

FACULTAD DE INGENIERIA MECANICA Y ELECTRICA
PROCESAMIENTO DE LENGUAJE NATURAL

COMPONENTES DEL PROCESAMIENTO DE LENGUAJE NATURAL

PROFESOR: OSWALDO MORAN CARRILLO

ALUMNO:

ULISES YAREL GARCIA DE LOS SANTOS

6-D

26/05/2022

QUÉ ES EL PROCESAMIENTO DEL LENGUAJE NATURAL (PLN O NLP)

El procesamiento del lenguaje natural (PLN o NLP) es un campo dentro de la inteligencia artificial y la lingüística aplicada que estudia las interacciones mediante uso del lenguaje natural entre los seres humanos y las máquinas. Más concretamente se centra en el procesamiento de las comunicaciones humanas, dividiéndolas en partes, e identificando los elementos más relevantes del mensaje. Con la Comprensión y Generación de Lenguaje Natural, busca que las máquinas consigan entender, interpretar y manipular el lenguaje humano.

Los asistentes virtuales o chatbots son una de las utilidades más conocidas de la PLN, pero no son la única. Además, es importante entender que el PNL no dota de inteligencia a un chatbot, sólo le da la capacidad de procesar y generar lenguaje humano. En caso de querer dotar de inteligencia a un asistente virtual, habría que utilizar sistemas como reglas o redes neuronales.

Muchas veces cuando se habla de procesamiento de lenguaje natural, algunas personas únicamente lo relacionan con los chatbots, por eso vamos a ver otros usos del PLN.

Para qué se utiliza el Procesamiento del Lenguaje Natural (PLN o NLP)

El procesamiento del lenguaje natural (PLN o NLP), se utiliza actualmente en diferentes áreas y para distintas funciones, como por ejemplo:

Comprensión del lenguaje natural (CLN o NLU)

La comprensión del lenguaje natural (CLN o NLU) es la parte del procesamiento del lenguaje natural que se encarga de interpretar un mensaje y entender su significado e intención, tal y como haría una persona. Para que el sistema funcione necesita datasets en el idioma específico, reglas de gramática, teoría semántica y pragmática (para entender el contexto e intencionalidad), etc.

Generación del lenguaje natural (GLN o NLG)

La generación del lenguaje natural (GLN o NLG) dota a la máquina de la capacidad de crear un nuevo mensaje en lenguaje humano de manera autónoma. De manera resumida, lo que hacen estos modelos es: escoger la información a reproducir (dependiendo de la interpretación del mensaje a contestar), decidir cómo organizarla y cómo reproducirla (léxico y recursos gramaticales, morfología, estructuras sintácticas, etc.). Estos modelos generan frases nuevas palabra a palabra y tienen que ser entrenados para que funcionen correctamente.

Recuperación de información (RI o IR)

La recuperación de información (RI) o en inglés Information Retrieval (IR), es el campo dentro de la informática que se encarga de procesar textos de documentos, para poder recuperar partes específicas en base a palabras clave. Por ejemplo técnicas como la extracción de información estructurada (permite obtener de un documento el trozo de texto en el que está lo que buscas) o los sistemas de respuesta a preguntas de usuarios (que devuelve ante una consulta, una respuesta de una batería de respuestas ya existentes, asociadas a palabras clave de la consulta). No genera nuevas frases, por lo que no necesita utilizar reglas gramaticales. No es tan “inteligente” como la Generación del Lenguaje Natural.

Reconocimiento y síntesis del habla

Los sistemas de reconocimiento de voz procesan los mensajes en voz humana, los transforman en texto, los interpretan y comprenden la intencionalidad de los mismos, y tras la generación de la respuesta en texto, se vuelve a transformar en voz humana a través de la síntesis de voz. La síntesis del habla o de voz, es la que capacita a la máquina para poder generar y reproducir habla en lenguaje natural.

Traducción automática

La Traducción automática o Machine Translation en inglés, es un campo de investigación dentro de la lingüística computacional que estudia los sistemas capaces de traducir mensajes entre diferentes lenguas o idiomas. Por ejemplo Google es una de las empresas que más ha invertido en sistemas de traducción automática, con su traductor que utiliza un motor estadístico propio. Los sistemas de autocorrección y autocompletado de texto, también utilizan Procesamiento del Lenguaje Natural (PLN o NLP).

Resumen y clasificación de textos

También se está utilizando el procesamiento del lenguaje natural para resumir textos de extensiones largas de manera automática o extraer palabras clave para clasificarlos. Muchas veces, debido a la gran cantidad de documentación o por la longitud de la misma, utilizar estos sistemas ayuda en sectores como el legal a encontrar partes dentro de las leyes, o resumir una gran cantidad de documentación.

Otro de los usos que se le da a esta función de clasificación, es la de detección de spam. Empresas como Google utilizan esta tecnología para clasificar los

textos de los correos electrónicos y detectar si se trata de spam o no. Para esto, toman palabras clave como “gratis” o “descuento”, la condición de palabras en mayúscula o las exclamaciones.

Detección de sentimientos o emociones

Uno de los usos más novedosos del PLN es el análisis de sentimientos. Cada vez más empresas y profesionales del marketing están utilizando esta tecnología para saber qué sienten los usuarios sobre una marca, producto o servicio, utilizando datos de entrada como mensajes, comentarios o reacciones en diferentes redes sociales.

COMPONENTES

Análisis morfológico

Su función consiste en detectar la relación que se establece entre las unidades mínimas que forman una palabra, como puede ser el reconocimiento de sufijos o prefijos. Este nivel de análisis mantiene una estrecha relación con el léxico.

El léxico es el conjunto de información sobre cada palabra que el sistema utiliza para el procesamiento. Las palabras que forman parte del diccionario están representadas por una entrada léxica, y en caso de que ésta tenga más de un significado o diferentes categorías gramaticales, tendrá asignada diferentes entradas.

En el léxico se incluye la información morfológica, la categoría gramatical, irregularidades sintácticas y representación del significado.

Normalmente el léxico sólo contiene la raíz de las palabras con formas regulares, siendo el analizador morfológico el que se encarga de determinar si el género, número o flexión que componen el resto de la palabra son adecuados.

Análisis sintáctico

Tiene como función etiquetar cada uno de los componentes sintácticos que aparecen en la oración y analizar cómo las palabras se combinan para formar construcciones gramaticalmente correctas. El resultado de este proceso consiste

en generar la estructura correspondiente a las categorías sintácticas formadas por cada una de las unidades léxicas que aparecen en la oración.

Análisis semántico

En muchas aplicaciones del PLN los objetivos del análisis apuntan hacia el procesamiento del significado. En los últimos años las técnicas de procesamiento sintáctico han experimentado avances significativos, resolviendo los problemas fundamentales.

Sin embargo, las técnicas de representación del significado no han obtenido los resultados deseados, y numerosas cuestiones continúan sin encontrar soluciones satisfactorias.

Definir qué es el significado no es una tarea sencilla, y puede dar lugar a diversas interpretaciones. A efectos funcionales, para facilitar el procesamiento, la modularidad es una de las propiedades más deseables. Haciendo uso de esta concepción modular es posible distinguir entre significado independiente y significado dependiente del contexto.

El primero, tratado por la semántica, hace referencia al significado que las palabras tienen por sí mismas sin considerar el significado adquirido según el uso en una determinada circunstancia. La semántica, por tanto, hace referencia a las condiciones de verdad de la frase, ignorando la influencia del contexto o las intenciones del hablante. Por otra parte, el componente significativo de una frase asociado a las circunstancias en que ésta se da, es estudiado por la pragmática y conocido como significado dependiente del contexto.

Atendiendo al desarrollo en el proceso de interpretación semántica, es posible optar entre múltiples pautas para su organización, tal como se determinan en los siguientes párrafos.

En referencia a la estructura semántica que se va a generar, puede interesarnos que exista una simetría respecto a la estructura sintáctica, o por el contrario que no se dé tal correspondencia entre ellas. En el primer caso, a partir del árbol generado por el análisis sintáctico se genera una estructura arbórea con las mismas características, sobre la cual se realizará el análisis semántico. En el segundo caso, en la estructura generada por la sintaxis se produce un curso de transformaciones sobre las cuales se genera la representación semántica.

Cada una de las dos opciones anteriores puede implementarse de forma secuencial o paralela. En la interpretación secuencial, después de haber finalizado la fase de análisis sintáctico, se genera el análisis semántico. En cambio, desde un procedimiento en paralelo, el proceso de análisis semántico no necesita esperar a que el analizador sintáctico haya acabado toda su tarea, sino que puede ir realizando el análisis de cada constituyente cuando éste ha sido tratado en el proceso sintáctico.

Finalmente en combinación con cada una de las opciones anteriores, podemos escoger un modelo en el que exista una correspondencia entre reglas sintácticas y semánticas o, contrariamente, podemos optar por un modelo que no cumpla tal requisito. En caso afirmativo, para cada regla sintáctica existirá una regla semántica correspondiente.

El significado es representado por formalismos conocidos por el nombre de knowledge representation. El léxico proporciona el componente semántico de cada palabra en un formalismo concreto, y el analizador semántico lo procesa para obtener una representación del significado de la frase.

Análisis pragmático

Añade información adicional al análisis del significado de la frase en función del contexto donde aparece. Se trata de uno de los niveles de análisis más complejos, la finalidad del cual es incorporar al análisis semántico la aportación significativa que pueden hacer los participantes, la evolución del discurso o información presupuesta.

Incorpora así mismo información sobre las relaciones que se dan entre los hechos que forman el contexto y entre diferentes entidades.