ISSN 1870-4069

Algoritmo basado en reglas de asociación para la extracción de relaciones no taxonómicas en corpus de dominio

Irvin Yair Cabrera Moreno, Mireya Tovar Vidal, José de Jesús Lavalle Matínez, Meliza Contreras Gonzalez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Puebla, México
yair.cb@gmail.com, mtovar@cs.buap.mx, jlavallenator@gmail.com, mcontreras@cs.buap.mx

Resumen. La identificación de relaciones no taxonómicas es una tarea que se realiza con el aprendizaje y la creación de ontologías. Además, la construcción manual de ontologías para expertos e ingenieros de conocimiento es una tarea costosa y lenta, por lo que es necesario crear algoritmos automáticos y/o semiautomáticos que agilicen el procedimiento. En esta investigación se propone un algoritmo para la extracción de relaciones no taxonómicas en una ontología de Inteligencia Artificial (IA), las cuales son evaluadas a través de una técnica de minería de datos: *reglas de asociación*, que cuenta con medidas estadísticas que determinan la probabilidad de ocurrencia entre los conceptos y el verbo conector relacionados. Los resultados experimentales indican que el 72 % de las relaciones obtenidas en el algoritmo existen en la ontología de IA.

Palabras clave: relaciones no taxonómicas, extracción de información, ontología.

An Algorithm Based on Association Rules for the Extraction of Non-Taxonomic Relationships in the Domain Corpus

Abstract. The identification of non-taxonomic relationships is a task that is carried out with learning and the creation of ontologies. Also, the manual construction of ontologies for experts and knowledge engineers is a costly and slow task, which is why it is necessary to create automatic or semi-automatic algorithms that speed up the procedure. In this research we propose an algorithm for the extraction of non-taxonomic relationships in an ontology of Artificial Intelligence (AI), evaluated through a data mining technique: *association rules*, which has statistical measures that determine the probability of occurrence between the concepts and the related connector verb. The experimental results indicate that 72 % of the relationships obtained in the algorithm exist in the ontology of AI.

Keywords: non-taxonomic relationships, information extraction, ontology.

1. Introducción

La creación automática de ontologías y su representación en un lenguaje formal comúnmente se realiza por ingenieros de conocimiento junto con expertos de dominio. Este proceso implica la extracción, conceptualización, evaluación y formalización del conocimiento del dominio. La construcción manual de ontologías se ha identificado en gran medida como una tarea costosa, tediosa y con tendencia a errores [1]. Además de dificultades técnicas, como la falta de estándares para reutilizar las ontologías existentes y la ausencia de métodos de extracción automática de conocimientos, son problemas que dificultan la creación de ontologías [2].

La definición más popular de ontología en computación es brindada por Gruber [3], quien define una ontología como: "*una especificación explícita de una conceptualización*", es decir, que proporciona una estructura y contenidos de forma explícita que codifica las reglas implícitas de una parte de la realidad; estas declaraciones explícitas son independientes del fin y del dominio de la aplicación en el que se usarán o se reutilizarán sus definiciones.

Actualmente, la creación de ontologías a partir de textos utilizando métodos de aprendizaje automático y de minería de datos se ha propuesto como un método que facilita el proceso de ingeniería ontológica. En este contexto, en [4] el aprendizaje ontológico se ha identificado como un campo que apunta a ayudar a los ingenieros del conocimiento, así como a los usuarios finales en la creación de ontologías. Puede verse como un campo multidisciplinario, con disciplinas como la ingeniería ontológica, el aprendizaje automático y el procesamiento del lenguaje natural. El uso de estas tecnologías se distribuye en tres tareas principales, extracción de entrada léxica, extracción de taxonomía y extracción de relaciones no taxonómicas [4]; todos juntos permiten construir una ontología desde cero o mejorar una ontología existente utilizando diferentes fuentes de información.

El aprendizaje automático de ontologías a partir de textos constituye un medio prometedor para que los ingenieros ontológicos aceleren la creación manual de ontologías, de tal forma que se han propuesto varios enfoques para cubrir las diferentes tareas. En este proceso, la fase de extracción de relaciones no taxonómicas ha sido reconocida como uno de los problemas con más dificultad [4] y menos cubiertos [5]. Esta fase se puede dividir en dos problemas diferentes: descubrir la existencia de una relación entre un par de conceptos y luego etiquetar esta relación de acuerdo con su significado semántico. La asignación de etiquetas a las relaciones también es difícil ya que son posibles varias relaciones entre instancias de los mismos conceptos generales [4].

En particular en esta investigación se usa un método de minería de datos llamado, *reglas de asociación*, en el cual gracias a medidas estadísticas se puede detectar la relación entre un par de conceptos de acuerdo a la ocurrencia de

ellos en las oraciones del texto. El objetivo de esta investigación es crear un algoritmo que extraiga de forma automática las relaciones no taxonómicas y evaluar su desempeño con respecto a las relaciones no taxonómicas existentes en una ontología del dominio de Inteligencia Artificial.

Este artículo está distribuido de la siguiente forma: en la Sección 2 se presentan algunos trabajos relacionados con la extracción automática de relaciones no taxonómicas. En la sección 3 se expone la teoría de las reglas de asociación; en la Sección 4 se presenta el algoritmo propuesto, en la Sección 5 se exponen los resultados obtenidos de la investigación y finalmente en la Sección 6 se presentan las conclusiones y el trabajo a futuro de esta investigación.

2. Trabajos relacionados

En esta sección se presentan trabajos de otros autores, estos trabajos están relacionados con la extracción de relaciones no taxonómicas, su evaluación y descubrimiento, así como también el aprendizaje automático de ontologías.

Serra y Girardi en [6] proponen un proceso semiautomático para extracción de relaciones no taxonómicas de fuentes de texto, haciendo uso de técnicas de procesamiento de lenguaje natural (PLN) para identificar las relaciones no taxonómicas y técnicas de minería de datos para sugerir un nivel alto dentro de la jerarquía de la ontología en inglés. El proceso se divide en tres fases. La primera fase consta de la extracción de las relaciones candidatas utilizando PLN. El objetivo es encontrar el verbo que indique la relación no taxonómica. En la segunda fase se hace un análisis del nivel jerárquico y por último, en la tercera fase se realiza una selección manual de relaciones por un experto.

Mäedche y Staab [7] establecen un nuevo enfoque para el descubrimiento de relaciones no taxonómicas en textos y facilitar la ingeniería de estas relaciones no taxonómicas. Usan reglas de asociación que no solo detectan relaciones entre conceptos. También, el apropiado nivel de abstracción el cual determina la relación. Se hace uso de métodos de procesamiento de texto plano para identificar pares de palabras relacionadas, esto comprende un etiquetador, lo que quiere decir, que crea abreviaciones, también hace uso de análisis léxico para el reconocimiento de entidades, la recuperación información en dominios específicos y el algoritmo de reglas de asociación. La salida de las reglas de asociación son pares de conceptos que son entregados al ingeniero de conocimiento para que las incluya en la ontología.

Weichselbraun, Scharl, Granitzer, Neidhart y Juffinger presentan [8] un enfoque automático para sugerencias en tipos de relaciones en ontologías. En él, crean un vector de verbos con las relaciones encontradas en el corpus, agrega centroides conocidos de relaciones ya identificadas que se encuentran en una base de conocimiento. Los vectores de verbos desconocidos son comparados con vectores ya conocidos y se crean sugerencias de relaciones. Las relaciones no taxonómicas son descubiertas en tres pasos a tráves de un análisis de los principales sustantivos, crear sinónimos a estos sustantivos y enriquecen WordNet. El

último paso es un análisis de subsunción, en el cual se asume que el documento está compuesto por oraciones de un conjunto de otros documentos.

Sánchez y Moreno [9] en su trabajo presentan un método automático que extrae conceptos de relaciones no taxonómicas y etiquetado de relaciones usando toda la web como corpus. Para el descubrimiento y extracción de las relaciones no taxonómicas en la red hacen uso de: *técnicas de análisis ligero* que se usa para tener una mejor escalabilidad en un entorno como la web, *análisis estadístico* que es aplicado en las tareas de adquisición del conocimiento. También se usan *patrones lingüísticos* que es una técnica muy efectiva para no consultar a expertos y para descubrir las relaciones no taxonómicas. Por último se usa *bootstrapping* que es utilizado para restringir las consultas hechas por el motor de búsqueda web para obtener corpora de documentos de dominios específicos.

Kavalec, Mäedche y Svátek presentan en su trabajo [10] una combinación entre análisis de texto plano, minería de datos y modelado de conocimiento. Además, para la extracción de relaciones no taxonómicas usan una técnica que se basa en el método de Text-to-Onto. El método usado para el descubrimiento de las relaciones está basado también en las reglas de asociación, donde dos o más léxicos pertenecen a una transacción si estos se encuentran juntos en un documento o en un texto definido, las transacciones más frecuentes son salidas como asociaciones entre sus objetos. Las relaciones no taxonómicas candidatas entre dos conceptos son aquellos verbos en el que el número de transacciones que se mantienen entre el verbo v, el concepto c_1 y el concepto c_2 aparecen en una ventana de n palabras a partir de una aparición de v.

Mäedche y Staab describen [11] un enfoque para extracción de relaciones no taxonómicas de un corpus creado con técnicas de procesamiento de texto plano, este enfoque está basado en el algoritmo de reglas de asociación para encontrar las relaciones y también para definir un nivel de abstracción en ellas. El algoritmo encuentra las relaciones que ocurren en un par de elementos. La regla de asociación básica consiste en un conjunto de transacciones, donde cada transacción tiene un conjunto de elementos y cada uno de los elementos forma parte de un conjunto de conceptos. Para descubrir las relaciones, se construye un esquema de aprendizaje donde modifican el conjunto de transacciones y las métricas para generar las reglas de asociación.

Nabila, Basir y Mamat en [12], presentan un método para extraer relaciones no taxonómicas usando la similitud entre relaciones que existen en más de una oración. El propósito del método es mejorar el proceso de recuperación de relaciones no taxonómicas en dominios específicos de texto. Consiste en extraer conceptos mediante preprocesamiento y se realiza un análisis estadístico para extraer los términos más importantes. Con eso, los conceptos se clasifican en dos listas una de temas y otra de objetos, después, se generan pares de conceptos usando producto cartesiano entre las dos listas y para evitar la existencia de relaciones taxonómicas se usan algunas restricciones. Finalmente, se realiza la extracción y asignación de buenas relaciones entre los pares de conceptos.

Villaverde en su documento [13] propone una técnica para el descubrimiento de relaciones no taxonómicas y la extracción de elementos léxicos que sirven

como conectores entre los conceptos relacionados. Su enfoque está basado en el análisis de estructuras sintácticas y dependencias entre conceptos. Para el descubrimiento de relaciones no taxonómicas su técnica toma como entrada un corpus de textos específicos de dominio y una jerarquía de conceptos que describe las relaciones taxonómicas entre conceptos y busca otras posibles relaciones en el texto. Se usa el etiquetador POS para asignar etiquetas a las palabras de contenido e identificar frases de verbos y nominales en cada oración. Este análisis es aplicado a todas las oraciones encontradas que contienen ambos conceptos candidatos que cumplan con el patrón: <*término*><*verbo*><*término*>. Una vez que las relaciones candidatas estén disponibles, son validadas bajo el criterio de las reglas de asociación.

En esta investigación se propone un algoritmo en el cual por medio de preprocesamiento de texto y a tráves de la técnica de minería de datos: reglas de asociación, se realiza la extracción automática de relaciones no taxonómicas en un corpus de dominio. En los resultados experimentales las relaciones no taxonómicas obtenidas son evaluadas bajo las medidas estadísticas llamadas soporte y confianza [14].

3. Reglas de asociación

Las reglas de asociación describen la relación entre los elementos de un conjunto de datos. Estas reglas nacen de la investigación de Agrawal, Imielinski y Swami [14] donde consideran la colección de datos que generan las compras en un supermercado, sirviendo como apoyo para saber que conjunto de productos se compran y generar por medio de las reglas de asociación promociones. En [14] definen las reglas de asociación como:

Dado C como el conjunto de conceptos y $T := \{t_i \mid i = 1 \dots n\}$ como la base de datos de transacciones, donde n es el número total de transacciones y cada transacción t_i consiste en un conjunto de elementos:

$t_i = \{a_{i,j} \mid j = 1 \dots m_i, a_{i,j} \in C\}$,

y cada elemento $a_{i,j}$ es un elemento del conjunto C y m el número total de elementos en t_i. El algoritmo calcula las *reglas de asociación* presentadas en la ecuación (1):

$$X_k \Rightarrow Y_k (X_k, Y_k \subset C, X_k \cap Y_k = \{\}). \tag{1}$$

Una regla de asociación es una implicación de la forma $X_k \Rightarrow Y_k$ donde X_k es un conjunto de algunos elementos de C también llamado *antecedente* y Y_k es un sólo elemento en C que no está presente en X, también llamado *consecuente*. La regla se satisface si las medidas de *soporte*, ecuación (2), y *confianza*, ecuación (3), sean iguales o mayores a las deseadas. El soporte de una regla $X_k \Rightarrow Y_k$ es el porcentaje de transacciones que contiene $X_k \cup Y_k$ como un subconjunto, y la confianza de una regla $X_k \Rightarrow Y_k$ está definida como el porcentaje de transacciones donde Y_k aparece si X_k se encuentra en una transacción:

$$Soporte(X_k \Rightarrow Y_k) = \frac{|\{t_i \mid X_k \cup Y_k \subseteq t_i\}|}{n}, \tag{2}$$

$$Confianza(X_k \Rightarrow Y_k) = \frac{|\{t_i \mid X_k \cup Y_k \subseteq t_i\}|}{|\{t_i \mid X_k \subseteq t_i\}|}. \quad (3)$$

Las medidas se pueden interpretar como: una regla con bajo soporte indicaría que habrá aparecido por casualidad. Sin embargo, una regla con baja confianza indicaría que no existe relación entre el antecedente y el consecuente. Además, existe una diferencia entre $X_k \Rightarrow Y_k$ y $Y_k \Rightarrow X_k$, debido a que las reglas comparten el mismo soporte pero su confianza tiende a ser distinta.

4. Algoritmo propuesto

En esta sección se presenta el algoritmo propuesto para la extracción automática de relaciones no taxonómicas. El funcionamiento general del algoritmo se detalla en el Algorimo 1.1. En el cual se identifican los acrónimos existentes en el corpus mediante la función *Acronimos*, la cuál por medio de una expresión regular, son seleccionados aquellos tokens de las oraciones que se deseen, especificando un patrón con la forma en que se desean encontrar las palabras. En este caso, se seleccionan aquellas palabras que cumplan con tener dos o más letras mayúsculas consecutivas sin que la palabra tenga minúsculas en ella. Después el corpus es dividido en oraciones en la función *Oracion*, esta división está dada por cada punto que aparezca en el corpus.

Algoritmo 1.1: Algoritmo de extracción de términos propuesto.

Datos: Corpus de dominio
Resultado: Corpus de dominio pre-procesado
1 a ← *Acronimos(corpus)*
2 sentences ← *Oracion(corpus)*
3 sentences, dic ← *prepro(sentences,a)*
4 resul ← *pos_tag*(sentences) inds ← *buscarNPS(resul)*
5 cand ← *relacionesCandidatas*
6 triple ← *reglasAsociacion*

Esto con la finalidad de que en la función *prepro* se realice una expansión de términos que es detallada en el Algoritmo 1.2. La función *prepro* recorre cada oración en busca de los acrónimos obtenidos en el paso anterior y los almacena en la variable *dic* que tiene función como diccionario de acrónimos. Cada oración es divida en tokens para realizar una comparación entre los tokens de cada oración y los acrónimos obtenidos, para comprobar qué acrónimos tiene la oración. Si el resultado de la comparación tiene elementos, la función continúa con verificar cada elemento encontrado en la comparación. Si el acrónimo no está en el diccionario de acrónimos, se busca su significado tomando el tamaño del acrónimo para buscar hacia la izquierda las palabras que formen el significado, el acrónimo es agregado al diccionario y se elimina de la oración dejando solo su significado. Si el acrónimo ya se encuentra en el diccionario, sólo se realiza la sustitución del acrónimo con su significado en la oración.

Algoritmo 1.2: Función *prepro.*

```
Datos: Oraciones del corpus, acrónimos
Resultado: Oraciones con acrónimos expandidos, diccionario de
           acrónimos
Para i = longitud-oraciones hacer:
  oracion ← oraciones[i]
  comp ← Para elemento en acronimos Si elemento está en oracion
  Si comp no está vacío hacer:
    Para palabra en comp hacer:
      indice← oracion.index(oracion)
      Si la palabra está en dic hacer:
        Para letra en palabra:
          Se toma el indice del acronimo, se elimina y se expande
        FinPara
    FinPara
      Sino:
        rango ← len(palabra)
        Para j<rango:
          Si oracion[ind-(rango-j)][0] == palabra[0] hacer:
            acronimo = acronimo + oracion[ind-(rango-j]
          Sino
            eliminar acronimo de acronimos
          FinSi
        FinPara
      FinSi
  FinSi
```

El siguiente paso es volver las oraciones en su forma minúscula para lograr una comparación entre las palabras más exacta. Después se realiza el etiquetado *Parte del discurso*, este etiquetado agrega una etiqueta al contenido y función de cada palabra logrando así identificar las clases de palabras en las oraciones. Mediante una gramática, las frases nominales de las oraciones se identifican. Esta gramática es muy parecida a una expresión regular, la diferencia es el uso de las etiquetas POS para crear el patrón de frase que se desea encontrar. Estas frases nominales solo son etiquetadas en el análisis y son devueltas como tipo árbol por cada oración, así que se hace una extracción de las frases nominales en la función *buscarNPS* que puede verse en el Algoritmo 1.3. En la función *buscarNPS* se recibe como entrada la variable *resul* la cuál lleva como contenido las oraciones en forma de árbol que fueron devueltas en el etiquetado POS. En *BuscarNPS* se recorre cada árbol generado en el paso anterior, con la intención de encontrar las frases nominales que se agruparon. Se considera que, al etiquetar una palabra con el analizador POS, cada palabra se vuelve una tupla[1], que consta de la palabra y su etiqueta de clase. Al crear la gramática, se agrupan ciertos tipos de palabras que conforman en grupo un mismo significado, dando

[1] Una tupla es un tipo de lista que solo consta de dos elementos

así la frase nominal. Para su extracción, en el algoritmo solo es necesario buscar conjuntos de tuplas, es decir, en el árbol resultado, se buscan nodos que sean de tipo árbol y se extrae cada palabra que lleven sus tuplas formando así la frase nominal.

Algoritmo 1.3: Función *buscarNPS*.

```
   Datos: resul
   Resultado: inds
 1 Para oracion en resul hacer:
 2    nps ←[ ]    Para nodo en oracion hacer:
 3      Si nodo es tipo árbol hacer:
 4         palabra ← ' '
 5        Para i < len(nodo) hacer:
 6           palabra ← palabra + nodo[i][0] + ' '
 7        FinPara
 8        palabra← palabra[ :len(palabra)-1]
 9        nps ← agrega(aux)
10      FinSi
11   FinPara
12    aux ← [index(oracion),nps]
13    inds ← aux
14 FinPara
```

Cuando se tiene la lista de frases nominales que ocurren en una oración, se guardan junto con el índice de la oración en el que se encuentran y se devuelve como salida esa nueva lista de índices con frases nominales. Una vez disponibles las frases nominales junto con el índice de la oración en la que se encuentran, se continúa con la extracción de las relaciones no taxonómicas. El algoritmo propuesto para la extracción de relaciones no taxonómicas consta de dos funciones: *relacionesCandidatas* y *reglasAsociacion*.

El procedimiento que sigue la función *relacionesCandidatas* se detalla en el Algoritmo 1.4. En esta función se observa que la entrada son las oraciones del corpus y la lista de frases nominales e índices de su oración de ocurrencia, la salida son las tripletas candidatas para la ontología. El algoritmo de la función *relacionesCandidatas* comienza recorriendo las oraciones del corpus. En este recorrido se pretende encontrar el texto entre dos frases nominales. En el texto entre ellos se puede hallar una frase verbal que conecte a dos frases nominales que ocurren en la oración.

De los pasos 3-5, se toman los índices donde comienzan cada frase nominal y así conocer el texto que hay entre ellas. Este texto es etiquetado por el análisis *POS* en la función *pos_tag* para conocer que tipo de palabras existen entre estas frases nominales. Nuevamente, se repite el procedimiento usado para encontrar las frases nominales, con la diferencia que aquí se buscan frases de verbos. De la misma forma, se crea una gramática en la función *parser* para identificar las frases verbales que ocurran en el texto obtenido espcificando etiquetas de tipo verbo y adverbio.

Si es encontrada una frase verbal entre las frases nominales, en la lista *cand*, son almacenadas la frase nominal de la izquierda, el verbo que conecta y la frase nominal de la derecha, formando una tripleta. Después de obtener las tripletas candidatas son evaluadas bajo las medidas de *soporte y confianza* de las *reglas de asociación* [14].

En este contexto, una transacción en T representa la ocurrencia de un par de conceptos con algún verbo de enlace en el cuerpo del texto. La fuerza de la asociación de ambos conceptos con el verbo estará dada por la regla de confianza. Se debe encontrar las frases nominales que tienen soporte de transacciones por encima del soporte mínimo. El soporte para un conjunto de elementos es el número de transacciones que contienen el conjunto de elementos.

El algoritmo fue codificado con la biblioteca NLTK de python [15]. En la siguiente sección se presentan los resultados obtenidos.

Algoritmo 1.4: Función *relacionesCandidatas*.

```
   Datos: oraciones, ind
   Resultado: tripleta
 1 Para i <len(oraciones) hacer:
 2   Para j < len(ind[i][1]-1) hacer:
 3     ind1 ← oraciones[i].index(ind[i][1][j])
 4     ind2 ← oraciones[i].index(ind[i][1][j+1])
 5     vp ← oraciones[i][len(ind[i][1][j])+ind1+1:ind2]
 6     vp2 ← pos_tag(vp)
 7     Si len(vp2) == 1 y vp2[0][1] == V hacer:
 8       aux← [ind[i][1][j],vp,ind[i][1][j+1]]
 9       tripleta← agrega(aux)
10     Sino Si 8 > len(vp2)>1 hacer:
11        vp2 ← parse(vp2)
12       Para k <len(vp2) hacer:
13         Si vp2[k] es tipo árbol hacer:
14           Para v en vp2[k] hacer:
15             verbo ← verbo + v[0] + ' '
16             verbo ← verbo[ :len(verbo)-1]
17             aux ← [ind[i][1][j],verbo,ind[i][1][j+1]]
18             cand ← agrega(aux)
19           FinPara
20         FinSi
21       FinPara
22     FinSi
23   FinPara
24 FinPara
```

5. Resultados experimentales

En esta sección se muestran los resultados obtenidos con el algoritmo propuesto para la extracción de relaciones no taxonómicas. Para la investigación se usó la ontología de inteligencia artificial propuesta en [16]. A continuación en la Tabla 1 se muestran los datos utilizados. La ontología consta de un número determinado de documentos, tokens o palabras, vocabulario y número de oraciones.

Tabla 1. Datos del corpus de dominio.

Ontología	Documentos	Tokens	Vocabulario	Oraciones
IA	8	10,805	1,510	460

Para la ontología IA, durante el procesamiento de sus datos, se obtuvieron los resultados que se presentan en la Tabla 2. En la Tabla 5 se muestran algunos de los resultados experimentales de la extracción automática de relaciones no taxonómicas usando las medidas de reglas de asociación. En nuestra investigación, el soporte y confianza de una regla candidata $C \Rightarrow v$, donde C indica ambos conceptos y v el verbo que los conecta. Una transacción en T representa la ocurrencia de un par de conceptos con algún verbo que los una.

Tabla 2. Total de conceptos y relaciones no taxonómicas en la ontología de dominio de IA.

Ontología	Conceptos	Relaciones no taxonómicas
IA	276	61

Tabla 3. Resultados obtenidos para pares de conceptos con medidas de Soporte y Confianza.

$Concepto_1$	$Concepto_2$	$Verbo$	Soporte	Confianza
neural network	agents	based	0.01949	1
temporal agent	time	may use	0.01949	1
intelligent agent	enviroment	perceives	0.01949	1

Gracias a estas medidas estadísticas se logra segmentar los resultados para obtener las relaciones que tengan un mayor grado de relación. Los resultados obtenidos se evaluaron con los de la ontología provista en [16] y se calculó la

exactitud, la cuál se refiere a la evaluación del sesgo de las predicciones, es decir, responde a la pregunta: *¿Cuál es el promedio de las predicciones correctas?* [17]. La fórmula de exactitud se presenta en la ecuación 4:

$$Exactitud = \frac{CantidadDeCasosCorrectos}{TotalDeCasos}. \tag{4}$$

donde nuestro sistema extrajo un total de casos correctos de 44 y el total de casos en la ontología es igual a 61, lo que representa un total de 72 % de exactitud.

6. Conclusiones

En esta investigación se implementó un algoritmo en Python utilizando la biblioteca de NLTK para el procesamiento del lenguaje natural y para la extracción de relaciones no taxonómicas en un corpus de dominio de inteligencia artificial mediante la técnica de reglas de asociación. Esta técnica describe la probabilidad de que exista una relación entre objetos, en nuestro caso entre un par de conceptos y un verbo que los conecta en una oración en el corpus de dominio. Posteriormente, los resultados fueron comparados con las relaciones no taxonomicas de una ontología de dominio de IA.

Con base en los resultados experimentales, se logró obtener el 72 % de relaciones no taxonómicas existentes en la ontología de IA [16]. Con base en estos resultados se observó que al contar con un corpus pequeño, el soporte que hay entre los conceptos y el verbo es muy bajo, es decir, menor al 2 %, ya que esta medida representa la probabilidad de encontrar al par de conceptos y el verbo que los conecta dentro del dominio. Sin embargo, esto provoca que la confianza sea mayor al 50 %, ya que describe la probabilidad de que esta relación sea verdadera, es decir, que al encontrar los dos conceptos, el verbo se encuentre en la misma oración. Además, cabe mencionar que algunas de las relaciones que no detectó el algoritmo, y existen en la ontología, son aquellas que la localización del verbo está al final de la oración y no de manera intermedia.

Como trabajo a futuro se propone implementar una propuesta de solución que identifique relaciones no taxonómicas en diferentes tipos de estructuras de la oración en inglés. Así mismo, aplicar el enfoque a otras ontologías y comparar los resultados.

Agradecimientos. Esta investigación es apoyada por el Fondo Sectorial de Investigación para la Educación, con el proyecto CONACyT CB/257357 bajo el número de becario 28617 y por el proyecto VIEP-BUAP 100409344-VIEP2019.

Referencias

1. Shamsfard, M., Abdollahzadeh Barforoush, A.: The state of the art in ontology learning: a framework for comparison. The Knowledge Engineering Review 18(4), 293–316. https://doi.org/10.1017/S0269888903000687 (2003)

2. Shamsfard, M., Abdollahzadeh Barforoush, A.: Learning ontologies from natural language texts. International Journal of Human-Computer Studies, No. 60, pp. 17–63. https://doi.org/10.1016/j.ijhcs.2003.08.001 (2004)
3. Gruber, T.R.: Toward Principles for the Design Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies 43(5-6), 907–928 (1992)
4. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems 16(2), 72–79. https://doi.org/10.1109/5254.920602 (2001)
5. Sánchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. Data and Knowledge Engineering 64(3), 600–623. https://doi.org/10.1016/j.datak.2007.10.001 (2008)
6. Serra, I., Girardi, R.: A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text. Intelligent Information Management 3, pp.124–119. https://doi.org/10.4236/iim.2011.34014 (2011)
7. Maedche, A. Staab, S.: Mining Non-Taxonomic Conceptual Relations from Text. In: Ekaw-00 – European knowledge acquisition workshop, Juan-les-pins, LNAI, Springer. https://doi.org/10.1.1.41.4860 (2000)
8. Weichselbraun, A., Wohlgenannt, G., Scharl, A., Granitzer, M., Neidhart, T., Juffinger, A.: Discovery and evaluation of non-taxonomic relations in domain ontologies. International Journal of Metadata, Semantics and Ontologies 4(3), 212–222 (2009)
9. Sánchez, D., Moreno, A.: Discovering non-taxonomic relations from the Web. In: 7th International Conference on Intelligent Data Engineering and Automated Learning. LNCS 4224. https://doi.org/10.1.1.83.3546, pp. 636–629 (2006)
10. Kavalec, M., Maedche, E., Svátek, V.: Discovery of Lexical Entries for Non–Taxonomic Relations in Ontology Learning. In: Proceedings of SOFSEM 2004: Theory and Practice of Computer Science, LNCS 2932. https://doi.org/10.1.1.10.2718, pp. 256–249 (2004)
11. Maedche, A., Staab, S.: Semi-Automatic Engineering of Ontologies from Text. In: Proceedings of the 12th Internal Conference on Software and Knowledge Engineering. https://doi.org/10.1.1.453.2051, pp. 239–231 (2000)
12. Nabila, N.F., Basir, N., Mamat, A.: Synonymous Non-Taxonomic Relations Extraction. ARPN Journal of Engineering and Applied Sciences 10(2) (2015)
13. Villaverde, J., Persson, A., Godoy, D., Amandi, A.: Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Data and Knowledge Engineering 36(7), 10288–10294. https://doi.org/10.1016/j.eswa.2009.01.048 (2009)
14. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data 22(2), 207–216, Washington DC (USA) https://doi.org/10.1145/170035.170072 (1993)
15. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia: Association for Computational Linguistics (2002)
16. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Proceedings of the 3rd International Conference on Ontology Patterns 929, pp. 61–72. CEUR-WS.org (2012)
17. Vazques, K.: Monitoreo de opiniones en redes sociales sobre la calidad del servicio. Tesis Licenciatura, Benemérita Universidad Autónoma de Puebla (2017)