Q-SENECA (Un sistema pregunta respuesta en castellano)

M. F. Verdejo - Facultad de Informática, Universidad del País Vasco

Presentación

Desde la perspectiva de la inteligencia artificial, el estudio del lenguaje natural tiene dos objetivos:

1) Facilitar la comunicación con el ordenador para que accedan al mismo usuarios no especialistas

2) Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas linguísticas complejas (traducción, resúmenes de textos, etc.)

Hay problemas en los que interesa fundamentalmente el primer objetivo. Lo que se desea es conseguir un intérprete para una clase de aplicaciones en un dominio restringido, que haga de traductor entre el ordenador y el usuario. El intérprete realiza dos tareas: una, de reconocimiento de la pregunta del usuario, otra de generación de una expresión equivalente en el lenguaje formal que utilice el ordenador para la aplicación. Este enfoque modela el lenguaje como una herramienta de comunicación (usuario-sistema) sobre conjuntos de información tipificada y restringida y, por tanto, sólo tratará el subconjunto del lenguaje natural que describirá los aspectos significativos en ese dominio. La comprensión de una consulta se plantea de este modo como un proceso de "detección" de un texto, de la información relevante.

El segundo objetivo se plantea el lenguaje como objeto de estudio, y la comprensión como un proceso complejo en el que intervienen grandes cantidades de conocimiento de naturaleza diferente (morfología, sintaxis, semántica, pragmática) y mecanismos de tratamiento variados (de comparación, búsqueda, inferencia aproximada, deducción...).

Algunos de estos aspectos son también objeto de estudio por parte de linguistas, psicólogos, lógicos y filósofos, y por ello las aportaciones de unos y otros son fundamentales en la elaboración de teorías que den paso a la formalización de algunos fenómenos del proceso de comprensión. Desde la Inteligencia Artificial se aborda el problema con una óptica propia: la teoría debe permitir construir un sistema automático, que no tiene por qué estar basado necesariamente en una simulación del procesador humano.

El propósito de nuestro trabajo es la construcción de un sistema pregunta-respuesta que acepta consultas en castellano y es capaz de responderlas.

Un sistema pregunta-respuesta se compone de dos fases. La primera es un proceso de análisis o comprensión de la pregunta. La segunda, es un proceso de resolución o construcción de la respuesta.

La potencia de un sistema pregunta-respuesta depende fundamentalmente del modelo de representación del conocimiento sobre el que está basado. Nuestro sistema se basa en un modelo general en forma de red semántica: SENECA (G. Camarero & G. Sanz & M. F. Verdejo 79 a,b). Para cada aplicación permite representar el conocimiento específico, dotándole de una estructura que facilita la búsqueda y deducción de información, procesos esenciales en la comprensión de la pregunta y generación de la respuesta.

Este método de representación es aplicable a diferentes dominios y en particular lo hemos utilizado para construir un sistema automático de consulta a un banco de datos formado por textos literarios, de los que se quiere resultados estadísticos, usualmente requeridos en linguística computacional.

En este dominio, el conocimiento se descompone en las siguientes unidades conceptuales:

Objetos: obras literarias, y en general, textos que representamos por nodos 

Propiedades: de una obra literaria: tener título, autor, ficha, editorial, etc. De un texto: su codificación, índice... Las representaremos por nodos 

Acciones: operaciones que pueden efectuarse sobre los objetos, por ejemplo, hallar frecuencias, contextos ... y las representamos por nodos 

Estas unidades están relacionadas entre sí, formando la red de la figura 1, que constituye el conocimiento conceptual que el sistema tiene acerca del dominio.

Caracterización

El subconjunto del castellano que tratamos comprende las frases que se utilizan normalmente en consultas a bases de datos. Consideramos tres tipos. El primero, está formado por las consultas acerca de los objetos del dominio, ya sea para preguntar sobre su existencia, sobre el número de objetos que verifican una propiedad o relación o bien la especificación de un objeto del que se da una descripción parcial. Ejemplos son las preguntas:

Pregunta 1. ¿Hay alguna obra literaria del Siglo de Oro?

Pregunta 2. ¿Cuáles son los adjetivos que aparecen en las Soledades?

Pregunta 3. ¿Hay alguna palabra de frecuencia mayor de 100 en la Realidad y el Deseo?

Pregunta 4. ¿Cómo está codificado el libro del Buen Amor?

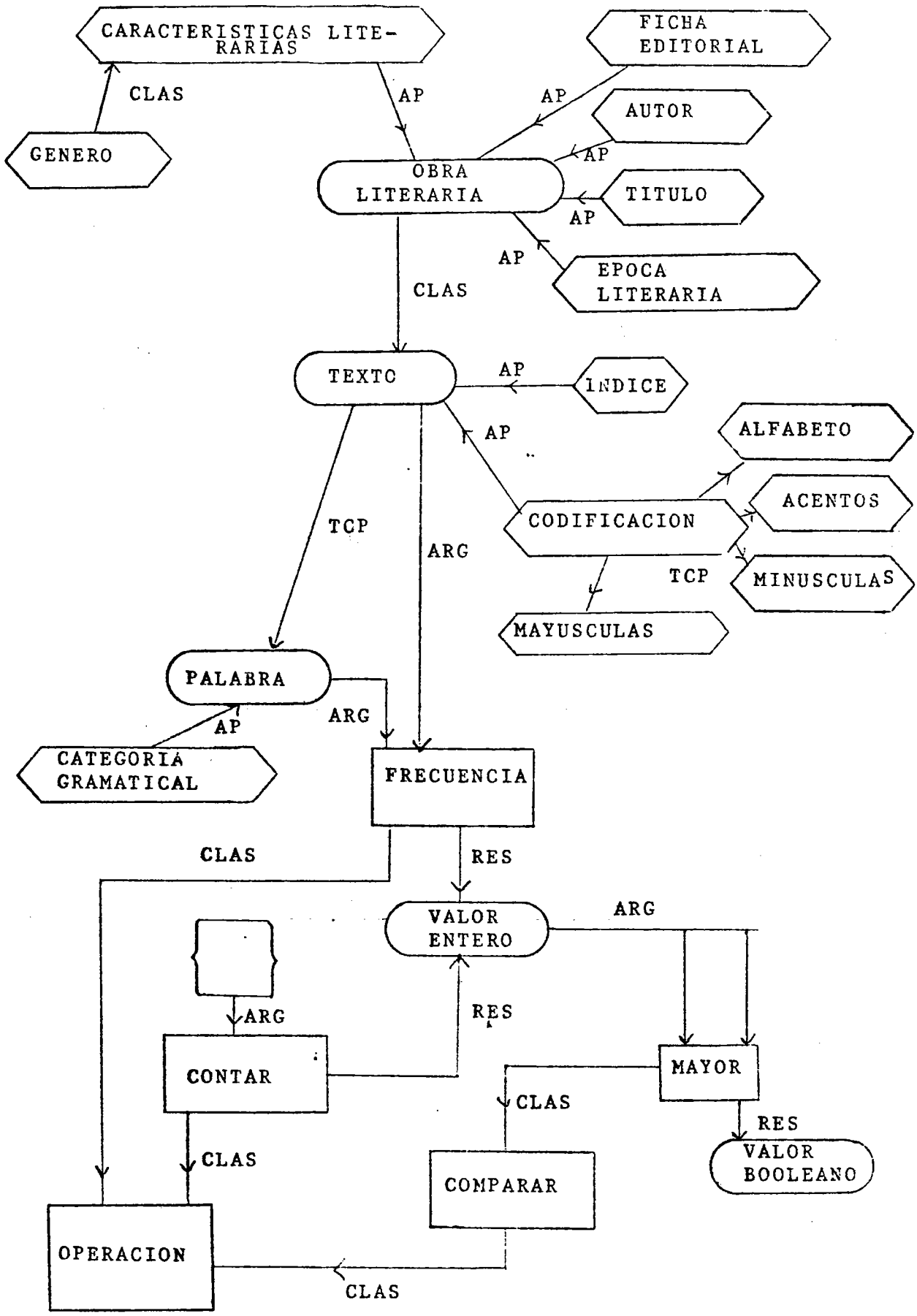
CARACTERISTICAS LITE-
RARIAS
FICHA
EDITORIAL
CLAS
AP
AP
AUTOR
GENERO
AP
OBRA
LITERARIA
TITULO
AP
AP
EPOCA
LITERARIA
CLAS
TEXTO
AP
INDICE
AP
ALFABETO
ACENTOS
TCP
ARG
CODIFICACION
TCP
MINUSCULAS
MAYUSCULAS
PALABRA
ARG
AP
CATEGORIA
GRAMATICAL
FRECUENCIA
CLAS
RES
VALOR
ENTERO
ARG
ARG
RES
CONTAR
MAYOR
CLAS
RES
CLAS
VALOR
BOOLEANO
COMPARAR
OPERACION
CLAS

Figura 1.

Análisis de la consulta

El analizador acepta una consulta cuando es capaz de representarla en forma de red semántica. En este proceso entran en juego diferentes tipos de conocimiento: sobre el léxico, reglas sintácticas y semántica del lenguaje, por una parte. Sobre la propia representación por otra. Nuestro analizador tiene como modelo semántico el plano conceptual de la red del dominio problema. Una parte del léxico está asociado mediante un diccionario a elementos (nodos o subredes) de dicha red.

El objetivo del analizador es extraer el significado de la consulta, es decir descomponerla en términos de la red del dominio y representarla en forma de red semántica.

En la consecución de esta tarea el papel que juega el conocimiento que se tiene sobre el dominio es fundamental, no solo para verificar que la pregunta que se formula es coherente con él mismo, sino también para levantar posibles ambiguedades. Por esta razón el analizador interacciona continuamente con la red dominio, su modelo semántico, que le permite dirigir y orientar el análisis.

El proceso se basa sobre criterios semánticos, en el sentido de que se intenta detectar el elemento conceptual sobre el que se pregunta, y una vez localizado en la red dominio, su entorno sirve de predictor sobre las posibles referencias que pueden darse en la descripción de un objeto concreto ligado a ese concepto.

La sintaxis, se utiliza en la medida en la que provee información importante desde el punto de vista semántico, y en particular para agrupar palabras en UNIDADES semánticas, para eliminar posibles ambiguedades, y sobre todo para determinar el ámbito de los cuantificadores permitiendo una imbricación correcta de las subredes generadas durante el análisis.

El analizador es una ATN semántica, ya que lo que se obtiene es la representación (en forma de red) de la consulta, y lleva a cabo únicamente las acciones ligadas a este fin.

Conclusión

Nuestro trabajo (Verdejo 80 a,b,c) se enmarca en un proyecto de construcción de un sistema pregunta-respuesta cuyo lenguaje de consulta es el castellano. El interés del proyecto presenta un doble aspecto:

Teórico - Estudio de la representación del conocimiento y los procesos de comprensión del lengueje natural.

Práctico - Aplicar el estudio teórico a la problemática de la comunicación con el ordenador en dominios habitualmente tratados en bases de datos.

En esta perspectiva, el analizador cubre una primera etapa, ya que acepta un subconjunto del castellano, que comprende las frases normalmente utilizadas en este tipo de sistemas.

Resaltamos que su comportamiento es independiente del dominio en que se trabaja y está ligado al formalismo de representación (red semántica) con sus mecanismos asociativos y de inferencia.

Actualmente, el sistema recibe la representación del dominio como un dato inicial. Estamos ampliando el subconjunto del lenguaje, para aceptar frases declarativas de forma que la red dominio pueda crearse de forma interactiva en un diálogo usuario-sistema.

Referencias

G. Camarero & G. Sanz & M. F. Verdejo (79 a)
Representación del conocimiento mediante redes semánticas
CIL 79. Pg 457-466

G. Camarero & G. Sanz & M. F. Verdejo (79 b)
Resolución de sistemas de ecuaciones con variables en el conjunto de los contornos de un grafo semántico
4 Congreso AEIA, Madrid. Pg. 711-717

M. F. Verdejo (80 a)
Un analizador semántico para un sistema pregunta respuesta en lenguaje natural. Tesis doctoral. UCM.

M. F. Verdejo (80 b)
Un sistema de producción para la imbricación de redes semánticas. Revista de Informática y Automática N. 45

M. F. Verdejo (80 c)
Análisis predictivo y representación del conocimiento
Convención Informática Latina. Barcelona 1980