ISSN 1870-4069

Comparación de dos modelos para resolver el problema de implicación textual multilingüe

Darnes Vilariño, Saul Leon Silverio, Beatriz Beltrán

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

{darnes, bbeltran}@cs.buap.mx

Resumen. En el presente trabajo de investigación se presentan dos modelos uno basado en anclas y otro basado en grafos para resolver el problema de implicación textual multilingüe. Los datos con los que se ha trabajado son los ofrecidos en la Conferencia Internacional Semeval 2012 y Semeval 2013. Al probar ambos modelos quedó claro que el modelo basado en grafos supera al modelo basado en anclas, para los 3 corpus utilizados. Se desarrolló una base de conocimientos utilizando ConceptNet5, OpenOffice Thesaurus y WordNet. Esta base de conocimientos se representó mediante un grafo y se diseñó un algoritmo para detectar si dos términos están relacionados dentro del grafo.

Palabras clave: Implicación textual, modelo basado en anclas, modelo basado en grafo, multilingüe.

Comparison of Two Models for Solving the Textual Entailment in Cross Lingual Problem

Abstract. In the present research work, two models are presented, one based on anchors and the other based on graphs, to solve the problem of cross-lingual textual entailment. The data with which we have worked are those offered at the Semeval 2012 and Semeval 2013 International Conference. By testing both models, it was clear that the graph-based model exceeds the anchor-based model, for the 3 corpus used. A knowledge base was developed using ConceptNet5, OpenOffice Thesaurus and WordNet. This knowledge base was represented by a graph and an algorithm was designed to detect if two terms are related within the graph.

Keywords: Textual entailment, model based on anchors, model based on graphs, cross-lingual.

1. Introducción

La vasta cantidad de información disponible en Internet hoy en día supera nuestra capacidad de almacenarla, procesarla y por ende aprovecharla. En los últimos años grandes repositorios de información, como enciclopedias, periódicos, blogs, revistas electrónicas, multimedia y otras fuentes digitales de información han generado contenidos que pueden ser útiles para algún usuario en particular [6,8], sin embargo la disponibilidad de información en varios idiomas, así como la variación léxico-sintáctica del lenguaje suponen un reto a los sistemas de Procesamiento de Lenguaje Natural (PLN) que existen actualmente.

Para explicar los fenómenos de la disponibilidad de información en varios idiomas y la variación léxico-semántica, se analizan los siguientes escenarios:

A. Alger, un usuario alemán, desea buscar información acerca de la cultura Maya de México, hay una gran cantidad de artículos en idioma inglés que tratan de la cultura Maya, no obstante, los artículos que son del interés de Alger se encuentran en español. Alger traduce del idioma alemán al idioma inglés sus inquietudes, pero como se ha notado no haya nada útil y desiste de ser arqueólogo de Chichen Itza. Al igual que Alger, miles de usuarios son limitados por la cantidad de información que encuentran en su idioma.
B. Roberto es un niño que tiene por tarea definir la lluvia, así que busca definiciones de ”Lluvia”, como buen estudiante, busca en más de una fuente para redactar una definición extensa, él se percata de que todas las definiciones denotan lo mismo, pero las palabras, y la forma en las que están escritas son diferentes. Así como Roberto, muchos usuarios buscan información concreta, pero la manera en que está escrita (léxico y sintaxis) impide que sea de fácil acceso. Roberto imagina un Google mejorado que pueda interpretar los textos que hablan de lo mismo, pero con diferentes palabras y escritos de manera diferente.

Estas dos problemáticas habían sido investigadas por separado hasta hace un par de años, ahora se estudian juntas en un problema denominado ”Implicación Textual Multilingüe” (ITML o CLTE por sus siglas en inglés), cuyo fin es solucionar los escenarios dados en A y en B.

En el año 2012, dentro del marco de la conferencia internacional SemEval-2012, se propone a la comunidad científica internacional resolver el problema CLTE[1], éste consiste en determinar si un texto T y una hipótesis H, escritos en diferentes idiomas, se puede inferir el significado de H a partir del significado de T. Formalmente:

Dado un par de fragmentos de texto, tópicamente relacionados, T1 y T2 escritos en diferentes idiomas, la tarea consiste en asignar automáticamente uno de los siguientes juicios de implicación textual:

- Bidirectional: $(T1 \rightarrow T2 \,\&\&\, T2 \rightarrow T1)$ equivalencia semántica.
- Forward: $(T1 \rightarrow T2 \,\&\&\, !T2 \rightarrow T1)$ implicación unidireccional de $T1$ a $T2$.
- Backward: $(!T1 \rightarrow T2 \,\&\&\, T2 \rightarrow T1)$ implicación unidireccional de $T2$ a $T1$.
- No-Entailment: $(!T1 \rightarrow T2 \,\&\&\, !T2 \rightarrow T1)$ sin implicación entre $T1$ y $T2$.

[1] Una extensión multilingüe del problema de Implicación Textual

En esta tarea se asume que tanto T1 y T2 son declaraciones verdaderas (TRUE) y que no existen pares contradictorios [3].

Para desarrollar modelos que solucionan el problema de Implicación Textual Multilingüe, los organizadores ponen a disposición de la comunidad de investigadores a nivel internacional un conjuntos de datos, para entrenar y probar los mismos (train / test). Se ofrecen las siguientes combinaciones de idiomas:

– Español / Inglés,
– Alemán / Inglés,
– Italiano / Inglés,
– Francés / Inglés.

Dotar a los sistemas de un módulo capaz de inferir cuando dos textos expresan lo mismo, enriquece las posibilidades de satisfacer totalmente a la consulta expresada por el usuario. Es importante destacar que esto aún se vuelve más relevante, si la consulta está en un idioma determinado y lo que se está buscando está en otro [4].

Otro resultado implícito que se obtiene con esta investigación, es la posibilidad de desarrollar corpus paralelos de manera no supervisada, que ayude a los sistemas que desarrollan modelos de traducción automática.

Los modelos desarrollados hasta el año 2014 no consideran el juicio de implicación textual contradicción, que en muchas ocasiones es muy difícil de detectar ya que solamente la presencia o ausencia de una palabra hace que dos textos sean totalmente contradictorios. El desarrollo de modelos para resolver este problema, puede ser incorporado a un módulo de procesamiento de lenguaje natural [5].

Detectar el juicio de implicación textual cross lingüe sin el desarrollo de modelos de traducción permite dar solución al problema sin necesidad de disponer de un corpus paralelo totalmente heterogéneo que permita el desarrollo de diccionarios estadísticos de dominio general y dominio particulares.

La mayoría de los modelos desarrollados para resolver este problema son supervisados, lo que obliga a disponer de un corpus de entrenamiento desarrollado por expertos. En muchas ocasiones no es tan fácil disponer de un corpus categorizado, por lo que es importante desarrollar modelos no supervisados, que ofrezcan buen comportamiento independientemente del corpus con el que se está trabajando [10].

El presente trabajo de investigación se encuentra estructurado de la siguiente manera en la sección 2 se explican los modelos que se desarrollaron para realizar las pruebas. En la sección 3 se tiene una breve descripción de las colecciones de datos y de cómo están conformadas dichas colecciones. En la sección 4 se muestran los resultados obtenidos con las pruebas realizadas y finalmente se dan las conclusiones a las que hemos llegado.

2. Modelos desarrollados

Antes de comenzar con la reseña de las 2 metodologías, es necesario definir algunos criterios generales de las mismas. El CLTE asigna juicios de implicación a un par de oraciones en idiomas diferentes, en la Figura 1 se muestran 4 oraciones de la colección español / inglés, cada una etiquetada con su respectivo juicio de implicación.

Ejemplo de CLTE

```
<entailment-corpus languages="spa-eng">

 <pair id="1" entailment="bidirectional">
  <t1>Mozart nació en la ciudad de Salzburgo.</t1>
  <t2>Mozart was born in Salzburg.</t2>
 </pair>

 <pair id="2" entailment="forward">
  <t1>Mozart nació el 27 de enero de 1756 en Salzburgo.</t1>
  <t2>Mozart was born in 1756 in the city of Salzburg.</t2>
 </pair>

 <pair id="3" entailment="backward">
  <t1>Mozart nació en la ciudad de Salzburgo.</t1>
  <t2>Mozart was born on 27th January 1756 in Salzburg.</t2>
 </pair>

 <pair id="4" entailment="no_entailment">
  <t1>Mozart nació el 27 de enero de 1756 en Salzburgo.</t1>
  <t2>Mozart was born to Leopold and Anna Maria Pertl Mozart.</t2>
 </pair>

</entailment-corpus>
```

Fig. 1. Ejemplo de CLTE con idiomas Inglés / Español.

El CLTE ha sido parcialmente resuelto utilizando algoritmos basados en reglas, que toman patrones repetitivos a niveles léxicos y/o sintácticos. Algunas de estas aproximaciones utilizan el conjunto de entrenamiento para poner a punto las reglas que evalúan sobre el conjunto de prueba [2].

Otra vertiente, quizás la más explotada, utiliza algoritmos de aprendizaje automático, que extraen características léxicas y/o sintácticas para descubrir patrones repetitivos en los 4 diferentes juicios de implicación. En este enfoque se construye un modelo para los datos de entrenamiento, que posteriormente permite evaluar los datos de prueba.

De este hecho se pueden distinguir dos perspectivas, la primera toma los cuatro juicios de implicación como clases, es decir que los modelos perciben 4 clases y por ello se nombra a este enfoque multiclase. El segundo punto de vista descompone las oraciones en sus formas más primitivas, reconociendo sólo 2 clases (implicación y no-implicación), esto quiere decir que para obtener el juicio final se realiza una composición de dos clases como exhibe la Tabla 1, este ajuste es denominado como composición.

Tabla 1. Tabla de composición de juicio de implicación.

Juicio de T1 ⇒ T2	**Juicio de T2 ⇒ T1**	**Juicio Final**
Entailment	Entailment	Bidirectional
Entailment	No_entailment	Backward
No_entailment	Entailment	Forward
No_entailment	No_entailment	No_entailment

Independientemente de lo anterior, se tienen dos posibles tratamientos derivados del lenguaje del par de oraciones, el primero busca utilizar métodos multilingüe que rescatan las técnicas de las investigaciones del área de traducción automática (MT, por sus siglas en inglés) [1]. Mientras que el segundo enfoque utiliza un idioma como pivote para hacer que las dos oraciones se encuentren en el mismo idioma, una vez hecho esto, el problema pasa de ser CLTE a implicación (TE) a ello se puede operar con metodologías del TE. Por su parte TE data del año 2005, por lo que utilizar el pivote supone una ventaja en comparación al CLTE.

A continuación, se explica con detalle cada una de las técnicas desarrolladas.

2.1. Modelo de inferencia basado en anclas

Si existe una gran cantidad de información compartida, quiere decir que las oraciones comunican la misma idea. Esta asunción no es del todo correcta, ya que puede presentarse el caso de que dos oraciones que compartan una gran cantidad de información no transmitan la misma idea, por ejemplo el par de oraciones: ”La becerra de Manuel está en la calle” y ”Esta es la calle de Manuel Becerra” es un par equivalente a nivel de tokens, pero al momento de interpretar ambas oraciones, se descubre que hablan de cosas completamente diferentes.

Ante esta nueva evidencia, es necesario proponer un nuevo modelo, el cual sea capaz de detectar los tokens que comparten un par de oraciones, y descubrir si estos tokens se relacionan de manera similar en ambas oraciones. Como preprocesamiento sólo se lematizan las oraciones.

Términos Anclas Consideremos el par de plantillas ”X es obra de Y” y ”X escribió Y”, veamos que en ambas se repiten las variables X y Y, estas variables pueden ser sustituidas por cualquier par de tokens, como por ejemplo Tokio Blues es obra de Murakami”, ”Murakami escribió Tokio Blues”. Ahora bien, estos tokens son denominados anclas, ya que, al estar presentes en ambas oraciones, nos permiten detectar la posición de los elementos en cada oración, de esta manera se pueden analizar cómo es que se relaciona una ancla con la otra dentro de cada par de oraciones. Siguiendo la metodología del empatamiento de plantillas, una vez detectada la posición de las anclas se recurre a buscar un conjunto de plantillas que contenga el texto que está entre las anclas, si se encuentra un par de plantillas que satisfacen las posiciones de las anclas, así como su relación (el texto que hay entre ellas), se dice que hay implicación textual.

Ahora pensemos de manera inversa. A partir de una oración se buscan sus tokens anclas, si estos distan entre sí, a no más de dos tokens en cada oración, quiere decir que se encuentran relativamente cerca, y sospechamos que sostienen alguna relación. Entonces, si en lugar de buscar la relación en plantillas, pudiéramos inferir qué similitud existe entre ambas relaciones, estaríamos descubriendo la implicación textual sin el uso de plantillas, pero con el mismo principio.

En el siguiente apartado se explica cómo se puede llevar a cabo el proceso de inferencia.

Proceso de inferencia La inferencia de dos relaciones, se puede abordar con diferentes técnicas, en los siguientes incisos se describen cada una de ellas:

- Relaciones Directas: Las relaciones directas corresponden a sinonimia, hiperonimia o hiponimia. Son relaciones tan obvias que basta con sustituir la relación con su sinónimo para detectar si son relaciones iguales.
- Similitud Semántica: Estas relaciones son afines, no necesariamente son sinónimos, pero son empleadas bajo contextos similares. Al tomar dos relaciones y detectar una relación semántica, entonces decimos que son iguales.

Para detectar el juicio de implicación que sostienen dos oraciones, se recurre al concepto de eliminación de información común, pero a diferencia de la eliminación de tokens, ahora se eliminarán relaciones de términos anclas, esta nueva variante garantiza que la información eliminada en ambas oraciones corresponde a un nivel interpretativo, más allá del léxico.

Se plantea el algoritmo de la siguiente manera, primero se obtienen todos los términos anclas, que por simplicidad serán los n-gramas más largos presentes en ambas oraciones, posteriormente se tokenizan ambas oraciones y se verifica la distancia a la que se encuentran las anclas, si los términos anclas no exceden un máximo de dos tokens entre ellos en ambas oraciones, se verifican las relaciones que sostienen en una y en otra. Si las relaciones son similares, se eliminan las relaciones y las anclas involucradas en cada oración, al final se asigna el juicio de implicación, en dependencia del porcentaje de eliminación.

2.2. Modelo de interpretación de oraciones basado en grafos

En general se han propuesto algunas metodologías que intentan resolver el problema de la implicación textual a través de propuestas supervisadas y no supervisadas [11]. Recordando que en un lenguaje natural, las oraciones tienen tres niveles de composición, léxico, sintáctico y semántico. En cada nivel se evalúan aspectos diferentes, al descomponer una oración en sus tres niveles, se puede apreciar ciertas igualdades y diferencias.

Por ejemplo, si se toman dos oraciones que comparten poco léxico, que una oración se encuentra escrita en voz pasiva y la otra en voz activa, y que en cuestión semántica difieren; para un modelo computacional que se basa en similitudes, este tipo de oraciones es claramente un no_entailment, y sin embargo puede ser bidirectional. El humano es capaz de interpretar el contenido de cada oración, y gracias a su sentido común, es capaz de relacionar si las ideas son las mismas, de esta manera oraciones como ”Un hombre sostiene un aparato electrónico” y ”Una cámara digital es aguantada por un transeúnte” son oraciones que reflejan la misma idea.

Hasta el desarrollo de esta investigación, no se ha encontrado reportado en la bibliografía ningún sistema que resuelva TE y mucho menos CLTE a través de una interpretación de oraciones, por ello se considera de suma importancia proponer un modelo no supervisado que interprete las oraciones para identificar qué entidades se encuentran en juego y determinar las relaciones que sostienen en cada oración, posteriormente se debe detectar qué entidades son equivalentes y verificar si sus relaciones también son equivalentes, para así determinar si son la misma idea.

El proceso de interpretación de oraciones que se presenta en esta metodología, tiene su origen en las investigaciones de Preguntas y Respuestas [9]. Dentro de los sistemas de preguntas y respuestas se encuentran sistemas que procesan preguntas factuales y preguntas complejas. Las preguntas factuales, corresponden a información verdadera que se encuentra escrita textualmente en algún lugar, mientras que las preguntas complejas, son preguntas que necesitan valerse de información extra para poder ser respondidas, por ejemplo, la pregunta ”¿Dónde nació Albert Einstein?” es una pregunta factual, y la pregunta ”¿Qué medicamento le puedo dar a un paciente enfermo de gripe si es alérgico al paracetamol?” es una pregunta compleja.

Como en el descubrimiento de la implicación textual involucra un par de oraciones, es posible ver una oración como la fuente de conocimiento y la otra como las preguntas que se le hacen a un sistema de preguntas factuales. En un sistema de preguntas factuales se generan árboles sintácticos de la información que se desea indexar, esta estructura permite generar estructuras atómicas del tipo Sujeto-Verbo-Objeto, o lo que es lo mismo Entidad_1-Accion-Entidad_2, una vez generadas estas tripletas son agregadas al índice, cuando una pregunta es lanzada al sistema, este identifica las entidades y acciones presentes en la pregunta y lanza un buscador con operador OR sobre su índice de tripletas, de modo que todas las tripletas recuperadas son respuestas de la pregunta, por ejemplo consideremos el siguiente ejemplo:

Sean los siguientes incisos, el índice de tripletas de un sistema de preguntas y respuestas:

1. (”Los lobos”, ”comen”, ”conejos”),
2. (”Las tortugas”, ”comen”, ”ranas”),
3. (”Las tortugas”, ”viven”, ”el mar”).

Cuando se le pregunta al sistema ”¿Dónde viven las tortugas?”, este identifica los términos ”viven” y ”tortugas”, los cuales son buscados en sus índices y encuentra que el índice 3 contiene ambos términos, finalmente el sistema asume que ”el mar” es la respuesta de la pregunta. Esta investigación retoma este mecanismo de validación de información para interpretar las oraciones.

Este modelo es una aproximación que emplea parsers sintácticos para interpretar las oraciones, después con ayuda de una serie de reglas se extraen hechos de cada oración, estos hechos se utilizan en la confección de grafos. Cada nodo de un grafo generado corresponde a términos de la oración y las aristas son las relaciones entre los términos. Una vez que se construyen los grafos de cada oración, se aplica un algoritmo que busca empatar subestructuras de los mismos, las cuales denominamos hechos-extendidos, que, con ayuda de una fuente de conocimiento, se puede determinar si los nodos son equivalentes o si tendrían alguna relación. Finalmente, la cantidad de subestructuras compartidas, con respecto a la cantidad de subestructuras presente en cada oración nos proporciona el juicio de implicación textual.

A continuación, se exponen cada una de las etapas involucradas en la confección de este modelo.

Extracción de hechos Para entender que es un hecho, se considera la oración ”El lobo feroz derribó la casa de los cerditos, se comió al cazador y huyó a las Vegas”, a partir

de ella podemos extraer la siguiente información: el lobo es feroz, el lobo derribó una casa, la casa es de los cerditos, el lobo se comió al cazador, el lobo huyó a las Vegas. Toda esta información obtenida es denominada hechos, y se define como información verdadera generada a partir de una oración.

El proceso de extracción de hechos consiste en tomar una oración y generar toda la información relevante a partir de ella, para realizar esta tarea se emplea el parser sintáctico de Stanford[2], quien genera el árbol sintáctico. A partir del árbol sintáctico se visita cada nodo para poder generar nuevos nodos, a continuación, se listan los nombres de los nuevos nodos:

- **ENTIDAD**: Son todos los términos de los nodos hijos que contengan algún sustantivo, se recuperan todos los términos asociados a las etiquetas de PoS del tipo: NN, NNS, NNP, NNPS y NAC.
- **ACTIVIDAD**: Son todos los términos de los nodos hijos que contengan algún verbo, se recuperan todos los términos asociados a las etiquetas de PoS del tipo: VB, VBG, VBZ, VBP, VBN, VBD y MD.
- **CALIDAD**: Son todos los términos de los nodos hijos que contengan algún adjetivo calificativo, se recuperan todos los términos asociados a las etiquetas de PoS del tipo: JJ y JJS.
- **PREPOSICIÓN**: Son todos los términos de los nodos hijos que contengan alguna proposición, se recuperan todos los términos asociados a las etiquetas de PoS del tipo: IN y TO.

Los nombres de los nodos reflejan los roles de la información recuperada para relacionar la información recabada, se buscan los siguientes hechos:

- **Sujeto**: Este hecho es obtenido cuando una entidad está asociada a una, o más, actividades. Se asume que la ENTIDAD es el sujeto de la ACTIVIDAD.
- **Objeto**: Este hecho es obtenido cuando una actividad está asociada a una, o más, entidades. Se dice que la ENTIDAD es el objeto de la ACTIVIDAD.
- **Califica**: Este hecho es obtenido cuando se detecta una cualidad asociada a una entidad, entonces se dice que esa CALIDAD califica a la ENTIDAD.
- **Extensión**: Este hecho es obtenido cuando una ENTIDAD está asociada a una PREPOSICIÓN.
- **Complemento**: Este hecho es obtenido cuando una PROPOSICIÓN está asociada a una ENTIDAD.

Estas reglas son extraídas al recorrer el árbol sintáctico, que contienen las nuevas etiquetas, y en cada nodo se aplican las reglas de la Tabla 2 para extraer los hechos.

Interpretación de hechos sobre grafos Se ha dicho que la estructura empleada por los sistemas de preguntas y respuestas, para representar la información indizada, consiste en una tripleta de la forma Entidad _1-Actividad-Entidad_2. Ahora bien, bajo el contexto de los hechos se aprecia que corresponde a un par de hechos: sujeto(Entidad_1,

[2] https://nlp.stanford.edu/software/lex-parser.shtml

Tabla 2. Reglas para la extracción de hechos.

Regla	Hecho
$X \Rightarrow$ ENTIDAD $\&$ $X \Rightarrow$ ACTIVIDAD	sujeto (ENTIDAD, ACTIVIDAD)
ACTIVIDAD $\Rightarrow$ ENTIDAD	objeto (ACTIVIDAD, ENTIDAD)
$X \Rightarrow$ ENTIDAD $\&$ $X \Rightarrow$ CALIDAD	califica (CALIDAD, ENTIDAD)
ENTIDAD $\Rightarrow$ PREPOSICIÓN	extensión (ENTIDAD, PREPOSICIÓN)
PREPOSICIÓN $\Rightarrow$ ENTIDAD	complemento (PREPOSICIÓN, ENTIDAD)

Actividad) & objeto(Actividad, Entidad_2), de modo que es necesario representar los hechos a través de una estructura que permita preservar tanto la información de los hechos como las relaciones entre ellos. Por ello se recurre a una estructura de grafo.

Al representar los hechos en un grafo surge el concepto de hechos ampliados, un hecho ampliado es la manera en la que se relacionan los términos de alguna oración, bajo el contexto de preguntas y respuestas corresponde a una tripleta, y para el contexto de grafos corresponden a subestructuras del grafo.

Tabla 3. Subestructuras del grafo con sus equivalencias en tripletas.

Subestructura	Tripleta
sujeto(Entidad_1, Actividad) & objeto(Actividad, Entidad_2)	(Entidad_1, Actividad, Entidad_2)
califica(Entidad, Calidad)	(Entidad, califica, Calidad)
extensión(Entidad_1, Preposición) & complemento(Preposición, Entidad_2)	(Entidad_1, Preposición, Entidad_2)

Para representar los hechos a través de grafos, basta con tomar a los textos y relacionarlos mediante sus hechos.

Empatamiento de grafos para el descubrimiento de la implicación textual Dentro de la teoría de grafos, el mecanismo que denota que dos grafos tienen la misma estructura es denominado isomorfismo de grafos, que es una metodología rígida que da mayor peso a la forma del grafo que a su contenido [7]. Por otro lado, en algunas ocasiones los grafos que se desean empatar varían en el número de nodos, y no es posible realizar un estudio de isomorfismo, en estos casos lo ideal sería aplicar un empatamiento de hipergrafos. Los hipergrafos son grafos que en cada nodo puede contener otro grafo, desafortunadamente esta última propiedad provoca que el empatamiento de grafos sea un problema NP-Completo.

Como empatar el grafo de forma matemática no es una opción viable, se propone un algoritmo ávido que recorre cada grafo para generar subestructuras, llamadas hechos-extendidos, que al final estas son validadas con otro conjunto de subestructuras. Para validar una subestructura contra otra es necesario decidir si dos segmentos de textos son equivalentes, llegados a este punto siempre se ha hablado de sinonimia, hiperonimia, hiponimia y similitud semántica.

Sin embargo, para este modelo se propone el uso de una fuente de información que busca capturar el sentido común del mundo real en una base de conocimientos, esta base es llamada Conceptnet[3], adicionalmente se utiliza el tesauro de OpenOffice y WordNet, para conformar una base de conocimiento lo más sólida posible.

El empatamiento de subestructuras, que ahora se denomina empatamiento de hechos extendidos, consiste en visitar cada nodo de ambos grafos, generar una subestructura, y verificar si son equivalentes, para ello se emplea la base de conocimientos, se debe asegurar que los hechos extendidos, que se están analizando, posean la misma información para decir que son hechos equivalentes.

3. Colecciones de datos y análisis de resultados

En esta sección se presentan los resultados alcanzados por ambos modelos. Con dos colecciones extras (RTE1 y SICK) que ayudarán al análisis de resultados de las metodologías desarrolladas.

A pesar de que los modelos desarrollados en esta investigación se basan en dos conjuntos de colecciones, es posible extrapolar las metodologías propuestas a otros conjuntos de datos, esto gracias a que los enfoques de las propuestas toman inspiración de modelos que tratan de resolver TE.

Para probar el desempeño de los enfoques expuestos en la secciones anteriores, se han considerado 3 colecciones de datos:

3.1. Corpus CLTE

La colección de datos del CLTE fue liberada a la comunidad científica en el foro internacional de evaluación semántica SemEval, una primera parte en su edición 2012, y una segunda en su edición 2013.

Como las colecciones del CLTE constan de 4 idiomas, se ha decidido solo reportar la subcolección de Inglés/Español, ya que es la colección en la que mejor se comportan todos los algoritmos del estado del arte. La distribución de esta colección puede apreciarse en la Tabla 4.

Tabla 4. Información de la subcolección Inglés/Español del Corpus CLTE.

Conjunto	Pares de oraciones	Juicio Bi-directional	Juicio Backward	Juicio Forward	Juicio No_Entailment
CLTE12-Train	500	125	125	125	125
CLTE12-Test	500	125	125	125	125
CLTE13-Test	500	125	125	125	1225

[3] http://conceptnet5.media.mit.edu

3.2. Corpus RTE1

La colección de datos del RTE1 (Recognising Textual Entailment 1 [2]) fue liberada en el marco del Pattern Analysis, Statistical Modelling and Computational Learning (SemEval), donde se utilizó por primera vez el término Textual Entailment. El corpus RTE1 consta de 2 clases, las cuales se encuentran balanceada. La Tabla 5 muestra la distribución de las colecciones que conforman al corpus RTE1.

Tabla 5. Información del Corpus RTE1.

Conjunto	Pares de Oraciones	Pares de oraciones con Juicio Entailment	Pares de oraciones con Juicio No-Entailment
RTE1-Train1	287	143	144
RTE1-Train2	280	140	140
RTE1-Test	800	400	400

A continuación, se exponen los resultados alcanzados por cada una de las metodologías propuestas en cada uno de los corpus.

4. Resultados

4.1. Resultados de modelo de inferencia basado en anclas

Es importante comentar que para este modelo el tokenizador que se use es determinante, ya que las anclas cambian totalmente, para todos los corpus el tokenizador que permitió detectar las mejores anclas fue el linguistic, este modelo para todos los corpus, supera los resultados obtenidos por los dos modelos que utilizan el principio de eliminación.

4.2. Resultados del modelo de interpretación de oraciones basados en grafos

Para este modelo solo se maneja una sola configuración, las tablas que a continuación se presentan, corresponden a los resultados obtenidos.

Tabla 6. Información de la subcolección Inglés/Español del Corpus CLTE.

Conjunto	Pares de oraciones	Juicio Bi-directional	Juicio Backward	Juicio Forward	Juicio No_Entailment
CLTE12-Train	500	125	125	125	125
CLTE12-Test	500	125	125	125	125
CLTE13-Test	500	125	125	125	125

La colección de datos del RTE1 (Recognising Textual Entailment 1 [2]) fue liberada en el marco del Pattern Analysis, Statistical Modelling and Computational Learning

(SemEval), donde se utilizó por primera vez el término Textual Entailment. El corpus RTE1 consta de 2 clases, las cuales se encuentran balanceada. La Tabla 5 muestra la distribución de las colecciones que conforman al corpus RTE1.

Tabla 7. Resultados del Modelo de Interpretación de Oraciones con el Corpus CLTE.

Conjunto	Precisión de Bidirectional	Precisión de Backward	Precisión de Forward	Precisión de NoEntailment	Precisión Global
CLTE12-Test	0.336	0.168	0.136	0.728	0.342
CLTE13-Test	0.200	0.096	0.080	0.768	0.286

Tabla 8. Resultados del Modelo de Interpretación de Oraciones con el Corpus RTE1.

Conjunto	Precisión de Entailment	Precisión de NoEntailment	Precisión Global
RTE1-Test	0.380	0.667	0.523

En general para los dos corpus reportados el comportamiento ha sido similar. Debido a que la metodología es muy novedosa, quizá haga falta realizar un estudio posterior, para detectar que debilidades pudiera presentar, ya que no logra empatar de manera correcta los grafos asociados a cada uno de los textos. A continuación, se realiza una comparativa entre todos los modelos desarrollados, presentando el mejor comportamiento de cada uno con respecto a cada colección de datos procesada.

5. Conclusiones

Se desarrollaron 163,622 experimentos diferentes, variando las medidas de similitud, el tokenizador, los modelos, el tipo de corpus, los umbrales para detectar la similitud de tokens y los umbrales empleados por el proceso de eliminación, lo que nos ha permitido llegar a las siguientes conclusiones:

1. Las características del corpus influyen completamente en los resultados de cada uno de los modelos.
2. El modelo basado en anclas está también directamente relacionado al proceso de tokenización, porque de este depende la selección de las anclas de cada parte de los textos y permite la extracción de patrones de manera automática.
3. El modelo de grafos puede utilizarse como una herramienta para extraer información.
4. Se desarrolló una base de conocimientos utilizando ConceptNet5, OpenOffice Thesaurus y WordNet. Esta base de conocimientos se representó mediante un grafo y se diseñó un algoritmo para detectar si dos términos están relacionados dentro del grafo.

Referencias

1. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005)
2. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. pp. 177–190. MLCW'05, Springer-Verlag, Berlin, Heidelberg (2005)
3. Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L.: UAlacant: using online machine translation for cross-lingual textual entailment (2012-06)
4. Jimenez, S., Becerra, C., Gelbukh, A.: Soft cardinality + ML: Learning adaptive similarity functions for cross-lingual textual entailment. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 684–688. Association for Computational Linguistics, Montréal, Canada (7-8 Jun 2012)
5. Mehdad, Y., Negri, M., C. de Souza, J.G.: FBK: Cross-lingual textual entailment without translation. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 701–705. Association for Computational Linguistics, Montréal, Canada (7-8 Jun 2012)
6. Och, F.J., Ney, H.: Improved Statistical Alignment Models. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. pp. 440–447. ACL '00, Association for Computational Linguistics, USA (2000)
7. Pazienza, M.T., Pennacchiotti, M.: Textual Entailment as Syntactic Graph Distance: a Rule Based and a SVM Based Approach. In: In Proceedings PASCAL RTE challenge. pp. 528–535 (2005)
8. Turchi, M., Negri, M.: ALTN: Word alignment features for cross-lingual textual entailment. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 128–132. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013)
9. Vilariño, D., Pinto, D., León, S., Alemán, Y., Gómez, H.: BUAP: N-gram based Feature Evaluation for the Cross-Lingual Textual Entailment Task. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 124–127. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013)
10. Vilariño, D., Pinto, D., Tovar, M., León, S., Castillo, E.: BUAP: Lexical and semantic similarity for cross-lingual textual entailment. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 706–709. Association for Computational Linguistics, Montréal, Canada (7-8 Jun 2012)
11. Wäschle, K., Fendrich, S.: HDU: Cross-lingual Textual Entailment with SMT Features. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 467–471. Association for Computational Linguistics, Montréal, Canada (7-8 Jun 2012)