Extracción automática de hechos en libros de texto basada en estructuras sintácticas

Honorato Aguilar-Galicia[1], Grigori Sidorov[1] y Yulia Ledeneva[2]

[1] Centro de Investigación en Computación, Instituto Politécnico Nacional, México
aguilargh@hotmail.com, sidorov@cic.ipn.mx

[2] Universidad Autónoma del Estado de México, México
yledeneva@yahoo.com

Resumen Las oraciones se conforman de fragmentos de texto que pueden separarse y poseer independencia semántica, pero normalmente se encuentran fusionados en la oración, enunciando algo de una forma más amplia. Las oraciones de forma conjunta le dan sentido a un tema, o lo que llamamos información. A estos fragmentos de texto que obtenemos de relacionar entidades a través de un solo verbo, en nuestra investigación les llamamos "*hechos*". Por ejemplo en la oración "*Los mesopotámicos nos legaron la rueda y la escritura*", podemos identificar los siguientes: "*Los mesopotámicos legaron la rueda*" y "*Los mesopotámicos legaron la escritura*". En estos hechos las entidades son "*mesopotámicos*", "*rueda*" y "*escritura*"; y el verbo es "*legaron*". En el presente artículo exponemos lo qué es un hecho, la importancia de extraerlos, y proponemos un método para extraerlos de manera automática con base al análisis de estructuras sintácticas.

Palabras clave: Recuperación de información, extracción de información, extracción de hechos, estructuras sintácticas.

1. Introducción

Desde tiempos antiguos la humanidad ha generado información la cual ha guardado en forma de lenguaje natural, ya fuera en arcilla o papiros. En la actualidad se continúa generando información la cual se guarda en libros, revistas, periódicos.

Con el invento de las computadoras y su desarrollo intenso, hemos logrado que la información actualmente se guarde en forma de bits, es decir, la información se guarda en formato electrónico. Y con el invento de la red Internet y la Web, la cantidad de información crece casi de forma exponencial, lo cual en ocasiones dificulta su administración, pues con tanta información es difícil encontrar la que se busca.

Áreas como la "Recuperación de información" y la "Extracción de información" nos ayudan a encontrar la información que necesitamos. Dentro de esta última podemos ubicar otra área que tratamos en este artículo: "La extracción de información semántica" o lo que llamamos aquí, de forma resumida, "La extracción de hechos".

Un texto está conformado por párrafos y cada párrafo es un conjunto de oraciones. Estas últimas se conforman de pequeñas unidades, que de forma conjunta le dan sentido a la oración, pero que podrían separarse y poseer independencia semántica, pero

normalmente se encuentran fusionadas en la oración, enunciando algo de una forma más amplia.

A estas pequeñas unidades las llamamos "hechos". La Real Academia Española define *Hecho* como: "*m. Acción u obra, m. Cosa que sucede, m. Asunto o materia de que se trata*" [RAE22] y [Joo07] nos dice que existen dos tipos de hechos: los simples y los complejos. Por ejemplo dos hechos simples pueden ser: Pedro y 1989, también llamados entidades. Y un hecho complejo es la relación que se establece entre estas entidades por medio de un verbo, entonces tenemos: *Pedro nació en 1989*. En [Her11] encontramos que "*se considera como hecho a una porción de texto, generalmente más pequeña que una oración, la cual tiene independencia semántica y que contiene un solo verbo asociado a entidades*".

Veamos un ejemplo, en la oración "*Benito Juárez nació en San Pablo Guelatao, Oaxaca, en 1806*", podemos identificar dos hechos:

- *Benito Juárez nació en San Pablo Guelatao, Oaxaca.*
- *Benito Juárez nació en 1806.*

Observamos claramente que son fragmentos de la oración; cada uno nos enuncia algo de forma independiente, o sea, ninguno de ellos necesita al otro para transmitir su información; en cada hecho notamos a un solo verbo, un sujeto y un complemento que juntos forman a una pequeña unidad con sentido completo.

2. La utilidad de los hechos

Contar con una base de datos de hechos es muy importante para otras áreas del procesamiento de lenguaje natural [Mon01], ya que estos pueden ser consultados para realizar sus tareas. Áreas como sistemas de pregunta-respuesta y creación de resúmenes automáticos aprovechan los hechos ya almacenados.

- Sistemas de Pregunta-Respuesta

  IMB ha estado construyendo Watson, un proyecto llamado DeepQA [Watson], es un sistema de preguntas y respuestas. Watson la computadora se enfrenta con gran éxito a humanos en el famoso juego de TV en EEUU llamado Jeopardy. Para la adquisición de conocimientos Watson visita bases de datos, taxonomías, ontologías; y los datos recolectados los almacena en sus bases de datos nombradas "fuentes de respuestas y evidencias" [Dav10]. Así que nuestro propósito de extraer hechos es para almacenarlos en una base de datos que posteriormente puedan ser empleados por un sistema como Watson.

- Resúmenes Automáticos

  Una base de datos de hechos puede ocuparse para crear resúmenes automáticos abstractivos principalmente. Por ejemplo, si en la base de datos tenemos los siguientes hechos "Gato es una mascota", "Canario es una mascota", "Pedro compró un gato", y "María compró un canario". Los algoritmos del método que está creando el resumen abstractivo, pueden aprovechar estos hechos para inferir, generalizar o parafrasear conocimiento; obteniendo como resultado: "Pedro y María compra-

ron una mascota cada uno". La extracción de hechos de un texto también crea un resumen extractivo de ese texto, pero es un resumen muy sencillo ya que en él puede haber redundancia de información.

3. Estado del arte

3.1. Extracción de hechos con intervención de usuario y entrenamiento

En [Joo07] la extracción de hechos lo hace marcando dentro de un texto lo que llama entidades o hechos simples (ejemplo: Williams y 1980) y luego en una ontología se establece una relación entre estos hechos, a lo que llaman hecho complejo o "hecho" (Williams was born on 1980). Para esto ocupan un software de anotación y una ontología.

3.2. Extracción y fusión de hechos de documentos múltiples

En este trabajo se extrae hechos realizando un cruce y fusión de información de diferentes documentos [Mon08], aprovechando la redundancia de datos. Se trabaja con software de anotación y ontologías [Man06].

Los dos trabajos anteriores trabajan con textos en inglés. Nuestra investigación es con textos en español, además el enfoque es con base al análisis de estructuras sintácticas, y los hechos se guardan en una base de datos relacional. No ocuparemos software de anotación, ni una ontología para establecer relaciones que identifiquen los hechos.

3.3. Un sistema de extracción automática de información semántica de los libros de textos estructurados

En [Her11] el trabajo de tesis extrae hechos basándose en análisis de estructuras sintácticas. El analizador sintáctico que se ocupa es "Connexor", un software que tiene un costo para adquirirlo, además de que es un software que tiene un desempeño poco regular en análisis de oraciones en español.

Aquí se extrajeron hechos de 50 oraciones, ya que Connexor en ocasiones arrojaba árboles mal formados. Los cuales se utilizaron para aplicar heurísticas en la búsqueda de hechos. Los hechos extraídos se almacenaron en una base de datos relacional.

En nuestra investigación utilizamos FreeLing-3.0 como analizador sintáctico, un software que no tiene ningún costo por usarlo [FrLi30]. Además fue creado especialmente para analizar las lenguas de Europa, incluyendo el idioma Español.

3.4 Un esquema de evaluación semiautomática

En muchas tareas de procesamiento de lenguaje natural, nos encontramos con el problema de determinar el nivel de granularidad adecuado para las unidades de información. Por lo general, usamos frases para modelar piezas individuales de información. Sin embargo, más aplicaciones de Procesamiento de Lenguaje Natural (PLN)

nos obligan a definir las unidades de texto más pequeñas que frases, esencialmente descomponiendo sentencias dentro de una colección de frases.

Cada frase lleva una pieza independiente de información que puede ser utilizado como una unidad independiente. Estas unidades de información de mayor granularidad son usualmente llamadas como "nuggets" [Lia07].

Trabajos anteriores muestran que las personas pueden crear nuggets de una manera relativamente sencilla. Sin embargo, la creación automática de los nuggets no es trivial.

Con base a un análisis manual y modelado computacional de nuggets, se definen como sigue:

- Un nugget es un evento o una entidad.
- Cada nugget consiste en dos partes: el ancla y el contenido.

El ancla es:

- El sustantivo principal de la entidad.
- El verbo principal del evento, más el sustantivo núcleo de su entidad asociada (si más de una entidad está unida al verbo, entonces su sujeto).

El contenido es una pieza única y coherente de la información asociada con el ancla. Cada anclaje debe tener varios contenidos distintos.

Cuando un nugget contiene sentencias anidadas, esta definición es aplicada recursivamente.

4. Método propuesto

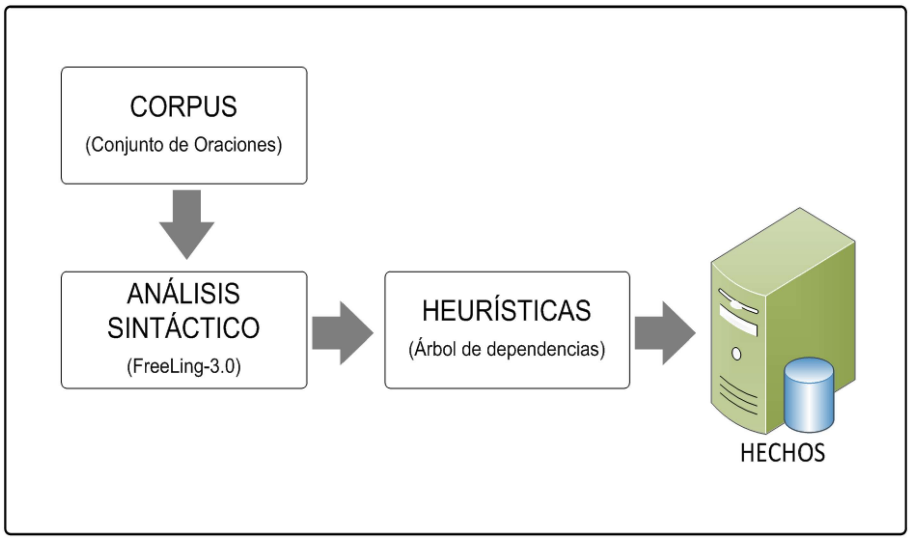

Fig. 1. Método propuesto.

La Figura 1 muestra las partes que conforman al método propuesto y como se relacionan.

4.1. Corpus

Se utilizó un corpus compuesto por 55 oraciones en español, este conjunto de oraciones fueron tomadas de libros de texto [Her11] y [SEP-H6-10], pues al ser creados para fines educativos contienen gran cantidad de conceptos, y por lo tanto contienen una gran cantidad de hechos. Además, estos libros tratan sobre temas muy específicos.

Los libros de texto, por su naturaleza, están estructurados por temas y capítulos. En donde cada capítulo está compuesto por un título, un conjunto de párrafos y cada párrafo por un conjunto de oraciones. A esta composición le llamamos un texto estructurado. Y fue esta característica la razón principal de seleccionar los libros de texto, pues las oraciones de un capítulo pueden separarse y ser procesadas por el método propuesto como una unidad.

Estas 55 oraciones fueron seleccionadas porque contiene una secuencia gramatical aceptable para el método, esto es, contienen sujeto y predicado. Oraciones formuladas como interjección o pregunta no fueron elegidas.

4.2 Análisis sintáctico

Estructuras sintácticas.

La extracción automática de información semántica que realizamos en nuestra investigación, se basa en el análisis de estructuras sintácticas, específicamente en el análisis de lo que se conoce en gramática como oración, el objeto de análisis en nuestro método.

Una "estructura", en el diccionario de la lengua española de la Real Academia Española [RAE22] encontramos que "estructura" es "1. f. Distribución y orden de las partes importantes de un edificio", y "2. f. Distribución de las partes del cuerpo o de otra cosa". De acuerdo a estas definiciones podemos deducir que una "estructura" para nuestro objeto de análisis sería: "la distribución y orden de las palabras en una oración" y considerando la sintaxis, agregamos que "las palabras son funciones unas de otras", como dice [Mun00] "las palabras adquieren un significado preciso y cumplen una función sintáctica determinada":

- Se lastimó la muñeca izquierda mientras jugaba a la pelota.
- La muñeca que le regalé a mi hija cierra los ojos.

Aisladamente, la palabra *muñeca* tiene varias acepciones, pero en cada oración toma una de ellas; además, esta misma palabra cumple una función distinta, en la primera oración es objeto directo y en la segunda, es sujeto [Rio09].

Dentro de las distintas partes de la gramática, la sintaxis es la que se dedica al estudio de la oración. Su estudio se basa en las diferentes funciones que desempeñan los componentes de la oración. [Fue10]

[Mun00] nos dice que "la sintaxis es la parte de la gramática que estudia la manera como se combinan y ordenan las palabras para formar oraciones; analiza las funciones que aquéllas desempeñan, así como los fenómenos de concordancia que pueden presentarse entre sí".

Herramienta FreeLing.

Para cada una de las oraciones realizamos un análisis sintáctico de forma automática, para ello utilizamos FreeLing-3.0, configurado con la opción "Dependency Parsing", la cual realiza un análisis morfosintáctico y nos proporciona un árbol de dependencias.

Cada nodo del árbol de dependencias representa una palabra de la oración, que contiene información sintáctica y morfológica de cada una de ellas, organizada de forma jerárquica.

Para el etiquetado morfosintáctico, FreeLing-3.0 utiliza un conjunto de etiquetas propuesto por el grupo [EAGLES]. Sirve para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas, y por lo tanto para el español [Eagv20].

Características de FreeLing-3.0.

Es una suite de código abierto que sirve para realizar análisis morfosintáctico de texto, publicado bajo la licencia GNU General Public License de la Free software Foundation [FrLi30].

Algunas características:

- Tokenización de texto.
- Análisis morfológico.
- Reconocimiento de fechas, números, monedas y medidas físicas (velocidad, peso, temperatura, densidad, etc.)
- PoS tagging.
- Análisis de dependencia basado en reglas.
- Resolución de referencia nominal.

4.3. Heurísticas

El método propuesto utiliza heurísticas para identificar los hechos. Estas trabajan sobre los árboles de dependencia que nos proporciona [FrLi30], en donde cada nodo cuenta con las siguientes etiquetas de la palabra que contiene. La descripción de las etiquetas es parte de la documentación de [FrLi30].

- Func: Indica la función que tiene la palabra, por ejemplo sujeto (subj), dobj (objeto directo).
- Synt: Indica la relación sintáctica, por ejemplo sintagma nominal (sn), grupo preposicional (grup-sp).
- Form: La forma real de la palabra, tal como está escrita en la oración.
- Lemma: Lemma de la palabra.
- Tag: Etiqueta de acuerdo a las etiquetas EAGLES [Eagv20].

A continuación describimos las heurísticas empleadas.

Heurística 1: Es la heurística principal, se aplica en el árbol en busca de la siguiente secuencia gramatical: Sujeto – Verbo – Complemento.

1. Ubicar verbo principal en el árbol o subárbol.

2. Buscar y extraer al sujeto. Ubicaremos al sujeto en los nodos dependientes del verbo obtenido en el paso 1, con la siguiente etiqueta: {func = subj} y {synt = sn}. O {func = subj-pac} y {synt = sn}
3. Buscar y extraer al complemento. Estará compuesto por los nodos hermanos siguientes del sujeto, es decir, se toma los nodos posteriores al nodo sujeto.
4. Construir hecho. El hecho estará formado por: (Sujeto extraído en el paso 2) + (Verbo extraído en el paso 1) + (Complemento extraído en el paso 3).

Heurística 2: Aplicada cuando la estructura contiene conjunción copulativa o disyuntiva de sustantivos.

1. Aplicar la heurística 1, para obtener el sujeto y el verbo.
2. Ubicar los sustantivos coordinados. Con las etiquetas {func = co-n} y {synt = sn},
3. Construir hechos. Se construyen varios hechos: (Sujeto obtenidos en el paso 1) + (Verbo obtenido en el paso 1) + (Complemento *1*, *2*, *n* obtenido en el paso 2).

Heurística 3: Aplicada cuando la estructura contiene conjunción copulativa o disyuntiva de verbos.

1. Ubicar los verbos coordinados. Con las etiquetas {func = co-v} y {synt = grup-verb}.
2. Aplicar heurística 1 para obtener sujeto y complemento.
3. Construir hechos. Se construye un hecho por cada verbo encontrado en el paso 1: (Sujeto obtenido en paso 2) + (Verbo *1*, *2*, *n* obtenido en paso 1) + (Complemento obtenido en paso 2).

Heurística 4: Aplicada cuando la estructura contiene un pronombre relativo.

1. Ubicar pronombre relativo. Con las etiquetas {func = subord-mod}, {synt = subord-rel} y {tag = PRxxxxxx}.
2. El pronombre relativo depende de un nodo sintagma nominal etiquetado con {synt = sn} y {tag = Nxxxxxx}, este será el sujeto para el hecho.
3. El pronombre relativo tiene como descendiente a un verbo. Este será el verbo para el hecho.
4. El complemento lo conformarán los elementos dependientes del verbo del paso 3.
5. Construir hecho con (Sujeto obtenido en paso 2) + (Verbo obtenido en paso 3) + (Complemento obtenido en paso 4).

Heurística 5: Aplicada cuando en la estructura encontramos las preposiciones *desde* y *hasta*.

1. Aplicar la heurística 1 para extraer el sujeto y verbo.
2. Ubicamos la preposición *desde* y la guardamos junto con sus nodos dependientes. Será el complemento de un hecho. El nodo tiene las siguientes etiquetas: {func = cc}, {synt = grup-sp} y {tag = SPS00}.
3. Construimos un hecho: (Sujeto obtenido en paso 1) + (Verbo obtenido en paso 1) + (Complemento obtenido en paso 2).
4. Ubicamos la preposición *desde* y la guardamos junto con sus nodos dependientes. Será el complemento de otro hecho. El nodo tiene las siguientes etiquetas: {func = ador}, {synt = grup-sp} y {tag = SPS00}.

5. Construimos un hecho: (Sujeto obtenido en paso 1) + (Verbo obtenido en paso 1) + (Complemento obtenido en paso 4).

Heurística 6: Aplicada cuando en la estructura encontramos las preposición *en*.

1. Aplicar la heurística 1 para extraer el sujeto y verbo.
2. Ubicamos la preposición *en* y la guardamos junto con sus nodos dependientes. Será el complemento de un hecho. El nodo tiene las siguientes etiquetas: {func = sp-obj}, {synt = grup-sp} y {tag = SPS00}.
3. Construimos un hecho: (Sujeto obtenido en paso 1) + (Verbo obtenido en paso 1) + (Complemento obtenido en paso 2).
4. Ubicamos la preposición *en* y la guardamos junto con sus nodos dependientes. Será el complemento de otro hecho. El nodo tiene las siguientes etiquetas: {func = sp-obj}, {synt = grup-sp} y {tag = SPS00}.
5. Construimos un hecho: (Sujeto obtenido en paso 1) + (Verbo obtenido en paso 1) + (Complemento obtenido en paso 4).

4.4 Base de datos de hechos

Los hechos obtenidos son almacenados en una base de datos relacional. Esta cuenta con tres tablas: Tabla de entidades, Tabla de verbos y Tabla de relaciones. En la Tabla de entidades guardamos las entidades (Sustantivos, Adjetivos, Preposiciones) de la oración, en la Tabla de verbos guardamos los verbos, y en la tercera tabla guardamos las relaciones entre las entidades y verbos, es decir, lo que forma un hecho.

5. Resultados

Para probar el método propuesto primero se proporcionó a una persona una guía para identificar hechos en una oración y se le encargó que identificara y extrajera los hechos de cada una de las oraciones de nuestro corpus de análisis. Esta persona extrajo un total de 119 hechos. Posteriormente se aplicó el método propuesto a cada una de las oraciones y comparamos ambos resultados.

5.1 Guía para un humano para extraer hechos

A continuación describimos el algoritmo, para que una persona pueda identificar y extraer hechos desde una oración. La estructura de un hecho es: Sujeto – Verbo – Complemento. El algoritmo nos ayuda a identificar los tres elementos y nos indica la estructura del hecho.

1. Localizar el sujeto en la oración. Para hacerlo contestamos a una de las siguientes preguntas: ¿De qué o de quién se habla? o ¿Quién o qué realiza la acción?

2. Localizar el predicado en la oración. Para hacerlo contestamos a una de las siguientes preguntas: ¿Qué se dice, de quien se habla o de lo que se habla? o ¿Qué se dice, de quien o lo que realiza la acción

3. Buscar hechos. La estructura del hecho es Sujeto - Verbo – Complemento, así que se construye el hecho de acuerdo a ella. Identifiquemos las siguientes características en la oración, de forma específica en el predicado:

(a) PRIMER Hecho. Extraemos el verbo principal, normalmente se encuentra al inicio del predicado. Con el sujeto obtenido en el paso 1, este verbo y el resto del predicado como complemento se crea el primer hecho. Revisamos los siguientes incisos para verificar si la oración en proceso contiene estas características.

(b) Revisar perífrasis. Generalmente la perífrasi tiene la forma: (verbo auxiliar) + (preposición o conjunción) + (infinitivo, gerundio o participio). Cuando la oración contenga esta característica, se tomará el último elemento como el verbo para el hecho.

(c) CONJUNCION.Copulativa. Conjunciones copulativas son las "que coordinan dos o más palabras las cuales desempeñan una misma función. También pueden unir oraciones. Las conjunciones copulativas son *y*, *e*, *ni*" [Mun00]. Ejemplo: "El domingo compré discos de música hindú, turca *y* rusa". Si la oración contiene conjunción copulativa, obtendremos un hecho por cada término coordinado por la conjunción. Para el ejemplo anterior los hechos son: "El domingo compré discos de música hindú", "El domingo compré discos de música turca", "El domingo compré discos de música rusa".

(d) CONJUNCION.Disyuntiva. Las conjunciones disyuntivas "son conjunciones que enlazan palabras u oraciones para expresar posibilidades alternativas, distintas o contradictorias. Las conjunciones disyuntivas son *o*, *u*" [Mun00]. Ejemplo: "Pedro se hospedará en una pensión u hotel cualquiera". Cuando en la oración se tenga conjunción disyuntiva, obtendremos un hecho por cada término coordinado por la conjunción. Para el ejemplo anterior los hechos son: "Pedro se hospedará en una pensión", "Pedro se hospedará en un hotel cualquiera".

(e) PRONOMBRE.Relativo. "Los pronombres relativos hacen referencia a alguien o a algo que se ha mencionado antes en el discurso o que ya es conocido por los interlocutores. Los pronombres relativos, funcionan, en la mayor parte de los casos, como elementos de subordinación de oraciones. Los pronombres relativos son *que*, *quien*, *quienes*, *cual*, *cuales*, *cuanto*, *cuantos*, *cuanta*, *cuantas*." [Mun00]. Ejemplo: "Pedro conoció a un estudiante que sabe hablar chino". Como el pronombre relativo hace referencia a alguien o a algo que se ha mencionado antes, entonces buscaremos al sujeto (sustantivo, pronombre personal) para un siguiente hecho, en la parte inmediata que antecede al pronombre relativo. El verbo de este hecho se encuentra localizado después del pronombre relativo. Los hechos del ejemplo, son: "Pedro conoció a un estudiante", "Estudiante sabe hablar Chino".

(f) PREPOSICIÓN: *desde* y *hasta*. "*Desde*. Denota inicio de una acción en el tiempo o en el espacio. *Hasta*. Expresa el fin de algo o límite de lugar, de número o de tiempo" [Mun00]. Ejemplo: "La primavera comprende desde el mes de marzo hasta el mes de junio". Si una oración contiene estas preposiciones, se

formará un hecho por cada una de ellas. Para el ejemplo tenemos los siguientes hechos: "La primavera comprende desde el mes de marzo", "La primavera comprende hasta el mes de junio".

(g) PREPOSICIÓN: *en*. La preposición en "indica tiempo, expresa lugar, señala modo, significa ocupación o actividad, indica medio o instrumento, forma locuciones adverbiales" [Mun00]. Ejemplo: "Pedro estudia en la mañana y en la noche". Cuando una oración que contenga una o más de una preposición en, se formará un hecho por cada una de ellas. Los hechos del ejemplo son: "Pedro estudia en la mañana", "Pedro estudia en la noche".

4. Fin.

5.2 Evaluación del método propuesto

Abajo presentamos unos ejemplos de oraciones analizadas.

Algunas oraciones que fueron analizadas con nuestro método. Las dos primeras fueron analizadas con éxito, pues cumplen una cierta estructura sintáctica en la cual se basa el método. En la tercera y cuarta nuestro método presenta dificultades.

Oración 1: "*La numeración arábiga procede de India*". De aquí se obtuvo el hecho: "*Numeración arábiga procede de India*", se tuvo éxito porque se identifica claramente el sujeto, el predicado incluyendo al verbo.

Oración 2: "*La civilización China nos heredó el papel, la pólvora, una forma de imprenta rudimentaria, y la brújula*". Sus hechos: "*Civilización china heredó el papel*", "Civilización china *heredó pólvora*", "*Civilización china heredó forma de imprenta rudimentaria*", "*Civilización china heredó brújula*". Aquí se obtiene el mismo sujeto y verbo para todos los hechos, y el complemento de los hechos son términos coordinados por una conjunción.

Ahora veamos las oraciones en donde nuestro método tiene complicaciones.

Oración 3: "*Roma no imponía ideas políticas o credos en sus territorios*". El método propuesto no incluye en los hechos la palabra "*no*", entonces al no contemplarse nos da como resultados hechos con un significado contrario a lo que expresa la oración: "*Roma imponía ideas políticas*", "*Roma imponía credos*".

Oración 4: "*El mamut era un animal de gran tamaño al que se cazaba mediante diversas técnicas*". Los hechos obtenidos por la persona son: "*El mamut era un animal*", "*El mamut era de gran tamaño*", "*El mamut se cazaba mediante diversas técnicas*". El método tiene dificultad para manejar la preposición "*de*" pues no la toma como otro hecho, los hechos que obtiene son: "*El mamut era un animal de gran tamaño*", "*El mamut se cazaba mediante diversas técnicas*", o sea dos.

Usamos las siguientes medidas de evaluación**.**

En la evaluación del método con el corpus, este ha obtenido un total de 116 hechos, de los cuales 82 fueron correctos y 34 incorrectos. A continuación las medidas de evaluación.

Precisión del método = Hechos correctos obtenidos por el método / Total de hechos obtenidos por el método (1)

Recall = Hechos correctos obtenidos por el método / Total de hechos existentes en el texto (extraídos por un humano) (2)

F1 = (2 * Precisión del método * Recall) / (Precisión del método + Recall) (3)

Los resultados que hemos obtenido hasta el momento son los siguientes:

Tabla 1. Resultados de la evaluación.

Característica	Resultado
Precisión del método	70.6%
Recall	68.9%
F1	69.7%

6. Conclusiones

Es importante decir que nuestra investigación está en proceso, así que los resultados hasta el momento no son definitivos pero son buenos para continuar en esta línea. Hemos identificado ciertos patrones en la estructura sintáctica de las oraciones cuando nos enuncian algo. Una unidad semántica sólo expresa información cuando nos dice algo de alguien. Estos patrones los hemos plasmado en heurísticas en el método, ayudándonos a obtener información semántica desde las oraciones y por ende de los textos.

Es importante mencionar que nuestra investigación también comprende otras áreas de estudio, como es la generación automática de resúmenes, pues nos encontramos que al extraer las unidades mínimas de información en una oración, realmente estamos extrayendo las unidades que informan algo en un texto. Si unimos todas estas unidades informativas o lo que nosotros llamamos hechos, obtenemos un resumen del texto de donde se hayan obtenido las oraciones.

Finalmente, también es importante mencionar que los resultados obtenidos hasta el momento son con base al análisis de nuestro corpus actual, corpus que acrecentaremos para identificar más patrones, y de esta forma mejorar nuestros resultados y nuestras heurísticas.

Agradecimientos. Trabajo realizado conel apoyo parcial del gobierno de México (proyectos CONACYT 50206-H y 83270, SNI) e Instituto Politécnico Nacional, México (proyectos SIP 20111146, 20113295, 20120418, COFAA, PIFI), Gobierno del DF (ICYT-DF proyecto PICCO10-120) y la Comisión Europea (proyecto 269180).

Referencias

1. [Dav10] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project. Association for the Advancement of Artificial Intelligence. ISSN 0738-4602 (2010)
2. [EAGLES] Expert Advisory Group on Language Engineering Standards <http://www.ilc.cnr.it/EAGLES96/home.html> [Consulta: 30/04/2012]
3. [Eagv20] Introducción a las etiquetas eagles (v. 2.0). <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html> [Consulta: 30/04/2012]
4. [FrLi30] FreeLing 2.2. An Open Source Suite of Language Analyzers. <http://nlp.lsi.upc.edu/freeling/index.php> [Consulta: 30/04/2012]
5. [Fue10] Fuentes de la Corte, Juan Luis. Gramática Moderna de la legua española. Limusa, México (2010)
6. [Gel09] Gelbukh, A. & Sidorov, G. Procesamiento automático del español, con enfoque en recursos léxicos grandes. Instituto Politécnico Nacional. México (2009)
7. [Her11] Herrera de la Cruz, Juve Andrea. Sistema de extracción automática de información semántica de los libros de textos estructurados. Tesis de maestría, CIC-IPN, México (2011)
8. [Hov06] Hovy, Eduard. Learning by Reading: An Experiment in Text Analysis. Springer-Verlag Berlin Heidelberg (2006)
9. [Joo07] Joosse, W. User Trainable Fact Extraction. Masters Thesis, Universiteit Twente de ondernemende universiteit (2007)
10. [Lia07] Liang Zhou, Namhee Kwon, and Eduard Hovy. A Semi-Automatic Evaluation Scheme: Automated Nuggetization for Manual Annotation. Proceedings of NAACL HLT 2007, Companion Volume, pages 217–220, Rochester, NY. Association for Computational Linguistics (2007)
11. [Man06] Mann, Gideon S. Multi-Document Statistical Fact Extraction and Fusion. Doctor Thesis, The Johns Hopkins University, Baltimore, Maryland (2006)
12. [Mor04] De la Mora Alejandro. Las partes de la oración. Trillas. México (2004)
13. [Mun00] Munguía Zatarain Irma, Munguía Zatarain Martha Elena, Rocha Romero Gilda. Gramática Lengua Española. Reglas y ejercicios. Ediciones Larousse, S.A. de C.V. México, D. F. (2000)
14. [SEP-H6-10] Secretaría de Educación Pública. Historia Sexto grado. ISBN: 978-607-469-414-7. México, D.F. (2010)
15. [Watson] Watson Research Center. IBM, <www.watson.ibm.com/index.shtml> [Consulta: 30/04/2012]
16. [Mon01] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López López. Mining the news: trends, associations, and deviations. Computación y Sistemas, Vol. 5 N 1, pp. 14-24 (2001)
17. [Rio09] Miguel Ángel Ríos Gaona, Alexander Gelbukh, Sivaji Bandyopadhyay: Web-based Variant of the Lesk Approach to Word Sense Disambiguation. In: MICAI 2009. Proc. of 2009 Eighth Mexican International Conference on Artificial Intelligence, IEEE CS Press, pp. 103–107 (2009)
18. [Mon08] Alfredo Monroy, Hiram Calvo, Alexander Gelbukh: Using Graphs for Shallow Question Answering on Legal Documents. In: MICAI 2008. Lecture Notes in Artificial Intelligence N 5317, Springer, 165-173 (2008)