Un Modelo Fundamentado en Análisis de Dependencias y *WordNet* para el Reconocimiento de Implicación Textual

Jesús Herrera de la Cruz

Departamento de Lenguajes y Sistemas Informáticos

Escuela Técnica Superior de Ingeniería Informática

Universidad Nacional de Educación a Distancia

24 de mayo de 2005

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática
Universidad Nacional de Educación a Distancia

Memoria de Trabajo del Período de Investigación
del Programa de Doctorado en
Lenguajes y Sistemas Informáticos
curso 2004-2005

Un Modelo Fundamentado en Análisis de Dependencias y *WordNet* para el Reconocimiento de Implicación Textual

Jesús Herrera de la Cruz
Ingeniero en Informática
Universidad Complutense de Madrid

Director: Dr. Anselmo Peñas Padilla
Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática
Universidad Nacional de Educación a Distancia

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción, Hipótesis de Partida y Objetivos

La presente memoria compila el trabajo realizado por el autor, durante el curso 2004-2005, para el Período de Investigación del Programa de Doctorado en Lenguajes y Sistemas Informáticos de la Universidad Nacional de Educación a Distancia. Se trata del planteamiento, desarrollo y prueba de un sistema de reconocimiento de implicación textual (RTE, Recognising Textual Entailment) cuyo funcionamiento sigue un modelo fundamentado en el análisis de dependencias y consultas a *WordNet*.

1.1. Motivación

El reconocimiento de la implicación textual es un tema sobre el que, hasta el primer RTE *Challenge*, que ha sido una propuesta de la Red de Excelencia PASCAL [1] concretada en junio de 2004, no se había realizado un trabajo de investigación unificado ni en abundancia. Sin embargo, proliferan los sistemas de Recuperación de Información y Procesamiento del Lenguaje Natural que, de un modo u otro, necesitan abordar la resolución de inferencias entre textos. Ésto abre una puerta al interés en desarrollar sistemas que de manera genérica aborden esta tarea y puedan actuar como subsistemas de otros más específicos.

Nos encontramos, pues, ante una necesidad de trabajo investigador en un campo de sumo interés con posibilidades de aportar soluciones. El PASCAL RTE *Challenge* es una fuente de motivación, unificación de criterios y foro de investigadores, por lo que participar en él supone la posibilidad de contribuir al avance del RTE en un marco común.

[1] Pattern Analysis, Statistical Modeling and Computational Learning. `http://www.pascal-network.org/`

1.2. Reconocimiento de Implicación Textual

El RTE es la tarea de decidir cuándo la verdad de un texto en lenguaje natural implica la verdad de otro texto (o hipótesis). La resolución de esta tarea puede involucrar un amplio rango de tipos de inferencia como, por ejemplo, las relativas a medidas, fechas, lugares, nombres propios, las que incluyan negación, forma pasiva, nominalización, reordenamiento de constituyentes, correferencias, paráfrasis, etcétera. Sean los dos siguientes textos:

1. El software de *iTunes* ha experimentado importantes ventas en Europa.

2. Importantes ventas de *iTunes* en Europa.

Es evidente que la verdad del segundo se puede inferir de la verdad del primero; se dice, pues, que existe implicación textual entre ambos textos (el primero implica al segundo).

La utilidad del RTE estriba en que puede formar parte de ciertas aplicaciones de Procesamiento del Lenguaje Natural o Recuperación de Información, a modo de motor de resolución de inferencias entre textos. Esta es la propuesta de futuro asociada al primer RTE *Challenge*, es decir, se espera que se llegue a disponer de motores de inferencia genéricos que puedan ser utilizados de manera independiente a las distintas aplicaciones que los requieran (Dagan et al., 2005). Así pues, la noción de implicación textual es independiente del tipo de tarea a la que se aplique. En el primer RTE *Challenge* la organización compiló una serie de tareas para las que podría ser de utilidad la utilización de módulos de RTE. A partir de esa lista se crearon los corpora que fueron utilizados en el *Challenge*, realizando un conjunto de casos apropiados para cada una de las componentes de la lista. Para ello se seleccionaron pares de textos de características tales que fuesen apropiados para uno u otro tipo de las aplicaciones elegidas. A continuación se revisan las aplicaciones consideradas por la organización del *Challenge*, junto con ejemplos ilustrativos – dándose implicación en todos – de pares de texto extraídos de los corpora del PASCAL RTE *Challenge*:

Tipos de aplicaciones tratados y ejemplos

- Recuperación de Información:

  1. *Phish disbands after a final concert in Vermont on Aug. 15*
  2. *Rock band Phish holds final concert in Vermont.*

- Documentos Comparables:

  1. *A two-day auction of property belonging to actress Katharine Hepburn brought in 3.2 million pounds.*
  2. *A two-day auction of property belonging to actress Katharine Hepburn brought in £3.2m.*

- Lectura Comprensiva:

  1. *A senior coalition official in Iraq said the body, which was found by U.S. military police west of Baghdad, appeared to have been thrown from a vehicle.*
  2. *A body has been found by U. S. military police.*

- Pregunta-Respuesta:

  1. *The incident in Mogadishu, the Somali capital, came as U.S. forces began the final phase of their promised March 31 pullout.*
  2. *The capital of Somalia is Mogadishu.*

- Extracción de Información:

  1. *There can be no doubt that the Administration already is weary of Aristide, a populist Roman Catholic priest who in December, 1990, won an overwhelming victory in Haiti's only democratic presidential election*
  2. *Aristide became president of Haiti in 1990.*

- Traducción Automática:

  1. *The American Defense Department reported today, that the bombing runs against Iraq are being especially carried out via cruise missiles launched from American aircraft carriers and B52 bombers.*
  2. *The Pentagon reported today, that the bombing runs against Iraq are being especially carried out via cruise missiles launched from American aircrafts.*

- Adquisición de Paráfrasis:

  1. *The Yellowstone Park Foundation recognizes the following organizations for their generous support in helping to protect the wonders and wildlife of Yellowstone National Park.*
  2. *The Yellowstone Park Foundation would like to acknowledge and thank the following organizations for their generous support.*

En el capítulo 2 (página 15) se describe con detalle el modo en que fueron obtenidos los pares de texto para cada una de las aplicaciones consideradas.

1.3. Hipótesis de Partida y Objetivos

El estudio de la bibliografía existente sobre trabajos previos relacionados con la temática elegida, así como la necesaria temporización de la tarea

a desarrollar, aconsejaron el desarrollo de un sistema relativamente simple. Para ello, el enfoque que otros investigadores habían dado a la resolución del problema, orientado a la representación del texto mediante árboles de análisis de dependencias (Lin and Pantel, 2001), resultaba suficientemente fundamentado y accesible con los medios disponibles en el tiempo establecido.

Hipótesis

La hipótesis de partida era que se podía obtener cierta cantidad de información semántica a partir de la estructura sintáctica de los textos, mediante el análisis de dependencias de los mismos. Además, la extracción de contenido semántico podía enriquecerse mediante la obtención de semántica léxica, motivo por el cual se decidió utilizar *WordNet* como fuente de información semántica a partir de unidades léxicas. La información semántica obtenida se utilizaría para determinar la existencia de implicaciones entre textos, que era el objetivo del PASCAL RTE *Challenge*. Debido a la naturaleza de las técnicas elegidas, sería de esperar un mejor comportamiento del sistema a la hora de determinar la existencia o no de implicación entre textos con alta similitud estructural o léxica.

Objetivos

Los objetivos de este trabajo de investigación fueron los que a continuación se enumeran:

- Estudiar el estado en que se encuentra actualmente la investigación en RTE.
- Proponer una metodología que incluyese elementos innovadores para abordar el problema de RTE.
- Desarrollar un prototipo válido que implementase la metodología propuesta – concretamente, orientado a resolver la tarea de RTE del PASCAL *Challenge* –.
- Estudiar los resultados obtenidos en la participación en el PASCAL RTE *Challenge* para así investigar cuáles han sido los puntos débiles de la metodología propuesta.
- Plantear líneas de trabajo futuro plausibles en función del análisis de resultados.

1.4. Estructura de la memoria

En el capítulo 2 (página 15) se describe pormenorizadamente la tarea de RTE propuesta en el PASCAL *Challenge*.

En el capítulo 3 (página 21) se describen tanto los trabajos relacionados con RTE que se han realizado antes del PASCAL RTE *Challenge* como aquellos que han sido realizados para dicho *Challenge*; dada la escasez de trabajos previos, la más amplia fuente de información al respecto son las

actas del *Workshop* del *Challenge*, constituyendo por sí solas la mayor parte del estado actual del RTE.

En el capítulo 4 (página 35) se define el modelo propuesto en el presente trabajo de investigación, así como el sistema que lo implementa.

En el capítulo 5 (página 47) se describen los experimentos que se realizaron con el sistema propuesto, así como la evaluación de resultados.

En el capítulo 6 (página 51) se establecen las conclusiones a las que sobre este trabajo se ha llegado.

En el capítulo 7 (página 55) se indica el trabajo que en el futuro cercano se debería realizar para mejorar el sistema propuesto.

En el capítulo 8 (página 57) se citan las actividades realizadas para difundir el trabajo realizado entre la comunidad científica.

El capítulo 9 (página 59) contiene los agradecimientos del autor tanto a instituciones como a personas.

Finalmente, las últimas páginas se dedican a la enumeración de las referencias bibliográficas consultadas.

Capítulo 2

Descripción de la Tarea del Primer PASCAL RTE *Challenge*

Si bien no todos los sistemas que abordan el RTE son los participantes en el primer PASCAL RTE *Challenge*, sí que conforman la gran mayoría de los existentes. Debido a ello, se dedica el presente capítulo a una descripción pormenorizada de la tarea propuesta en dicho *Challenge*, ya que será objeto de continuas referencias a lo largo de esta memoria.

2.1. Motivación del PASCAL RTE *Challenge*

La Red de Excelencia PASCAL ha propuesto el RTE *Challenge* como respuesta al advenimiento que, en los últimos años, se ha producido en la investigación sobre aplicaciones de procesamiento de texto que realizan inferencias orientadas a la semántica de significados concretos del texto y sus relaciones. Aunque muchas aplicaciones afrontan problemas semánticos similares, a estos problemas se les suele buscar una solución *ad-hoc* según la aplicación de que se trate. En consecuencia, es difícil comparar bajo un marco común de evaluación los métodos semánticos desarrollados en las diferentes aplicaciones. El PASCAL RTE *Challenge* presenta la implicación textual como una tarea y un marco de evaluación comunes para los investigadores en Procesamiento del Lenguaje Natural, Recuperación de Información y Traducción Automática, cubriéndose así un amplio rango de inferencias semánticas necesarias para aplicaciones prácticas. Esta tarea es, pues, apropiada para la evaluación y comparación de modelos semánticos de una manera genérica. Inclusive, el trabajo sobre implicación textual puede promover el desarrollo de motores semánticos genéricos, que jugarían un papel análogo al que los analizadores sintácticos tienen en mútiples aplicaciones

Tarea y marco de evaluación comunes

actualmente.

El PASCAL RTE *Challenge* tenía como meta proporcionar una primera oportunidad para presentar y comparar diferentes aproximaciones para modelizar la implicación textual. Por ello, se invitó a los participantes a que no lo considerasen un ambiente competitivo sino exploratorio. El evento fue organizado por Ido Dagan, Oren Glickman (ambos de la Universidad Bar Ilan de Israel) y Bernardo Magnini, del ITC-irst, Centro per la Ricerca Scientifica e Tecnologica (Italia) (Dagan et al., 2005).

2.2. Descripción de la tarea

Detección automática de implicación entre textos

La tarea que habían de resolver los sistemas en este *Challenge* era la detección automática de implicación semántica entre parejas de textos en lenguaje natural (monolingüe inglés). Para ello, los organizadores proporcionaron a los participantes sendos corpora de entrenamiento y de prueba, compuestos por pares de textos cortos en lenguaje natural pertenecientes al dominio de las noticias de prensa. Los componentes de cada par de fragmentos textuales se denominaron "texto" e "hipótesis", respectivamente. Los sistemas debían detectar si el significado de la hipótesis se podía inferir del significado del texto; es decir, el sentido de la implicación estaba predeterminado.

2.3. Los corpora

Cada uno de los corpora estaba realizado en formato XML [1] y lo componían pares de texto e hipótesis manualmente anotados. En el caso del corpus de entrenamiento, en cada par se especificaba si existía o no implicación entre el texto y la hipótesis. Además, los pares habían sido elegidos de modo que cubriesen características propias de diferentes aplicaciones de procesamiento de texto. De este modo, cada par llevaba una etiqueta indicadora del tipo de aplicación al que las inferencias necesarias para detectar la implicación le eran más propias; como se puede ver en las figuras 2.2 (página 19) y 2.1 (página 18), en las etiquetas XML el atributo `task` indicaba la clase de aplicación a la que se adscribía el par <texto, hipótesis> mediante un código de dos letras que se explica después; en el corpus de entrenamiento (ver figura 2.1, página 18), además, el atributo `value` indicaba si existía implicación (`TRUE`) o no (`FALSE`). En la figura la figura 2.2 (página 19) se puede ver un ejemplo de cada uno de los tipo de aplicaciones que la organización consideró a la hora de crear los corpora, que fueron:

Tipos de aplicaciones tratados

- Recuperación de Información (código **IR**, Information Retrieval): los anotadores generaron hipótesis que pudiesen corresponder a consultas

[1] XML, eXtensible Markup Language: `http://www.w3.org/XML/`

significativas de Recuperación de Información, tales que expresasen algunas relaciones semánticas concretas (generalmente más largas y específicas que una consulta estándar por palabras clave, representando de este modo una variante orientada a la semántica en Recuperación de Información). Las hipótesis se seleccionaron de entre frases significativas de noticias de prensa, con las que se alimentó un motor de búsqueda. Los textos candidatos se seleccionaron de entre los documentos recuperados por el motor de búsqueda, tomándose tanto textos que implicaban las hipótesis como otros que no.

- Documentos Comparables (código **CD**, Comparable Documents): los anotadores seleccionaron los pares de textos e hipótesis buscando entre artículos de prensa comparables con un argumento común, identificando pares de frases "alineadas" según cierto solapamiento léxico. Podía haber o no implicación semántica.

- Lectura Comprensiva (código **RC**, Reading Comprehension): esta tarea se corresponde con el típico ejercicio de lectura comprensiva en el que a los estudiantes se les pide que emitan un juicio sobre si un determinado aserto se puede inferir de la historia leída. Los anotadores crearon hipótesis con estas características a partir de noticias de prensa.

- Pregunta-Respuesta (código **QA**, Question Answering): utilizando un corpus creado a partir de noticias de prensa especialmente para tareas de Pregunta-Respuesta, los anotadores seleccionaron algunas preguntas y las convirtieron a forma afirmativa, que pasaron a ser las hipótesis. Tras esto, eligieron fragmentos relevantes de texto susceptibles de contener la respuesta correcta a la pregunta para construir los pares <texto, hipótesis>.

- Extracción de Información (código **IE**, Information Extraction): esta tarea está inspirada en la Extracción de Información, pero realizando una adaptación para disponer de pares de textos en lugar de pares formados por un texto y una plantilla estructurada. Dado un conjunto de relaciones interesantes de Extracción de Información, los anotadores identificaron como texto frases candidatas de noticias de prensa en las que la relación dada podía (o no) existir. Como hipótesis crearon formulaciones en lenguaje natural de la relación de Extracción de Información, fácilmente identificables por un sistema actual de Extracción de Información.

- Traducción Automática (código **MT**, Machine Translation): dos traducciones del mismo texto, una automática y otra manual, se comparaban y modificaban con el fin de obtener pares <texto, hipótesis>. Las traducciones automáticas eran corregidas gramaticalmente, para obtener textos en lenguaje correcto.

- Adquisición de Paráfrasis (código **PP**, Paraphrase Acquisition): significados similares se pueden expresar de diferentes maneras, donde no sólo varía el léxico sino también la estructura sintáctica de las expresiones. Los sistemas de adquisición de paráfrasis intentan obtener pares (o conjuntos) de expresiones que parafraseen a las demás. Los anotadores utilizaron pares candidatos de expresiones parafraseadas mediante sistemas automáticos para crear los pares <texto, hipótesis>.

```
...
<pair id=‘‘78’’ value=‘‘FALSE’’ task=‘‘IR’’>
<t>Clinton’s new book is not big seller here.</t>
<h>Clinton’s book is a big seller.</h>
</pair>
...
<pair id=‘‘96’’ value=‘‘TRUE’’ task=‘‘IR’’>
<t>The Massachusetts Supreme Judicial Court has cleared the
way for lesbian and gay couples in the state to marry, ruling
that government attorneys ‘‘failed to identify any
constitutionally adequate reason’’ to deny them the right.
</t>
<h>U.S. Supreme Court in favor of same-sex marriage</h>
</pair>
...
<pair id=‘‘128’’ value=‘‘TRUE’’ task=‘‘IR’’>
<t>Hippos do come into conflict with people quite often.
</t>
<h>Hippopotamus attacks human.</h>
</pair>
...
<pair id=‘‘781’’ value=‘‘TRUE’’ task=‘‘CD’’>
<t>Voting for a new European Parliament was clouded by
concerns over apathy.</t>
<h>Voting for a new European Parliament has been clouded
by apathy .</h>
</pair>
...
```

Figura 2.1: Ejemplo del corpus de entrenamiento.

El corpus de entrenamiento, compuesto por 567 pares <texto, hipótesis> se puso a disposición de los participantes durante el período de desarrollo de sus sistemas, para utilizarlo en el ajuste de los mismos. El corpus de

```
...
<pair id=‘‘276’’ task=‘‘IR’’>
<t>Aristide was educated in the Vatican, and therefore h’s
more fluent in Italian, Greek and Hebrew than English.</t>
<h>Aristide speaks Italian, Greek, Hebrew and English.</h>
</pair>
...
<pair id=‘‘822’’ task=‘‘CD’’>
<t>Satomi Mitarai died of blood loss.</t>
<h>Satomi Mitarai bled to death.</h>
</pair>
...
<pair id=‘‘164’’ task=‘‘RC’’>
<t>AOL has more than 33 million paying customers.</t>
<h>33 million customers pay to use AOL.</h>
</pair>
...
<pair id=‘‘2039’’ task=‘‘QA’’>
<t>The Ploce mayor Josko Damic spoke of the first Croatian
president.</t>
<h>Josko Damic was the first Croatian president.</h>
</pair>
<pair id=‘‘1697’’ task=‘‘IE’’>
<t>In 1999 Ford bought the apartment in Manhattan that he
shares with his new girlfriend, Calista Flockhart.</t>
<h>Calista Flockhart lives in Manhattan.</h>
</pair>
...
<pair id=‘‘1325’’ task=‘‘MT’’>
<t>In turn, the Editor-in-Chief of Al Jumhoria Newspaper was
appointed Ambassador of Iraq to India</t>
<h>Al Jumhoria is the Iraqi Ambassador to India.</h>
</pair>
...
<pair id=‘‘1987’’ task=‘‘PP’’>
<t>The girl was found in Drummondville earlier this
month.</t>
<h>The girl was discovered in Drummondville.</h>
</pair>
...
```

Figura 2.2: Ejemplo del corpus de Prueba.

prueba se publicó una semana antes de la fecha establecida para el envío de resultados; este corpus estaba compuesto por 800 pares de texto e hipótesis. Los sistemas debían emitir un juicio sobre la existencia o no de implicación para todos los pares del corpus de prueba o bien parcialmente para algunas de las tareas. Además, los equipos podían enviar una o dos ejecuciones.

2.4. Participación y evaluación

Al *Challenge* se presentaron 16 equipos, más uno que realizó una ejecución manual como análisis del corpus para determinar si los pares podían ser correctamente clasificados por un sistema "ideal" que se basase sólo en consideraciones sintácticas y, opcionalmente, usando un tesauro (ver la sección 3.2.1, página 26). Los sistemas automáticos cubrieron un amplio rango de técnicas: solapamiento de palabras, relaciones léxicas estadísticas, consultas a *WordNet*, coincidencia sintáctica, conocimiento del mundo e inferencia lógica.

Los resultados (Dagan, 2005) de los sistemas fueron evaluados, oficialmente, según su precisión (fracción de juicios acertados entre los juicios emitidos), que coincide con la cobertura cuando se responde a todos los ejemplos del corpus, como medida principal y según Confident-Weighed Score (CWS) (Voorhees, 1999) como medida secundaria, ya que los sistemas podían dar una autovaloración de la confianza en cada uno de sus juicios. Extraoficialmente, también se sometieron los resultados a la medida F (con igual peso para la precisión y la cobertura) sobre ejemplos positivos, es decir, sobre ejemplos que sí contenían implicación y la respuesta del sistema había sido correcta.

Capítulo 3

Reconocimiento de Implicación Textual: Estado Actual

El reconocimiento de implicación textual ha sido empezado a tener en cuenta internacionalmente como tarea de interés y desarrollo independiente en el *Recognising Textual Entailment Challenge* organizado por primera vez entre junio de 2004 y abril de 2005 por la Red de Excelencia PASCAL [1]. Aún así, de un modo u otro el reconocimiento de la implicación entre textos en lenguaje natural ha sido objeto de estudio en los últimos años, bien como parte de sistemas más complejos bien como aplicación independiente.

Los tres primeros ejemplos que se revisan, en las secciones 3.1.1 (página 22), 3.1.2 (página 23) y 3.1.3 (página 24), son lo principales trabajos que sobre implicación textual (búsqueda de inferencias entre textos) se han realizado previamente al PASCAL RTE *Challenge*. Éstos están caracterizados por tener orígenes y motivaciones diversos, no tratando el problema de manera focalizada.

Gracias al primer (y hasta ahora único) PASCAL RTE *Challenge*, actualmente existe una serie de grupos en todo el mundo que han desarrollado proyectos de investigación en el ámbito del reconocimiento de implicación textual. Además, es prometedor el interés que muestran estos grupos en continuar sus investigaciones al respecto, así como el que otros grupos puedan mostrar en el futuro cercano para incorporarse a los actualmente involucrados. En este último año, la práctica totalidad de la investigación mundial sobre reconocimiento de implicación textual la ha aglutinado el PASCAL RTE *Challenge*.

El presente trabajo de investigación se inscribe dentro de los participantes en el PASCAL RTE *Challenge* con los que, conjuntamente, define el panorama mundial actual en lo tocante a reconocimiento de implicación textual. La característica fundamental del trabajo de todos los grupos participantes es la orientación

[1] Pattern Analysis, Statistical Modeling and Computational Learning. `http://www.pascal-network.org/`

del mismo a la consecución de objetivos comunes que dirigen el desarrollo de la investigación.

3.1. Trabajos en RTE previos al PASCAL *Challenge*

Los principales trabajos que versan sobre RTE realizados antes de que se propusiese el primer PASCAL *Challenge* se describen a continuación.

3.1.1. Extracción de reglas de equivalencia textual

Dekang Lin y Patrick Pantel (Lin and Pantel, 2001) propusieron un método no supervisado de extracción de reglas de inferencia a partir de texto, del tipo *"X es autor de Y" = "X excribió Y"*, *"X resolvió Y" = "X encontró una solución a Y"* o *"X causó Y" = "Y lo provocó X"*. Su algoritmo está basado en una versión extendida de la Hipótesis de Distribución de Harris (Harris, 1985), que expone que las palabras que ocurren en los mismos contextos tienden a ser similares; en lugar de utilizar esta hipótesis aplicada a palabras, la aplicaron a caminos en árboles de dependencias extraídos de un corpus. El tipo de relaciones anteriormente expuestas (*"X es autor de Y" = "X excribió Y"*, etcétera) habían sido calificadas hasta el trabajo de Lin y Pantel como paráfrasis o variantes ; ellos utilizan la terminología "regla de inferencia", debido a que también incluían en esta categoría relaciones que no eran exactamente paráfrasis y que resultaban útiles para los sistemas de recuperación de información; por ejemplo, *"X causó Y" = "Y es por culpa de X"* es una regla de inferencia, porque existe una relación semántica a pesar de que ambas componentes no significan exactamente lo mismo.

Hipótesis de Distribución de Harris

Paráfrasis, variantes y reglas de inferencia

El trabajo de Lin y Pantel estaba dirigido a simplificar el trabajo de creación de bases de conocimiento de ese tipo de reglas , que habitualmente se realiza de manera manual y es muy laboriosa. También realizan una clasificación de los campos en los que había sido de utilidad el reconocimiento de variantes y las paráfrasis, identificando los siguientes:

Bases de conocimiento de reglas de inferencia

Campos de aplicación de variantes y paráfrasis.

- Generación de lenguaje: donde se han focalizado los esfuerzos básicamente en las transformaciones de texto basadas en reglas, para satisfacer restricciones externas como la longitud y la legibilidad.
- Resumen multidocumento: en el que el parafraseo es importante para evitar redundancias en los resúmenes.
- Recuperación de información: donde es común generar variantes de los términos de la consulta para la expansión de la misma.

- Minería de textos: en la que se se intenta encontrar reglas de asociación entre términos.

Lin y Pantel proponen un algoritmo no supervisado al que llaman DIRT (Discovery of Inference Rules from Text), que es una generalización de otros algoritmos de búsqueda de palabras similares. Los algoritmos de búsqueda de palabras similares se fundamentan en la Hipótesis de Distribución de Harris, que establece que las palabras que acontecen en contextos iguales tienen significados similares. La generalización de Lin y Pantel consiste en hipotetizar que si dos caminos pertenecientes a árboles de dependencias de dos textos tienden a unir conjuntos iguales de palabras, los significados de ambos textos son similares. DIRT realiza una búsqueda de caminos en función de ciertas restricciones (como no considerar relaciones de dependencia que conecten palabras sin contenido) y mide la similitud entre dos caminos utilizando una medida propuesta previamente por Lin (Lin, 1998), basada en la información mutua [2], utilizada también por otros autores como Alshawi y Carter (Alshawi and Carter, 1994).

Hipótesis de Harris extendida

Lin y Pantel concluyen que su trabajo con la creación de DIRT es el primero realizado para encontrar automáticamente reglas de inferencia a partir de corpora textuales, señalan la utilidad de la detección de inferencias y exponen tipos de reglas que en trabajos futuros sería interesante poder detectar.

3.1.2. Búsqueda de implicación textual sobre el modelo de bolsa de palabras

Christof Monz y Maarten de Rijke (Monz and de Rijke, 2001), defendiendo un tratamiento superficial de los textos en lugar de utilizar representaciones complejas a la hora de determinar implicaciones entre textos, construyeron un modelo basado en bolsas de palabras para tratar este problema. Su trabajo se justificaba en la gran utilidad del reconocimiento de implicaciones entre textos para aplicaciones fundamentadas en semántica computacional.

Para probar la eficacia de su modelo frente a la tradicional pobreza de resultados de los modelos basados en representaciones semánticas complejas, diseñaron un experimento fuertemente influenciado por la tarea de eliminar pasajes redundantes en resumen automático. Su metodología era la siguiente:

1. Una vez seleccionados conjuntos de documentos relacionados entre sí, calculaban para cada conjunto el peso de todas las palabas que lo

[2] La información mutua relaciona la probabilidad $P_1(a)P_2(b)P_3(c)$ de la tripleta (a, b, c) asumiendo independencia entre los tres campos, donde $P_p(x)$ es la probabilidad de observar x en la posición p, con la probabilidad A estimada a partir de las observaciones de tripletas derivadas de los análisis mejor valorados del corpus de entrenamiento; concretamente, se utiliza $ln[A/(P_1(a)P_2(b)P_3(c))]$

componían basándose en el *idf* [3] (cociente del total de pasajes entre el número de pasajes en que aparecía cada palabra); de este modo, los pesos obtenidos dependían del conjunto de documentos para el que se calculaba y, por ende, de los contextos, ya que cada conjunto de documentos respondía a una temática común.

Textos como bolsas de palabras con peso asociado

2. Dados dos documentos , establecían la similitud entre dos pasajes – uno de cada documento – mediante la razón de la suma de pesos de términos comunes entre los pasajes con respecto a la suma de pesos de los términos de uno de los pasajes. Este último pasaje sería el implicado en caso de obtenerse suficiente valor de similitud, ya que la medida no es simétrica. De este modo, los pasajes eran tratados como bolsas de palabras con un peso asociado.

3. Para determinar cuándo se podía considerar que existía implicación entre un par de pasajes, establecieron empíricamente un valor umbral para la medida de similitud, superado el cual se suponía que uno de los pasajes implicaba al otro. Aunque el coste de la medida de similitud entre todos los pares de pasajes de una colección de documentos es exponencial con respecto al número de pasajes, el cálculo es asequible temporalmente con un computador de sobremesa.

Conjuntos temáticos de documentos

Para evaluar su método, construyeron varios conjuntos temáticos de documentos y les aplicaron el algoritmo, habiendo previamente determinado manualmente la existencia de implicación entre cada par de pasajes. Los juicios conseguidos con este algoritmo tenían una precisión media cercana al 30 % y una cobertura media del 48 %, con significativas variaciones según el conjunto temático de documentos.

Semántica léxica con *WordNet*

Para mejorar su método propusieron enriquecerlo con el uso de semántica léxica , mediante sinónimos e hipónimos/hiperónimos de *WordNet*; también sugirieron realizar un estudio con diferentes tipos de representaciones y así entender de qué manera las inferencias y las representaciones estaban conectadas.

3.1.3. Implicación textual aplicada a sistemas de pregunta-respuesta mediante árboles de dependencias

Hristo Tanev, Milen Kouylekov y Bernardo Magnini (Tanev et al., 2004) desarrollaron un sistema de búsqueda de implicación textual para utilizarlo como subsistema de otro de pregunta-respuesta que presentaron a la edición de 2004 del TREC; el objetivo a largo plazo era disponer de un sistema de pregunta-respuesta cuyo núcleo fuese un motor de implicaciones, ca-

Motor de implicaciones

[3] *idf (inverse document frequency)* es el logaritmo de la inversa de la frecuencia de un documento para un determinado término, o sea, $idf_i = log\frac{N}{n_i}$, donde N es el total de documentos a considerar por el sistema y n_i el número de documentos en los que aparece

paz de realizar inferencias a partir de una gran base de datos de reglas de implicación. En esta línea de investigación estaban en colaboración con la Universidad israelí Bar Ilan, lo que dio como resultado conjunto un módulo para la extracción automática de la *web* de reglas de implicación (Szpektor et al., 2004).

Extracción automática de reglas de implicación

El primer experimento que realizaron se fundamentaba en la capacidad de obtener inferencias textuales a partir de una representación sintáctica del texto – árboles de dependencias, concretamente –. Tras obtener los árboles de dependencias de una pareja de textos, comprobaban el grado de solapamiento entre ambos árboles para determinar la existencia o no de implicación entre los textos. De esta manera, se evaluaba el solapamiento entre el árbol de dependencias de la pregunta en versión afirmativa y los árboles de dependencias de cada uno de los textos recuperados por un motor de búsqueda como candidatos a contener la respuesta. Este solapamiento se cuantificaba según la coincidencia de palabras, evaluando para ello que tuviesen el mismo lema y categoría gramatical o perteneciesen a la misma clase en un tesauro. En caso de alto solapamiento se entendía que se había encontrado una respuesta. Las frases que supuestamente contenían la respuesta a la pregunta se clasificaban según una métrica propuesta por Hristo Tanev y Milen Kouylekov (Tanev et al., 2004); dicha métrica consistía en asignar pesos a los subárboles de dependencias que mostraban solapamiento; el peso de un subárbol se calculaba mediante el sumatorio de productos de los pesos de cada par de palabras y el peso de la relación que las une; todos los pesos se calculaban en base al *idf*.

Árboles de dependencias

Solapamiento léxico ⇒ respuesta

El otro experimento llevado a cabo consistía en una evaluación de la contribución de las reglas de implicación textual para encontrar respuestas a un conjunto limitado de preguntas; estas reglas se habían obtenido de manera semiautomática basándose en la aproximación dada por Deepak Ravichandran y Eduard Hovy (Ravichandran and Hovy, 2002), en forma de plantillas que se intentaban encajar en los documentos recuperados.

3.2. Trabajos participantes en el primer PASCAL RTE *Challenge*

Al primer primer PASCAL RTE *Challenge* se presentaron 17 equipos de todo el mundo, proponiendo modelos para abordar el problema de RTE según la tarea especificada por la organización.

Dieciseis equipos realizaron sistemas que se ajustaban a la tarea propuesta en el *Challenge*, pero un equipo presentó un análisis manual del corpus de entrenamiento, como seguidamente se indica.

---

el término k_i (Baeza-Yates and Ribeiro-Neto, 1999).

3.2.1. Análisis manual del corpus de entrenamiento del PASCAL RTE *Challenge*

El equipo de Microsoft Research formado por Lucy Vanderwende, Deborah Coughlin y Bill Dolan (Vanderwende et al., 2005) no participaron en el PASCAL RTE *Challenge* como establecían las normas, si no que realizaron un análisis manual (no desarrollaron ningún sistema) de los pares <texto, hipótesis> del corpus de prueba. El análisis consistió en comprobar si, teniendo en cuenta únicamente consideraciones sintácticas – enriquecidas de manera opcional con el uso de un tesauro léxico –, un mecanismo de decisión "ideal" podía emitir un juicio correcto sobre la implicación de la hipótesis por el texto. Dos evaluadores humanos con altos conocimientos en lingüística revisaron el corpus, dando uno de los siguientes juicios para cada par <texto, hipótesis>:

Juicios para evaluación manual

- Cierto (el texto implica la hipótesis), según indicios sintácticos.
- Falso (el texto no implica la hipótesis), según indicios sintácticos.
- No hay indicios sintácticos que permitan emitir un juicio de implicación.
- No se puede decidir.

Tras su análisis del corpus identificaron una serie de capacidades que debían poseer los sistemas de RTE fundamentados en análisis sintáctico; de este modo, sería deseable que tales sistemas detectasen:

Funcionalidad deseable

- **Alternancias sintácticas**. La alternancia más frecuente era una construcción en el texto resultante de realizar una aposición en la hipótesis. En la figura 3.1 (página 26) se muestra un ejemplo (Vanderwende et al., 2005):

```
<pair id=‘‘760’’ value=‘‘TRUE’’ task=‘‘CD’’>
<t>The Alameda Central, west of the Zocalo, was created in
1952.</t>
<h>The Alameda Central is west of the Zocalo.</h>
</pair>
```

Figura 3.1: Par 760 del corpus de entrenamiento.

- **Ejemplos negativos**. Se trata de detectar cuándo no puede haber implicación por motivos sintácticos. Encontraron dos clases principales de casos en los que, según consideraciones sintácticas, se podía determinar que no había implicación semántica:

1. Cuando existe algún tipo de incorrelación sintáctica entre el texto y la hipótesis – aunque mostraban su inseguridad sobre que los sistemas automáticos pudiesen abordar este caso –. En el ejemplo de la figura 3.2 (página 27) (Vanderwende et al., 2005) se puede observar alineación entre los verbos y los sujetos del texto y la hipótesis, pero no entre los objetos:

```
<pair id=‘‘103’’ value=‘‘FALSE’’ task=‘‘IR’’>
<t>The White House ignores Zinni’s opposition to the Iraq War
</t>
<h>White House ignores the threat of attack.</h>
</pair>
```

Figura 3.2: Par 103 del corpus de entrenamiento.

2. Cuano no existe una estructura sintáctica compartida para la mayor parte de los elementos que componen el par. Por ejemplo, véase la figura 3.3 (página 27) (Vanderwende et al., 2005):

```
<pair id=‘‘2179’’ value=‘‘FALSE’’ task=‘‘RC’’>
<t>An ambulance crew responding to an anonymous call found a
3-week-old baby girl in a rundown house Monday, two days after
she was snatched from her mother at a Melbourne shopping mall.
</t>
<h>A baby girl bought an ambulance at a Melbourne
shopping mall.</h>
</pair>
```

Figura 3.3: Par 2179 del corpus de entrenamiento.

3.3. ¿Cómo Son Actualmente los Sistemas de RTE?

El estado de la cuestión actual refleja la existencia de dos vertientes por las que discurren los desarrollos: la aproximación lingüística y la aproximación empírica. A continuación se revisan ambas aproximaciones, realizando un recorrido por las diversas técnicas, recursos y herramientas – extraídos del conjunto de los utilizados por los sistemas que actualmente abordan el RTE – que podrían conformar un sistema de RTE que aglutinase a todas las propias de la aproximación a la que pertenecen.

3.3.1. La Aproximación Lingüística

Un sistema de RTE que basase su funcionamiento en el Procesamiento del Lenguaje Natural podría verse como una conjunto de módulos en los que, en cada una de ellos, se desarrollase una técnica determinada para tratar la información en su nivel correspondiente. De este modo, se puede establecer la siguiente serie de técnicas aplicables: análisis morfológico-léxico, reconocimiento de multipalabras, reconocimiento de entidades, análisis sintáctico y análisis semántico.

Cada sistema queda caracterizado por el uso concreto que realiza de las técnicas disponibles.

Preprocesamiento.

Aparte de la necesaria identificación de *tokens*, existen sistemas que realizan un preprocesamiento de los textos antes de aplicarles el análisis morfológico-léxico, que se sitúa en el primer nivel dentro del procesamiento lingüístico. Este preprocesamiento se corresponde, en los casos en los que se da, con la segmentación. Esta ha sido utilizada bien como preparación para el análisis morfológico bien para la creación de estructuras de representación de los textos.

El MITRE [4] (Bayer et al., 2005), previamente al análisis morfológico efectuado por su sistema, aplica a los textos y las hipótesis un proceso de segmentación de frases.

El sistema de la Universidad de Concordia (Andreevskaia et al., 2005) no llega a realizar un análisis morfológico, utilizando la segmentación de sintagmas nominales como apoyo para crear estructuras de predicados con argumentos. Se crea una estructura por cada texto e hipótesis y se establece la similitud entre las estructuras de cada par de fragmentos textuales para determinar si existe implicación entre ambos.

Análisis morfológico.

Dentro de éste se diferencian: la extracción de lemas o *stems*, el etiquetado de categorías gramaticales, el uso de analizadores morfológicos y la extracción de relaciones impuestas por la morfología derivacional.

El análisis morfológico se ha utilizado como un primer procesamiento de los textos para obtener información con la que alimentar etapas posteriores que permitan evaluar la implicación entre textos.

La **extracción de lemas** es una técnica que se ha utilizado con cierta profusión y que, en algunos casos, supone gran parte del procesamiento total que realiza el sistema de RTE. La lematización ha sido utilizada con tres

[4] The MITRE Corporation, Estados Unidos.

fines diferentes: para evaluar su coincidencia en medidas de similitud tratando los textos como bolsas de palabras, como atributos de representaciones en forma de grafo de los textos y para ajustar parámetros en algoritmos de evaluación de similitud. Así, por ejemplo, en el sistema de las uiversidades de Edimburgh y Leeds (Bos and Markert, 2005) la lematización es el procesamiento del lenguaje más elaborado que se realiza y, tras él, sólo se aplica una medida de solapamiento entre lemas de la hipótesis y el texto para determinar la existencia de implicación entre ambos. El sistema de la Universidad de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005) utiliza los lemas como parte de los atributos asociados a los nodos de árboles conceptuales con los que representaban tanto los textos como las hipótesis. La Universidad "Tor Vergata" de Roma, asociada con la Universidad de Milano-Bicocca (Pazienza et al., 2005), ha desarrollado un sistema en el que se aplica el análisis morfológico para la extracción de lemas que, junto con los *tokens* y otros elementos, se utilizan para ajustar – mediante un algoritmo de aprendizaje SVM – los parámetros de la medida global de similitud entre los dos grafos con los que representan el texto y la hipótesis.

La **extracción de *stems*** ha sido una técnica utilizada básicamente para alimentar otros módulos del sistema. El uso de *stems* en el caso monolingüe inglés está justificado por el buen comportamiento que han demostrado, dada la simplicidad de la morfología del inglés; en el futuro, cuando se desarrollen sistemas de RTE para otras lenguas será necesario evaluar la posibilidad de trabajar sólo con *stems* o, por el contrario, se habrá de hacer uso de los lemas. Como ejemplo de uso de *stems*, el sistema de las universidades "Tor Vergata" y Milano-Bicocca (Pazienza et al., 2005) evalúa su coincidencia en una medida de subsunción de nodos. Con esta medida, junto con otra de subsunción de vértices, determinan la subsución global entre los grafos que representan el texto y la hipótesis; la medida de subsunción global sirve para establecer la implicación entre el texto y la hipótesis.

El **etiquetado de categorías gramaticales** ha sido utilizado de dos maneras: el sistema del MITRE (Bayer et al., 2005) y el de las Universidad Ca' Foscari y el ITC-irst [5] (Delmonte et al., 2005) lo incluyen como uno de los módulos de análisis lingüístico en cascada; pero la Universidad de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005) utiliza las categorías gramaticales como parte de los atributos asociados a los nodos de árboles conceptuales con los que representan tanto los textos como las hipótesis.

El **uso de analizadores morfológicos** como tales sólo se ha dado en el sistema del MITRE, que aplica un analizador morfológico (Minnon et al., 2001) cuya acción se suma a la del etiquetado de categorías gramaticales, para alimentar con sus resultados a las etapas siguientes (un analizador gramatical, un analizador de dependencias y un generador de proposiciones lógicas).

[5] ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Italia

La **extracción de relaciones impuestas por la morfología derivacional** es otra técnica poco utilizada, que encuentra un ejemplo en el sistema de la Language Computer Corporation (Fowler et al., 2005), que extrae relaciones entre palabras a partir de la morfología derivacional contenida en *WordNet*.

Reconocimiento de multipalabras.

No es una técnica muy extendida, dándose sólo dos casos: el sistema de la UNED [6] (Herrera et al., 2005) la utiliza para detectar implicación entre unidades léxicas; para ello busca en el texto y la hipótesis – mediante un solapamiento borroso aplicando la distancia de Levenshtein – multipalabras de *WordNet*. El otro caso es el de uno de los módulos en cascada del sistema de la Universidad Ca' Foscari y el ITC-irst.

Reconocimiento de entidades.

Es otra técnica todavía poco utilizada, con tan sólo dos ejemplos: la Universidad de Stanford (Raina et al., 2005) y la de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005). El sistema de Stanford detecta entidades nombradas mediante resolución de correferencias, con el fin de encontrar dependencias entre nodos de los grafos con los que representa los textos. En el caso de la Universidad de Illinois, las entidades nombradas se utilizan como atributos de los nodos de los grafos que representan los textos.

Análisis sintáctico.

El **análisis de dependencias** es una de las técnicas más utilizadas, situación propiciada probablemente en el caso del inglés por la disponibilidad pública de analizadores de dependencias de gran eficiencia temporal y alta cobertura, como el desarrollado por Dekang Lin (Lin and Pantel, 2001). Generalmente se obtiene el árbol asociado como representación del texto analizado, aunque también como auxiliar para llegar a una representación en forma lógica. Ejemplos de estos dos tipos de usos son: el sistema de la UNED (Herrera et al., 2005), que evalúa la existencia de implicación entre texto e hipótesis mediante el solapamiento entre los árboles de dependencias de ambos fragmentos de texto. El otro caso lo ejemplifica el MITRE (Bayer et al., 2005), con su implementación en cascada de sistemas clásicos de análisis lingüístico, que incluye una etapa de análisis de dependencias; antes del analizador de dependencias hay un analizador gramatical y, después, un generador de predicados lógicos.

El **análisis de constituyentes**, por contra, es una técnica poco común. La Universidad "Tor Vergata", asociada a la de Milano-Bicocca, (Pazienza et

[6] Universidad Nacional de Educación a Distancia, España.

al., 2005) utiliza los constituyentes para extender grafos de dependencias. La Universidad Ca' Foscari, junto al ITC-irst (Delmonte et al., 2005) realizan análisis de constituyentes como parte de un análisis sintáctico híbrido.

Análisis semántico.

La etiquetación de **roles semánticos** la usaron las universidades de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005), la de Stanford (Raina et al., 2005) y la Ca' Foscari en asociación con el ITC-irst (Delmonte et al., 2005). En todos los casos las etiquetas se aplicaban a los nodos de los grafos con los que representaban los fragmentos de texto.

El estado conjunto de la aproximación lingüística.

La mayor parte de los sistemas fundamentados en análisis lingüístico hacen un tratamiento superficial; por ejemplo, el de la Universidad de Amsterdam (Jijkoun and de Rijke, 2005). Pero unos pocos realizan un análisis profundo, como el del grupo formado por la Universidad Ca' Foscari y el ITC-irst (Delmonte et al., 2005), llegando hasta el etiquetado de roles semánticos.

A grandes rasgos, se pueden identificar unas tendencias básicas en el desarrollo de este tipo de sistemas:

- Los que tratan los textos como bolsas de palabras y la extracción de lemas es el análisis lingüístico más profundo que realizan.
- Los que se fundamentan en una representación sintáctica de los textos, que utilizan algunos procesamientos morfológico-léxicos de manera accesoria para incrementar la capacidad del sistema.
- Los que realizan un tratamiento lingüístico profundo, lo que implica hacer un análisis clásico por etapas, recorriendo un amplio rango de niveles de análisis: morfológico-léxico, sintáctico y semántico.

Esta variedad de implementaciones permite realizar estudios sobre la viabilidad de cada tipo, sus puntos fuertes y sus carencias. Un análisis profundo permitiría identificar cuál es el camino a seguir para mejorar cada metodología propuesta para resolver el RTE.

3.3.2. La Aproximación Empírica

Los sistemas de RTE que han optado por esta aproximación son minoritarios y utilizan como base técnicas estadísticas. Actualmente, los resultados obtenidos son muy similares a los que realizan Procesamiento del Lenguaje Natural e, incluso, mejores (Dagan et al., 2005). El MITRE y la Universidad Bar Ilan proponen dos aproximaciones diferentes dentro de este campo.

El MITRE ha desarrollado un **sistema inspirado en los modelos de traducción automática estadística** (Bayer et al., 2005), siguiendo el siguiente proceso:

1. Entrenamiento del sistema de traducción automática mediante el *Gigaword newswire corpus* (Graff, 2003), buscando implicaciones entre titulares y subtítulos.

2. Estimación manual de la fiabilidad del entrenamiento anterior.

3. Refinamiento del corpus obtenido anteriormente mediante el entrenamiento de el clasificador de documentos SVMlight (Joachims, 2002).

4. Inducción de modelos de alineamiento en el subconjunto seleccionado del *Gigaword newswire corpus*, mediante las herramientas GIZA++ (Och and Ney, 2003).

5. Utilización de un clasificador de los vecinos más próximos a distancia k con cada uno de los pares <texto, hipótesis> del corpus de prueba, para elegir el valor dominante de verdad de entre los vecinos más próximos a distancia 5 en el corpus de desarrollo.

La Universidad Bar Ilan propone un **sistema en el que se define un marco probabilístico** para modelizar la noción de implicación textual (Glickman et al., 2005); así mismo, recurren a una representación de bolsa de palabras para describir un modelo de implicación léxica a partir de estadísticas de coocurrencia en la *web*. Se dice que un texto implicaba probabilísticamente a una hipótesis si el texto incrementaba la probabilidad de que el valor de verdad asignado a la hipótesis sea *cierto*. Equivalentemente, un texto implica probabilísticamente a una hipótesis si la información mutua punto a punto es mayor que 1. Para tratar la implicación léxica, establecen un modelo probabilístico según el cual se espera que cada palabra de la hipótesis sea implicada por alguna palabra del texto; ésto se puede ver, alternativamente, como la inducción de una alineación entre términos de la hipótesis y el texto, de manera similar a como se realiza en traducción automática estadística (Brown et al., 1993). Así pues, la implicación probabilística entre texto e hipótesis la calculan en función de la implicación léxica referida. Las probabilidades de implicación léxica las estiman de manera empírica mediante un proceso no supervisado fundamentado en coocurrencias en la *web*.

3.3.3. Técnicas Auxiliares

Además de las técnicas básicas anteriormente reseñadas, la mayor parte de los sistemas implementan una o varias técnicas que completan su funcionalidad. Las que utilizan los sistemas actuales de RTE se tratan a continuación.

Aprendizaje automático.

Algunos sistemas han hecho uso de este tipo de algoritmos como, por ejemplo, el de las universidades "Tor Vergata" y Milano-Bicocca (Pazienza et al., 2005), que aplicaban un SVM para evaluar los parámetros de una medida de evaluación.

Uso de tesauros, grandes corpora y *WordNet*.

Una parte significativa de los sistemas obtiene conocimiento léxico y morfológico de tesauros y *WordNet*. Las consultas a *WordNet* han sido realizadas buscando bien la obtención de relaciones entre unidades léxicas a partir de relaciones de *WordNet* – como es el caso del sistema la UNED, que busca relaciones de sinonimia, hiperonimia e implicación de *WordNet* – bien la obtención de relaciones a partir de cadenas léxicas, como el sistema de la Universidad de Concordia (Andreevskaia et al., 2005). Los tesauros han sido utilizados para extraer conocimiento de algún campo concreto, como el conocimiento geográfico que extraen las universidades de Edimburgh y Leeds (Bos and Markert, 2005) a partir del "CIA factbook". Grandes corpora como la *web* o el *Gigaword newswire corpus* han sido utilizados para extraer propiedades léxicas (Bayer et al., 2005) o estadísticas de coocurrencia (Glickman et al., 2005).

Paráfrasis.

El uso de paráfrasis se centra en la obtención de reglas de reescritura con las que poder mejorar el comportamiento al intentar determinar si dos frases son equivalentes. Tal es el caso del sistema de la Universidad de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005).

Detección de expresiones numéricas y temporales.

Una técnica muy deseable y poco extendida es la detección de ciertos tipos de expresiones. La Universidad de Stanford (Raina et al., 2005) realiza un tratamiento de expresiones numéricas, capaz de determinar inferencias del tipo *"2113 es más que 2000"*. La Universidad Ca' Foscari y el ITC-irst (Delmonte et al., 2005) efectúan una detección de expresiones temporales.

Demostradores automáticos.

Los sistemas que, tras realizar un análisis lingüístico, representan en forma lógica los textos entre los que se tiene que resolver la posible inferencia, utilizan demostradores automáticos para llevarlo a cabo, como es el caso del de la Universidad de Macquaire (Akhmatova, 2005).

3.4. Generalidades

A pesar de la variedad de técnicas implementadas para el desarrollo actual de sistemas de RTE, existe una preponderancia de algunas de ellas. Por ejemplo, la determinación de la implicación a partir del solapamiento de árboles de dependencias, que representan los textos candidatos, está muy extendida. Por contra, son escasos los ejemplos de tratamientos estadísticos o de sistemas que realicen un análisis lingüístico profundo.

Los resultados obtenidos en el PASCAL RTE *Challenge* no son indicativos sobre la idoneidad de las técnicas empleadas, ya que todos los participantes – en un marco común – obtuvieron valores de las medidas de evaluación muy similares. De este modo, el estado actual no permite decidir si es más efectivo el tratamiento lingüístico superficial o profundo, o bien un tratamiento exclusivamente estadístico. La experiencia del *Challenge* sugiere una revisión profunda del funcionamiento de los sistemas con respecto a los corpora utilizados, para detectar qué tipos de inferencias no se han llevado a cabo y así proponer nuevas vías para mejorar.

El estudio de la posible redefinición de la tarea propuesta en el primer PASCAL RTE *Challenge*, incluidas las clasificaciones a que se han sometido los textos que conforman los corpora, es una tarea interesante para el futuro cercano.

Capítulo 4

Propuesta de un Modelo Fundamentado en Análisis de Dependencias y *WordNet* para RTE

En este capítulo se concreta la propuesta que en el presente trabajo de investigación se ha desarrollado para comprobar la validez de la hipótesis de partida. Esta propuesta se apoya básicamente en dos metodologías: la búsqueda de implicación léxica, obtenida a partir de consultas a *WordNet*, y el solapamiento entre árboles de dependencias. La aportación novedosa del modelo que aquí se presenta se resume en los siguientes puntos:

- La combinación de dos técnicas existentes: la obtención de semántica léxica utilizando consultas a *WordNet* y la representación mediante árboles de dependencias[1].

- La metodología de evaluación del solapamiento de árboles de dependencias mediante la similitud de ramas léxicamente solapantes, conceptos que han sido definidos *ex-profeso*.

El modelo propuesto, que a continuación se describe, se refleja en el desarrollo de un sistema cuyo funcionamiento ha sido evaluado mediante la participación en el primer PASCAL RTE *Challenge*.

[1]Si bien en la actualidad esta combinación de técnicas no resulta novedosa, sí lo era en el momento de acometer el trabajo; paralelamente, otros grupos participantes en el PASCAL RTE *Challenge* consideraron también adecuado utilizar la misma combinación; estas cuestiones se hicieron de dominio público con la publicación de las actas del *Workshop*.

4.1. Descripción del sistema

El sistema desarrollado es una primera propuesta hacia la resolución del RTE. La presente aproximación se basa en técnicas superficiales de análisis léxico y sintáctico, explorando las posibilidades del solapamiento entre árboles de dependencias de un texto y una hipótesis. Se trata de un enfoque generalista, que no aborda el tratamiento específico de las diferentes tareas consideradas en el PASCAL RTE *Challenge* (Documentos Comparables, Pregunta Respuesta, etcétera) ni de las diferentes clases de inferencia (fechas, medidas, nominalizaciones, etcétera). Se pretende así probar la validez y el alcance de estas técnicas según el tipo de tarea específica al que esté orientada la detección de implicación. Los componentes del sistema, cuya representación gráfica se puede ver en la figura 4.1 (página 36) son los siguientes:

Minipar

1. Un analizador de dependencias, basado en el sistema *Minipar* de Dekang Lin (Lin, 1998), que normaliza las palabras, lleva a cabo el análisis de dependencias y crea en memoria las estructuras adecuadas para representar el árbol de dependencias.

2. Un módulo de implicación léxica, que toma la información generada por el analizador y devuelve los nodos de la hipótesis que son léxicamente implicados por nodos del texto.

3. Un módulo de evaluación de solapamiento, que busca ramas en el árbol de dependencias de la hipótesis conformadas por nodos léxicamente implicados por nodos del texto.

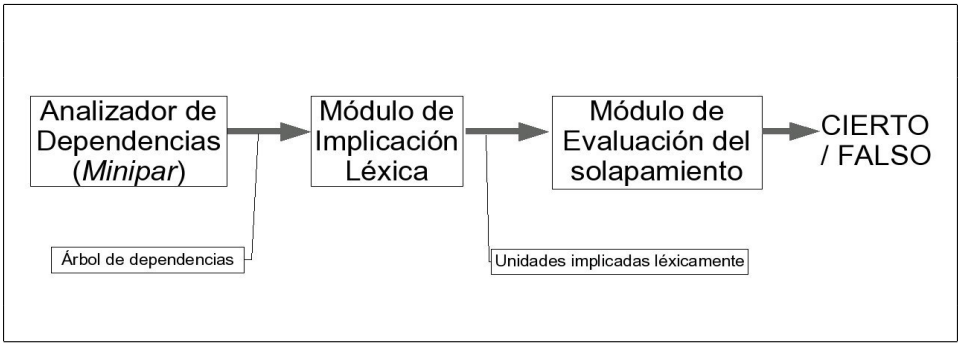

Figura 4.1: Arquitectura del sistema propuesto.

4.2. Recursos Externos

El sistema propuesto hace uso de dos herramientas: el analizador de dependencias *Minipar* (Lin, 1998) y la base de datos léxica *WordNet* (Miller

et al., 2004). A continuación se describen ambas.

4.2.1. *Minipar*

Minipar es un analizador de dependencias automático desarrollado por Dekang Lin (Lin, 1998) y caracterizado por obtener unas altas cobertura y precisión.

Evaluado frente al corpus SUSANNE (que es un subconjunto del Corpus *Brown* del Inglés Americano) (Sampson, 1995), dio unos resultados del 79 % de cobertura y el 89 % de precisión; la cobertura es el porcentaje de relaciones de dependencia encontradas por el analizador del conjunto de relaciones de dependencia anotadas manualmente en el corpus; la precisión es el porcentaje de relaciones de dependencia encontradas por el analizador que también se encuentran entre las relaciones de dependencias anotadas manualmente en el corpus.

Además, *Minipar* se ejecuta muy eficientemente por lo que resulta muy útil como subsistema.

La entrada de *Minipar* es un fichero de texto con el discurso en lenguaje natural (inglés) que se pretende analizar y su salida es un fichero de texto en el que, mediante tuplas, se indican las relaciones de dependencia entre las palabras del discurso analizado. Estas tuplas contienen la siguiente información:

- la palabra considerada,
- su categoría léxica,
- el núcleo de la palabra considerada (la palabra del discurso de la que depende) y
- el tipo de relación de dependencia (por ejemplo: sujeto, adjunto, complemento, especificador, etcétera).

El sistema descrito en la presente memoria normaliza los fragmentos de texto (textos e hipótesis) proporcionados en el corpus para alimentar al analizador *Minipar*. Cuando éste ha actuado, toma los ficheros de texto que contienen las tuplas que representan el análisis y construye con esa información árboles de dependencias en memoria.

4.2.2. *WordNet*

WordNet es una base de datos léxica del idioma inglés, organizada semánticamente, cuyo diseño está inspirado en las actuales teorías psicolingüísticas de la memoria léxica humana.

WordNet ha sido desarrollada, bajo la dirección de George A. Miller[2], en el Laboratorio de Ciencia Cognitiva[3], de la Universidad de Princeton, EE.UU.

Las palabras están organizadas en una jerarquía de unos 12 niveles, que permite la herencia (basada en las ideas de las redes semánticas).

Considera cuatro categorías gramaticales: verbo, sustantivo, adjetivo y adverbio.

La primera versión (Miller et al., 2004) contenía 95.600 palabras o multipalabras, organizadas en 70.100 significados diferentes o conjuntos de sinónimos, llamados "synsets". Actualmente contiene 152.059 palabras o multipalabras (114.648 sustantivos, 11.306 verbos, 21.436 adjetivos y 4.669 adverbios), organizadas en 115.424 *synsets*.

Las palabras representadas están relacionadas entre sí por distitntas relaciones semánticas, que son:

- sinonimia, antonimia
- "es un": hiponimia, hiperonimia
- "parte de": meronimia, holonimia

WordNet se puede consultar a través de la *web* o bien se puede obtener una implementación a la que se pueden realizar consultas en *Prolog*, que fue la utilizada en el desarrollo del sistema aquí propuesto.

4.3. Análisis de dependencias

Gramáticas de dependencias

La estructura lingüística puede ser definida en términos de dependencias entre palabras. Para ello se utilizan las gramáticas de dependencias ; en una gramática de dependencias una palabra es el núcleo de una frase y el resto de las palabras de la frase son o bien dependientes del núcleo o bien dependen de otra palabra de la frase que, a su vez, depende del núcleo (Manning and Schütze, 1999). Estas dependencias son del tipo: "el verbo principal de la frase es el núcleo", "un sustantivo del sujeto depende del verbo principal del predicado", "un adjetivo depende del sustantivo al que afecta", etcétera; estas reglas que establecen las dependencias son las reglas de producción de la gramática. Las dependencias en una frase se pueden mostrar gráficamente mediante aristas dirigidas (estructura en forma de árbol), como en el ejemplo mostrado en la figura 4.2 (página 39).

Árbol de dependencias

El análisis de dependencias tiene como objetivo obtener un árbol de análisis de una frase según una gramática de dependencias, llamado árbol de dependencias.

[2] `http://wordnet.princeton.edu/ geo/`

[3] Cognitive Science Laboratory, `http://www.cogsci.princeton.edu/`

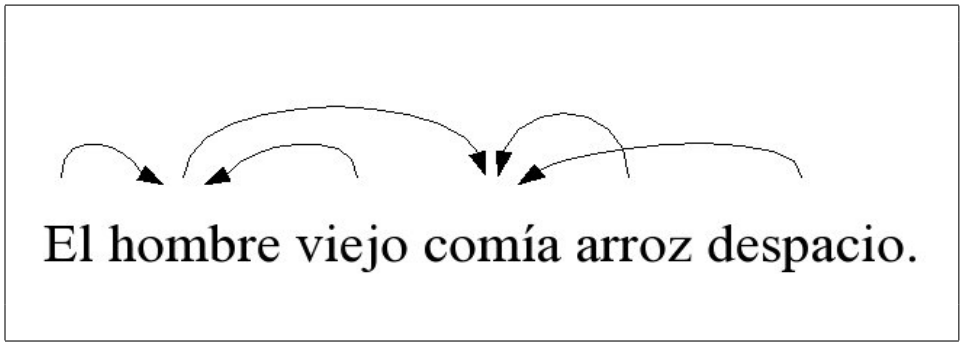

Figura 4.2: Dependencias en una frase (Manning and Schütze, 1999).

En el sistema propuesto, el análisis de dependencias se delega al sistema *Minipar* de Dekang Lin. Previamente, el sistema desarrollado para el presente trabajo extrae los textos y las hipótesis del corpus de entrada, normaliza las palabras y alimenta al analizador de dependencias para que éste genere un árbol de dependencias para cada texto e hipótesis.

El sistema propuesto recoge la salida del analizador *Minipar* y crea en memoria las estructuras adecuadas para que las siguientes etapas actúen.

4.4. Implicación léxica

Tras el análisis de dependencias, se ejecuta un módulo de implicación léxica sobre los nodos del texto y la hipótesis. La salida de este módulo es una lista de pares $<T, H>$, donde T es un nodo del árbol de dependencias del texto cuya unidad léxica implica a la unidad léxica del nodo H del árbol de dependencias de la hipótesis. Esta implicación a nivel léxico se determina teniendo en cuenta relaciones de *WordNet* (secciones 4.4.1, página 39, y 4.4.2, página 39), detección de multipalabras de *WordNet* (sección 4.4.3, página 40) y la negación (sección 4.4.4, página 40), según se indica a continuación:

4.4.1. Sinonimia y Similitud

La unidad léxica T implica la unidad léxica H si ambas son sinónimas según *WordNet* o si existe una relación de similitud entre ellas. En el corpus de entrenamiento del PASCAL Challenge se pueden entrontrar ejemplos de este tipo, como: *discover* y *reveal*, *obtain* y *receive*, *lift* y *rise*, *allow* y *grant*, etcétera.

4.4.2. Hiponimia e Implicación de *WordNet*

La hiponimia y la implicación son relaciones que gozan de la propiedad transitiva, entre *synsets* de *WordNet*. El predicado lógico que denota la

implicación entre dos *synsets* se ha implementado como la búsqueda de un camino desde el *synset* S_T al synset S_H, en el que las relaciones de hiponimia e implicación de *WordNet* entre *synsets* intermedios se consideran en la dirección desde S_T hasta S_H, únicamente en dirección ascendente en la jerarquía de *WordNet*. Algunos ejemplos, obtenidos tras el procesamiento del corpus de entrenamiento del PASCAL Challenge, son:

- De hiponimia: *glucose* implica *sugar*.
- De implicación de *WordNet*: *crude* implica *oil*, *death* implica *kill*.

4.4.3. Multipalabras

El reconocimiento de multipalabras ha de ser posterior a la lematización y al análisis sintáctico (de dependencias). Para evitar errores es necesario realizar un pre y post procesamiento; por ejemplo, el reconocimiento de la multipalabra *came_down* requiere la obtención previa de lema *come*, porque la multipalabra presente en *WordNet* es *come_down*. El sistema utiliza las multipalabras contenidas en *WordNet*, por lo que se pueden reconocer algunas entidades como, por ejemplo, *Hamas* al ser confrontada con la multipalabra *Islamic_Resistance_Movement*.

Reconocimiento de entidades

La variación léxica en las multipalabras no es debida sólo a la lematización; a veces existen caracteres que cambian como, por ejemplo, un punto en un acrónimo o un nombre propio con diferentes formulaciones. Por este motivo se implementó un reconocimiento difuso de la coincidencia entre candidatos a multipalabras, mediante la distancia de edición de Levenshtein (Levensthein, 1966); si dos cadenas difieren en menos del 10 %, entonces se considera que hay coincidencia. Por ejemplo, en el par 345 del corpus de entrenamiento, que se muestra en la figura 4.3 (página 41), la multipalabra *Japanise_capital* de la hipótesis se tradujo como la multipalabra de *WordNet* *Japanese_capital*, permitiendo la implicación entre dicha multipalabra y la palabra *Tokyo* del texto. Otros ejemplos de implicación léxica resueltos por el sistema tras el reconocimiento de multipalabras en el corpus de entrenamiento son, en cuanto a sinonimia, *blood_glucose* y *blood_sugar*, *Hamas* e *Islamic_Resistance_Movement*, *Armed_Islamic_Group* y *GIA* y, en cuanto a hiponimia, *war_crime* implica *crime*, *melanoma* implica *skin_cancer*.

Distancia de Levenshtein

4.4.4. Negación y antonimia

La negación se detecta tras la búsqueda de hojas en el árbol de dependencias de las que parte una relación de negación con su nodo padre. Esta relación de negación se propaga a todos los antecesores de la hoja hasta el núcleo correspondiente. Por ejemplo, en las figuras 4.4 (página 41) y 4.5 (página 42) se muestra un extracto de los árboles de dependencias para los ejemplos de entrenamiento 74 y 78, respectivamente.

Propagación de la negación

```
<pair id=‘‘345’’ value=‘‘TRUE’’ task=‘‘RC’’>
<t>Ahern, who was travelling to Tokyo for an EU-Japan summit
yesterday, will consult with other EU leaders by telephone
later this week in an effort to find an agreed candidate.</t>
<h>A summit between Europe and Japan is taking place in the
Japanise capital.</h>
</pair>
```

Figura 4.3: Par 345 del corpus de entrenamiento.

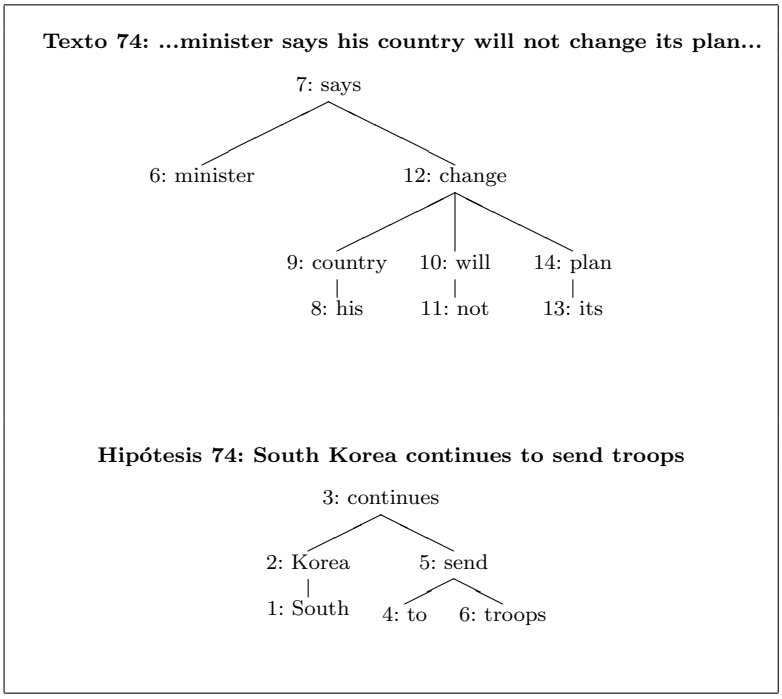

Figura 4.4: Árboles de dependencias del par 74 del corpus de entrenamiento.

La negación en el nodo 11 del texto 74 se propaga al nodo 10 (*neg(will)*) y al nodo 12 (*neg(change)*). La negación en el nodo 6 del texto 78 se propaga al nodo 5 (*neg(be)*).

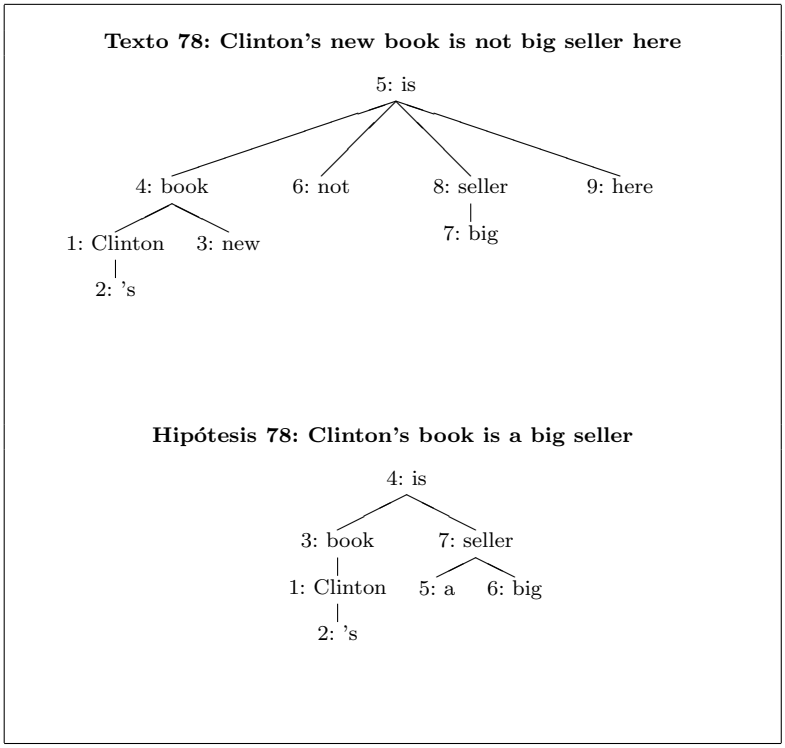

Figura 4.5: Árboles de dependencias del par 78 del corpus de entrenamiento.

La implicación no es posible entre una unidad léxica y su negación. Por ejemplo, antes de tener en cuenta la negación, el nodo 5 del texto 78 (*be*) implica al nodo 4 de la hipótesis 78 (*be*); sin embargo, tras la propagación de la negación anteriormente enunciada no es posible la implicación.

Antonimia

La implicación entre nodos afectados por negación se implementó considerando la relación de antonimia de *WordNet*, aplicándoles posteriormente el procesamiento indicado en las secciones 4.4.1 (página 39), 4.4.2 (página 39) y 4.4.3(página 40). Por ejemplo, como el nodo 12 del texto 74 está negado (*neg(change)*), se consideran los antónimos de *change* en las relaciones de implicación entre el texto y la hipótesis; se tiene que *stay* es un antónimo de *change*; como, a su vez, *stay* es sinónimo de *continue*, entonces se tiene que *neg(change)* implica *continue* en la hipótesis.

4.5. Solapamiento entre árboles de dependencias

Los árboles de dependencias son una representación estructurada de los textos y las hipótesis. El solapamiento entre árboles de dependencias puede

dar una idea sobre la similitud semántica de dos fragmentos de texto, debido a que cierta información semántica está implícitamente contenida en los árboles de dependencias. La técnica utilizada para evaluar el solapamiento entre árboles de dependencias está inspirada en la propuesta de Lin (Lin and Pantel, 2001), pero simplificada lo más posible. De este modo, se utilizó un algoritmo muy sencillo de detección de solapamiento, que se apoya en la siguiente definición:

Definición 1 (Rama coincidente de la hipótesis) *Es la rama del árbol de dependencias de la hipótesis tal que todos sus nodos son implicados léxicamente por nodos de una rama del árbol de dependencias del texto correspondiente.*

Ejemplo 1 (Ramas coincidentes de la hipótesis) *En la figura 4.6 (página 43) se muestran de color rojo, amarillo y verde las ramas coincidentes del árbol de dependencias de una hipótesis abstracta y de los mismos colores las ramas del árbol de dependencias de un texto abastracto que contienen nodos que implican a los de las ramas coincidentes correspondientes de la hipótesis. Nótese que no todos los nodos de una rama del árbol de dependencias del texto han de implicar a algún nodo de la hipótesis, mientras que una rama de la hipótesis no se considera coincidente si alguno de sus nodos no está involucrado mediante implicación léxica con otro de la rama correspondiente del árbol de dependencias del texto.*

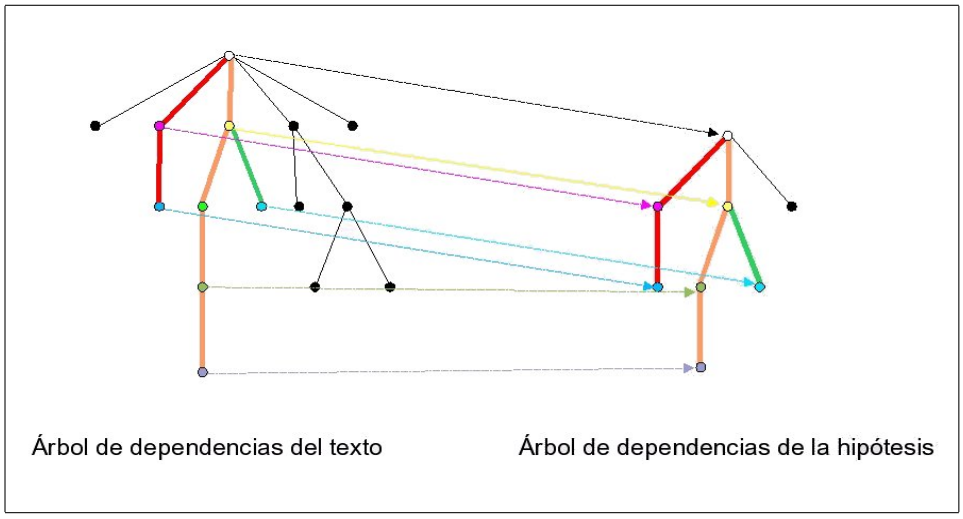

Figura 4.6: Ejemplo de ramas coincidentes de la hipótesis.

Así pues, el algoritmo de detección de solapamiento entre el árbol de la hipótesis y el del texto es:

Algoritmo 1 (Detección de solapamiento entre árboles) *Buscar todas las ramas coincidentes de la hipótesis y contabilizar los nodos pertenecientes a dichas ramas coincidentes.*

Al aplicar este algoritmo se dispone del número de nodos del árbol de dependencias de la hipótesis que pertenecen a ramas coincidentes, lo que permite realizar los cálculos necesarios para decidir si existe o no implicación entre el texto correspondiente y la hipótesis, como se indica seguidamente.

4.6. Decisión sobre la existencia de implicación

Como el sistema ha de decidir automáticamente cuándo existe o no implicación entre un par $<T,\ H>$, se ha de establecer un mecanismo que tome dicha decisión. La existencia o no de una relación de implicación desde un texto hacia su correspondiente hipótesis se determina mediante la similitud, definida como sigue:

Definición 2 (Similitud) *Es la proporción de nodos del árbol de dependencias de la hipótesis pertenecientes a ramas coincidentes.*

Ejemplo 2 (Cálculo de la similitud) *En la figura 4.7 (página 44) se observan los árboles de dependencias de un texto y una hipótesis abstractos. El número de nodos de la hipótesis es $N_H = 8$. El número de nodos pertenecientes a ramas coincidentes de la hipótesis es $N_H^{coincidentes} = 7$. Por lo tanto, la similitud en este ejemplo es:*

$$similitud = \frac{N_H^{coincidentes}}{N_H} = \frac{7}{8}$$

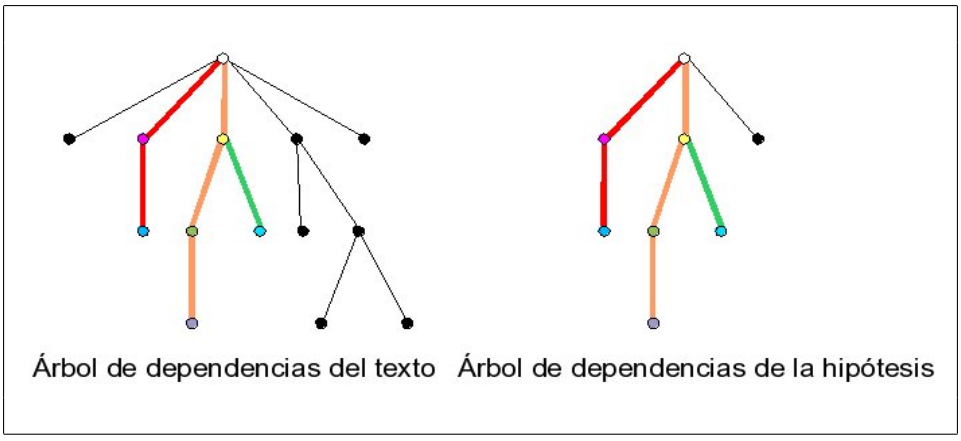

Figura 4.7: Ejemplo de similitud entre árboles de dependencias.

Como el trabajo se fundamenta en que un alto solapamiento entre árboles indica la existencia de una relación semántica, se determinó que el sistema había de emitir un juicio de existencia de implicación siempre que la similitud entre los árboles del texto y la hipótesis superase un cierto umbral, y un juicio de inexistencia de implicación cuando dicho umbral no fuese alcanzado.

A partir de los resultados obtenidos del corpus de entrenamiento, se estableció empíricamente un umbral para el valor de similitud. El sistema presentaba su mejor valor de cobertura/precisión cuando dicho umbral se fijaba en el 50 %. Así pues, cuando el árbol de dependencias de una hipótesis mostraba un porcentaje de solapamiento de nodos mayor o igual al 50 %, se consideraba que había implicación entre el texto y la hipótesis; cuando dicho porcentaje era menor, se admitía que no existía implicación entre ambos.

Umbral de decisión

Capítulo 5

Experimentos Realizados y Evaluación del Sistema Propuesto

El sistema que implementa el modelo propuesto en el capítulo 4 (página 35) fue desarrollado *ex-profeso* para participar en el PASCAL RTE *Challenge*. Debido a ello, su evaluación final ha sido la realizada por la organización del *Challenge*. Las evaluaciones durante el tiempo de desarrollo, necesarias para los sucesivos refinamientos del sistema, se realizaron también siguiendo el modelo de la evaluación oficial del *Challenge*, utilizando el corpus de entrenamiento que dispusieron al efecto.

5.1. Experimentos

Durante el tiempo de desarrollo del sistema propuesto se llevaron a cabo diferentes experimentos para obtener información que realimentase el proceso de implantación de sucesivas mejoras. Para ello se desarrollaron varios sistemas de referencia, cuyos resultados se compararon con el corpus de entrenamiento.

5.1.1. Sistemas de referencia

Se generaron dos "baselines" para analizar el comportamiento del sistema propuesto contra el corpus de entrenamiento. Como la implicación léxica se determina antes del solapamiento entre árboles de dependencias, se desarrollaron dos sistemas más simples para obtener dichas "baselines":

- El sistema "baseline" I calculaba la *ratio* de palabras de la hipótesis que aparecían en el texto.

- El sistema "baseline" II calculaba la *ratio* de lemas de la hipótesis implicados por algún lema del texto.

En todos los casos, el valor umbral para la clasificación de muestras se fijó en el 50 %, ya que con este valor se obtenía la mejor precisión/cobertura.

5.1.2. Ejecuciones enviadas al *Challenge*

Como se podían enviar hasta dos ejecuciones por participante, se decidió preparar un tercer "baseline" para comparar con el sistema propuesto frente al corpus de prueba. Para este sistema "baseline" III, se eliminaron las consultas a *WordNet*, teniéndose en cuenta sólo la coincidencia entre lemas del texto y la hipótesis. Una de las ejecuciones enviadas fue generada por este último sistema "baseline".

El sistema se refinó para su ejecución contra el corpus de prueba. La última versión se enriqueció haciendo que buscabase también solapamiento de relaciones de sujeto y objeto entre la hipótesis y el texto tras encontrarlas a lo largo de las ramas coincidentes de la hipótesis.

5.2. Resultados del sistema propuesto

A continuación se tabulan los resultados de los experimentos contra el corpus de entrenamiento y los obtenidos en la participación en el PASCAL RTE *Challenge*.

5.2.1. Resultados contra el corpus de entrenamiento

El sistema propuesto y los "baselines" mostraron resultados similares. Se calculó la precisión[1] para cada tipo de tarea presente en el corpus, variando entre el 46,67 % y el 55,56 %, excepto para la tarea de Documentos Comparables (CD, Comparable Documents), para la que los resultados fueron del 76,29 %, 71,13 % y el 80,41 % para el sistema "baseline" I, el sistema "baseline" II y el sistema propuesto, respectivamente. Los resultados globales fueron del 54,95 %, 55,48 % y el 56,36 % para el sistema "baseline" I, el sistema "baseline" II y el sistema propuesto, respectivamente. En el cuadro 5.1 (página 49) se tabulan los valores de precisión de la ejecución de los tres sistemas, detallados para cada tipo de tarea [2].

[1] Fracción de los juicios acertados entre los juicios emitidos. Ver sección 2.4, página 20

[2] CD = Documentos Comparables, IE = Extracción de Información, IR = Recuperación de Información, MT = Traducción Automática, PP = Adquisición de Paráfrasis, QA = Pregunta-Respuesta, RC = Lectura Comprensiva. Véase la sección 2.3, página 16

Tipo	Baseline I	Baseline II	Sistema Propuesto
CD	76,29 %	71,13 %	80,41 %
IE	47,14 %	50,00 %	47,14 %
IR	51,43 %	52,86 %	51,43 %
MT	53,70 %	53,70 %	55,56 %
PP	52,44 %	52,44 %	54,88 %
QA	51,11 %	54,44 %	46,67 %
RC	48,54 %	50,49 %	53,40 %
Global	54,95 %	55,48 %	56,36 %

Cuadro 5.1: Valores de precisión frente al corpus de entrenamiento

5.2.2. Resultados oficiales en el *Challenge*

La precisión se calculó para cada tipo de tarea presente en el corpus, variando entre el 42,55 % y el 55,83 %, excepto para CD, para la que los resultados fueron del 79,33 % y del 78,67 % para el sistema "baseline" III y el sistema propuesto, respectivamente. Los resultados globales fueron del 56,75 % y del 54,88 % para el sistema "baseline" III y el sistema propuesto, respectivamente.

El resultado de ambos sistemas es similar al de los ejecutados contra el corpus de entrenamiento; sin embargo, la consideración de relaciones de sujeto y objeto provocó un ligero descenso de la precisión.

En el cuadro 5.2 (página 49) se tabulan los valores de precisión de la ejecución de los dos sistemas, detallados para cada tipo de tarea.

Tipo	Baseline III	Sistema Propuesto
CD	79,33 %	78,67 %
IE	52,50 %	55,00 %
IR	51,77 %	51,77 %
MT	55,83 %	54,17 %
PP	48,94 %	42,55 %
QA	48,46 %	45,38 %
RC	47,86 %	47,14 %
Global	55,75 %	54,88 %

Cuadro 5.2: Valores de precisión frente al corpus de prueba.

5.3. Resultados generales del *Challenge*

A continuación, se muestran los resultados de todas las ejecuciones enviadas al PASCAL RTE *Challenge*, incluida la manual de Microsoft Research, ordenadas por número de juicios correctos.

Equipo	Pares tratados	Juicios correctos	Precisión	Cobertura	F	CWS
MITRE	800	469	0,5863	0,5863	0,5863	0,6174
Bar Ilan	800	469	0,5863	0,5863	0,5863	0,5718
Dublin	800	452	0,5650	0,5650	0,5650	0,6000
Dublin	800	450	0,5625	0,5625	0,5625	0,5917
Edimburgh-Leeds	800	450	0,5625	0,5625	0,5625	0,5931
Stanford	800	450	0,5625	0,5625	0,5625	0,6207
UIUC	800	449	0,5613	0,5613	0,5613	0,5691
ITC-irst	800	447	0,5588	0,5588	0,5588	0,6068
ITC-irst	800	447	0,5588	0,5588	0,5588	0,5849
UNED	800	446	0,5575	0,5575	0,5575	0,5750
Edimburgh-Leeds	800	444	0,5550	0,5550	0,5550	0,5864
Stanford	800	442	0,5525	0,5525	0,5525	0,6860
Amsterdam	800	442	0,5525	0,5525	0,5525	0,5593
LCC	800	441	0,5513	0,5513	0,5513	0,5602
UNED	800	439	0,5488	0,5488	0,5488	0,5710
Amsterdam	800	429	0,5363	0,5363	0,5363	0,5526
Bar Ilan	800	424	0,5300	0,5300	0,5300	0,5349
Roma-Milano	797	418	0,5245	0,5225	0,5235	0,5574
Concordia	800	415	0,5188	0,5188	0,5188	0,5154
Macquaire	800	415	0,5188	0,5188	0,5188	0,5067
Concordia	800	413	0,5163	0,5163	0,5163	0,5200
Roma-Milano	797	413	0,5182	0,5163	0,5172	0,5591
Hong-Kong	800	410	0,5125	0,5125	0,5125	0,5497
Hong-Kong	800	404	0,5050	0,5050	0,5050	0,5361
UAM	800	396	0,4950	0,4950	0,4950	0,5168
Microsoft	391	374	0,9565	0,4675	0,6280	0,9513
MITRE	587	303	0,5162	0,3788	0,4369	0,5033
Ca' Foscari	498	250	0,5020	0,3125	0,3852	0,6637
Microsoft	238	227	0,9538	0,2838	0,4374	0,9462
UAM	150	105	0,7000	0,1313	0,2211	0,7824

Cuadro 5.3: Resultados generales del PASCAL RTE *Challenge*

Se observa que precisión (medida oficial del *Challenge*) y cobertura coinciden cuando se aborda la resolución de todos los pares. El CWS (Voorhees, 1999) fue la medida secundaria del *Challenge*, ya que los juicios podían ir acompañados de una autovaloración en la confianza de la respuesta dada; aunque no todos los sistemas emitieron dicha autovaloración, la organización había previsto para lo que no lo hicieran que diesen, a cambio, un valor constante igual a 1, de este modo pudieron calcular el CWS para todos los participantes, aunque en algunos casos resulte un valor ficticio. La medida F sobre ejemplos positivos – es decir, sobre ejemplos que sí contenían implicación y la respuesta del sistema había sido correcta – se calculó de manera extraoficial (con igual peso para la precisión y la cobertura) y da cuenta del equilibrio de los sistemas en precisión y cobertura.

Capítulo 6

Conclusiones

A partir de los objetivos inicialmente planteados y tras el desarrollo del trabajo de investigación, se concluye la consecución de los mismos: En el capítulo 3, página 21, se expone el estudio realizado sobre el estado en que se encuentra la investigación en RTE actualmente; se detectaron las técnicas implicadas en el total de sistemas existentes para, mediante una labor de síntesis, describir las diferentes aplicaciones que se dieron a cada una de ellas. En el capítulo 4, página 35, se detalla la propuesta metodológica para abordar el problema de RTE que fundamenta el presente trabajo de investigación. En el capítulo 7, página 55, se plantean líneas de trabajo futuro en función del análisis de resultados, que – además de unos breves apuntes en la sección 3.4, página34 – se muestra a lo largo del presente capítulo.

A pesar de que el sistema propuesto obtuvo unos resultados destacados en la evaluación comparativa, éstos muestran que la aproximación basada en solapamiento léxico-sintáctico en general no es suficiente para abordar de manera adecuada el problema. Bien es cierto que, según se planteó en la hipótesis de partida (ver sección 1.3, página 11), los mejores resultados se consiguieron en la detección de implicación entre textos con alta similitud estructural o léxica, como es el caso de la aplicación de Documentos Comparables (CD). De todas maneras, el resto de los sistemas presentados al PASCAL RTE *Challenge* obtuvieron generalmente mejores resultados – incluso con notables diferencias con respecto al resto de aplicaciones – para la aplicación de Documentos Comparables, independientemente de la aproximación utilizada. Esto es, probablemente, debido a la alta tasa de coincidencia léxica que suelen mostrar los pares <texto, hipótesis> de esta aplicación, característica que todos los sistemas presentados abordaban de un modo u otro.

Tras un análisis de casos, se puede deducir que un alto solapamiento léxico no significa que exista implicación semántica y que un bajo solapamiento léxico no significa que la semántica sea diferente, lo que indica que se ha de tender a modelos más complejos que el análisis léxico-sintáctico super-

ficial. Hay modelos bastante complejos (por ejemplo, el de la Universidad de Stanford) que, sin embargo, han dado peores resultados que "simples" modelos estadísticos (por ejemplo, el de la Universidad Bar Ilan). Así pues, la cuestión más difícil de concretar estriba en determinar en qué aspectos hay que aplicar un procesamiento más complejo, o qué aspectos hasta ahora no considerados conviene tener en cuenta, además de cómo combinarlos adecuadamente. Hasta ahora no hay suficientes indicios para tomar una decisión a ese respecto, porque los resultados de todos los sistemas presentados (en cobertura) apenas diferían, distando sólo 9,1 puntos porcentuales entre el mejor y el peor resultado, dándose casos en los que sistemas de concepción dispar arrojaban resultados distinguibles en el orden de las décimas en tanto por ciento.

Si bien se pretende que en el futuro se pueda disponer de "motores de implicación" de amplia cobertura, como actualmente se dispone de analizadores sintácticos, posiblemente la mejor manera de llegar a esa generalidad sea pasar por el tratamiento en profundidad de casos particulares, con los que se puedan distinguir las técnicas apropiadas para resolverlos y descartar las aproximaciones que no valgan. Así, conjugando múltiples sistemas especializados en determinados tipos de inferencias, se podría construir en el futuro un sistema de uso general. En cierto modo, en el PASCAL RTE *Challenge* existió la posibilidad de focalizar los trabajos en la resolución de uno o varios tipos de apliaciones propuestos, ya que se podían presentar ejecuciones parciales. De todos modos, los participantes en general prefirieron evaluar sus sistemas frente al total de casos, opción lógica teniendo en cuenta que era una primera aproximación y así se podían formar una visión de conjunto.

En cuanto al sistema aquí descrito, tras un primer análisis posterior a la obtención de resultados, se han detectado dos situaciones léxico-sintácticas bastante frecuentes cuyo tratamiento ha de ser mejorado y que a continuación se apuntan:

- Algún tipo de parafraseo entre n-gramas sería de utilidad; por ejemplo, en el par 96 del corpus de entrenamiento (véase figura 6.1, página 53) es necesario detectar la equivalencia entre *same-sex* y *gay* o *lesbian*; o, en el par 128 (véase figura 6.1, página 53), *come into conflict with* y *attacks* deben detectarse como expresiones equivalentes. Este problema hace presente la necesidad de estudiar la inclusión de algún subsistema que sea capaz de detectar paráfrasis de longitud corta.

- Otro problema es que, en ciertos casos, se da un alto grado de solapamiento entre los nodos de la hipótesis y los nodos del texto pero, simultáneamente, las ramas de la hipótesis solapan con ramas dispersas en el árbol de dependencias del texto; entonces, las relaciones sintácticas entre subestructuras del texto y de la hipótesis, en términos del grado de dispersión de las ramas del árbol de análisis que

```
...
<pair id=‘‘96’’ value=‘‘TRUE’’ task=‘‘IR’’>
<t>The Massachusetts Supreme Judicial Court has cleared the
way for lesbian and gay couples in the state to marry, ruling
that government attorneys ‘‘failed to identify any
constitutionally adequate reason’’ to deny them the right.
</t>
<h>U.S. Supreme Court in favor of same-sex marriage</h>
</pair>
...
<pair id=‘‘128’’ value=‘‘TRUE’’ task=‘‘IR’’>
<t>Hippos do come into conflict with people quite often.
</t>
<h>Hippopotamus attacks human.</h>
</pair>
...
```

Figura 6.1: Pares 96 y 128 del corpus de entrenamiento.

las determinan, deben ser analizadas para determinar la existencia de implicación semántica. Este hecho sugiere que se ha de llevar a cabo un tratamiento en profundidad de las relaciones sintácticas; concretamente, el análisis actual podría ser enriquecido con el uso simultáneo de gramáticas de constituyentes, que permitan analizar ciertos rasgos que con gramáticas de dependencias, por su propia definición, no se pueden abordar.

Así pues, se observa que para el problema RTE es necesario abordar un amplio conjunto de fenómenos lingüísticos de manera específica, tanto en el nivel léxico como en el sintáctico.

A la vista de los resultados generales obtenidos, es claro que el RTE es un problema abierto, de sumo interés por sus amplias aplicaciones como base de múltiples tipos de tareas y con prometedores avances en los próximos años. Ante tal amplitud de posibilidades, la tarea definida en el primer PASCAL RTE *Challenge* no parece madura. Se trata de la primera vez que se realiza una tarea de este estilo y, aunque como prospectiva inicial ha sido suficientemente satisfactoria, sería conveniente redefinirla de manera oportuna. Si bien es cierto que se pretende obtener en el futuro sistemas de implicación textual genéricos, tras esta primera experiencia parece adecuado que las investigaciones podrían discurrir de manera que se partiese de sistemas especializados para, paulatinamente, irlos integrando. En algunos casos sólo se puede llegar a la detección de la implicación entre textos tras un análisis

profundo de determinados tipos de implicaciones, como relaciones numéricas o temporales. De este modo, una posible redefinición para la tarea sería que, tras el estudio de la casuística de tipos de inferencias que se presentan en los corpora hasta ahora utilizados, se proponga que cada sistema se especialice en la resolución de uno o varios (pero pocos) de esos tipos de inferencias, que obtuviesen un alto rendimiento. Una vez conseguida una colección de sistemas capaces de resolver tipos particulares de inferencias con un alto grado de precisión, el siguiente paso sería conjuntarlos convenientemente con el objetivo de obtener un sistema genérico.

Capítulo 7

Trabajo Futuro

A la luz del análisis de los resultados obtenidos por el sistema descrito en la presente memoria, así como los de los demás participantes en el PASCAL RTE *Challenge*, se evidencian una serie de actividades de realización recomendable para avanzar en el desarrollo del modelo propuesto:

- Hacer un estudio detallado de los corpora, con el objeto de determinar qué tipos de inferencias son necesarias para abordar con éxito la detección de implicación entre pares <texto, hipótesis> según el tipo de aplicación al que estén adscritos. Por ejemplo: relaciones temporales, relaciones espaciales, relaciones numéricas, relaciones entre entidades nombradas, detección de paráfrasis, etcétera.

- Desarrollar subsistemas que aborden la resolución de tipos particulares de inferencias, como las enumeradas en el punto anterior.

- Estudiar las diferencias posibles entre los tipos de implicación según el origen de los pares <texto, hipótesis> como, por ejemplo, Documentos Comparables (CD), Pregunta-Respuesta (QA), Traducción Automática (MT), etcétera.

- Buscar una representación de los textos más rica que los árboles de dependencias, añadiendo representaciones conceptuales de medidas numéricas, espaciales, temporales, de entidades nombradas, etcétera. Añadir, así mismo, más conocimiento del mundo, a partir de bases de datos, ontologías, tesauros y adquisición de conocimiento a partir de la *web*.

Al desarrollo de este trabajo debería acompañar la participación activa en foros internacionales, como las sucesivas ediciones del PASCAL RTE *Challenge*.

Capítulo 8

Difusión del Trabajo

Para dar a conocer el trabajo realizado y sus resultados a la comunidad científica internacional, así como para conocer el resto de los trabajos participantes en el evento y entablar contacto con científicos interesados en la materia, el autor asistió al *Workshop* del primer PASCAL RTE *Challenge*, que tuvo lugar el 12 de abril de 2005 en la Universidad de Southampton (Reino Unido), realizando la presentación titulada *Textual Entailment Recognition Based on Dependency Analysis and WordNet*. Dicha presentación dio lugar a la publicación en las actas del *Workshop* del artículo[1] que se refiere a continuación:

- Jesús Herrera, Anselmo Peñas y Felisa Verdejo. *Textual Entailment Recognition Based on Dependency Analysis and WordNet*. Proceedings of the First PASCAL Recognizing Textual Entailment Challenge Workshop, Southampton, UK, April 2005.

[1] Disponible en: `http://cs.biu.ac.il/~glickmao/rte05/herrera-et_al.pdf`

Capítulo 9

Agradecimientos

9.1. Agradecimientos institucionales

El presente trabajo ha sido financiado en parte por el Ministerio de Ciencia y Tecnología, con cargo a los presupuestos del proyecto SyEMBRA (Sistemas y Evaluación Multilingües de Busqueda de Respuestas) TIC-2003-07158-C04-02.

9.2. Agradecimientos personales

La primera persona a quien debo estar agradecido es al director de este trabajo, Anselmo Peñas, que con su encomiable labor de guía ha conseguido que la tarea llegue a este punto. He de destacar su claridad de ideas, su capacidad de intuir el camino a seguir, su flexible exigencia y el *savoir faire* en los momentos difíciles.

A Felisa Verdejo, directora del Departamento de Lenguajes y Sistemas Informáticos de la UNED, he de agradecer la disposición de dependencias y materiales del Departamento para que realizase el trabajo, así como sus recomendaciones.

A Ido Dagan y Oren Glickman, de la Universidad Bar Ilan de Israel, agradezco todo su apoyo prestado como organizadores del primer PASCAL RTE *Challenge*.

Gracias a los compañeros de trabajo de la UNED, por hacer que el ambiente de trabajo sea tan agradable y que no falte nunca esa pequeña ayuda que te permita continuar: Víctor Peinado (mención especial por la plantilla de LaTeXpara realizar esta memoria), Juan Mascarell (que me prestó sus libros de LaTeX), Iñaky Cárdenas (que también me ayudó con el LaTeX), Emilio Lorenzo, Teresa Sastre, Alberto Ruiz, Cova Rodrigo, Tim Read (qué poco inglés sé), Juan Cigarrán, Javier Artiles, Enrique Amigó, Fernando López, Nacho Mayorga, Valentín Sama, Julio Gonzalo, Carlos Celorrio, Javier Vélez, Elena Ruiz y Yoli Calero.

Este tipo de trabajos, que tienen características de carrera de fondo, te imbuyen de tal manera que acabas haciendo partícipes de él a los que más cerca te rodean. Por ello, fuera del ámbito académico, durante estos meses ha sido vital la compañía, la comprensión y los momentos – algunos de ellos inolvidables – que me han proporcionado: Raúl Domínguez, Θεοδώρα Παναποπούλου, Rosa García-Gasco, Mónica Durán y, por supuesto, Alma Eyjólfsson, que está plagada de sonrisas.

Bibliografía

E. Akhmatova. Textual Entailment Resolution via Atomic Propositions. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 61–64, April 2005.

H. Alshawi and D. Carter. Training and Scaling Preference Functions for Disambiguation. In *Computational Linguistics 20(4)*, pages 635–648, December 1994.

A. Andreevskaia, Z. Li, and S. Bergler. Can Shallow Predicate Argument Structures Determine Entailment? In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 45–48, April 2005.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Series/Addison Wesley, New York, 1999.

S. Bayer, J. Burger, L. Ferro, J. Henderson, and A. Yeh. MITRE's Submissions to EU PASCAL RTE Challenge. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 41–44, April 2005.

J. Bos and K. Markert. Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 65–68, April 2005.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation. In *Computational Linguistics 19(2)*, 1993.

I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 1–8, April 2005.

R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons. Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 29–32, April 2005.

R. Delmonte, S. Tonelli, M. A. Picollino Boniforti, A. Brsitot, and E. Pianta. VENSES – a Linguistically-Based System for Semantic Evaluation. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 49–52, April 2005.

A. Fowler, B. Hauser, D. Hodges, I. Niles, A. Novischi, and J. Stephan. Applying COGEX to Recognize Textual Entailment. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 69–72, April 2005.

O. Glickman, I. Dagan, and M. Koppel. Web Based Textual Entailment. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 33–36, April 2005.

D. Graff. English Gigaword. http://www.ldc.upenn.edu/Catalog/, 2003.

Z. Harris. Distributional Structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–37, 1985.

J. Herrera, A. Peñas, and F. Verdejo. Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 21–24, April 2005.

V. Jijkoun and M. de Rijke. Recognizing Textual Entailment Using Lexical Similarity. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 73–76, April 2005.

T. Joachims. *Learning to Classify Text Using Support Vector Machines.* Kluwer Academic Publishers, Massachusetts, 2002.

V. I. Levensthein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics - Doklady*, volume 10, pages 707–710, 1966.

D. Lin. Extracting Collocations from Text Corpora. In *Workshop on Computational Terminology. Montreal, Canada*, 1998.

D. Lin and P. Pantel. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328, 2001.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts, 1999.

G. A. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, number 3(4), pages 235–244, 2004.

C. Monz and M. de Rijke. Light-Weight Entailment Checking for Computational Semantics. In P. Blackburn and M. Kohlhase, editors, *Inference in Computational Semantics (ICoS-3)*, pages 59–72, 2001.

J. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics 29(1)*, 2003.

M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto. Textual Entailment as Syntactic Graph Distance: a Rule Based and a SVM Based Approach. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 25–28, April 2005.

R. Raina, A. Haghighi, C. Cox, J. Finkel, J. Michels, K. Toutanova, B. MacCartney, M. C. de Marneffe, C. D. Manning, and A. Y. Ng. Robust Textual Inference using Diverse Knowledge Sources. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 57–60, April 2005.

D. Ravichandran and E. Hovy. Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 2002 ACL Conference*, 2002.

G. R. Sampson. *English for the Computer*. Oxford University Press, 1995.

I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling Web-Based Acquisition of Entailment Relations. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-04)*, 2004.

H. Tanev, M. Kouylekov, and B. Magnini. Combining Linguistic Processing and Web Mining for Question Answering. In *Proceedings of the 2004 Edition of the Text Retrieval Conference*, 2004.

L. Vanderwende, D. Coughlin, and B. Dolan. What Syntax can Contribute in Entailment Task. In *Proceedings of the First PASCAL Recognizing Textual Entailment Workshop, Southampton, UK*, pages 13–16, April 2005.

E. M. Voorhees. The TREC-8 Question Answering Track Report. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eigth Text Retrieval Conference (TREC 8)*, number 500-246, pages 77–82. NIST Special Publication, 1999.