PLN
PROCESAMIENTO DEL LENGUAJE NATURAL EN LA RECUPERACIÓN DE INFORMACIÓN

Paula Andrea Benavides Cañón
Sandra Rodríguez Correa
UNIVERSIDAD DE LA SALLE – COLOMBIA
SISTEMAS DE INFORMACIÓN Y DOCUMENTACIÓN

RESUMEN

El lenguaje natural se entiende como el lenguaje hablado y escrito con el objetivo de que exista comunicación entre una o varias personas. La interpretación del lenguaje natural lo hace el cerebro, empieza a interpretar determinadas entradas sensoriales, tal como ver u oír alguna señal de alarma; el cerebro convierte la información codificada en un conjunto simbólico o lenguaje. La razón principal del procesamiento del lenguaje natural es construir sistemas y mecanismos que permitan la comunicación entre personas y máquinas por medio de lenguajes naturales.

PALABRAS CLAVES: Lenguaje natural, procesamiento, recuperación de información.

ABSTRACS

The natural language is understood as the language spoken and corresponded with the aim that communication exists between one or several persons. The interpretation of the natural language makes it the brain, starts interpreting certain sensory income, like to see or to hear some sign of alarm; the brain turns the information codified in a symbolic set or language. The principal reason of the processing of the natural language is to construct systems and mechanisms that allow the communication between persons and machines by means of natural languages.

KEY WORDS: natural language, processing, information retrieval.

Para poder hablar de procesamiento del lenguaje natural, se debe hablar en primer lugar de lenguaje natural y su dimensión en un entorno social. El lenguaje natural se entiende como el lenguaje hablado y escrito con el propósito que exista comunicación entre una o varias personas, es más directo para expresar lo que se quiere comunicar, el lenguaje natural es menos susceptible a malas interpretaciones por el empleo de términos con un solo significado. La comunicación es importante en el lenguaje natural debido a que este proceso involucra la transmisión y recepción de información.

La interpretación del lenguaje natural lo hace el cerebro, empieza a interpretar determinadas entradas sensoriales, tal como ver u oír alguna señal de alarma; el cerebro convierte la información codificada en un conjunto simbólico o lenguaje. El aprendizaje del lenguaje comienza con la repetición de palabras entendidas por su asociación con experiencias.

El lenguaje natural es un fenómeno muy complejo, pero ha sido sobradamente demostrado que las expresiones del lenguaje humano están organizadas a través de un conjunto de reglas. Todas nuestras expresiones tienen una clara organización: las palabras en una oración se asocian para describir objetos y acciones, posiblemente complejas. El objetivo de un analizador sintáctico es precisamente descubrir estas asociaciones entre palabras, lo que se conoce como estructura sintáctica. Un analizador sintáctico es un programa que toma como entrada una oración y trata de descubrir la estructura sintáctica que explica las relaciones entre las palabras de esa oración. Los analizadores buscan la estructura correcta dentro de un conjunto de análisis posibles, este conjunto esta usualmente definido por una gramática. El modelo de lenguaje en el cual se basa el analizador sintáctico decide cuáles son los componentes sintácticos de las oraciones y como éstos están relacionados. [1]

Procesamiento del Lenguaje Natural (PLN) es una subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional; ahora bien, la razón principal del PLN es construir sistemas y mecanismos que permitan la comunicación entre personas y máquinas por medio de lenguajes naturales.

El logro que una computadora aprenda a interpretar el lenguaje natural se debe a dos caminos, uno epistemológico y otro heurístico:

- El epistemológico: define el espacio de conceptos que el programa puede aprender.

- El heurístico: define los algoritmos para el aprendizaje.

El primer avance obtenido en el PLN se dio en el área del acceso a las bases de datos con el sistema lunar (1973) construidos en la NASA por William Woods.

El PLN busca poder crear programas que puedan analizar, entender y generar lenguajes que los humanos utilizan habitualmente, de manera que el usuario pueda llegar a comunicarse con la máquina o computador de la misma forma que lo haría con un ser humano.

La relación entre el PLN y la recuperación de información es evidente; su objetivo es la conversión del lenguaje natural al lenguaje maquina. El resulta en muchas ocasiones de esta relación es crear un buscador web en el que el

[1] **http://www.cs.famaf.unc.edu.ar/~pln/Proyectos/Parsing/parsing.html**

usuario pueda preguntar cualquier cosa, y el sistema sea capaz de responder de manera adecuada y correctamente, como lo haría un ser humano común y corriente, este proceso se denomina con el nombre sistemas de información question-answering.

El PLN es de manera general, un conjunto de instrucciones que un sistema recibe en un lenguaje de programación dado, que permita comunicarse con un humano en su propio lenguaje, este procesamiento presenta diversas aplicaciones:

- Corrección de textos.
- Traducción automática.
- Recuperación de la Información.
- Extracción de información y resúmenes.
- Búsqueda de documentos.
- Sistemas inteligentes para la educación y el entrenamiento.

Además de su utilidad en el campo del procesamiento y la recuperación de información, el PLN se aplica a otros aspectos como el reconocimiento del habla o la corrección ortográfica de textos.

EL PLN Y LA RECUPERACIÓN DE INFORMACIÓN

Como ya se ha dicho el PLN es lenguaje entre hombre y máquina, con el objetivo que la maquina le responda satisfactoriamente a la necesidad manifestada, ahora bien, existen diversos modelos asociados a esta recuperación de información que constituyen una herramienta que permite diferenciar una consulta previa y una seria de respuestas para dicha consulta.

En el funcionamiento de estos modelos se han implementado múltiples tipos de búsquedas con la razón fundamental, que los motores de búsqueda o la maquina pueda modelar las preguntas a lenguaje de maquina.

Hoy en día los modelos de recuperación de información tienen una alta importancia debido a los metabuscadores que existen en Internet, que por esta razón es necesario por lo menos entender como se estructuran internamente estos modelos de recuperación de información.

A continuación se relacionan los principales modelos de recuperación de información:

- **Modelo de recuperación booleano**: Es un modelo de recuperación basado en la teoría de conjuntos y el álgebra booleana de gran simplicidad. Su principal algoritmo de recuperación está fundamentado en un criterio de decisión binario sin ninguna noción de escala de medida ni ningún emparejamiento parcial en las condiciones de la consulta. En este modelo el método de representación, consiste en especificar los documentos como un conjunto de términos de indexación o keywords. El algoritmo utilizado en el modelo de recuperación booleano tiene como entrada dos listas ordenadas ascendentemente y como salida una lista ordenada con la mezcla de las dos listas de entrada. El método de ordenación es el número de identificación de los documentos que agrupan los términos a recuperar. Para ello se requiere una función que devuelva los identificadores de los documentos que contienen el término de la búsqueda, para lo cual se busca en el archivo invertido y luego se mezclan las listas.[2]

- **El modelo de recuperación vectorial:** se basa en la construcción de una matriz de términos y documentos, donde las filas contienen los documentos almacenados en una base de datos y las columnas se corresponden con los términos que se incluye en cada documento. De esta manera cada fila representa un vector que contiene los términos que aparecen en cada documento. De modo que un documento se expresa en forma de vector de esta forma doc1=(3,4,0,0,1..1) siendo cada uno de estos valores numéricos el número de veces que aparece cada término en el documento. La longitud del vector de documentos se corresponde con el número total de términos.[3]

- **El modelo de recuperación probabilístico:** se fundamenta en el cálculo de la probabilidad de que un documento sea relevante a la consulta proporcionada.

Características Principales:

- Los términos son independientes los unos de los otros
- Los pesos que se asignan a cada termino son binarios (1 o 0)

Para el cálculo de las probabilidades se toma la siguiente formula:

$$prob=n/N$$

[2] http://modelosderecuperacioni.iespana.es/

[3] http://modelosderecuperacioni.iespana.es/

Donde n es el número de documentos que son relevantes a una consulta y N el número de documentos totales del sistema.

La determinación de la relevancia se hace en base al número de términos coincidentes entre la consulta y los documentos de la base de datos del sistema. Para almacenar estos documentos se suelen emplear bases de datos nativas.[4]

Sin embargo, aunque se tengan identificados ciertos modelos en la recuperación de información, este proceso exige que el sistema comprenda en cierta medida el contenido del mismo, este razonamiento se ve apoyado en el hecho que el mayor problema en la recuperación de información es en la variación lingüística del lenguaje; un gran apoyo al PLN es la evolución del desarrollo tecnológico y su adaptación a diferentes lenguajes.

EL estudio, evolución y madurez del PLN tienes dos objetivos fundamentales que son:

- Facilitar la comunicación con la maquina para que puedan acceder diferentes usuarios desde aquel que posee mínimos conocimientos de consulta hasta el que especializado.
- Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje natural para diseñar sistemas que realicen tareas lingüísticas complejas (traducción, resúmenes de texto, etc.).

El objetivo del PLN en un sistema es extraer el significado de la consulta, es decir descomponer en términos de la red del dominio y representarla en forma de red semántica. El papel del conocimiento que juega en esta consecución de recuperación de información es fundamental, no solo para verificar que la pregunta que se formula es coherente, sino también para examinar y eliminar ambigüedades. Por esta razón el PLN interactúa constantemente con la red de dominio, su modelo semántico para permitirle orientar un adecuado procesamiento del lenguaje.

[4] http://modelosderecuperacioni.iespana.es/

CONCLUSIONES

El procesamiento del lenguaje natural tiene como objetivo fundamental lograr una comunicación máquina-humano similar a la comunicación humano-humano. El empleo del lenguaje le permite al hombre transmitir sus conocimientos, sentimientos sensaciones, emociones y estado de ánimo.

A lo largo de la historia los lenguajes naturales han ido evolucionando de forma paralela al desarrollo y evolución de la especie humana. Han sido varios los sistemas informáticos inteligentes que se han desarrollado que emplean el procesamiento del lenguaje natural.

BIBLIOGRAFÍA

Negrete J., De la filosofía a la Inteligencia Artificial. Grupo Noriega Editores, 1992. p. 199

Rusell S. Inteligencia Artificial: Un enfoque moderno. Prentice Hall, 1996. p. 729 – 740
http://www.cs.famaf.unc.edu.ar/~pln/Proyectos/Parsing/parsing.html

CARNAP, R., Logical Foundations of Probability. Chicago, University of Chicago Press, 1949.

HINTIKKA, J., «Game-theoretical semantics: insights and prospects», Notre Dame Journal of Formal Logic 23 (1982), 219-241.

JANSSEN, T., «Compositionality and the form of rules in Montague grammar», en Proceedings of the Second Amsterdam Symposium on Montague Grammar and Related Topics, Groenendijk, J. and Stokhoff, M. (eds.), 1978.

MC CAWLEY, J. D., Everything that Linguists have Always Wanted to Know about Logic. Chicago, University of Chicago Press, 1981.

NIINILUOTO, I., Truthlikeness. Sysnthese Library, vol. 185, Riedel, Holanda, 1987.

TARSKI, A., «The Semantic Conception of Truth and the Foundations of Semantics», en Philosophy and Phenomenological Research 4 (1944), 341-376.

ZADEH, L. A., «Test-score semantics for natural languages and meaning-representation via PRUF», Tech Note 247, AI Center, SRI International, Menlo Park, CA, 1981. También en: Empirical Semantics, Rieger, B. B. (ed.). Bochum, Brockmeyer, 1981, 281-349.

ZADEH, L. A., «Precisation of meaning via translation into PRUF», en Cognitive Constraints on Communication, Representation and Processes, Vaina, L. and Hintikka, J. (eds.), Dordrecht, Reidel, 1984.

ZADEH, L. A., «Test-score Semantics as a Basis for a Computational Approach to the Representation of Meaning», Literary and Linguistic Computing, vol. 1, núm. 1 (1986), 24-35.

http://modelosderecuperacioni.iespana.es/

http://www.cs.famaf.unc.edu.ar/~pln/

http://procesamientolenguajeresuperacion.50webs.org/

http://personales.com/ecuador/quito/intart/introduccion.htm