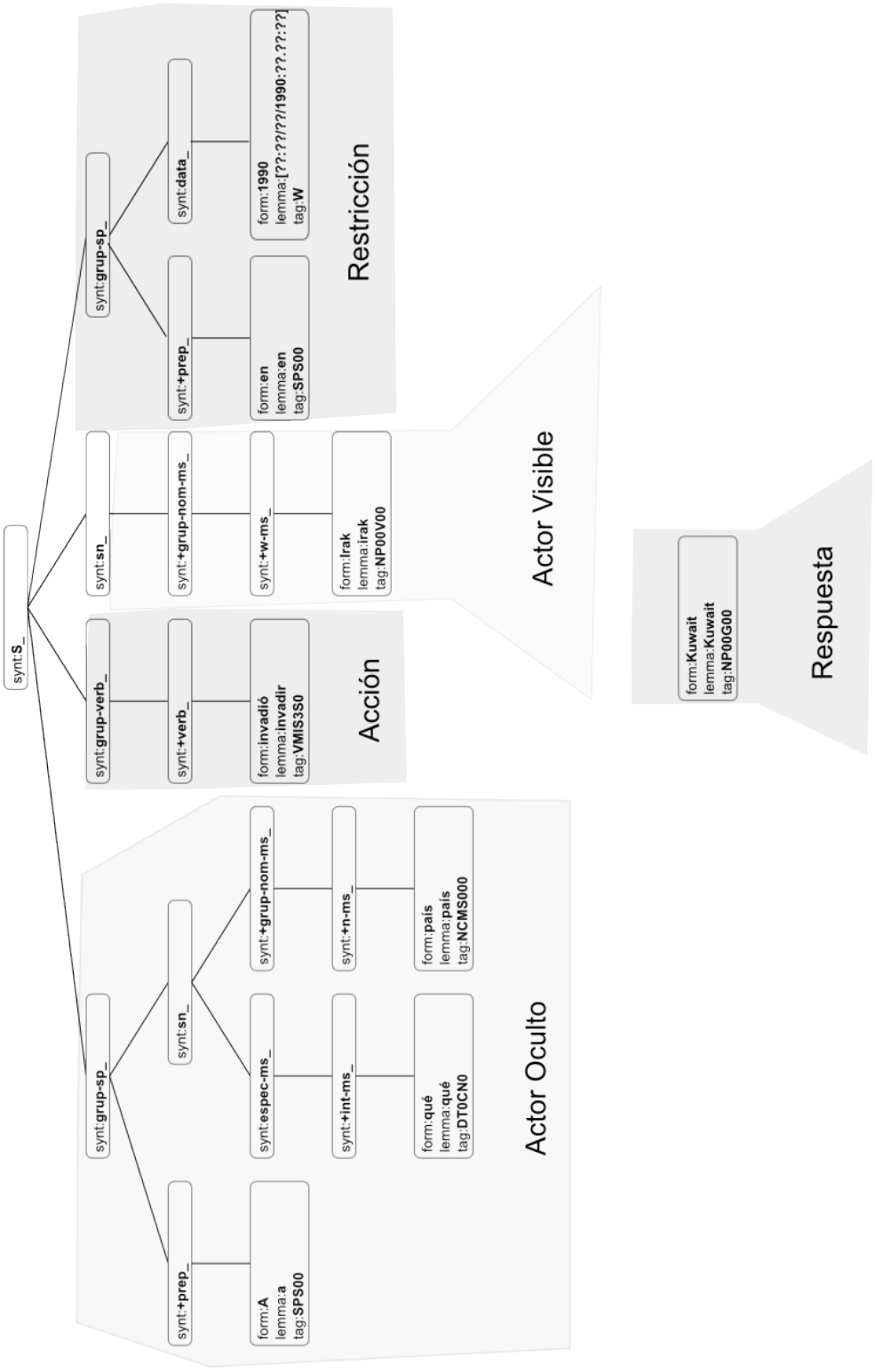

Figura 4.2: Constituyentes de la pregunta y la respuesta.

contexto de la pregunta y que puede perjudicar el análisis de similitud entre T y H en la validación de respuestas.

Con el propósito de evitar los problemas causados por un pasaje de soporte excesivo, como parte del método se propuso reducir al pasaje de soporte a un fragmento de texto mínimamente útil (de aquí en adelante llamado el *núcleo del pasaje de soporte*). Para realizar esta reducción el proceso es el siguiente:

- Primero, al pasaje de soporte le aplicamos un análisis sintáctico superficial, el resultado es su árbol sintáctico PS_{sint}.

- Segundo, los términos de contenido (sustantivos, verbos, adjetivos y adverbios) de los constituyentes de la pregunta (la acción, la restricción y los actores) son coincididos, a nivel de lemas, con los términos en PS_{sint}. Para evitar que algunas diferencias mínimas de escritura causen el problema de no poder coincidir términos referentes a un mismo concepto y que además no es resuelto con el análisis morfológico, en la comparación de los términos evaluamos la *distancia de edición* de sus lemas. De tal manera que si el valor de la distancia de edición de los términos comparados es menor que un umbral predeterminado, estos términos coinciden[1]. Además, en la comparación de las *expresiones temporales* se aplica un proceso especial de normalización que permite hacer coincidir casos de tipo fecha (*p. ej.,* `el 3 de Agosto de 1996` coincide con `en Agosto de 1996`) y casos de tipo periodo (*p. ej.,* `entre 1990 y 1998` coincide con `en 1996`), esto es especialmente útil para evaluar respuestas para preguntas con alguna restricción temporal.

- Tercero, en base en el número de términos de contenido que coinciden, los constituyentes de la pregunta son alineados con los sintagmas en PS_{sint}.

- Cuarto, los términos en el constituyente de la respuesta también son alineados con los sintagmas en PS_{sint}. La idea es encontrar todas las ocurrencias de la respuesta en su pasaje de soporte.

[1]En los experimentos con un conjunto de entrenamiento un umbral igual a 0.4 para la distancia de edición fue el que proporcionó los mejores resultados en la validación de respuestas (utilizando una validación cruzada de 10 pliegues en la evaluación).

- Quinto, en el pasaje de soporte determinamos el mínimo contexto de la respuesta como aquel que contiene todos los sintagmas alineados (*i.e.,* que contiene los constituyentes de la pregunta y el de la respuesta). Este contexto mínimo (representado por una secuencia de palabras alrededor de la respuesta) es lo que llamamos el núcleo del pasaje de soporte (denotado por T'). En caso de que el pasaje de soporte incluya varias ocurrencias de la respuesta, aquella con el contexto más pequeño es la seleccionada.

Aplicando el procedimiento antes descrito podemos determinar que el pasaje de soporte para la primera respuesta candidata presentada en la tabla 3.1 tiene como núcleo el fragmento de texto `de la invasión iraquí de Kuwait`. De esta manera se evita que las otras coincidencias de los términos de H con T sobreestimen su análisis de similitud.

4.1.2. Generación de la hipótesis

Para generar la hipótesis necesaria en el estudio de la implicación textual para la validación de respuestas, como parte del método propuesto se incluye una forma de construir automáticamente una H afirmativa que conserva la semántica que se espera de una respuesta válida. Esta forma automática consiste en sustituir el mínimo sintagma que incluye la partícula interrogativa en la pregunta por el constituyente de la respuesta.

El sintagma sustituido puede ser todo o una parte del constituyente del actor oculto en la pregunta, identificado durante el pre-proceso del texto. Por ejemplo, en la figura 4.3 se muestra el árbol sintáctico de la pregunta en la tabla 3.1, en este árbol los nodos frente a la sombra representan el sintagma a sustituir (una parte del constituyente del actor oculto); por lo tanto, omitiendo los símbolos de interrogación la hipótesis para la respuesta candidata `Kuwait` es H: `A Kuwait invadió Irak en 1990`.

Es importante hacer notar que permutando los constituyentes más generales de la H resultante es posible construir diversas formas de la misma (los constituyentes de H a permutar coinciden con los constituyentes de la pregunta y de la respuesta previamente detectados). Por ejemplo, para la pregunta de la tabla 3.1 y la respuesta `Kuwait` podemos tener las 4! (24) hipótesis mostradas en la tabla 4.1, donde 4 es el número de constituyentes que están presentes. Esta situación es aprovechada por el método aquí descrito en el cual en lugar de construir todas las H's posibles mediante

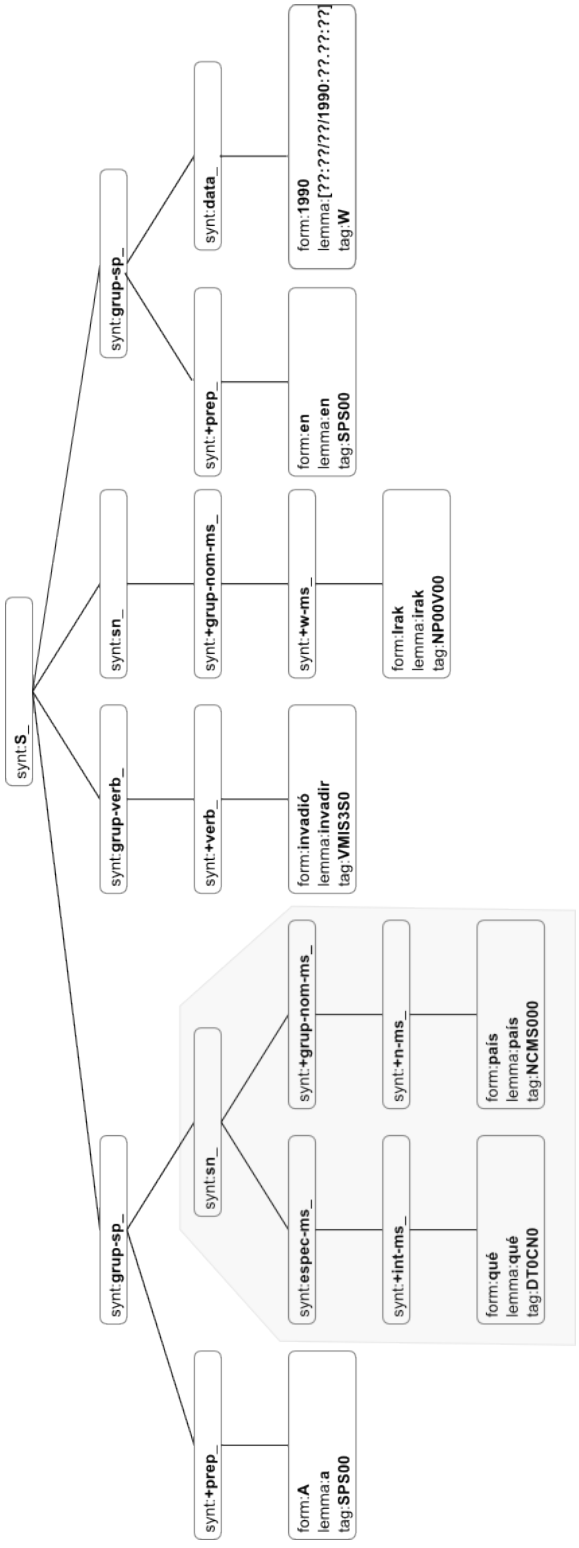

Figura 4.3: Análisis sintáctico superficial de una pregunta.

Tabla 4.1: Hipótesis construidas automáticamente para validar respuestas.

A Kuwait invadió Irak en 1990	A Kuwait Irak invadió en 1990
invadió A Kuwait Irak en 1990	invadió Irak A Kuwait en 1990
Irak A Kuwait invadió en 1990	Irak invadió A Kuwait en 1990
A Kuwait invadió en 1990 Irak	A Kuwait en 1990 invadió Irak
invadió A Kuwait en 1990 Irak	invadió en 1990 A Kuwait Irak
en 1990 A Kuwait invadió Irak	en 1990 invadió A Kuwait Irak
A Kuwait Irak en 1990 invadió	A Kuwait en 1990 Irak invadió
Irak A Kuwait en 1990 invadió	Irak en 1990 A Kuwait invadió
en 1990 A Kuwait Irak invadió	en 1990 Irak A Kuwait invadió
invadió Irak en 1990 A Kuwait	invadió en 1990 Irak A Kuwait
Irak invadió en 1990 A Kuwait	Irak en 1990 invadió A Kuwait
en 1990 invadió Irak A Kuwait	en 1990 Irak invadió A Kuwait

las permutaciones, en este caso se prefiere tomar como H al conjunto formado por los sintagmas que se deben permutar. Es decir, el método no parte de una única forma de la hipótesis, en lugar de eso su conjunto de sintagmas le permiten adaptarse a la forma más cercana del texto T para beneficio del análisis de similitud.

4.1.3. Análisis de similitud entre T y H

Este proceso incluye un conjunto de subprocesos que le permiten extraer diversos atributos que miden la similitud entre T y H. Estos atributos pueden ser divididos en dos grupos: *i*) atributos que miden las *coincidencias* entre el pasaje de soporte y la hipótesis; y *ii*) atributos que denotan las discrepancias entre esos dos componentes. A continuación se describen a ambos grupos de atributos y se explica la manera en

que éstos son calculados. Pero antes cabe recordar que, a diferencia de otros métodos de validación de respuestas, en el método propuesto no utilizamos el pasaje de soporte completo; en lugar de eso, en este método sólo usamos su fragmento núcleo T'. Adicionalmente, en el método propuesto no usamos un texto hipótesis, en su lugar utilizamos como H el conjunto de constituyentes pregunta-respuesta (*i.e.*, la unión de P' y R', cuyo resultado es H').

Características relacionadas a las coincidencias

Estas características expresan el grado de coincidencia —en número de términos— entre T' y H'. En particular, un atributo de coincidencia es calculado por cada uno de los cuatro términos de contenido (sustantivos, verbos, adjetivos y adverbios) así como por cada uno de los cuatro tipos de entidades nombradas (nombres de personas, lugares, organizaciones y otros) y los dos tipos de expresiones (expresiones numéricas y temporales). Estos diez diferentes atributos de coincidencia son generados por separado para cada uno de los cinco constituyentes de H' (constituyente de la acción, de la restricción, del actor oculto, del actor visible y de la respuesta). De esta manera, un total de cincuenta atributos con valores entre 0 y 1 son generados para representar las coincidencias (un 0 indica la falta total de coincidencia y el 1 que todo coincide).

En el caso de que alguno de los constituyentes de P' esté vacío (*p. ej.*, cuando no hay restricción temporal en la pregunta), los valores de coincidencia para este *constituyente ausente* son establecidos a uno; es decir, por defecto todos los constituyentes en H' coinciden en T'.

Es importante mencionar que, en este subproceso, reutilizamos los cálculos de la coincidencia hechos para encontrar el núcleo del pasaje de soporte (ver Sección 4.1.1). Entonces en este caso, también se utiliza la distancia de edición y la coincidencia de expresiones temporales para evaluar la similitud entre los términos de T' y H'.

Cabe mencionar que en este método omitimos del análisis de similitud las palabras funcionales (preposiciones, conjunciones, determinativos, etc.), además de que evaluamos sólo una parte del texto de soporte (su núcleo). Todo esto para evitar que en respuestas erróneas exista un alto traslape causado por información irrelevante para contestar a la pregunta.

Características relacionadas a las discrepancias

Tomando en cuenta los casos de alta coincidencia entre T' y H' aún en casos donde el texto coincidente es relevante para la pregunta pero la respuesta es errónea, en el análisis de similitud se incluye un nuevo conjunto de características como son la evaluación de las discrepancias. Estas nuevas características indican el número de términos que no coinciden en el núcleo del pasaje de soporte con H'. En otras palabras, los valores de estos atributos indican el número de términos en T' que no están presentes en ninguno de los constituyentes de la pregunta-respuesta.

Principalmente, el cálculo de estos valores es hecho para los términos no traslapados entre el constituyente de la respuesta y cada uno de los cuatro constituyentes de la pregunta. Entonces, para cada uno de los cuatro pares de constituyentes (respuesta–acción, respuesta–actor_oculto, respuesta–actor_visible, respuesta–restricción), un atributo es generado por cada uno de los cuatro tipos de términos de contenido, los cuatro tipos de entidades nombradas y los dos tipos de expresiones. En total cuarenta diferentes atributos son generados en este subproceso.

Para el caso de que P' carezca de alguno de sus cuatro constituyentes o que éste no haya sido detectado en T', los valores para los atributos del par respuesta–*constituyente_ausente* son establecidos a uno; es decir, por defecto no existen términos sin traslapar entre los pares de constituyentes.

4.1.4. Análisis de la relación pregunta-respuesta

Para el estudio de la implicación textual ocurrida en la validación de respuestas, el análisis del par (T, H) puede ser complementado con el análisis de la pregunta y la respuesta. A continuación se describen los atributos obtenidos de este análisis complementario.

Características de la pregunta

Los atributos considerados de la pregunta son cuatro: la partícula interrogativa (`qué`, `dónde`, `cuándo`, etc.), la clase de la pregunta (*factual* o *definición*), el tipo de respuesta esperada (*expresión temporal*, *expresión numérica*, *entidad nombrada* u *otro*) y el tipo de restricción de la pregunta (*fecha*, *periodo*, *evento* o *nada*).

Con un emparejamiento de patrones léxico-morfológicos en la pregunta es posible determinar su partícula interrogativa, su clase y el tipo de respuesta que espera[1]. Algunos de estos patrones son mostrados a continuación, en cada uno de éstos es posible observar que se incluye información de la categoría de la pregunta y del tipo de respuesta esperado. Es importante mencionar que aunque en la actualidad existen diversas aproximaciones para la clasificación de preguntas basadas en aprendizaje automático (*p. ej.,* los trabajos reportados en (Zhang & Lee, 2003) y Blunsom *et al.* (2006)), en nuestra clasificación las clases asignadas a las preguntas son muy generales por lo que se prefiere continuar utilizando el método clásico basado en reglas hechas a mano (Lee *et al.*, 2000; Pasca & Harabagiu, 2001).

CUÁNTO [*cualquier cosa*]	→	*FACTUAL – EXP. NUMÉRICA*
CUÁNDO [*cualquier cosa*]	→	*FACTUAL – EXP. TEMPORAL*
DÓNDE [*cualquier cosa*]	→	*FACTUAL – ENT. NOMBRADA*
CÚAL `es el nombre de` [*cualquier cosa*]	→	*FACTUAL – ENT. NOMBRADA*
QUÉ `es` [*un sustantivo*]	→	*DEFINICIÓN – OTRO*
QUIÉN `es` [*una ent. nombrada*]	→	*DEFINICIÓN – OTRO*

Por otro lado, el valor del tipo de restricción de la pregunta depende de la forma que tiene el constituyente de la restricción. Si este constituyente contiene sólo una expresión temporal, entonces se le asigna el valor *fecha*. En el caso que el constituyente de la restricción incluya dos expresiones temporales, su valor es *periodo*. Si el constituyente de la restricción no contiene ninguna expresión temporal, a la restricción de la pregunta se le asocia el valor *evento*. Finalmente, cuando la pregunta tiene un constituyente de la restricción vacío, el valor de la restricción de la pregunta es *nada*. Enseguida mostramos algunos ejemplos de preguntas que tienen diferentes valores en su tipo de restricción.

`¿Dónde se celebraron los Juegos Olímpicos de Invierno` *`de 1994`*`?`	→	*FECHA*
`¿Quién fue el presidente de Perú` *`entre 1985 y 1990`*`?`	→	*PERIODO*
`¿Quién era el presidente de Francia` *`durante las pruebas de armas nucleares en el Pacífico Sur`*`?`	→	*EVENTO*
`¿Cuándo fue la coronación oficial de Isabel II?`	→	*NADA*

[1]Estos patrones son una extensión de los utilizados para clasificar preguntas en el sistema de BR descrito en (Téllez-Valero *et al.*, 2007).

Compatibilidad de la respuesta

Este atributo indica si los tipos de la pregunta y la respuesta son compatibles. La idea de evaluar esta característica es capturar la situación donde la clase semántica de la respuesta evaluada no coincide con el tipo de respuesta esperada por la pregunta. En particular, este atributo ayuda a clasificar como erróneas todas las respuestas candidatas que no son compatibles con la pregunta pero que pueden provocar confusión durante el análisis de la similitud.

Las categorías utilizadas para clasificar las respuestas candidatas son los mismos cuatro tipos de las respuestas esperadas en el análisis de la pregunta (*expresión temporal*, *expresión numérica*, *entidad nombrada* u *otro*). Entonces, si la respuesta candidata contiene una expresión temporal esa es la categoría que se le asigna. Lo mismo ocurre si ésta contiene una expresión numérica o una entidad nombrada. Para asignar la categoría *otro* basta que la respuesta candidata contenga un término de contenido que no forme parte de alguna de las expresiones o entidades que asignan las otras categorías. Bajo este esquema de clasificación, una respuesta candidata puede tener asignado más de un tipo; por ejemplo, la respuesta candidata `galardonado con el premio Nobel de la Paz en 1983` (para la pregunta `¿Quién es Danuta Walesa?`) tiene asignado todos los tipos excepto el de *expresión numérica*.

El valor de este atributo va a ser igual a 1 cuando algún tipo de la respuesta candidata corresponda con el tipo de la respuesta esperada por la pregunta. En caso contrario, el valor del atributo es 0 (*i.e.*, un atributo binario). Por ejemplo la pregunta `¿Cuál es el record del mundo de salto de altura?` y la respuesta candidata `Javier Sotomayor` no son compatibles; por lo tanto, el valor del atributo que evalúa su compatibilidad es 0 para indicar que la respuesta es errónea a pesar de que el análisis de similitud indique lo contrarío (*p. ej.,* lo que sucede cuando se analiza la similitud con `T:el record del mundo de salto de altura es de Javier Sotomayor`).

4.1.5. Clasificación de la respuesta

En la etapa de validación de respuestas, este proceso final es utilizado para catalogar la respuesta candidata como válida o errónea por medio de un método de aprendizaje automático supervisado. El clasificador construido le asigna a la respuesta candidata su categoría basado en los noventa-y-cinco atributos descritos en las secciones previas

(en la tabla 4.2 se muestra un resumen de tales atributos). Adicionalmente, un valor de confianza β entre 0 y 1 es estimado para la respuesta (un valor de 0 indica la total desconfianza y 1 la máxima confianza). Este valor de confianza es la distribución de probabilidad asociada por el clasificador a la categoría asignada a la respuesta[1], dicho valor es útil para la selección de la respuesta final en la segunda etapa del método.

4.2. Selección de respuestas

Tomando como entrada un conjunto de respuestas etiquetadas como válidas durante la primera etapa del método, el objetivo de esta segunda etapa es seleccionar a partir de este conjunto de respuestas una para contestar a la pregunta. Los elementos utilizados para llevar a cabo la selección son el valor de confianza β que le fue asignado a la respuesta en su validación y la redundancia de la respuesta en el conjunto de respuestas válidas.

Entonces, la respuesta final es aquella clasificada como válida y con la mayor redundancia. En el caso de que dos o más respuestas tengan la misma máxima redundancia, entonces se prefiere de entre éstas aquella con el mayor valor de confianza β.

La idea de agregar la redundancia como criterio de selección se basa en el trabajo de Dalmas & Webber (2007) aplicado a la comparación de respuestas, en el cual se propone considerar a los candidatos como aliados más que como competidores. Pero con la diferencia de que en lugar de utilizar directamente la frecuencia de ocurrencia de las respuestas, en el método propuesto la redundancia corresponde a la suma de los valores de similitud calculados mediante la distancia de edición de cada una de las respuestas válidas al resto de las respuestas en el conjunto de respuestas clasificadas también como válidas. De esta manera la respuesta `telescopio espacial` puede contribuir a la redundancia de `telescopio espacial Hubble` o de `telescopio` para responder a la pregunta `¿Qué es el Hubble?`

[1]Durante la clasificación de una instancia el esquema de aprendizaje calcula un valor entre 0 y 1 para cada una de las clases que le puede asignar a dicha instancia; entonces, la clase con el máximo valor es la categoría finalmente asignada. Al valor calculado para cada clase se le conoce como la *distribución de probabilidad* de dicha categoría, tal que la suma de las distribuciones de probabilidad de todas las clases es igual a uno.

Tabla 4.2: Resumen de los atributos propuestos para validar las respuestas

Número/Tipo de atributo	Descripción
4 características de la pregunta	Atributos nominales que representan la partícula interrogativa, la clase de la pregunta, el tipo de respuesta esperada y el tipo de restricción de la pregunta.
1 característica de la compatibilidad pregunta-respuesta	Atributo booleano que indica la correspondencia entre el tipo de la respuesta candidata y el tipo de la respuesta esperada por la pregunta.
50 características de coincidencia	Atributos numéricos que describen el grado de coincidencia entre H' y T'. Estos atributos son calculados para cada uno de los cuatro tipos de términos de contenido y por cada una de las seis entidades nombradas. Además, estos diez diferentes atributos de coincidencia son calculados por cada uno de los cinco constituyentes en H'.
40 características de discrepancia	Atributos numéricos que indican el número de términos en T' que no coinciden con los de H' y que ocurren entre el constituyente de la respuesta y cada uno de los cuatro constituyentes de la pregunta. Estos cuatro diferentes atributos de discrepancia son calculados para cada uno de los cuatro tipos de términos de contenido así como para cada uno de los seis tipos de entidades nombradas.

Capítulo 5

Evaluación intrínseca:

La validación de respuestas

Para evaluar la funcionalidad del método propuesto (descrito en el capítulo 4) se implementó el sistema llamado VR-INAOE. En la evaluación se plantearon dos escenarios de prueba para dicho sistema, estos escenarios son una evaluación intrínseca y una evaluación extrínseca. En el caso de la evaluación intrínseca el objetivo fue evaluar la capacidad del método para distinguir las respuestas válidas de las erróneas. Mientras que en la evaluación extrínseca el objetivo fue medir el impacto que produce el método en el desempeño de los sistemas de BR. En este capítulo se muestra la evaluación intrínseca, la evaluación extrínseca se presenta en el siguiente capítulo.

5.1. Definición del experimento

5.1.1. Etapa de entrenamiento

Respecto al entrenamiento del sistema VR-INAOE, la colección de preguntas y respuestas en español llamada SPARTE fue utilizada como conjunto de datos de entrenamiento. Esta colección contiene los datos obtenidos durante las evaluaciones de

la búsqueda de respuestas en español de los años 2003, 2004 y 2005 (Peñas *et al.*, 2006). Todas las preguntas en la colección pertenecen a una complejidad de nivel uno (*i.e.,* preguntas del tipo factual y definición), esto significa que el sistema VR-INAOE está configurado para validar las respuestas a preguntas de ese nivel. En total el conjunto de entrenamiento contiene 2962 instancias, las cuales están formadas por una pregunta y una respuesta junto con su pasaje de soporte. Además, cada instancia tiene asignada una etiqueta (válida o errónea) que indica su valor en la validación. De las instancias un 23 % (695) corresponden a respuestas válidas mientras que el restante 77 % (2267) son respuestas erróneas. Esta diferencia en proporción de respuestas válidas y erróneas es un reflejo de los resultados del 2003 al 2005, así como del estado actual de la búsqueda de respuestas en español.

Es importante explicar que debido al desbalance en el conjunto de entrenamiento el máximo valor asignado por el clasificador a la distribución de probabilidad de la categoría errónea supera al de la categoría válida (las dos clases para las que fue entrenado). Este sesgo en el entrenamiento provoca que el sistema tienda a catalogar a la mayoría de las respuestas como erróneas. El resultado es un sistema con una poca cobertura para identificar las respuestas válidas, lo cual es un problema debido a que el propósito del sistema es filtrar las respuestas válidas para ayudar a mejorar los resultados de los sistemas de BR.

Con el propósito de incrementar la cobertura de respuestas válidas se incluyó en el sistema una *relajación de la clasificación*, esto de manera similar a lo propuesto en el sistema de VR descrito en (Moldovan *et al.*, 2007) donde se relaja una demostración lógica. En este caso, en lugar de sólo clasificar como respuestas válidas a aquellas donde el clasificador asigna una distribución de probabilidad mayor que 0.5 para esa clase (el umbral por default en un clasificador binario); también etiquetamos como respuestas válidas a las que tienen una distribución de probabilidad superior al valor que resulta de dividir entre dos la máxima distribución de probabilidad asignada a una respuesta válida durante la etapa de entrenamiento. En otras palabras, para clasificar cada respuesta se aplica una combinación de clasificación automática con aplicación de restricciones y valores por defecto (dos de los tres métodos aplicados en el estado del arte para clasificar la respuesta, los cuales fueron descritos en la sección 3.2.3).

En este caso, en el conjunto de entrenamiento la máxima distribución de probabilidad asignada por el clasificador es 1.0 para la clase errónea y 0.82 para la clase

válida, esta diferencia en valores es un reflejo del sesgo del clasificador hacia clasificar las respuestas como erróneas. Entonces, después de relajar la clasificación se establece que una respuesta va a ser etiquetada como válida si su distribución de probabilidad en esa categoría supera el umbral de 0.41 (la mitad de 0.82). Por lo tanto, el valor de confianza final para las respuestas válidas va a ser un valor normalizado entre 0 y 1 con respecto a la máxima distribución de probabilidad asignada por el clasificador para una respuesta en esa categoría (*i.e.,* como la máxima distribución de probabilidad para la clase válida es 0.82, una respuesta con dicha distribución de probabilidad en esa clase va a obtener una confianza de 1.0).

Cabe mencionar que en los experimentos en la etapa de entrenamiento además del algoritmo SVM basado en núcleos (el algoritmo con que trabaja el sistema), también se evaluaron algoritmos probabilísticos como el NaïveBayes, de árboles de decisión como el C4.5, basados en memorización de instancias como el KNN, así como sus ensambles usando algoritmos de Boosting y Bagging[1]. Donde a pesar de que el algoritmo SVM arrojó el mejor resultado de clasificación (aplicando una validación cruzada de 10 pliegues), este no supero significativamente a los otros métodos[2]. Esto muestra que los atributos propuestos para discriminar las respuestas son independientemente del algoritmo de aprendizaje utilizado. Para todos los procesamientos del lenguaje necesarios, llámese, lematización, etiquetado de las partes de la oración, reconocimiento y clasificación de entidades nombradas y el análisis sintáctico superficial, se utilizó la herramienta de libre distribución Freeling[3] (Atserias *et al.*, 2006).

5.1.2. Conjunto de prueba

Con el propósito de mostrar la capacidad del sistema para distinguir las respuestas válidas de las erróneas, en este experimento se utilizó como conjunto de prueba una colección de preguntas y respuestas en español que contiene un total de 2286 instancias. Dichas instancias corresponden a datos obtenidos durante la evaluación de la búsqueda

[1]La implementación de todos estos algoritmos está disponible en la librería para Java llamada Weka (Witten & Frank, 1999).

[2]De acuerdo a los resultados de una prueba de significancia estadística para comparar la diferencia de la proporción pA con la proporción pB como se describe en (Dietterich, 1998), donde H_0 indica que pA es igual a pB utilizando una región de rechazo $|z| > Z_{\alpha=0,05}$ =1.96

[3]Disponible en `http://www.freeling.com`

de respuestas en el CLEF del año 2006. Lo cual permite retomar las evaluaciones humanas de la búsqueda de respuestas para la evaluación de la validación de respuestas. En el conjunto un 29 % (671) de las instancias son respuestas válidas y el restante 71 % (1615) de las instancias corresponden a respuestas erróneas. En (Peñas *et al.*, 2007) se dan detalles de cómo fue formado este conjunto de prueba.

La decisión de utilizar este conjunto como de prueba es su compatibilidad con el conjunto utilizado como de entrenamiento, es decir, ambos contienen respuestas a preguntas del nivel uno en la búsqueda de respuestas (*i.e.,* preguntas del tipo factual y definición). Además de que es en esta colección de prueba donde se han evaluado la mayor parte de los métodos de VR aplicados al español, lo cual permite comparar directamente los resultados obtenidos por el sistema basado en el método propuesto con los obtenidos por otros sistemas de VR.

5.1.3. Medidas de evaluación

Para la evaluación intrínseca se utilizaron las medidas propuestas por (Peñas *et al.*, 2007) para evaluar a sistemas de VR. Estas medidas de evaluación se basan en las medidas tradicionales de recuperación de información: precisión, cobertura y la medida-F. Aunque con la característica de que dada la naturaleza del problema donde actualmente no existe un balance entre respuestas válidas y erróneas, las medidas se calculan sólo para las respuestas que son clasificadas como válidas por los sistemas de VR (ver las fórmulas 5.1, 5.2 y 5.3). De otra manera un sistema que rechaza todas las respuestas podría obtener resultados demasiado altos e inútiles para propósitos de evaluación. Por ejemplo en el español, donde los sistemas de búsqueda de respuestas producen más de un 70 % de respuestas erróneas.

$$Precisión = \frac{\#\_respuestas\_válidas\_clasificadas\_correctamente}{\#\_respuestas\_clasificadas\_como\_válidas} \tag{5.1}$$

$$Cobertura = \frac{\#\_respuestas\_válidas\_clasificadas\_correctamente}{\#\_respuestas\_válidas\_en\_la\_colección\_de\_prueba} \tag{5.2}$$

$$Medida\text{-}F = \frac{2 \times Precisión \times Cobertura}{Precisión + Cobertura} \tag{5.3}$$

Todas estas medidas de evaluación generan valores entre 0 y 1, siendo uno el mejor resultado posible. En el caso de la precisión un valor igual a uno significa que el sistema nunca se equivoca cuando clasifica una respuesta como válida, mientras que una precisión de cero significa que todas sus respuestas clasificadas como válidas en realidad son erróneas. En la cobertura un valor igual a uno sucede cuando el sistema clasifica correctamente todas las respuestas válidas en la colección de prueba. Entre más respuestas válidas sean mal clasificadas por el sistema su cobertura decrece.

La medida-F es utilizada para tener un valor único en la evaluación de un sistema. Esta medida es la combinación lineal de la precisión y la cobertura, el número 2 en la formula sirve para indicar que la precisión y la cobertura tienen la misma importancia en el cálculo de esta medida.

Es importante hacer notar que estas medidas están diseñadas para evaluar la capacidad de los sistemas para localizar las respuestas válidas. Por lo tanto, para garantizar su correcta aplicación es necesario que el conjunto de prueba contenga al menos una respuesta válida y que el sistema a evaluar clasifique por lo menos una respuesta como válida en el conjunto de prueba. Además, en el caso de que ninguna de las respuestas clasificadas como válidas por el sistema sea correcta, la medida-F por default es cero.

5.2. Resultados y comparación

La tabla 5.1 muestra los resultados de validar las respuestas del conjunto de prueba. Esta tabla, además, incluye los resultados alcanzados por sistemas de VR del estado del arte sobre la misma colección de prueba. Cabe mencionar que todos los sistemas de VR comparados en la tabla utilizan la colección de datos de entrenamiento SPARTE ya sea para ajustar o entrenar su sistema, por lo tanto la diferencia de los resultados sólo depende de los métodos empleados por cada uno de los sistemas. Respecto a tales métodos, la tabla 5.2 resume las características generales de cada uno de los sistemas de VR comparados (estas características fueron descritas en la sección 3.2).

La evaluación muestra que el resultado de VR-INAOE supera al de los otros sistemas de VR evaluados sobre el mismo conjunto de prueba. Además, cabe hacer notar que una medida-F igual a 0.61 es el máximo resultado alcanzado por un sistema de VR en español. Ahora sobre la misma colección de prueba (y utilizando la misma colección

Tabla 5.1: Resultados de la evaluación intrínseca

Sistema de VR	Medida-F	Precisión	Cobertura
Tatu *et al.* (2007)	0.61	0.53	0.71
Herrera *et al.* (2006)	0.57	0.47	0.72
Herrera *et al.* (2006)	0.56	0.47	0.71
Rodrigo *et al.* (2007a)	0.53	0.44	0.68
Kozareva *et al.* (2007)	0.53	0.41	0.76
Bosma & Callison-Burch (2007)	0.47	0.48	0.46
Bosma & Callison-Burch (2007)	0.43	0.55	0.36
Kozareva *et al.* (2007)	0.43	0.47	0.39
VR-INAOE	0.75	0.74	0.76

Tabla 5.2: Características de los sistemas de VR comparados

Sistema de VR	Generación de la hipótesis	Análisis de similitud entre T y H	Clasificación de la respuesta
Tatu *et al.* (2007)	Patrones	Semántico	Demostración lógica y restricciones
Herrera *et al.* (2006)	Patrones	Léxico-morfológico	Clasificación automática
Herrera *et al.* (2006)	Patrones	Léxico-morfológico	Clasificación automática
Rodrigo *et al.* (2007a)	Patrones	Léxico-morfológico	Restricciones
Kozareva *et al.* (2007)	Patrones	Léxico-morfológico	Clasificación automática
Bosma & Callison-Burch (2007)	Patrones	Semántico	Clasificación automática
Bosma & Callison-Burch (2007)	Patrones	Sintáctico	Clasificación automática
Kozareva *et al.* (2007)	Patrones	Léxico-morfológico	Clasificación automática
VR-INAOE	Automático	Sintáctico	Clasificación automática y restricciones

de entrenamiento) el resultado del sistema implementado con el método propuesto fue capaz de superar significativamente ese máximo resultado[1]. En particular, el sistema propuesto comparado con los otros sistemas obtiene una mayor precisión. Sin embargo, este resultado aún se encuentra lejos de ser perfecto, en especial la falta de cobertura muestra que varias de las respuestas válidas en el conjunto de prueba son incorrectamente clasificadas por el sistema. Lo cual muestra la incapacidad del método para filtrar todas las respuestas válidas de los sistemas de BR. En la siguiente sección se discute al respecto.

5.3. Análisis de los resultados

De los resultados en la tabla 5.1 se puede ver que el sistema VR-INAOE en contraste con los otros sistemas suele mejorar la precisión aunque carece de cobertura. Tomando en cuenta que el análisis de la seccion 3.3.1 reveló que en una muestra de la colección de prueba hay aproximadamente un 34 % de casos de respuestas válidas que requieren de conocimiento de la semántica del lenguaje para ser evaluadas correctamente, debido a que el sistema VR-INAOE no incluye el conocimiento semántico necesario no puede etiquetar correctamente dichas respuestas válidas lo que produce la pérdida de cobertura.

Sin embargo, la relajación de la clasificación en el sistema VR-INAOE en algunos casos le permitió clasificar respuestas válidas correctamente a pesar de no contar con todo el conocimiento necesario. La tabla 5.3 muestra el resultado del sistema utilizando y omitiendo la relajación. En estos resultados se puede ver que hay una ganancia de un 19 % en la cobertura cuando se incluye la relajación, aunque esto también conlleva una pérdida del 7 % en la precisión (la relajación provoca que algunas respuestas erróneas sean etiquetadas como válidas).

Entre los casos de respuestas válidas que favorece la relajación se encuentran principalmente aquellos donde se detectó la coincidencia de los actores oculto y visible así como la discrepancia de algún término entre estos dos constituyentes, pero que

[1]De acuerdo a los resultados de una prueba de significancia estadística para comparar la diferencia de la proporción pA con la proporción pB, donde H_0 indica que pA es igual a pB con una región de rechazo $|z| > Z_{\alpha=0,05}$ =1.96; en este caso las proporciones comparadas son pA =0.75 y pB =0.61 con un $n = 2286$ y dando como resultado una z =10.15 lo que permite rechazar la hipótesis nula.

Tabla 5.3: La relajación en el sistema VR-INAOE

	Medida-F	Precisión	Cobertura
Con relajación	0.75	0.74	0.76
Sin relajación	0.71	0.79	0.64

además el constituyente de la acción no tuvo coincidencia. Tal como ocurre en el caso de *H:* `Irak invadió el país de Kuwait` y *T:* `liberó el país de Kuwait de la ocupación iraquí`, donde los términos que coinciden son `país-país`, `Kuwait-Kuwait` e `Irak-iraquí`; aquí aunque no coincide en H el verbo `invadió` existe en T un término sin coincidencia (con discrepancia) (`ocupación`) que hace que aumente la confianza en reconocer la implicación textual entre (T,H), este aumento en el valor de confianza es suficiente para que con la relajación su etiqueta sea la de válida.

Respecto a la aportación de los nuevos atributos incluidos en este sistema, en la tabla 5.4 se muestra la *ganancia de información* (GI) máxima y promedio obtenida para los diversos grupos de atributos empleados por el sistema para clasificar las respuestas en la colección de entrenamiento y prueba. La medida GI calcula que tan bien un cierto atributo a separa un conjunto de datos D conforme a las categorías C dadas a las instancias, las formulas 5.4 y 5.5 muestran como se calcula esta medida. La diferencia en el valor de la GI a favor de los nuevos atributos muestra la importancia de tales aportaciones.

$$GI(D,a) = Entropía(D) - \sum_{V_i \in Valores(a)} \frac{|D_{v_i}|}{|D|} Entropía(D_{v_i}) \quad (5.4)$$

$$Entropía(D) = \sum_{i=1}^{|C|} -P(c_i) log_2 P(c_i) \quad (5.5)$$

En la tabla 5.4 también se puede ver que para cada grupo de atributos (de coincidencia, de discrepancia y de la pregunta-respuesta) los más discriminativos son, respectivamente, la coincidencia de sustantivos en el actor visible, la discrepancia de sustantivos entre el par respuesta-actor visible y la clase de la pregunta. Esto significa que, según nuestros resultados, los sustantivos son de los términos más importantes para validar respuestas. Por lo tanto, en una respuesta válida esta clase de términos deben de coincidir en H y no mostrar discrepancias en el núcleo de T (*i.e.*, T'). Este

Tabla 5.4: Ganancia de información de los atributos del sistema VR-INAOE

Características	GI promedio	GI máxima	Atributo con GI máxima
	Conjunto de entrenamiento		
De coincidencias	0.003	0.024	Sustantivos en actor visible
De discrepancias	0.013	0.059	Sustantivos entre respuesta-actor visible
De la pregunta-respuesta	0.020	0.027	Clase de la pregunta
	Conjunto de prueba		
De coincidencias	0.006	0.046	Sustantivos en actor visible
De discrepancias	0.041	0.114	Sustantivos entre respuesta-actor visible
De la pregunta-respuesta	0.016	0.032	Clase de la pregunta

conocimiento es especialmente útil para identificar respuestas erróneas a pesar de que cuenten con pasajes de soporte relevantes (*p. ej.*, la respuesta `ONU` en la tabla 3.1 donde todos los sustantivos de H coinciden pero el sustantivo `Irak` muestra la discrepancia en T').

Por otro lado, la información de la clase de la pregunta le sirve al sistema para aprender a clasificar correctamente respuestas válidas donde existe información implícita en H' que debe coincidir con información explícita en T'. Por ejemplo, creando un árbol de decisión con el algoritmo C4.5 para el conjunto de entrenamiento, encontramos que para preguntas del tipo definición se tiene en una de las ramas del árbol la regla de decisión siguiente: si en H' no se puede hacer coincidir la acción y en T' entre la respuesta y el actor visible no hay discrepancias de verbos o sustantivos, entonces se trata de una respuesta válida. Esta regla de decisión refleja los casos de aposición y asignación de adjetivos comunes en el lenguaje para definir algún concepto o persona (*p. ej.*, para la pregunta `¿Qué es el Hubble?` se tiene la respuesta válida `telescopio espacial` con T' como *i*) `el telescopio espacial Hubble`, *ii*) `el Hubble, el telescopio espacial,` y *iii*) `Hubble (el telescopio espacial)`).

En resumen, de los resultados de la evaluación intrínseca se puede decir que el reconocimiento de la implicación textual ha resultado efectivo para validar respuestas de sistemas de BR. En particular, para aquellas respuestas donde *i*) el conocimiento de las convenciones de uso del lenguaje es suficiente para reconocer la implicación textual en las respuestas válidas y donde *ii*) la información propia del problema —la validación de respuestas— sirve para descartar la implicación textual en respuestas erróneas.

Capítulo 6

Evaluación extrínseca: El impacto en la búsqueda de respuestas

Con el propósito de mostrar el impacto que tiene el método propuesto en la búsqueda de respuestas, en este capítulo se presenta la evaluación extrínseca del sistema VR-INAOE. Las siguientes secciones detallan esta evaluación y los resultados alcanzados.

6.1. Definición del experimento

Tomando en cuenta los enfoques actuales para integrar la validación de respuestas con los sistemas de BR (descritos en la sección 2.1.3), en esta evaluación se tienen dos objetivos: *i*) mostrar que el sistema VR-INAOE puede exitosamente combinar las respuestas de diversos sistemas de BR que se complementan; y *ii*) mostrar que se puede mejorar el resultado de un sistema de BR por aplicar el sistema VR-INAOE a sus respuestas. En comparación con la evaluación intrínseca, en este caso el sistema además de validar las respuestas también debe seleccionar una respuesta final para cada pregunta.

Para evaluar el primer objetivo se utilizaron como datos de prueba una colección de 190 preguntas y la respuesta de 17 diferentes sistemas de BR para cada pregunta

(obteniendo un total de 2369 respuestas diferentes de *nil* de las cuales sólo el 28 % son respuestas válidas). Las preguntas y respuestas en el conjunto de prueba fueron evaluadas en la búsqueda de respuestas en español del CLEF del año 2006. Dichas preguntas, al igual que en el conjunto de entrenamiento, pertenecen a un nivel uno en la clasificación de la complejidad de la búsqueda de respuestas (el entrenamiento del sistema fue descrito en la sección 5.1.1).

Respecto al segundo objetivo, como datos de prueba se utilizaron las respuestas del sistema de BR descrito en (Juárez-González *et al.*, 2006) para responder a las mismas preguntas de la colección utilizada para evaluar el primer objetivo. En este caso al sistema de BR se le solicitaron 50 respuestas por pregunta (obteniendo un total de 1308 respuestas diferentes de *nil* de las cuales un 33 % son respuestas válidas). En ambos objetivos el sistema VR-INAOE tiene la tarea de filtrar del conjunto de respuestas (provenientes de uno o varios sistemas de BR) una válida, si existe, para contestar a cada pregunta.

Cabe mencionar que una correcta validación de respuestas además de permitir filtrar una respuesta válida de entre todas las respuestas candidatas a una pregunta, también debe ser capaz de identificar cuando todas esas respuestas proporcionadas a la pregunta son erróneas. Por lo tanto, para evaluar el impacto de los sistemas de VR en el desempeño de los sistemas de BR, las medidas de evaluación a utilizar son las descritas en (Rodrigo *et al.*, 2008) para evaluar el desempeño de sistemas de BR multi-flujo. Estas medidas son la *exactitud*, la *exactitud de rechazo* y el *desempeño estimado en BR*, las fórmulas 6.1, 6.2 y 6.3 muestran la forma de calcular estas medidas.

$$exactitud = \frac{\#\_preguntas\_contestadas\_con\_una\_respuesta\_válida + \#\_preguntas\_nil\_sin\_contestar}{\#\_preguntas\_en\_la\_colección\_de\_prueba} \tag{6.1}$$

$$Exactitud\_de\_rechazo = \frac{\#\_preguntas\_contestadas\_nil\_correctamente\_con\_VR}{\#\_preguntas\_en\_la\_colección\_de\_prueba} \tag{6.2}$$

$$desempeño\_estimado\_en\_BR = exactitud + exactitud\_de\_rechazo \times exactitud \tag{6.3}$$

En el caso de la exactitud, esta medida muestra la proporción de preguntas contestadas correctamente por el sistema del total de preguntas a contestar. Los valores que toma la exactitud son de 0 a 1, una exactitud de uno indica que el sistema contestó todas las preguntas correctamente y de cero que en todas se equivocó. Cabe hacer notar que una pregunta es contestada correctamente si se le otorga una respuesta válida, o bien si el sistema correctamente dice que no existe respuesta alguna en la colección de documentos de prueba para contestar a dicha pregunta (el caso de las preguntas *nil*).

Respecto a la exactitud de rechazo, esta medida muestra la capacidad del sistema para reconocer que entre las respuestas que pretende combinar no existe alguna válida para responder a la pregunta. El valor de esta medida es 1 sólo si se cumplen las dos condiciones siguientes: *i*) para todas las preguntas de prueba las respuestas a combinar son erróneas, y *ii*) el sistema de validación rechaza todas esas respuestas y contesta con un *nil* a las preguntas. En caso de que existan respuestas válidas que combinar para todas las preguntas, el valor de la exactitud de rechazo siempre es 0. Cuando sólo algunas de las preguntas de prueba tienen respuestas válidas que combinar, el valor de esta medida obtiene valores menores que 1 pero no inferiores a 0.

Por lo tanto, con la medida del desempeño estimado en BR se obtiene un valor que refleja la capacidad de los sistemas tanto para responder correctamente a las preguntas, así como para evitar dar respuestas erróneas a las preguntas cuando los sistemas de BR son incapaces de obtener al menos una respuesta válida. El valor de esta medida va de 0 a 1, siendo un valor de uno cuando todas las preguntas son contestadas correctamente y cero cuando ninguna de las preguntas pudo ser contestada.

6.2. Resultados y comparación

La tabla 6.1 muestra el desempeño estimado alcanzado con el sistema VR-INAOE tanto para combinar sistemas de BR complementarios (BR multi-flujo) así como para validar las respuestas de un único sistema de BR (BR típica). La tabla también muestra los resultados de los diecisiete sistemas de BR empleados para la BR multi-flujo. Además, de estos sistemas el descrito en (Juárez-González *et al.*, 2007) es el sistema al que se le incorporó de manera individual el sistema VR-INAOE.

Estos resultados muestran que el método propuesto permite *i*) crear un sistema de BR multi-flujo exitoso superando el máximo resultado alcanzado por el mejor de sus flujos (los sistemas de BR); y *ii*) mejorar el desempeño actual de un sistema de BR por medio de validar sus respuestas.

A pesar de que aún se está lejos de una validación de respuestas perfecta, el resultado en la combinación de múltiples sistemas de BR muestra que es posible mejorar con una significancia estadística el mejor resultado actualmente alcanzado en la búsqueda de respuestas en español (un desempeño estimado de 0.53)[1], esto por utilizar VR-INAOE para la combinación.

Respecto a incorporar el sistema VR-INAOE a un único sistema de BR, el resultado de evaluación muestra que principalmente la validación de respuestas ayuda a eliminar muchas de las respuestas erróneas producidas por el sistema de BR (véase el valor de la exactitud de rechazo); pero, al mismo tiempo la correcta selección de respuestas válidas permite responder a varias de las preguntas (véase el valor de la exactitud). Esta combinación de resultados le permiten al sistema propuesto incrementar el desempeño estimado del sistema descrito en (Juárez-González *et al.*, 2007).

Como se describió en la sección 2.1.3, la validación de respuestas no es la única opción para crear sistemas de BR muti-flujo. Con el propósito de comparar el sistema VR-INAOE —y en general los sistemas que utilizan una validación de respuestas— contra los otros enfoques existentes para crear un sistema de BR multi-flujo, en la tabla 6.2 se presenta el desempeño estimado alcanzado por estos otros enfoques en la colección de prueba. De los resultados podemos reafirmar la conclusión de Jijkoun & de Rijke (2004), la cual indica que la combinación de coro-caballo negro supera a todos los otros enfoques tradicionales. Pero, además podemos agregar que el enfoque que utiliza validación de respuestas supera con una significancia estadística a todos los enfoques tradicionales que no la utilizan[2].

[1] En la comparación de los resultados se utilizó la prueba de significancia estadística para comparar la diferencia de la proporción pA con la proporción pB, donde H_0 indica que pA es igual a pB con una región de rechazo $|z| > Z_{\alpha=0,05}$ =1.96; en este caso las proporciones comparadas son pA =0.74 y pB =0.53 con un $n = 190$ y dando como resultado una z =4.25 lo que permite rechazar la hipótesis nula.

[2] Para comparar los resultados se utilizó la prueba de significancia estadística para comparar la diferencia de la proporción pA con la proporción pB, donde H_0 indica que pA es igual a pB con una región de rechazo $|z| > Z_{\alpha=0,05}$ =1.96; en este caso las proporciones comparadas son pA =0.74 y pB =0.58 con un $n = 190$ y resultando una z =3.29 que permite rechazar la hipótesis nula.

Tabla 6.1: Resultados de la evaluación extrínseca

Sistema de BR descrito en	Desempeño estimado	Exactitud	Exactitud de rechazo
(Cassan *et al.*, 2007)	0.53	0.53	0.00
(Juárez-González *et al.*, 2007)	0.51	0.51	0.00
(Pérez-Coutiño *et al.*, 2007)	0.42	0.42	0.00
(Ferrández *et al.*, 2007)	0.37	0.37	0.00
(Buscaldi *et al.*, 2007)	0.34	0.34	0.00
(Cassan *et al.*, 2007)	0.30	0.30	0.00
(Buscaldi *et al.*, 2007)	0.30	0.30	0.00
(Tomás & González, 2007)	0.23	0.23	0.00
(Tomás & González, 2007)	0.23	0.23	0.00
(Ferrández *et al.*, 2007)	0.22	0.22	0.00
(de Pablo-Sánchez *et al.*, 2007)	0.21	0.21	0.00
(Costa, 2007)	0.20	0.20	0.00
(Bowden *et al.*, 2007)	0.19	0.19	0.00
(de Pablo-Sánchez *et al.*, 2007)	0.15	0.15	0.00
(Bos & Nissim, 2007)	0.14	0.14	0.00
(Bos & Nissim, 2007)	0.10	0.10	0.00
(Bos & Nissim, 2007)	0.06	0.06	0.00
VR-INAOE en BR multi-flujo	0.74	0.65	0.15
Validación perfecta en BR multi-flujo	0.95	0.77	0.23
VR-INAOE en BR típica	0.57	0.47	0.23
Validación perfecta en BR típica	0.81	0.57	0.44

Tabla 6.2: Validación de respuestas contra enfoques típicos en BR multi-flujo

	Desempeño estimado	Exactitud	Exactitud de rechazo
Ordenamiento ligero	0.52	0.52	0.00
Caballo negro	0.52	0.52	0.00
Coro	0.53	0.53	0.00
Coro Web	0.17	0.17	0.00
Coro-Caballo negro	0.58	0.58	0.00
VR-INAOE	0.74	0.65	0.15
Mejor sistema de BR en los flujos	0.53	0.53	0.00
Sistema de BR multi-flujo perfecto	0.95	0.77	0.23

6.3. Análisis de los resultados

Los resultados de la evaluación extrínseca dejaron ver que, aplicando VR-INAOE es posible mejorar significativamente el desempeño actual en la búsqueda de respuestas en Español. Sin embargo, aún se está lejos del incremento máximo que se puede lograr con esta tecnología. Un análisis de los resultados muestra que la etapa de selección de las respuestas no es un problema con el método propuesto, esto tomando en cuenta que en un 98 % de las veces que se clasifica correctamente al menos una respuesta como válida por cada pregunta, una de éstas es seleccionada por el sistema como la respuesta final.

Cabe mencionar que, debido al análisis aplicado en la validación de respuestas, el método propuesto para seleccionar las respuestas siempre elige aquella considerada válida y que tiene el contexto más simple. Por ejemplo, para la pregunta `¿A qué país invadió Irak en 1990?` y la respuesta correctamente seleccionada `Kuwait`, el sistema prefiere la que está soportada por el fragmento de texto `Kuwait, país invadido por Irak en 1990` en lugar de la soportada por `Irak desde que sus tropas invadieran Kuwait en agosto de 1990`, mientras que prefiere rechazar la soportada por el fragmento de texto `Irak sigue en posesión de unos 9.000 equipos militares incluidos misiles y cohetes que robó a Kuwait durante su invasión y ocupación del emirato en 1990`.

Tabla 6.3: El sistema VR-INAOE en preguntas factuales

	Desempeño estimado	Exactitud	Exactitud de rechazo
VR-INAOE	0.71	0.62	0.14
Mejor sistema de BR en los flujos	0.46	0.46	0.00
Sistema de BR multi-flujo perfecto	0.94	0.75	0.25

Tabla 6.4: El sistema VR-INAOE en preguntas de definición

	Desempeño estimado	Exactitud	Exactitud de rechazo
VR-INAOE	0.86	0.74	0.17
Mejor sistema de BR en los flujos	0.83	0.83	0.00
Sistema de BR multi-flujo perfecto	0.97	0.83	0.17

Con el propósito de tener una idea de cómo se comporta el sistema VR-INAOE en los diferentes tipos de preguntas evaluadas. En las tablas 6.3 y 6.4 se muestran, respectivamente, los resultados para las preguntas factuales y de definición en la colección de prueba. En estos resultados se puede ver que en las preguntas factuales es donde realmente se está haciendo una mejora importante al resultado del mejor sistema de BR (obteniendo una ganancia de un 54 %). Mientras que en el caso de las preguntas de definición la ganancia es mínima (aproximadamente un 4 %), esto se debe a que actualmente los sistemas de BR ya alcanzan altos resultados en este tipo de preguntas y por lo tanto un proceso extra de validación de respuestas puede no ser requerido. Esto muestra que los esfuerzos futuros en la validación de respuestas se deben enfocar principalmente a las preguntas del tipo factual.

Para finalizar este análisis de los resultados en la tabla 6.2 se puede ver que con la validación de respuestas es posible construir un sistema de BR multi-flujo con una exactitud de rechazo diferente de cero. En contraste, en los enfoques tradicionales siempre que exista una respuesta candidata a la pregunta éstos van a otorgar una respuesta final así sea errónea. Por lo tanto, la validación de respuestas es la mejor opción para combinar las respuestas de sistemas de BR con bajos resultados, como ocurre en el español.

6.4. Evaluación con los conjuntos de prueba del AVE 2007 y AVE 2008

Como parte de la evaluación extrínseca del sistema VR-INAOE también se realizaron experimentos utilizando las colecciones de preguntas y respuestas de prueba del AVE (el ejercicio de validación de respuestas del CLEF, por sus siglas en ingles de *Answer Validation Exercise*) realizado en los años 2007 y 2008. Los cuales corresponden a preguntas y respuestas evaluadas en el CLEF de esos mismos años. Respecto a la colección de prueba del 2007, ésta contiene 564 respuestas generadas por cinco sistemas de BR para responder a 170 preguntas; de las respuestas sólo el 23 % (127) son respuestas válidas. Mientras que en la colección de prueba del 2008, existen 1528 respuestas generadas por diez sistemas de BR para responder a 136 preguntas (cada sistema de BR extrajo tres respuestas por pregunta en lugar de sólo una); en esta colección sólo el 10 % (153) de las respuestas son válidas. La tabla 6.5 muestra un resumen de los resultados alcanzados por el sistema VR-INAOE y otros sistemas evaluados sobre las mismas colecciones de prueba.

Como se puede ver en la tabla, los resultados del sistema VR-INAOE son menores en comparación con los presentados en la sección 6.2. Esta baja en los resultados se debe en parte al menor número de respuestas válidas en los conjuntos de prueba, pero principalmente a que en el conjunto de entrenamiento (y en el conjunto de prueba del experimento descrito en la sección 6.2) los textos analizados son de noticias de periódicos mientras que en estos nuevos conjuntos de prueba las respuestas fueron extraídas en su mayoría de la enciclopedia llamada Wikipedia[1] (en las colecciones del AVE 2007 y del AVE 2008 un 67 % y un 63 % del total de las respuestas, respectivamente, fueron extraídas de esta enciclopedia).

En Wikipedia la forma de presentar la información difiere de como es presentada en las noticias de periódicos (donde fue entrenado el sistema), por ejemplo para responder a la pregunta `¿Quién fue Marco Pantani?` en las noticias es común encontrar el texto `el ciclista italiano Marco Pantani`, mientras que en Wikipedia esa misma respuesta aparece como `Marco Pantani (Cesenatico, Italia, 13 de enero de 1970 - Rimini, Italia, 14 de febrero de 2004) un ciclista italiano`. Más aún,

[1] `http://es.wikipedia.org`

Tabla 6.5: Evaluación extrínseca en el AVE: Respuestas de múltiples flujos

	Desempeño estimado	Exactitud	Exactitud de rechazo
	Colección del AVE del año 2007		
Rodrigo *et al.* (2007b)	0.49	0.42	0.18
Cumbreras *et al.* (2007a)	0.41	0.41	0.01
Cumbreras *et al.* (2007a)	0.41	0.41	0.01
VR-INAOE	0.59	0.47	0.25
Mejor sistema de BR individual	0.49	0.49	0.00
Sistema de BR multi-flujo perfecto	0.83	0.59	0.41
	Colección del AVE del año 2008		
Ferrández *et al.* (2008b)	0.37	0.32	0.14
Ferrández *et al.* (2008b)	0.33	0.27	0.21
Cumbreras *et al.* (2007a)	0.06	0.04	0.32
Cumbreras *et al.* (2007a)	0.03	0.02	0.35
VR-INAOE	0.38	0.36	0.06
Mejor sistema de BR individual	0.54	0.54	0.00
Sistema de BR multi-flujo perfecto	0.85	0.62	0.38

en Wikipedia es común la correferencia al título de cada documento; por ejemplo, el documento con título `Templo de Debod` en su fragmento de texto `En su nuevo emplazamiento, fue inaugurado en julio de 1972 por Carlos Arias Navarro, alcalde de Madrid` presenta la respuesta a la pregunta `¿Quién inauguró el Templo de Debod en Madrid?`, esto tomando en cuenta que el adjetivo posesivo "su" hace referencia al concepto representado por el título.

Esta diferencia en los textos además de limitar aún más a las herramientas que se utilizan para el análisis sintáctico necesario en el sistema, provocan una falta de precisión en la validación de respuestas (en estas colecciones de prueba el sistema en promedio valida las respuestas con una precisión de 0.38 y una cobertura de 0.71). El resultado es una mala selección de respuestas debido al alto número de respuestas erróneas mal clasificadas. Además, en la colección de prueba del AVE 2008 la frecuencia de las respuestas fue eliminada; es decir, cada respuesta se repite sólo una vez en la colección por lo que la redundancia de respuestas válidas no puede ser aprovechada por el método propuesto en su etapa de selección de respuestas (descrita en la sección 4.2).

Para finalizar cabe señalar como los bajos resultados se ven reflejados en la pobre ganancia de información que obtienen los nuevos atributos propuestos en estas nuevas colecciones de prueba. La tabla 6.6 muestra un resumen de los valores de GI obtenidos. Dichos valores muestran como en estas colecciones de prueba los atributos del sistema son menos discriminativos que en las colecciones de respuestas extraídas exclusivamente de textos de noticias (ver tabla 5.4). De estos nuevos experimentos se puede concluir que para validar correctamente respuestas extraídas de textos como los de Wikipedia es necesario tanto incluir en la etapa de entrenamiento ejemplos de esta fuente, así como nuevos recursos de procesamiento del lenguaje que realicen la resolución de correferencias. Algo que hasta el momento no incluye el sistema VR-INAOE.

Tabla 6.6: Ganancia de información de los atributos del sistema propuesto en las colecciones de prueba del AVE 2007 y 2008

Características	GI promedio	GI máxima	Atributo con GI máxima
	Conjunto de prueba AVE-2007		
De coincidencias	0.003	0.035	Fechas en restricción
De discrepancia	0.038	0.106	Sustantivos entre respuesta-actor visible
De la pregunta-respuesta	0.015	0.027	Clase de la pregunta
	Conjunto de prueba AVE-2008		
De coincidencias	0.002	0.014	Sustantivos en actor visible
De discrepancias	0.014	0.029	Sustantivos entre respuesta-actor visible
De la pregunta-respuesta	0.009	0.027	Clase de la pregunta

Capítulo 7

Síntesis y conclusiones

En síntesis esta tesis mostró un método de validación de respuestas basado en el enfoque de reconocer la implicación textual. Con este método se implemento el sistema llamado VR-INAOE, el cual utiliza un clasificador basado en aprendizaje automático supervisado para decidir cuándo cada una de las respuestas de los sistemas de búsqueda de respuestas son válidas o erróneas para contestar una pregunta. Adicionalmente, el método propuesto en una segunda etapa realiza una selección de respuestas, esto para elegir una respuesta final a partir de las respuestas que previamente fueron etiquetadas como válidas para una pregunta.

Tomando en cuenta métodos similares, el método propuesto difiere de éstos en lo siguiente:

- En comparación con otros métodos de validación de respuestas, el método propuesto incluye un pre-proceso del texto y la hipótesis que evita una alta coincidencia con información irrelevante al estudio; además, éste también incluye atributos de discrepancia y del análisis de la pregunta-respuesta que aumentan su precisión. También es importante mencionar que el método está basado únicamente en un análisis léxico-sintáctico del texto, donde no se utilizan recursos para un análisis profundo del lenguaje ni se consultan fuentes de conocimiento externo (elementos característicos de los métodos que intentan llevar el análisis a un nivel semántico).

- En contraste con los enfoques típicos para combinar las respuestas de sistemas de BR complementarios, el método propuesto permite combinar dichas respuestas sin considerar ninguna información acerca de la confianza en los sistemas de BR utilizados; además, la redundancia de las respuestas es un atributo secundario en el método para seleccionar una respuesta final mientras que en muchos de los mejores enfoques típicos es una característica básica.

Respecto a los resultados, los experimentos en preguntas y respuestas en español con el sistema VR-INAOE mostraron lo siguiente:

- En la validación de respuestas el sistema VR-INAOE logró resultados superiores o comparables a los mejores en el campo de estudio. Como sucede con el mejor resultado históricamente alcanzado por un sistema de validación de respuestas en español (una medida-F de 0.61), el cual es superado significativamente al ejecutar VR-INAOE sobre la misma colección de respuestas de prueba (logrando una medida-F de 0.75).
- Con su incorporación en sistemas de búsqueda de respuestas, el sistema VR-INAOE permitió tanto incrementar los resultados de un único sistema de BR así como combinar exitosamente las respuestas de diferentes sistemas de BR complementarios. Tal como se pudo ver con la colección de preguntas de prueba en español del CLEF del año 2006, donde con ayuda de VR-INAOE un sistema de BR logró incrementar su resultado (pasando de un desempeño estimado de 0.51 a un 0.57 en su evaluación). Además, utilizando VR-INAOE para combinar las respuestas de los sistemas de BR evaluados en el foro de ese año se consiguió mejorar el máximo desempeño estimado alcanzado por el mejor sistema en dicha evaluación (un 0.53), logrando obtener un resultado de 0.74 en el desempeño estimado.

Finalmente se puede concluir que:

- **El sistema VR-INAOE con un análisis a nivel léxico-sintáctico del texto logra reconocer la implicación textual en respuestas válidas para las preguntas.** Esta implementación del sistema hace que éste sea muy apropiado para interactuar con los sistemas de BR con un pobre desempeño, los cuales representan el estado actual para la mayoría de los idiomas donde se ha evaluado a

la búsqueda de respuestas y donde herramientas confiables para un procesamiento profundo del lenguaje aún no están disponibles, como es el caso del español.

- **Las características tradicionales relacionadas a evaluar coincidencias en el análisis de similitud entre el par (T, H) se complementan tanto con las nuevas características utilizadas para evaluar sus discrepancias, así como con los atributos obtenidos del análisis de la pregunta-respuesta.** Esta combinación de atributos le permiten al sistema VR-INAOE analizar correctamente situaciones donde existe un alto traslape entre los textos analizados, pero no necesariamente ocurre la implicación textual entre ellos.

- **La combinación de respuestas de sistemas de BR complementarios es más adecuada utilizando el sistema VR-INAOE que aplicando los enfoques típicos de combinación de respuestas.** Esto tomando en cuenta que en la evaluación de un nuevo sistema de BR se desconoce su confianza para contestar a las preguntas y que en lenguajes con sistemas de BR con pobres resultados —como el español— no existe una alta redundancia de respuestas válidas.

- **El máximo desempeño alcanzado por un sistema de BR en español puede ser mejorado utilizando el sistema VR-INAOE tanto para validar las respuestas de un único sistema de BR pero sobre todo para combinar las respuestas de sistemas de BR complementarios.** Tal como se mostró con los experimentos en la colección de prueba del CLEF del año 2006, donde hasta la fecha en esta colección de prueba es donde se ha reportado el mejor resultado en la búsqueda de respuestas en español.

- **En la validación de respuestas extraídas de Wikipedia se hace más evidente la necesidad de recursos sofisticados para el procesamiento del lenguaje natural.** Esto se pudo ver en los experimentos con las colecciones de prueba del AVE de los años 2007 y 2008, donde entre otras cosas la resolución de correferencias es indispensable tanto para la validación así como para la misma búsqueda de respuestas.

Por último en las siguientes secciones se muestra un resumen de las aportaciones de la tesis, las publicaciones obtenidas de la misma, así como el trabajo futuro y las líneas de investigación abiertas.

7.1. Aportaciones del trabajo de tesis

Las aportaciones de esta tesis radican principalmente en el desarrollo de un método de validación de respuestas basado en el enfoque de reconocer la implicación textual. A continuación presentamos un listado de tales aportaciones.

- **La restricción del análisis de similitud al núcleo del pasaje de soporte.** Comúnmente, en la validación de respuestas durante el análisis de la implicación textual se utiliza todo el pasaje de soporte de la respuesta. Aunque este fragmento de texto es pequeño (como máximo 700 bytes), en ocasiones puede contener información irrelevante que se traslapa con la hipótesis y que afecta el resultado final. Para reducir este problema en la sección 4.1.1 se describió un método que consiste en coincidir los sintagmas de la pregunta-respuesta con los sintagmas en el pasaje de soporte, de tal manera que toda aquella información que aparece fuera de la mínima secuencia de palabras que contiene a todos los sintagmas coincidentes en el pasaje de soporte es eliminada. La secuencia de palabras resultante es el llamado núcleo del pasaje de soporte el cual permite mejorar el análisis de similitud en la validación de respuestas.

- **La construcción automática de una hipótesis afirmativa para el estudio de la implicación textual, así como una manera de crear transformaciones de dicha hipótesis para mejorar el análisis de similitud.** Para la construcción de una hipótesis afirmativa que conserve la semántica de una respuesta válida (utilizada en el reconocimiento de la implicación textual), los métodos de validación de respuestas generalmente aplican un conjunto de patrones construidos manualmente. En la sección 4.1.2 se mostró que aplicando un análisis sintáctico a la pregunta es posible obtener de manera automática dicha hipótesis después de identificar y sustituir el menor sintagma que incluye a la partícula interrogativa en la pregunta por la respuesta a ser evaluada. Además, cuando los sintagmas principales de la pregunta son permutados es posible obtener trasformaciones de la hipótesis que facilitan analizar su similitud con T.

- **La introducción de atributos relacionados a las coincidencias que incluyen información de partes de la oración y de la posición de los términos en la pregunta.** El trabajo relacionado muestra que es mejor omitir

del análisis de similitud las palabras funcionales porque sobre-estiman los cálculos. En este trabajo además de confirmar esta suposición, los resultados también mostraron que ciertos términos de contenido dependiendo de su categoría morfológica y de su posición en la pregunta resultan más relevantes que otros para analizar las coincidencias, tal como sucede con los sustantivos que aparecen en el constituyente del actor visible cuyo atributo es el que obtienen la mayor ganancia de información de todos los atributos relacionados al traslape (ver tabla 5.4). La forma en que se calculan dichos atributos se describió en la sección 4.1.3.

- **La incorporación de atributos relacionados a las discrepancias en el análisis de similitud.** Tomando en cuenta que los sistemas de BR actuales suelen producir respuestas erróneas pero con pasajes de soporte relevantes, los métodos de validación de respuestas basados en el análisis de similitud típico generalmente se equivocan en clasificar como válidas a estas respuestas debido a la alta tasa de coincidencias que existe. Para tratar con este inconveniente en la sección 4.1.3 se presentó una forma de complementar la evaluación de las coincidencias durante el análisis de similitud. Dicha forma consiste en detectar el texto discrepante en el núcleo del pasaje de soporte respecto a la hipótesis. Entonces, de las discrepancias se obtienen un conjunto de atributos que reflejan el tipo de información no coincidente entre el constituyente de la respuesta y cada uno de los constituyentes de la pregunta. Sorprendentemente, los resultados de evaluar la ganancia de información mostraron que estos nuevos atributos para evaluar discrepancias en promedio son más discriminativos que los atributos tradicionales de coincidencias en la validación de respuestas (ver tabla 5.4).

- **Un análisis de la pregunta-respuesta para complementar el estudio de la implicación textual en la validación de respuestas.** Como se describe en la sección 4.1.4, este análisis permite obtener atributos que representan características de la pregunta y la respuesta así como de su compatibilidad. Al igual que los atributos propuestos para medir discrepancias, en la tabla 5.4 se muestra que los atributos del análisis pregunta-respuesta en promedio también son más discriminativos que los atributos típicos de coincidencias.

Otros resultados que pueden ser tomados como aportaciones secundarias de la tesis son los siguientes:

- **La relajación de un clasificador implementado con un esquema de aprendizaje.** En los conjuntos de entrenamiento y prueba existe un desbalance entre respuestas válidas y erróneas el cual es un reflejo del estado actual de la búsqueda de respuestas (en el español en promedio más del 70 % de las respuestas de los sistemas de BR son erróneas). Esto provoca que los sistemas de validación de respuestas que emplean algún tipo de aprendizaje supervisado contengan un sesgo hacia las respuestas erróneas afectando su cobertura (*i.e.,* casi no obtienen respuestas clasificadas como válidas). Tomando en cuenta el proceso de relajación utilizado en los enfoques de inferencia lógica, en la sección 5.1.1 se presentó un proceso similar pero aplicado al clasificador generado con un esquema de aprendizaje. Los resultados en la tabla 5.3 muestran como esta propuesta sirvió para mejorar la cobertura, la cual es necesaria para tener un impacto en la búsqueda de respuestas donde lo que interesa es encontrar las respuestas válidas.

- **La selección de una respuesta final tomando como aliadas a las respuestas clasificadas como válidas.** Posterior a la validación de respuestas, para responder a la pregunta es necesario seleccionar una respuesta de aquellas etiquetadas como válidas. Como se describió en la sección 4.2, en el sistema VR-INAOE además de tomar como respuesta final la respuesta válida que obtiene una mayor confianza en la validación, como comúnmente se hace, primero se prefiere aquella que es más redundante en el conjunto de respuestas válidas. De cierta forma con este criterio también se presenta un método de selección de respuestas híbrido que combina la validación de respuestas con el típico enfoque de coro mejorando con esto el proceso de selección de las respuestas.

Finalmente, como una aportación al procesamiento automático del idioma español cabe mencionar que el trabajo presentado en esta tesis es el primero de su tipo en América Latina. Además, en las evaluaciones en español del AVE, el foro internacional donde se evalúa a la validación de respuestas, el método propuesto es el único que ha logrado mejorar el desempeño estimado de los sistemas de BR por combinar todas sus respuestas. También, en el foro de evaluación de búsqueda de respuestas en español del CLEF, el único que incluye la evaluación de esta tarea en nuestro idioma, el sistema de BR del Laboratorio de Tecnologías del Lenguaje del INAOE es el primer sistema que

ha integrado en su arquitectura un módulo de validación de respuestas (implementado con el método aquí descrito), el cual le permitió mejorar sus resultados.

7.2. Publicaciones obtenidas del trabajo de tesis

Las publicaciones obtenidas durante el desarrollo de la tesis fueron las siguientes:

- Alberto Téllez-Valero, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Anselmo Peñas-Padilla. *Learning to Select the Correct Answer in Multi-Stream Question Answering.* Information Processing & Management. Special Issue on Question Answering. *(Por aparecer)*

- Alberto Téllez-Valero, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Anselmo Peñas-Padilla. *Towards Multi-Stream Question Answering using Answer Validation.* International Journal of Informatica. Special Issue on Computational Linguistics, Vol. 34, No. 1, 2010. *(Por aparecer)*

- Alberto Téllez-Valero, Antonio Juárez-Gonzales, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. *Analyzing the Use of Non-Overlap Features for Supervised Answer Validation.* Lecture Notes in Computer Science 5706, Springer, pp. 476-479, 2009. *(Por aparecer)*

- A. Téllez, A. Juárez, M. Montes, and L. Villaseñor. *INAOE at QA@CLEF 2008: Evaluating Answer Validation in Spanish Question Answering.* In Working notes for the Cross Language Evaluation Forum Worshop (CLEF 2008), Aarhus, Denmark, September 2008.

- A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda. *A Supervised Learning Approach to Spanish Answer Validation.* Lecture Notes in Computer Science 5152, Springer, pp. 391-394, 2008.

- A. Téllez, A. Juárez, G. Hernández, C. Denicia, E. Villatoro, M. Montes, and L. Villaseñor.*A Lexical Approach for Spanish Question Answering.* Lecture Notes in Computer Science 5152, Springer, pp. 328-331, 2008.

- A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda. *Improving Question Answering by Combining Multiple Systems via Answer Validation.* Lecture Notes in Computer Science 4919, Springer, pp. 544-554, 2008.
- A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda. *INAOE at AVE 2007: Experiments in Spanish Answer Validation.* In Working notes for the Cross Language Evaluation Forum Worshop (CLEF 2007), Budapest, Hungary, September 2007.
- A. Téllez, A. Juárez, G. Hernández, C. Denicia, E. Villatoro, M. Montes, and L. Villaseñor. *INAOE's Participation at QA@CLEF 2007.* In Working notes for the Cross Language Evaluation Forum Worshop (CLEF 2007), Budapest, Hungary, September 2007.
- A. Juárez-González, A. Téllez-Valero, C. Delicia-Carral, M. Montes-y-Gómez, and L. Villaseñor-Pineda. *Using Machine Learning and Text Mining in Question Answering.* Lecture Notes in Computer Science 4730, Springer, pp. 415-423, 2007.
- A. Téllez, M. Montes, L. Villaseñor. *Una propuesta para la Validación de Respuestas utilizando Implicación Textual.* 3er Taller Nacional de Tecnologías del Lenguaje Humano, Encuentro Internacional de Computación ENC-2006, San Luis Potosí, México, Septiembre 2006.
- A. Juárez-Gonzalez, A. Téllez-Valero, C. Denicia-Carral, M. Montes-y-Gómez, and L. Villaseñor-Pineda. *INAOE at CLEF 2006: Experiments in Spanish Question Answering.* In Working notes for the Cross Language Evaluation Forum Worshop (CLEF 2006), Alicante, Espain, September 2006.

7.3. Trabajo futuro y líneas de investigación abiertas

Claramente se puede ver que mejorando la validación de respuestas se puede mejorar también su impacto en la búsqueda de respuestas. Por lo tanto, el trabajo futuro se enfoca a mejorar la primera etapa —validación de respuestas— del método propuesto. En particular, para mejorar la precisión que es donde se tienen bajos resultados (*i.e.,* aún existen problemas para discriminar correctamente a las respuestas erróneas) en

un futuro se considera incluir más características relacionadas al análisis de discrepancias, en especial información de los términos funcionales hasta ahora descartados del proceso de validación. Esto tomando en cuenta que, en contraste con las coincidencias donde estos términos afectan su análisis, en el estudio de las discrepancias parece ser que éstos son relevantes. Por ejemplo, en la pregunta `¿Quién ganó la primera etapa de la XXI edición del Rally del Chaco?` y la respuesta `Ford Escort` que aparece en el pasaje de soporte `Martín María Massi ganó la primera etapa de la XXI edición del Rally del Chaco con un Ford Escort`; en este caso el pasaje de soporte contiene dos términos funcionales no traslapados entre las palabras de la respuesta y las de la pregunta (los términos `con` y `un`), los cuales parecen ser la clave para etiquetar correctamente como errónea a dicha respuesta.

Respecto a líneas de investigación abiertas, en esta tesis se evaluaron los alcances de una validación de respuestas basada únicamente en un análisis superficial del lenguaje. Sin embargo, en el estudio reportado en la sección 3.3.1 se mostró que existe un porcentaje de respuestas válidas (aproximadamente un 34 % de las estudiadas) donde se requiere un análisis más profundo del lenguaje, como lo es el semántico. Entonces, tomando en cuenta que el método de validación de respuestas propuesto no se aplica a estos casos pero que en el porcentaje restante de respuestas obtiene altos resultados, la línea de investigación que se propone seguir es tratar de validar correctamente las respuestas que requieren un mayor entendimiento del lenguaje. En particular, lo que se sugiere es proponer métodos que ayuden a determinar cuándo dos conceptos escritos con palabras diferentes son iguales y que además no presentan una relación de sinonimia o hiperonimia que pueda ser resuelta con diccionarios u ontologías (*p. ej.,* accidente nuclear y catástrofe, o guerra bolchevique y revolución rusa). Esto tomando en cuenta que en el estudio de la sección 3.3.1 se encontró que estos casos suceden en la validación de respuestas. Los cuales si se detectan y tratan correctamente entonces permitirían su adecuado proceso por parte de los métodos de validación de respuestas actuales enfocados a evaluar la similitud.

Bibliografía

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. & Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, 48–55, Genoa, Italy.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B. & Szpecktor, I. (2006). The second PASCAL recognising textual entailment challenge. In *Proceedings of Second Pascal Challenge Workshop on Recognizing Textual Entailment*, 1–9, Venice, Italy.

Belkin, N.J., Kantor, P., Fox, E.A. & Shaw, J.A. (1995). Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, **31**, 431–448.

Blunsom, P., Kocik, K. & Curran, J.R. (2006). Question classification with log-linear models. In E.N. Efthimiadis, S.T. Dumais, D. Hawking & K. Järvelin, eds., *Special Interest Group on Information Retrieval (SIGIR)*, 615–616, ACM.

Bos, J. & Nissim, M. (2007). Answer translation: An alternative approach to cross-lingual question answering. In Peters *et al.* (2007a), 290–299.

Bosma, W. & Callison-Burch, C. (2007). Paraphrase substitution for recognizing textual entailment. In Peters *et al.* (2007a), 502–509.

Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J. & Moldovan, D.I. (2007). Lcc's poweranswer at qa@clef 2006. In Peters *et al.* (2007a), 310–317.

BIBLIOGRAFÍA

Burger, J.D., Ferro, L., Greiff, W., Henderson, J., Mardis, S., Morgan, A. & Light, M. (2002). MITRE's qanda at TREC-11. In *Text REtrieval Conference (TREC) TREC 2002 Proceedings*.

Buscaldi, D., Soriano, J.M.G., Rosso, P. & Sanchis, E. (2007). -gram vs. keyword-based passage retrieval for question answering. In Peters *et al.* (2007a), 377–384.

Carbonell, J., Harman, D. & Hovy, E. (2000). Vision statement to guide research in question & answering (q&a) and text summarization. Tech. rep., http://www-nlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-v1a.doc.

Cassan, A., Figueira, H., Martins, A.F.T., Mendes, A., Mendes, P., Pinto, C. & Vidal, D. (2007). Priberam's question answering system in a cross-language environment. In Peters *et al.* (2007a), 300–309.

Chu-carroll, J., Czuba, K., Prager, J. & Ittycheriah, A. (2003). In question answering, two heads are better than one. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 24–31.

Clarke, C.L.A., Cormack, G.V., Kemkes, G., Laszlo, M., Lynam, T.R., Terra, E.L. & Tilker, P.L. (2002). Statistical selection of exact answers (multitext experiments for TREC 2002). In *Text REtrieval Conference (TREC) TREC 2002 Proceedings*.

Costa, L. (2007). Question answering beyond clef document collections. In Peters *et al.* (2007a), 405–414.

Cumbreras, M.A.G., Perea-Ortega, J.M., Santiago, F.M. & López, L.A.U. (2007a). Combining lexical information with machine learning for answer validation at qa@clef 2007. In Peters *et al.* (2007b), 381–386.

Cumbreras, M.A.G., Perea-Ortega, J.M., Santiago, F.M. & López, L.A.U. (2007b). Sinai at qa@clef 2007. answer validation exercise. In *Working notes for the CLEF Workshop*.

Dagan, I., Magnini, B. & Glickman, O. (2005). The PASCAL recognising textual entailment challenge. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment*, 1–8, Southampton, UK.

Dalmas, T. & Webber, B.L. (2007). Answer comparison in automated question answering. *J. Applied Logic*, **5**, 104–120.

de Chalendar, G., Dalmas, T., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I. & Vilnat, A. (2002). The question answering system QALC at LIMSI, experiments in using web and wordnet. In *Text REtrieval Conference (TREC) TREC 2002 Proceedings*.

de Pablo-Sánchez, C., González-Ledesma, A., Moreno-Sandoval, A. & Vicente-Díez, M.T. (2007). Miracle experiments in qa@clef 2006 in spanish: Main task, real-time qa and exploratory qa using wikipedia (wiqa). In Peters *et al.* (2007a), 463–472.

Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, **10**, 1895–1924.

Ferrández, Ó., Terol, R.M., Muñoz, R., Martínez-Barco, P. & Palomar, M. (2006). A knowledge-based textual entailment approach applied to the qa answer validation at clef 2006. In *Working notes for the CLEF Workshop*.

Ferrández, Ó., Muñoz, R. & Palomar, M. (2008a). Improving question answering tasks by textual entailment recognition. In E. Kapetanios, V. Sugumaran & M. Spiliopoulou, eds., *NLDB*, vol. 5039 of *Lecture Notes in Computer Science*, 339–340, Springer.

Ferrández, Ó., Muñoz, R. & Palomar, M. (2008b). A lexical-semantic approach to ave. In *Working notes for the CLEF Workshop*.

Ferrández, S., López-Moreno, P., Roger, S., Ferrández, A., Peral, J., Alvarado, X., Noguera, E. & Llopis, F. (2007). Monolingual and cross-lingual qa using aliqan and brili systems for clef 2006. In Peters *et al.* (2007a), 450–453.

BIBLIOGRAFÍA

Forner, P., Peñas, A., Alegria, I., Forascu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R. & Sang, E.T.K. (2008). Overview of the clef 2008 multilingual question answering track. In C.P. et al., ed., *Working Notes for the CLEF 2008 Workshop*, Springer, working Notes. Online Proc.

Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B. & Sutcliffe, R.F.E. (2007a). Overview of the clef 2007 multilingual question answering track. In Peters *et al.* (2007b), 200–236.

Giampiccolo, D., Magnini, B., Dagan, I. & Dolan, B. (2007b). The third PASCAL recognising textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 1–9, Prague, Czech Republic.

Glöckner, I. (2007). Answer validation through robust logical inference. In Peters *et al.* (2007a), 518–521.

Glöckner, I., Hartrumpf, S. & Leveling, J. (2007). Logical validation, answer merging and witness selection - a study in multi-stream question answering. In D. Evans, S. Furui & C. Soulé-Dupuy, eds., *RIAO*, CID.

Gómez-Soriano, J.M., y Gómez, M.M., Arnal, E.S., Villaseñor-Pineda, L. & Rosso, P. (2005). Language independent passage retrieval for question answering. In A.F. Gelbukh, A. de Albornoz & H. Terashima-Marín, eds., *MICAI*, vol. 3789 of *Lecture Notes in Computer Science*, 816–823, Springer.

Harabagiu, S. & Hickl, A. (2006). Methods for using textual entailment in open-domain question answering. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 905–912, Association for Computational Linguistics, Morristown, NJ, USA.

Harabagiu, S. & Moldovan, D. (1998). Knowledge processing on extended wordnet. In M.P. C. Fellbaum, ed., *WordNet: An Electronic Lexical Database and Some of its Applications*, 379–405.

Herrera, J., Rodrigo, A., Peñas, A. & Verdejo, F. (2006). Uned submission to ave 2006. In *Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, Alicante, Spain.

Jijkoun, V. & de Rijke, M. (2004). Answer selection in a multi-stream open domain question answering system. In S. McDonald & J. Tait, eds., *ECIR*, vol. 2997 of *Lecture Notes in Computer Science*, 99–111, Springer.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In C. Nedellec & C. Rouveirol, eds., *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 137–142, Springer, Heidelberg et al.

Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., y Gómez, M.M. & Pineda, L.V. (2006). Inaoe at clef 2006: Experiments in spanish question answering. In *Working notes for the CLEF Workshop*.

Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., y Gómez, M.M. & Pineda, L.V. (2007). Using machine learning and text mining in question answering. In Peters *et al.* (2007a), 415–423.

Kozareva, Z., Vázquez, S. & Montoyo, A. (2006). Adaptation of a machine-learning textual entailment system to a multilingual answer validation exercise. In *Working notes for the CLEF Workshop*.

Kozareva, Z., Vázquez, S. & Montoyo, A. (2007). University of alicante at qa@clef2006: Answer validation exercise. In Peters *et al.* (2007a), 522–525.

Lee & Ho, J. (1997). Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval*, Combination Techniques, 267–276.

Lee, K.S., Oh, J.H., Huang, J., Kim, J.H. & Choi, K.S. (2000). TREC-9 experiments at KAIST: QA, CLIR and batch filtering. In *Text REtrieval Conference (TREC) TREC-9 Proceedings*.

Magnini, B., Negri, M., Prevete, R. & Tanev, H. (2001). Is it the right answer?: exploiting web redundancy for answer validation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 425–432, Association for Computational Linguistics, Morristown, NJ, USA.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. & de Rijke, M. (2003). The multiple language question answering track at clef 2003. In C. Peters, J. Gonzalo, M. Braschler & M. Kluck, eds., *CLEF*, vol. 3237 of *Lecture Notes in Computer Science*, 471–486, Springer.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K.I. & Sutcliffe, R.F.E. (2004). Overview of the clef 2004 multilingual question answering track. In C. Peters, P. Clough, J. Gonzalo, G.J.F. Jones, M. Kluck & B. Magnini, eds., *CLEF*, vol. 3491 of *Lecture Notes in Computer Science*, 371–391, Springer.

Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B. & Sutcliffe, R.F.E. (2007). Overview of the clef 2006 multilingual question answering track. In Peters *et al.* (2007a), 223–256.

Maybury, M.T. (2004). Question answering: An introduction. In M.T. Maybury, ed., *New Directions in Question Answering*, 3–18, AAAI Press.

Moldovan, D. & Novischi, A. (2002). Lexical chains for question answering. In *Proceedings of COLING 2002*.

Moldovan, D. & Rus, V. (2001). Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of ACL 2001*, 394–401.

Moldovan, D.I., Clark, C. & Harabagiu, S.M. (2005). Temporal context representation and reasoning. In L.P. Kaelbling & A. Saffiotti, eds., *IJCAI*, 1099–1104, Professional Book Center.

Moldovan, D.I., Clark, C., Harabagiu, S.M. & Hodges, D. (2007). Cogex: A semantically and contextually enriched logic prover for question answering. *J. Applied Logic*, **5**, 49–69.

Pasca, M. & Harabagiu, S. (2001). High performance question answering. In W.B. Croft, D.J. Harper, D.H. Kraft & J. Zobel, eds., *Proceedings of the 24th Annual International ACM SIGIR (Special Interest Group on Information Retrieval) Conference on Research and Development in Information Retrieval (SIGIR-01)*, 366–374, ACM Press, New York.

Peñas, A., Rodrigo, Á. & Verdejo, F. (2006). Sparte, a test suite for recognising textual entailment in spanish. In A.F. Gelbukh, ed., *CICLing*, vol. 3878 of *Lecture Notes in Computer Science*, 275–286, Springer.

Peñas, A., Rodrigo, Á., Sama, V. & Verdejo, F. (2007). Overview of the answer validation exercise 2006. In Peters *et al.* (2007a), 257–264.

Peñas, A., Rodrigo, Á., Sama, V. & Verdejo, F. (2008). Testing the reasoning for question answering validation. *J. Log. Comput.*, **18**, 459–474.

Pérez-Coutiño, M.A., y Gómez, M.M., López-López, A. & Pineda, L.V. (2006). The role of lexical features in question answering for spanish. In Peters *et al.* (2006), 492–501.

Pérez-Coutiño, M.A., y Gómez, M.M., López-López, A., Pineda, L.V. & Pancardo-Rodríguez, A. (2007). Applying dependency trees and term density for answer selection reinforcement. In Peters *et al.* (2007a), 424–431.

Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B. & de Rijke, M., eds. (2006). *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, vol. 4022 of *Lecture Notes in Computer Science*, Springer.

Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M. & Stempfhuber, M., eds. (2007a). *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, vol. 4730 of *Lecture Notes in Computer Science*, Springer.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V. & Santos, D., eds. (2007b). *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, vol. 5152 of *Lecture Notes in Computer Science*, Springer.

Pizzato, L.A.S. & Molla-Aliod, D. (2005). Extracting exact answers using a meta question answering system. In *Proceedings of the Australasian Language Technology Workshop*, 105–112, Sydney, Australia.

Radev, D.R., Qi, H., Zheng, Z., Blair-Goldensohn, S., Zhang, Z., Fan, W. & Prager, J.M. (2001). Mining the web for answers to natural language questions. In *CIKM*, 143–150, ACM.

Rodrigo, Á., Peñas, A., Herrera, J. & Verdejo, F. (2007a). The effect of entity recognition on answer validation. In Peters *et al.* (2007a), 483–489.

Rodrigo, Á., Peñas, A. & Verdejo, F. (2007b). Uned at answer validation exercise 2007. In Peters *et al.* (2007b), 404–409.

Rodrigo, Á., Peñas, A. & Verdejo, F. (2008). Evaluating answer validation in multi-stream question answering. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*.

Rotaru, M. & Litman, D.J. (2005). Improving question answering for reading comprehension tests by combining multiple systems. In *Proceedings of the American Association for Artificial Intelligence (AAAI) 2005 Workshop on Question Answering in Restricted Domains, Pittsburgh, PA.*.

Roussinov, D., Chau, M., Filatova, E. & Robles-Flores, J.A. (2005). Building on redundancy: Factoid question answering, robust retrieval and the "other". In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2005)*, 15–18.

Tatu, M., Iles, B. & Moldovan, D.I. (2007). Automatic answer validation using cogex. In Peters *et al.* (2007a), 494–501.

Téllez-Valero, A., y Gómez, M.M. & Villaseñor-Pineda, L. (2006). Una propuesta para la validación de respuestas utilizando implicación textual. In *Memorias del 3er. Taller Nacional de Tecnologías del Lenguaje Humano, Encuentro Nacional de Ciencias Computacionales*, SLP, México.

Téllez-Valero, A., Juárez, A., Hernández, G., Denicia, C., Villatoro-Tello, E., y Gómez, M.M. & Pineda, L.V. (2007). A lexical approach for spanish question answering. In Peters *et al.* (2007b), 328–331.

Tomás, D. & González, J.L.V. (2007). Re-ranking passages with lsa in a question answering system. In Peters *et al.* (2007a), 275–279.

Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D. & Sutcliffe, R.F.E. (2006). Overview of the clef 2005 multilingual question answering track. In Peters *et al.* (2006), 307–331.

Vicedo, J.L. (2003). Recuperación de información de alta precision: Los sistemas de búsqueda de respuestas. *Colección de monografías de la Sociedad española para el procesamiento del lenguaje natural (SEPLN)*, **2**, 1–139.

Voorhees, E.M. (2004). Overview of the trec 2004 question answering track. In E.M. Voorhees & L.P. Buckland, eds., *TREC*, vol. Special Publication 500-261, National Institute of Standards and Technology (NIST).

Voorhees, E.M. & Dang, H.T. (2005). Overview of the trec 2005 question answering track. In *Text REtrieval Conference (TREC) TREC 2005 Proceedings*.

Witten, I.H. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Zaenen, A., Karttunen, L. & Crouch, R. (2005). Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 31–36, Association for Computational Linguistics, Ann Arbor, Michigan.

Zhang, D. & Lee, W.S. (2003). Question classification using support vector machines. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR (Special Interest Group on Information Retrieval) conference on Research and development in informaion retrieval*, 26–32, ACM, New York, NY, USA.