El rol de las predicaciones verbales en la extracción automática de conceptos

Rodrigo Alarcón
Gerardo Sierra
Instituto de Ingeniería, UNAM

In specialised texts, there are some recurrent typographical and syntactic patterns that authors use to define a term. Some of these syntactic patterns are verbal predications that function as connectors between terms and definitions. In this paper we analyse the role of the verbal predications involved in definitional contexts of specialised texts in Spanish, in order to develop a system capable of extracting definitional contexts.

Se ha observado que cuando un autor define un término en un texto especializado, utiliza una serie de patrones léxicos que ayudan a resaltar visual y gramaticalmente la presencia del término que define. Uno de los elementos más comunes en estos patrones son los verbos que conectan al término con su definición y que aquí definiremos como predicaciones verbales. En este artículo se analizan estas predicaciones con el fin de establecer reglas y restricciones que nos permitan desarrollar un sistema de recuperación automática de contextos definitorios.

Palabras clave: *contexto definitorio, terminótica, Precision & Recall, lingüística computacional, ingeniería lingüística.*
Fecha de recepción del manuscrito: marzo del 2003

Rodrigo Alarcón y Gerardo Sierra
Grupo de Ingeniería Lingüística, Instituto de Ingeniería, UNAM
Torre de Ingeniería, Circuito Interior, 04510 México D. F.
correos electrónicos: ralarconm@iingen.unam.mx, gsierram@iingen.unam.mx.

1. Introducción

La terminografía es la práctica de elaborar diccionarios de términos especializados, esto es, unidades léxicas que pertenecen a un área específica de conocimiento. Esta práctica puede realizarse en tres etapas, las cuales corresponden a la identificación de los términos característicos del dominio en cuestión, al análisis conceptual de los términos y al análisis del uso de los términos en su contexto. (Meyer, 2001)

Para llevar a cabo este trabajo se puede consultar a especialistas del área, directamente, o bien pueden consultarse textos correspondientes al dominio que se estudia. En este sentido, uno de los esfuerzos de la terminología computacional o terminiótica, radica en desarrollar herramientas que faciliten el análisis de textos con fines terminográficos.

En el Grupo de Ingeniería Lingüística, de la UNAM, se ha desarrollado una investigación con el fin de elaborar una herramienta para la identificación automática de los posibles conceptos de un texto especializado, esto es, los términos y sus definiciones.

Esta investigación se basa en esfuerzos previos como la identificación sistemática de definiciones a partir de patrones léxicos y metalingüísticos (Pearson, 1998), el análisis de *Operaciones Metalingüísticas Explícitas* (Rodríguez, 1999) y el análisis de *Contextos ricos en conocimiento* (Meyer, 2001).

Con estas investigaciones se ha podido determinar que en los textos especializados se utiliza una variedad de patrones que permiten identificar la presencia de un posible contexto definitorio. En esta investigación, por **contexto definitorio** entenderemos todo aquel fragmento textual donde se aporta la información necesaria para definir a un término.

2. Objetivos

El objetivo principal de este artículo es presentar un análisis de las predicaciones verbales encontradas en un corpus de documentos especializados en el área de ingeniería. Dichas predicaciones se encontraron al clasificar los distintos patrones léxicos en el corpus.

Ahora bien, el estudio y clasificación de las predicaciones verbales es el primer paso para desarrollar un sistema computacional basado en reglas y restricciones, que sea capaz de identificar los posibles contextos definitorios de un documento especializado.

Primero se expondrán los patrones recurrentes encontrados en dichos contextos. En seguida se presentará una metodología para expandir el paradigma inicial de predicaciones verbales. Finalmente se presentará una evaluación de los verbos encontrados en las predicaciones.

3. Patrones recurrentes en contextos definitorios

Las investigaciones previas de Meyer [2001] y Rodríguez [1999] nos permitieron delimitar qué elementos tendría que poseer como mínimo un fragmento textual para poder ser considerado como un contexto definitorio.

Se considera que la definición puede aportar información de varios tipos. Puede presentar la fórmula de una definición aristotélica: *Definición = Género próximo + Diferencia específica,* que en la fórmula de Meyer está dada como *X = Y + características distintivas*; puede aportar información que permita clasificar al término dentro de una clase general, esto es, su hiperónimo o merónimo; o bien puede proporcionar información acerca de la función del término.

Por otro lado, en los contextos definitorios es común encontrar elementos estilísticos y sintácticos empleados por lo autores. Estos elementos sirven para resaltar los constituyentes de los contextos definitorios. Nosotros definimos que pueden ser marcas tipográficas o bien predicaciones pragmáticas o predicaciones verbales. En algunos casos, las marcas tipográficas, al igual que las predicaciones verbales, funcionan como enlace entre el término y la definición.

3.1. El corpus de análisis

Se utilizó un corpus de 25 textos especializados en las áreas de logística, transporte, sistemas expertos y estructuras bioclimáticas, pertenecientes a la ingeniería. Los textos fueron provistos por académicos y estudiantes del Instituto de Ingeniería de la UNAM, y se compone esencialmente por tesis, informes a patrocinadores y artículos en congresos.

En estos documentos, por su naturaleza, se incluye un apartado (introducción, presentación o bien un capítulo específico) que funciona como marco conceptual y en el cual se definen los conceptos esenciales para la comprensión del texto. Esta fue una de las características principales que se consideraron como criterios básicos de selección, ya que nos permitió tener una mayor seguridad de encontrar no sólo términos especializados, sino también definiciones.

3.2. Metodología

Se determinó que los contextos definitorios pueden representarse mediante secuencias denominadas *patrones*. Estos patrones se clasificaron en cuatro formas distintas, que van de simples a complejas: *patrones tipográficos*, *patrones sintácticos*, *patrones mixtos* y *patrones compuestos*.

Para representar cada secuencia se utilizaron los siguientes símbolos: **T** (término)**,** **D** (definición), **mt** (marca tipográfica), **PV** (predicación verbal) y **PP** (predicación pragmática). **T** y **D** son los elementos mínimos de un contexto definitorio y se unen mediante una **PV** o una **mt**. A su vez, las **mt** pueden ser características de **T** y **D**. Para representar la unión de estos elementos en secuencias se utilizó el signo "+", en tanto la combinación de dos elementos se representa contiguamente. En la siguiente tabla se muestra un ejemplo.

Tabla 1. Ejemplo de patrones

Patrón	Contexto definitorio
T mt + PP + PV + D	**La energía primaria**, por definición, es aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción. En este caso se encuentran el petróleo crudo, el gas asociado...

3.3. Patrones tipográficos

La tipografía de un texto sirve al lector como ayuda visual para identificar fácilmente algún elemento importante y diferenciarlo del resto del texto común, por ejemplo los términos y sus definiciones.

En algunos casos se define un término sin la necesidad de emplear algún verbo que funcione como conector. Sintácticamente, los verbos son sustituidos por signos de puntuación tales como *dos puntos*, *punto y seguido* o *coma*.

De esta forma, el primer grupo de patrones que se consideró es aquél donde se utiliza alguna marca tipográfica para resaltar la presencia del término y/o la definición, y donde se sustituye al verbo que une a los elementos constitutivos por algún signo de puntuación. A este grupo le denominamos **patrones tipográficos** y constituyen las formas más simples encontradas, ya que se asemejan a un tipo de definición que se presenta en un diccionario.

Tabla 2. Ejemplos de patrones tipográficos

Patrón	Contexto definitorio
T mt + mt + D mt	**Diseño:** *Desarrollo de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones.*
T mt + mt + D mt	DESASTRE. *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves*.
T mt + mt + D	"Impactos agregados sociales" ¶ Los que impactan a la sociedad, produciendo, por ejemplo, la perturbación de las relaciones familiares

En los dos primeros contextos, el término aparece en negritas o en mayúsculas, y la definición en cursivas. En estos dos patrones se observa que el verbo es sustituido por dos puntos o punto y seguido. En el tercer ejemplo se presenta el término en comillas y la definición aparece después de un salto de párrafo, representado mediante el símbolo ¶. Para nuestra investigación, el salto de párrafo se consideró como un símbolo tipográfico importante ya que nos permitió definir aquellas estructuras donde el término se presenta en un título o una viñeta, y la definición en el párrafo siguiente.

3.4. Patrones sintácticos

Analizamos los contextos que no presentan ninguna marca tipográfica pero que utilizan una forma verbal o una forma pragmática para identificar que dicho fragmento corresponde a un contexto definitorio, y definimos a estas secuencias como **patrones sintácticos**.

Se identificó que en estos patrones suelen combinarse las predicaciones pragmáticas con las predicaciones verbales, dando como resultado, hasta el momento, las siguientes posibilidades de combinación:

(PP/PV) + T/D + (PP/PV) + D/T + (PP)

Tabla 3. Ejemplos de patrones sintácticos

Patrón	Contexto definitorio
PV + T + D	Se considera como protección civil a la actividad solidaria de los diversos sectores que integran a la sociedad…
T + PP + PV + D	Un soporte logístico de plataforma, de manera general, se define como un territorio equipado para el desarrollo de actividades logísticas…
PP + T + PV + D	De acuerdo con esta conceptualización, los daños probables se definen como el riesgo que corre el SA por ser expuesto al…

Podemos observar que cada contexto carece de marcas tipográficas para resaltar el término o la definición; sin embargo, los elementos que resaltan el carácter definitorio de estos fragmentos son de tipo sintáctico, esto es las predicaciones verbales *se considera como* y *se define como*, así como las predicaciones pragmáticas *de manera general* y *de acuerdo con*.

3.4.1. Predicaciones pragmáticas

Un contexto definitorio puede contener, además de la definición, otro tipo de información relevante para la comprensión del término, la que Rodríguez define como información semántico – pragmática.

Definimos que esta información pragmática, en general, nos permite distinguir que el fragmento textual corresponde a un contexto definitorio. Esta información está en relación con el uso y tratamiento del término, su introducción dentro del texto, y toda aquella información que nos proporciona una base para entender al término dentro del contexto en el cual aparece. Consideramos a esta información como predicaciones pragmáticas.

El estudio profundo de estas predicaciones se tiene contemplado en la siguiente etapa de la investigación, donde se expandirá y evaluará el paradigma correspondiente. Por el momento, cabe señalar que dentro de estas predicaciones se encuentran frases adverbiales (*de manera general*), frases prepositivas (*en términos* generales) y palabras simples (*definición*, *concepto*, *término*) que se identifican como palabras metalingüísticas.

Las predicaciones pragmáticas constituyen otro de los elementos que nos permiten reconocer la presencia de un contexto definitorio dentro del texto. Es de reconocer que estas formas pertenecen a un paradigma estructural amplio, ya que su composición puede variar de acuerdo a formas estructurales o estilísticas utilizadas por cada autor.

Tabla 4. Ejemplos de predicaciones pragmáticas

El término	El nombre de
De manera general	De acuerdo con
Definición	El concepto de
Un aspecto fundamental de	En su acepción más amplia

3.4.2. Predicaciones verbales

Las **predicaciones verbales** sirven para unir directamente al término con su definición. Esta característica funcional es la que las distingue de los demás elementos sintácticos de los contextos definitorios.

Los verbos y formas verbales que se emplean en los contextos corresponden a lo que se ha denominado como verbos metalingüísticos. Esta clasificación aplica comúnmente para verbos como *definir*, *describir*, *denominar*, etc., verbos que por su naturaleza se emplean para referirse al propio lenguaje.

Para nuestros fines, tomamos en cuenta la estructura de las predicaciones verbales y consideramos que se clasifican en dos grupos: las **formas simples** y las **formas compuestas**.

La característica principal de las formas simples es que en ellas existe un sujeto que define o predica *algo* sobre un término, o bien el término funciona como sujeto gramatical del contexto definitorio. Entre estas formas se encontraron las siguientes: *entendemos*, *ocurre*, *afirma que*, etc.

En las formas compuestas suele emplearse el pronombre *se* para construir formas verbales que permiten, de manera impersonal, predicar *algo* sobre un término. Estas formas se representan, generalmente, mediante el pronombre *se* + *verbo conjugado* + *partícula*, donde el orden puede ser aleatorio, y la partícula puede corresponder a una preposición, a una conjunción o a un adverbio. Las formas más comunes que se encontraron fueron: *se define como*, *se concibe como*, *se refiere a*, etc.

El paradigma verbal se expandió tomando en cuenta los criterios estructurales señalados, lo cual nos permitió determinar qué verbos y formas verbales ofrecen una mayor seguridad al momento de extraer automáticamente candidatos a contextos definitorios. La metodología y los resultados se presentan más adelante. La expansión del paradigma tomando en cuenta criterios semánticos se tiene contemplada dentro de la segunda etapa de la investigación.

Tabla 5. Ejemplos de predicaciones verbales

Formas personales	Formas impersonales
Afirma que	Se basa en
Comprende	Se concibe como
Consiste en	Se conoce (como/con)
Consta de	Se considera (como)
Constituye	Se define como
Corresponde a	Se denomina (como)
Define a	Se encarga de
Incluye	Se refiere a
Ocurre	Se utiliza (para/en)

3.5. Patrones mixtos

Hasta ahora hemos explicado los dos elementos que caracterizan a un contexto definitorio y que son la tipografía y la sintaxis. Los dos grupos de patrones mencionados anteriormente utilizan una de estas características por separado. Cuando estas características se mezclan en el contexto definitorio fueron denominados como **patrones mixtos**.

Estos patrones presentan una estructura más sólida, ya que utilizan elementos que permiten resaltar visual y gramaticalmente la presencia de un contexto definitorio.

Se observa que en los dos primeros ejemplos se utiliza una marca tipográfica en el término y además se emplean las predicaciones *ser* y *se definen como* para unir al término con la definición. El último ejemplo presenta al término en cursivas y se utiliza la predicación verbal *se entiende por*. En el tercer ejemplo se utiliza una predicación pragmática, *según + (autor)* y una predicación verbal *se define como*; la definición presenta dos marcas tipográficas: comillas y negritas. Cabe mencionar que aquí se considera el *autor* como un elemento característico del contexto definitorio

Tabla 6. Ejemplos de patrones mixtos

Patrón	Contexto definitorio
T mt + PV + D	**a. Canal de comercialización** es el conjunto de actores y actividades que interactúan para que un bien producido…
T mt + PV + D	- *Las actividades* se definen como los elementos principales de una acción…
PP + T + PV + D mt	Según G. Malagón (1996, p.18) un hospital se define como: "**una parte integrante de la organización médica, cuya función es la de proporcionar a la población**…
PV + T mt + D	Se entiende por *paradigma* una forma epistemológica que, como instrumento cognoscitivo, permite diferenciar la…

3.6. Patrones compuestos

El último grupo de patrones que se consideró fue aquel donde en un mismo contexto definitorio se definen dos o más términos. Se le nombró **patrones compuestos**.

Hasta el momento, se encontró que este grupo puede presentar dos formas distintas. En la primera, un contexto definitorio sirve para definir dos o más términos que por lo general se presentan en un orden que se señala mediante alguna predicación pragmática. La segunda forma es aquella donde la definición de un término sirve como un contexto definitorio para otro término y su correspondiente definición.

Estos patrones son las formas más complejas que se encontraron, ya que en su estructura se incluye un mayor número de referencias anafóricas. En total se presentaron 9 patrones mixtos distintos, en un total de 9 ocurrencias.

El primer ejemplo corresponde al primer grupo, donde en un mismo contexto definitorio se definen los términos *gestión correctiva* y *gestión planificada*. A continuación de la presentación de estos términos, en el contexto se emplean dos formas pragmáticas equivalentes a una referencia anafórica y que sirven para reconocer el término al cuál se refiere la definición. Estas predicaciones corresponden, en el ejemplo citado, a las formas *la primera* y *la segunda*.

Tabla 7. Ejemplos de patrones compuestos

T1 y T2 + PP + PV + D1 + PP + PV + D2	A su vez, en el proceso de gestión se distinguen dos modalidades polares y complementarias: la gestión correctiva y la planificada. La primera modalidad trata de mantener al objeto conducido en un estado dado o de optimizar su operación, (…) La segunda, se caracteriza por preestablecer un estado futuro deseado del objeto conducido, como objetivo…
PV + T1 + D1 + PP + T2 + D2	Se considera calamidad todo acontecimiento que pueda impactar el sistema afectable, en este caso la central y sus alrededores, incluyendo la mina Carbón II...

En el segundo ejemplo podemos observar que se define el término *calamidad*, y dentro de su definición se emplea otro término, *sistema afectable*, que a su vez recurre a la predicación pragmática *en este caso* para introducir su propia definición.

4. Expansión del paradigma de predicaciones verbales

Hasta ahora, hemos presentado una tipología para agrupar los distintos patrones encontrados en los textos de ingeniería. Dentro de estos patrones encontramos los sintácticos, aquellos que emplean una predicación verbal para unir al término con la definición.

Ahora bien, para determinar qué predicaciones verbales nos ofrecen una mayor seguridad al momento de buscar contextos definitorios, elaboramos una metodología donde el primer paso constituye la expansión del paradigma de dichas predicaciones.

Esta expansión se realizó tomando en cuenta que los verbos encontrados suelen unirse con determinadas partículas (preposiciones, artículos y adverbios), y nos permitió determinar cuáles de ellas se unen con ciertos verbos en contextos definitorios, y cuáles se unen con los mismos verbos en otro tipo de fragmentos textuales.

4.1. El Corpus de Referencia del Español Actual

En esta etapa se utilizó el Corpus de Referencia del Español Actual (CREA), de la Real Academia Española. Este corpus se seleccionó debido al criterio de representatividad de sus textos [1].

4.2. Metodología

Analizamos en el CREA las 33 predicaciones verbales encontradas. Como ejemplos, en este artículo presentamos la expansión y evaluación de 10 verbos: *denominar*, *definir*, *entender*, *conocer*, *referir*, *comprender*, *consistir*, *permitir*, *representar* e *incluir*.

La búsqueda se realizó mediante los siguientes operadores y criterios restrictivos que permite el CREA:

· Operadores

- ***Dist/ núm*.** Donde *núm* equivale al número máximo de distancia en palabras que puede haber entre cada elemento de la búsqueda. De esta forma, si la búsqueda es *describe* dist/3 *determina;* la palabra *describe* debe aparecer a una distancia no mayor de tres de la palabra *determina.*

- **Comodín (*)** Se utiliza para buscar cualquier número de caracteres unidos a una palabra flexionada. De esta forma, la búsqueda de la forma *describ** dará como resultado palabras como *describir, describe, describen, describiría, describimos,* etc.

- ***Y*, *O*, *Y NO*.** Estos operadores permiten buscar conjuntos de palabras, así como buscar una palabra sin que en el fragmento recuperado aparezca otra especificada. Por ejemplo: *definir* Y *describir* deberá recuperar ambas palabras; *definir* O *describir* deberá recuperar cualquiera de las dos palabras; *definir* Y NO *describir* deberá recuperar únicamente *definir*.

[1] Para mayor referencia véase la página del CREA: http://corpus.rae.es/creanet.html

· Criterios restrictivos

  – **Medio:** Libros y revistas

  – **Geográfico:** México

  – **Tema 1:** Ciencias y tecnologías

En todos los resultados de las consultas se analizaron los primeros 25 casos recuperados por el CREA. Esto se realizó tomando en cuenta los criterios de representatividad del corpus. Cuando los resultados obtenidos fueron notablemente inferiores a 25 casos, los criterios restrictivos se ampliaron a:

  – **Geográfico:** México y España

  – **Tema 1**.- Ciencias y tecnologías

  – **Tema 2**.- Ciencias sociales

Cuando el resultado era mayor a 300 casos se aplicó un filtro que permite disminuir el número de ejemplos de cada documento, conservando la representatividad de los resultados. Esto es, el filtro *1/10* recuperó uno de cada 10 ejemplos de un mismo texto.
Se utilizó la casilla *Mantener documentos*, la cual funciona para conservar al menos un ejemplo de cada texto donde se encontró la búsqueda, y se utilizó la casilla *Agrupaciones*, por medio de la cual se pueden encontrar las partículas adyacentes a la forma que se busca.

En resumen, en el CREA se buscó cada predicación verbal para determinar qué partículas se agrupan con los verbos en los contextos definitorios. Enseguida se analizó por separado cada agrupación y se determinó cuántos fragmentos recuperados correspondían a contextos definitorios.

4.3. Resultado

A continuación se presenta una tabla con ejemplos de las búsquedas y los resultados obtenidos. La casilla **FORMA** corresponde a la predicación verbal buscada. La casilla **CD's** corresponde al número de contextos definitorios encontrados sobre el total de fragmentos textuales recuperados.

Las formas impersonales se buscaron utilizando el operador *Dist/3*. Esta distancia se consideró ya que la estructura recurrente de las formas impersonales es: *pronombre se + verbo conjugado + partícula*; sin embargo, entre el pronombre *se* y el verbo conjugado pueden aparecer otros pronombres o verbos auxiliares.

Podemos observar que las predicaciones verbales que recuperaron un mayor número de contextos definitorios fueron *se denomina*, *se define*, y *se entiende*. Las formas que recuperaron un menor número de contextos definitorios fueron *representa* e *incluye*.

Tabla 8. Ejemplos de búsquedas realizadas en el CREA

FORMA	CD's
se dist/3 denomin*	23/25
se dist/3 (defin* Y NO definitiv*)	20/25
se dist/3 (entiend* O entend*)	10/25
se dist/3 conoc*	6/25
se dist/3 refier*	4/25
comprende*	3/25
consist* Y NO (consistente* O consistencia*)	3/25
permit*	2/25
represent* Y NO representant*	1/25
Inclu*	1/25

Una vez que se obtuvo estos resultados se realizó una búsqueda de las agrupaciones encontradas para cada predicación verbal. De igual forma se determinó cuántos contextos definitorios se encontraron sobre el total de fragmentos recuperados.

En la siguiente tabla se presentan los resultados de las agrupaciones encontradas.

Tabla 9. Ejemplos de agrupaciones encontradas

FORMA	CD's
se le denomina	15/15
se denomina + artindef.	0/1
se denomina + art def.	0/0
se define como	8/8
se define por	2/2
se puede definir	2/2
se define + art def.	1/2
se le define	1/1
se debe definir	0/1
se ha logrado definirlos	0/1
se entiende + art def.	6/7
se entiende como	2/3
se entiende por	1/1
se entiende cuando	1/1
se conoce como	24/25
se le* conoce	20/20
se conoce* con	5/9
se refiere* a	6/25
se refiere* + art def.	0/28
comprende + art	2/4
comprende desde	1/1
consiste en	19/25
Consiste básicamente en	5/6
Permite	16/25
representa a	3/5
incluye a	4/15

Podemos observar que en el caso del verbo *denominar* no se encontró ningún contexto definitorio cuando al verbo le sigue algún artículo indefinido o definido. Lo mismo ocurrió en el caso del verbo *referir*. Las formas *se debe definir* y *se ha logrado definirlos* tampoco dieron como resultado algún contexto definitorio.

Por otro lado, las formas *se le denomina*, *se define como*, *se entiende + art def.* y *se conoce como*, recuperaron un número alto de contextos definitorios.

Así, se pudo delimitar qué formas no ofrecen la seguridad de presentarse en un contexto definitorio, y qué formas son comunes para conectar a un término con su definición.

En resumen, con este análisis pudimos expandir el número de partículas que suelen unirse con ciertos verbos en los contextos definitorios, y pudimos eliminar aquellas formas que nos ofrecen otro tipo de fragmentos textuales.

5. Evaluación del paradigma de predicaciones verbales

Una vez que analizamos cada predicación verbal y determinamos qué formas son recurrentes en los contextos definitorios, evaluamos las predicaciones en un nuevo corpus de documentos especializados.

Con esta evaluación identificamos sistemáticamente qué predicaciones verbales nos ofrecen una mayor grado de confiabilidad al momento de buscar contextos definitorios.

5.1. El corpus de evaluación

El corpus de evaluación consistió en 10 documentos en formato electrónico igualmente provistos por investigadores y estudiantes del Instituto de Ingeniería. En este corpus se encuentran tesis, informes a patrocinadores y artículos en congresos.

5.2. Metodología

Para llevar a cabo esta evaluación se buscaron manualmente los contextos definitorios que presentaron alguna predicación verbal, de forma que se excluyeron aquellos contextos donde sólo se utilizaban marcas tipográficas.

Por otro lado, y tomando en cuenta los resultados de la búsqueda en el CREA, se desarrolló un "macro" en el programa Word. Esta función nos permitió buscar automáticamente las formas verbales expandidas en el CREA y recuperar párrafos donde se presentara alguna de estas predicaciones.

Así, pudimos obtener tres cifras que corresponden al número de contextos definitorios encontrados manualmente, al número de fragmentos textuales encontrados automáticamente donde se presenta alguna predicación verbal, y de estos fragmentos, el número de contextos definitorios encontrados.

De esta forma, se evaluaron las predicaciones utilizando unas medidas comunes para determinar la efectividad de un sistema de recuperación de información. Estas medidas se denominan **Recall** & **Precision**. En nuestra investigación pueden entenderse de la siguiente forma:

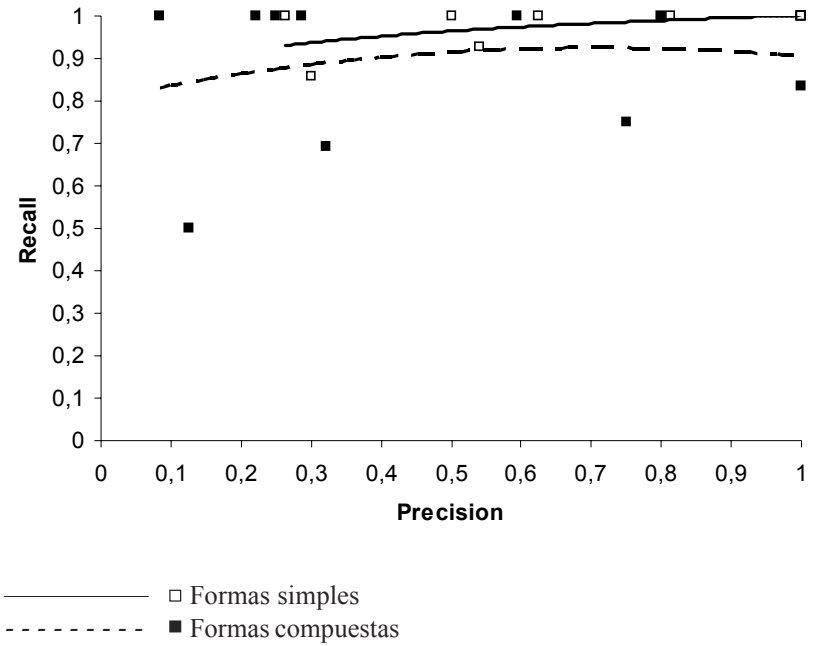
- **Recall** corresponde al número de PV *relevantes* recuperadas *automáticamente*, sobre el número de PV *relevantes* encontradas *manualmente*.
- **Precision** corresponde al número de PV *relevantes* encontradas *automáticamente* sobre el número total de PV encontradas *automáticamente*.

PV corresponde a predicaciones verbales, y *relevante* quiere decir que el fragmento recuperado dio como resultado un contexto definitorio. Estas dos medidas nos dan como resultado valores entre 0 y 1. Un valor más cercano a 1 indica que se ha obtenido un mejor resultado.

5.3. Resultados

A continuación se presenta una tabla con los resultados obtenidos en la evaluación de las predicaciones verbales.

Tabla 10. Resultados de precisión & recall



Podemos observar en la gráfica que se describen dos curvas correspondientes a las formas simples y compuestas. Estas últimas obtuvieron mejores resultados al acercarse al máximo valor de Recall, a pesar de que su Precision disminuyó considerablemente, dependiendo del verbo empleado en el contexto definitorio.

En general puede observarse que nuestra metodología nos permitió recuperar un gran número de contextos definitorios presentes en el corpus estudiado; sin embargo, se obtuvo una gran cantidad de "ruido", esto es, algunos verbos obtuvieron un buen resultado tanto de Recall como de Precision: *denominar*, *conocer*, *entender* y *permitir*, mientras que en otros casos, como *referir* e *incluir*, se encontró un valor alto de recall, aunque el valor de precision fue sumamente bajo. En estos últimos dos casos se recuperó automáticamente un número grande de los contextos definitorios encontrados manualmente; sin embargo, la mitad (0.54) o más de la mitad (0.32) de los párrafos recuperados automáticamente no son contextos definitorios.

Cabe señalar que hasta ahora se han buscado de manera aislada las distintas predicaciones que se encontraron y expandieron en el corpus de análisis y en el CREA. Si tomamos en cuenta que además de las predicaciones verbales existen otro tipo de elementos característicos en los contextos definitorios y se elabora un sistema que permita buscar una combinación de estos elementos, podremos entonces depurar los resultados. Con lo cual se tendrá una mayor seguridad de que los fragmentos textuales que se recuperen sean contextos definitorios.

6. Conclusiones y trabajo futuro

A través de esta investigación hemos determinado una parte esencial de las reglas y restricciones necesarias en la elaboración de un sistema de extracción conceptual. Las reglas corresponden a las secuencias que conforman los contextos definitorios, mientras que las restricciones corresponden a la estructura de las predicaciones verbales que deben tomarse en cuenta (es decir, las combinaciones válidas de verbos más partículas).

Si bien hemos definido parte de las reglas y restricciones, aún se necesitan cubrir algunos puntos esenciales para el desarrollo del sistema. Estos puntos corresponden a un análisis más profundo de los siguientes elementos constitutivos:

- **Predicaciones verbales**. A pesar de que en esta investigación se expandió y evaluó el paradigma de las predicaciones verbales, es necesario contar con un número mayor de verbos que se empleen a un nivel definitorio. Es indudable que tratar de encontrar todos los verbos que pueden funcionar como conectores entre un término y una definición es una tarea sumamente costosa, en cuanto a tiempo y esfuerzo humano se refiere. Sin embargo, para simplificar esta tarea se propone que la búsqueda de verbos definitorios se base en secuencias recurrentes; por ejemplo fórmulas como: T + verbo + partícula gramatical + D; T + *se* + partícula gramatical + verbo + D.

- **Predicaciones pragmáticas**. De la misma forma en que estudiamos el paradigma verbal, el paradigma de las predicaciones pragmáticas también deberá expandirse y evaluarse. Para ello, deberá realizarse un estudio profundo de las estructuras recurrentes de estas formas, con el fin de tratar de delimitar patrones sintácticos de formación.
- **Términos**. Se deberá realizar un análisis de estudios donde se determinen las estructuras sintácticas recurrentes de los términos. Si bien hemos encontrado algunas estructuras recurrentes como SN + SP, aún es necesario considerar un mayor número de patrones de formación que siguen estos elementos constitutivos.
- **Definiciones**. Al igual que en el caso de los términos, es necesario realizar una síntesis de estudios donde se muestren las estructuras sintácticas comunes en las definiciones. Asimismo, se debe tratar de delimitar lo más específicamente posible todos los tipos de definiciones que pueden existir. Esto último está en relación con la probabilidad de extraer automáticamente las relaciones conceptuales presentes en un contexto definitorio.

Cabe mencionar que nuestra investigación también comprende otras áreas de estudio del procesamiento del lenguaje natural, las cuales no impactan directamente sobre el desarrollo de nuestro sistema, pero resultan útiles para poder simplificar algunos problemas específicos. Tal es el caso de las relaciones anafóricas que se presentan en los patrones compuestos. Al respecto, Meyer menciona que un problema recurrente en la identificación automática de contextos ricos en conocimiento es que en los textos reales los términos no se repiten una y otra vez. En su lugar se utilizan pronombres, términos genéricos o variantes de los términos.

Esto nos lleva a considerar que si bien no necesitamos realizar estudios profundos sobre relaciones anafóricas, sí es indispensable estar al tanto de los avances en esta área de investigación, ya que en algún momento servirán para mejorar nuestra metodología.

Finalmente, debe tomarse en cuenta que para desarrollar un sistema de extracción conceptual se necesita un corpus que permita buscar no sólo unidades léxicas, sino también marcas tipográficas. En este sentido, el Grupo de Ingeniería Lingüística se encuentra desarrollando un Corpus de Ingeniería, donde se utilizarán etiquetas XML para representar las distintas marcas tipográficas encontradas en los contextos definitorios, tales como notas al pie de página, encabezado, autor, siglas, etc., lo cual nos permitirá desarrollar un sistema de búsqueda integral.

Referencias

DAVIDSON, L. (1997) *Knowledge extraction technology for terminology*. M.A. Thesis, Ot: Universidad de Ottawa: Universidad de Ottawa.

MEYER, I. (2001) "Extracting Knowledge-rich contexts for terminography". En *Recent advances in computational terminology*, Didier Bourigault (ed.). Amsterdam: John Benjamin's, 279-302.

PEARSON, J. (1998) *Terms in context*. Amsterdam: John Benjamin's.

RODRÍGUEZ, C. (1999) *Operaciones metalingüísticas explícitas en textos de especialidad.* Treball de Recerca. Instituto Universitario de Lingüística Aplicada, Barcelona: Universidad Pompeu Fabra.

RODRÍGUEZ, C. (2000) "Extraction of knowledge about terms from indications of metalinguistic activity in texts". En *Conference on intelligent text processing and computational linguistics*. *Proceedings*, Alexander Gelbuhk (ed.). México: Instituto Politécnico Nacional.

RODRÍGUEZ, C. (2002) "Automatic extraction of non-standard lexical data for a metalinguistic information database". En *Lecture Notes in Computer Science*, Alexander Gelbukh (ed.). Berlin: Springer.

Agradecimientos

Agradecemos al CONACyT (R37712-A) y a la DGAPA-UNAM (IN402900) por su apoyo para el desarrollo de este proyecto.