

Minería de Textos y Aprendizaje Automático en el Procesamiento del Lenguaje Natural

Vivian López Batista, Ramiro Aguilar Quispe

Departamento de Informática y Automática
Universidad de Salamanca
Plaza de la Merced S/N, 37008 Salamanca, Spain
{vivian, ramiro}@tejo.usal.es

Abstract. En este trabajo se brinda una metodología de minería de textos basada en la metodología de minería de datos tradicional. El enfoque emprendido busca comprender la sintaxis y la semántica de los textos y comprender la intención de los mismos logrando así disponer de otra alternativa para el procesamiento y la interpretación del lenguaje natural.

1 Introducción

Generalmente, en el procesamiento del lenguaje natural como en muchas otras áreas del saber el conocimiento se obtiene a partir de un conjunto de observaciones y de conocimientos previos, la intuición del investigador le conduce a formular una hipótesis que describe cierto contexto y comprensión del lenguaje. Sin embargo, esta “intuición” resulta inoperante para tratar enormes cantidades de datos almacenados en un soporte informático. Al hablar de ingentes cantidades de datos, diversas investigaciones se orientan a la minería de datos donde, se pueden aplicar técnicas predictivas y descriptivas para obtener conocimiento. Un enfoque particular de descubrimiento de conocimiento es la minería de textos, la cual servirá como alternativa para comprender el procesamiento del lenguaje natural. La minería de textos, además de usar las cualidades presentes en la minería de datos como la combinación de técnicas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, permitirá entender los aspectos relacionados con la identificación, organización y comprensión de la sintaxis y la semántica presentes en algún texto.

Considerando conjuntos contextuales y *textos básicos* sobre algún universo de discurso, se realiza el proceso de minería de datos para obtener una aproximación sintáctico-semántica coherentes respecto del conocimiento del texto base. Con esto, se desea mejorar el procesamiento del lenguaje natural “minando” datos textuales presentes en alguna fuente de datos. Así, no sólo, se pretende automatizar el descubrimiento de conocimiento, sino también, automatizar la generación de los datos para la minería.

2 Proceso de Minería de Textos

El objetivo es estructurar y obtener conocimiento semántico sobre algún texto escrito en lenguaje natural.

Para esto se necesitan realizar las siguientes tareas:

- Definir contextos dentro de algún “corpus” lingüístico.
- Seleccionar un texto como base para adquirir el conocimiento y codificar su léxico asociado (codificar su contenido).
- Descubrir los patrones secuenciales en el texto base codificado (minería de datos).
- Definición de la gramática del texto base a partir de los patrones secuenciales antes obtenidos (conocimiento sintáctico del texto).
- Etiquetar las reglas de producción de la gramática en base a diferentes contextos claves.
- Entrenamiento de una red neuronal con las reglas de producción etiquetadas para lograr el conocimiento semántico.

2.1 Definición de Contextos

Sea V un conjunto finito denominado vocabulario, cualquier secuencia de elementos en V se denomina una cadena en V y el conjunto de todas las cadenas en V es el *lenguaje universal* V^* generado por V . Cualquier subconjunto de V^* se denomina **lenguaje** (L). Si consideramos el problema lingüístico con respecto a L , cualquier cadena x en L se denomina una cadena correcta o *bien formada* siempre que los elementos x pertenezcan a V .

Sea Ξ el conjunto dinámico de contextos definido como $\Xi = \{C_i : 1 \leq i \leq n\}$; cada contexto C_i está definido como $C_i = \{c_1^{(i)}, c_2^{(i)}, \dots, c_m^{(i)}\}$. c_j representa a un símbolo terminal del lenguaje (elemento del vocabulario). Por ejemplo el contexto “animales” compuesto de los elementos $\{paloma, gallina, lechuza, alcón, zorro, perro, lobo, tigre\}$.

En lingüística el concepto de *representación de contexto*, se asocia con una cantidad de palabras adyacentes. La similitud entre los elementos puede ser reflejada a través de las similitudes en el contexto. Nótese que para ordenar los conjuntos de códigos arbitrarios deben expresarse similitudes invariantes, por ejemplo, en términos de elementos que ellos tienen en común.

Por otro lado es evidente que el significado (semántica) de un símbolo no sólo puede ser derivable de la probabilidad condicional de sus ocurrencias con otras codificaciones, sino que puede emerger del propio contexto.

En [7], para aprender la semántica de un texto, se generan ejemplos de entrenamiento de la forma $X = X_s + X_c$ donde, X_s el vector que representa la expresión simbólica de un elemento semántico y X_c la representación del contexto.

Entonces, algún modelo neuronal muy simple asume que X_s y X_c están concatenados a la misma unidad neuronal, digamos que representan el vector X definido como:

$$X = \begin{bmatrix} X_s \\ X_c \end{bmatrix} = \begin{bmatrix} X_s \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ X_c \end{bmatrix} \quad (1)$$

El fundamento central está en que las dos partes tienen su propio peso en el proceso pero predomina la norma del contexto, reflejando las relaciones métricas de los conjuntos asociados y la codificación inicial se realiza siguiendo un orden espacial que refleja similitudes semánticas. Con lo anterior, una red neuronal será entrenada sobre una base de oraciones, que reflejen los diferentes contextos semánticos con los vectores de entrada X descritos en la ecuación 1. Tras el entrenamiento, se verifica qué unidad del mapa neuronal se activa para cada vector de entrada y se procede al etiquetado que se corresponderá con el nombre de la clase. Como resultados los contextos similares aparecen proyectados en unidades cercanas al mapa, lo que permite crear un mecanismo y análisis de clase posibilitando la interpretación semántica.

2.2 Codificación del texto base

Con la definición de contextos o conjuntos de elementos se codifica el “texto base” con lo que se construye un código, compuesto de subcadenas de tres caracteres del tipo 1**, el carácter 1 denota que se trata de un contexto conocido y, los caracteres ** denotan el identificador del número correspondiente al contexto. Por ejemplo, si la palabra “amistad” está definida dentro del contexto C_2 , entonces la subcadena correspondiente a la palabra será 102, con lo que podemos disponer de hasta 99 contextos. Si la subcadena no corresponde a un contexto definido se codifica como 888 (ver figuras 1 y 2). Cada subcadena del código del texto base es un elemento de algún contexto (esto es análogo a los procesos en genómica, el texto base sería el cromosoma, la codificación sería el genoma y las subcadenas, los genes; secuenciar el genoma sería conocer la sintaxis y, comprender el genoma, comprender la semántica. Ver [1]).

En las subcadenas del código del texto base, la aparición de código 888, como sabemos, expresa desconocimiento. Para conocer la expresión de tales subcadenas se aproxima a la subcadena más cercana conocida. Por ejemplo, si tenemos la aparición de las subcadenas 104 888 108 y se tienen subcadenas 104 102 108 lo más seguro es que se sustituya la subcadena 888 por 102 a través de la expresión de la segunda subcadena.

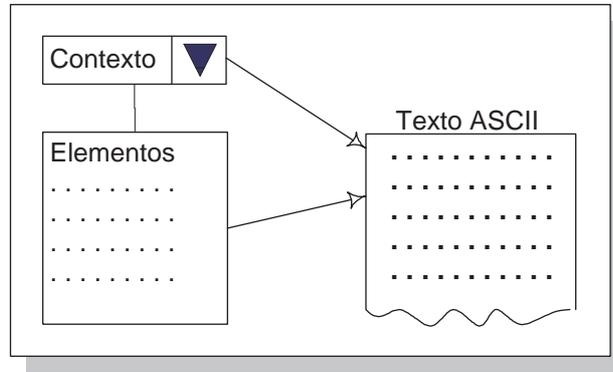


Fig. 1. Interrelación entre los elementos contextuales y el texto a analizar.

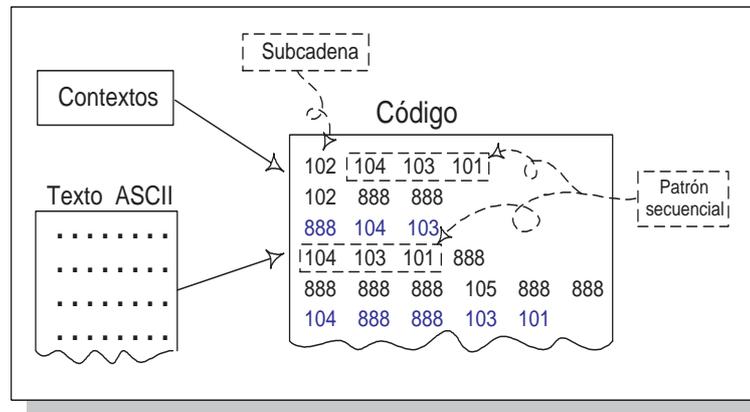


Fig. 2. Codificación del Cromosoma del texto original, se identifican algunos patrones secuenciales.

2.3 Descubrimiento de Patrones Secuenciales

A partir del cromosoma del texto base, se genera un conjunto de ejemplos o transacciones sobre los cuales se aplica el descubrimiento de patrones secuenciales (la técnica de descubrimiento de patrones secuenciales se describe susintamente en [1]). Con la obtención de patrones secuenciales se considera que cada uno de los mismos es una regla gramatical presente en el texto base. Agrupando el conjunto de reglas gramaticales o patrones secuenciales se forma una gramática G , a saber, coherente respecto de la expresión sintáctica del texto base (ver figura 3).

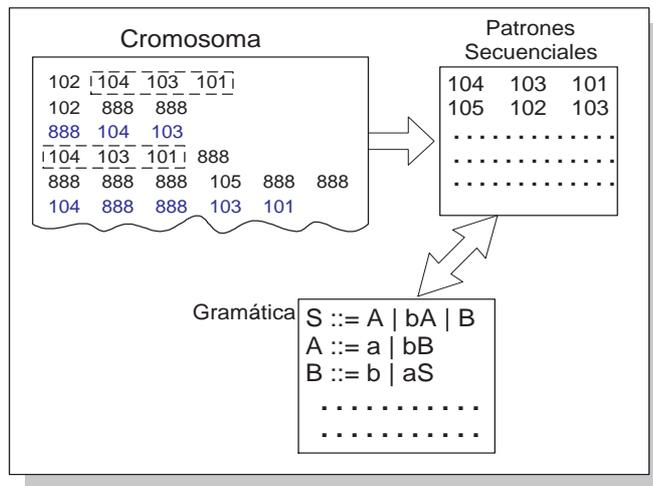


Fig. 3. Estructuración de la gramática generada por la agrupación de los patrones secuenciales.

2.4 Etiquetado de las Reglas de Producción de la Gramática

Sean los contextos claves C_1^* , C_2^* , ... C_N^* , donde, por ejemplo: $C_1^* = \{\text{atentato, bomba}\}$, $C_2^* = \{\text{fraude, dinero}\}$, ... Se etiquetan las reglas de producción de la gramática G en base a los elementos contextuales que describen. De esta manera, cualquier regla de producción asociada con algún elemento de los contextos clave denotará el significado o semántica de la regla (ver figura 4).

2.5 Entrenamiento de la Red Neuronal

Una vez, etiquetado el conjunto de ejemplos, se distinguen dos conjuntos uno de entrenamiento y otro de prueba. Con el conjunto de entrenamiento se aplica

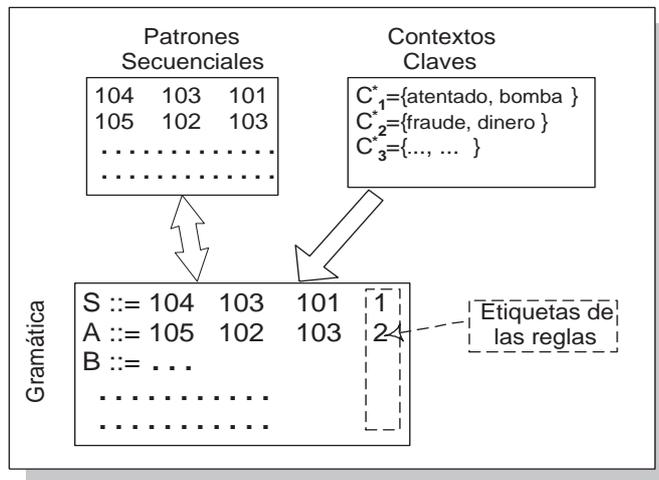


Fig. 4. Etiquetado de las reglas de producción en base a los predeterminados contextos claves.

una red neuronal de conexiones hacia adelante cuyas entradas son los patrones secuenciales y las salidas son las clases de contextos claves que denotan la intensidad o semántica de algún texto base. Después del aprendizaje de la red, con el conjunto de prueba se verifica la eficacia del aprendizaje de la red.

3 Conclusiones

La propuesta permite un enfoque de minería de textos orientada al descubrimiento de conocimiento sintáctico y semántico a partir del contenido de algún texto, posibilitando no sólo brindar una visión selectiva y perfeccionada de la información contenida en documentos escritos, sino también automatizar la generación de los datos para la minería de textos como medio de potenciar este tipo de herramientas.

Así, se realiza minería de textos que hace el análisis léxico de los textos y especialmente la construcción automática de estructuras de clasificación y categorización que se codifican en forma de tesauros, donde cada uno de sus términos, al menos en principio, se utiliza para denotar un concepto, la unidad semántica básica que permite expresar una idea.

El conocimiento sintáctico se obtiene de forma automática con un enfoque de descubrimiento de conocimiento sin el previo planteamiento de hipótesis. Es decir, éste se extrae automáticamente de los patrones contextuales y de las características lingüísticas de los propios textos que componen el córpora.

El conocimiento semántico se obtiene tras el entrenamiento de las reglas de producción previamente etiquetadas con el uso de contextos claves.

Con lo anterior, se tienen alternativas para saber los temas o conceptos principales a que se refiere un texto e incluso generar el resumen del mismo.

References

1. R. Aguilar (2002). *Análisis de Técnicas de Minería de Datos, Comparación de Resultados en Diferentes Dominios de Aplicación*. Memoria para el Grado de Salamanca de la Universidad de Salamanca. España.
2. R. Aguilar (2002). Pautas Para la Simbiosis: Minería de Datos y Lógica Borrosa. *XI Congreso Español sobre Tecnologías y Lógica Fuzzy*. Universidad de León. ESTYLF 2002.
3. J.C. Bezdek (1993). A review of Probabilistic, Fuzzy and Neural Models for Pattern Recognition. *Journal of Intelligent and Fuzzy Systems*.
4. M. Berry, G. Linoff (1999). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons.
5. U. Fayyad, et.al. (1996). From Data Mining to Knowledge Discovery: An Overview. *Knowledge Discovery in Databases*. MIT Press.
6. J. Han, M. Kamber (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.
7. V. López (1996). *Desambiguación Semántica Basada en Métodos Conexionistas para un Problema de Traducción Automática Alemán-Español*. Tesis Doctoral de la Universidad de Valladolid. España.
8. V. López, L. Alonso, M. Moreno (2000). Mapas Organizados para la Minería de Datos en Procesamiento del Lenguaje Natural. *Actas del II Taller Iberoamericano sobre Aplicaciones e Implementaciones de Redes Neuronales en Reconocimiento de Patrones*. Universidad de Salamanca. España.
9. S. Wallis, G. Nelson (2001). Knowledge Discovery in Grammatically Analysed Corpora. *Data Mining and Knowledge Discovery*. Kluwer Academic Publisher.