Contextualización del aprendizaje automático en procesamiento del lenguaje natural

Autores: Sergio Rafael Flores, Ilda Flavia Millán, Susana Ruiz

Instituto de Informática - Departamento de Informática

Facultad de Ciencias Exactas, Físicas y Naturales

Universidad Nacional de San Juan

Domicilio: Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan CPA: J5402DCS

Teléfono: 4260353 4260355 - 4260394 - 4264721 - 4234129, Fax 0264-4234980

Emails: sflores@iinfo.unsj.edu.ar flavia.millan@gmail.com sbruiz@yahoo.com.ar

Resumen

En la actualidad, la implementación de la gramática en forma de un analizador automático permite no solo verificar si la misma es completa o no, sino, además, medir cuantitativamente en qué grado es completa y exactamente que productividad tiene cada una de sus reglas. Cuando se habla de las aplicaciones de la Teoría de la Computación, se puede citar que la Computación le proporciona a la Lingüística un interlocutor, es decir la computadora, con características singulares. Por lo anterior resulta evidente los aportes que realiza la Teoría de la Computación a la Lingüística, en el Procesamiento Automático del Lenguaje Natural (PLN), que se ocupa más de los aspectos técnicos, algorítmicos y matemáticos con la aplicación de modelos de la Inteligencia Artificial para proponer el desarrollo de una arquitectura que posibilite la interacción entre el ser humano y la computadora contextualizando el discurso escrito.

Palabras Claves: gramática, analizador sintáctico automático, Procesamiento del Lenguaje Natural.

Contexto

La línea de investigación es Teoría de la Computación, más específicamente la construcción de un Modelo de Procesamiento del Lenguaje Natural y en este ámbito, la contextualización del mismo. Las Instituciones que coordinan el proyecto son el Instituto de Informática y el Departamento de Informática de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan, y se encuentra en vías de evaluación.

Introducción

Si bien es cierto que "enseñarle" a la máquina puede llegar a ser una tarea complicada, se debe reconocer que una vez que ésta "ha aprendido", la acción se convierte en un logro inapreciable. Esto se muestra, por ejemplo, en la forma en que puede facilitar el trabajo del lingüista gracias a su gran capacidad de memoria y de análisis [1]. Estas y otras tareas que antes hubieran requerido muchas horas de trabajo y muchas personas, en la actualidad y debido a la perfomance referida al procesamiento, son realizadas en unos instantes.

La ciencia que estudia el lenguaje humano se llama Lingüística. A su vez, esta ciencia se relaciona con otros campos, como la tecnología, la educación, la medicina, entre otros. Es importante aclarar que existe una relación muy estrecha entre la Lingüística y la Computación. Por un lado, el conocimiento lingüístico es la base teórica para el desarrollo de una amplia gama de aplicaciones tecnológicas, cada vez más importantes en la actualidad, como por ejemplo: aplicaciones en procesos educativos mediados por tecnologías, lograr interfaces entre el lenguaje natural y las computadoras, comunicación entre humanos y robots, por citar algunas.

Por otro lado, las Ciencias de la Computación pueden aportar a las ciencias de la Lingüística, en tiempo y capacidad de procesamiento. Noam Chomsky comenzó a estudiar la estructura del lenguaje natural. Estos estudios condujeron a la clasificación de los lenguajes de acuerdo con la complejidad de sus gramáticas, es decir, las reglas que especifican su estructura, y la potencia de los algoritmos necesarios para reconocerlas. Es por ello, que ahora resulta importante los beneficios de las Ciencias de la Computación en cuanto a la posibilidad de la verificación masiva de las teorías, las gramáticas y los diccionarios lingüísticos.

En la actualidad, la implementación de la gramática en forma de un analizador automático permite no solo verificar si una gramática es completa o no, sino, además, medir cuantitativamente en qué grado es completa y exactamente que productividad tiene cada una de sus reglas. Cuando se habla de las aplicaciones de la Teoría de la Computación, se puede citar que la Computación le proporcionó a la Lingüística un interlocutor, es decir la computadora, con características singulares, no conoce nada de antemano, cabe aclarar que tampoco posee intuición, ni sentido común y tan solo es capaz de interpretar y aplicar literalmente las descripciones del lenguaje que el lingüista le proporciona. Lo anterior permite que la computadora solicite al lingüista afinar y completar sus formulaciones a partir de la búsqueda de respuestas a preguntas tan difíciles de responder, que para el humano resulta “obvias”. Se puede llegar a pensar que la Computación convierte a la Lingüística, en una ciencia exacta, dándole una nueva motivación y nuevas propuestas de investigación [2].

Siguiendo con los aportes de la Computación a la Lingüística, es posible mencionar la intersección que existe entre la lingüística y la teoría de la computación. Esta última es una ciencia que trata la construcción de modelos de lenguaje para las computadoras, es decir, lenguajes más formales que los modelos tradicionales orientados a los interlocutores humanos. Otra área es el procesamiento automático de lenguaje natural (PLN), que se ocupa más de los aspectos técnicos, algorítmicos y matemáticos de la aplicación de dichos modelos. Estas dos disciplinas tienen el mismo objeto de investigación, aunque lo consideran desde enfoques diferentes. Un sistema de PLN simula de forma parcial el comportamiento lingüístico humano [3]. Para ello debe modelizar tanto las estructuras propias del lenguaje, como el conocimiento general acerca del universo del discurso, asimismo de establecer mecanismos de razonamiento.

Normalmente, el procesamiento del lenguaje natural se divide en seis etapas [4]:

1. fonética / fonología
2. morfología
3. sintaxis
4. semántica
5. pragmática
6. discurso

1. FONÉTICA / FONOLOGÍA

La fonética es la parte de la Lingüística que se dedica a la exploración de las características del sonido, que es un elemento substancial del lenguaje. Eso determina que los métodos de fonética sean en su mayoría físicos; por eso su

posición dentro de la Lingüística es bastante independiente.

La prosodia tiene por objeto un dominio muy amplio que comprende el estudio de diversos fenómenos asociados al acento, al ritmo y a la entonación, así como a sus manifestaciones físicas producto de las variaciones de la duración, de la frecuencia fundamental y de la intensidad

2. MORFOLOGÍA

El área de morfología es la estructura interna de las palabras (sufijos, prefijos, raíces, flexiones) y el sistema de categorías gramaticales de los idiomas (género, número, etc.). Hay lenguas que tienen muchas diferencias en relación con las reglas que se tiene en el español.

3. SINTAXIS

La sintaxis se dedica a analizar las relaciones entre las palabras dentro de la frase. Existen dos modelos principales para la representación de tales relaciones:
1) dependencias, donde las relaciones se marcan con flechas y una palabra puede tener varias que dependen de ella, y
2) constituyentes, donde las relaciones se pueden representar en forma de árbol binario.

4. SEMÁNTICA

El propósito de la semántica es «entender» la frase. ¿Pero qué significa «entender»? Hay que saber el sentido de todas las palabras e interpretar las relaciones sintácticas. Los investigadores están más o menos de acuerdo que los resultados del análisis semántico deben ser redes semánticas, donde se representan todos los conceptos y las relaciones entre ellos. Otra tarea de la semántica (o más bien, de sus subdisciplinas llamadas lexicología y lexicografía) es definir los sentidos de las palabras, lo que representa de por sí una tarea muy difícil, aún cuando se realiza manualmente. Los resultados de la definición de los sentidos de las palabras existen en forma de diccionarios.

Una aplicación importante del análisis semántico es la desambiguación automática de sentidos de palabras. Por ejemplo, un gato puede ser un felino, o una herramienta, o una persona. Para saber cuál de los sentidos se usa en un contexto dado se pueden aplicar diferentes métodos con el fin de analizar las demás palabras presentes en el contexto. Por ejemplo, en la frase “El gato se acostó en el sillón y estaba maullando”, las palabras acostarse y maullar indican que es un felino; mientras que en la frase “El mecánico usó un gato para subir el automóvil”, las palabras mecánico, subir y automóvil dan la preferencia al sentido una herramienta. Sin embargo, en la frase “El mecánico compró un gato y lo llevó en su carro”, no se puede definir el sentido, a menos que se amplíe el contexto.

5. PRAGMÁTICA

Usualmente se dice que la pragmática trata de las relaciones entre la oración y el mundo externo. Un ejemplo famoso es el siguiente: “usted y yo estamos comiendo juntos y yo le pregunto a usted si puede pasarme la sal, usted contesta que sí... y sigue comiendo.” Seguramente la respuesta es formalmente correcta, porque usted realmente puede pasarme la sal y eso es lo que contiene literalmente la pregunta, pero la intención fue pedir la sal y no preguntar sobre la posibilidad de pasarla. De otra manera, se puede decir que lo que interesa a la pragmática son las intenciones del autor del texto o del hablante.

6. DISCURSO

Normalmente no se habla con una oración aislada, sino con varias oraciones. Esas oraciones tienen ciertas relaciones entre sí. Las oraciones hiladas forman una nueva entidad llamada discurso.

En el análisis del discurso existe un problema muy importante: la resolución de correferencia. Las relaciones de correferencia también se llaman anafóricas. Por ejemplo, en el discurso «He visto una nueva casa ayer. Su cocina era excepcionalmente grande» (su = de la casa); o «Llegó Juan. Él estaba cansado» (él = Juan). Esas son relaciones de correferencia, y la computadora tiene que interpretarlas correctamente para poder construir las representaciones semánticas. Existen algoritmos de resolución de correferencia bastante buenos, donde se alcanza hasta 90% de exactitud, sin embargo, resolver el 10% restante todavía es una tarea difícil.

Líneas de Investigación

Para realizar una implementación de lo detallado en el punto anterior, se requiere de la modelización de cada una de las etapas citadas anteriormente, es decir la implementación de un procesador lingüístico. La estructura general del procesador lingüístico —el programa que hace el análisis de los textos— corresponde a los niveles del lenguaje; la excepción es el nivel fonético, porque el texto ya está representado con palabras escritas y no con sonidos. Las fases previstas serán:

Fase 1. Transformación morfológica entre las palabras. En este paso se resuelven las secuencias de letras en la llamada representación morfológica del texto: la secuencia de las estructuras de palabras en la forma del lema (que puede servir como una clave a una base de datos que guarda todas las propiedades de la palabra) y las propiedades específicas en el texto: ←fuera SER, subjuntivo, tercera persona, singular.

Fase 2. Transformación sintáctica entre la representación morfológica y la representación sintáctica. Ésta última es una secuencia de estructuras de oraciones, siendo una estructura de oración un árbol sintáctico que representa qué palabras están relacionadas sintácticamente a cuáles otras en la misma oración. Para este paso se usan los diccionarios sintácticos y los métodos matemáticos de gramáticas libres de contexto; los algoritmos que aplican tales gramáticas para el análisis del texto se llaman parsers [5].

Fase 3. Transformación semántica entre la representación sintáctica (la secuencia de árboles) y la representación semántica (la red semántica). En este paso se identifican las palabras que refieren a la misma entidad (o situación). Por ejemplo, en el texto: “Juan sacó 8 en el examen. Esto desanimó mucho al pobrecito”, se tiene que detectar que el desanimado es Juan, lo que técnicamente consiste en mapear las dos frases —Juan y el pobrecito— al mismo nodo (entidad) de la red semántica; también se debe mapear esto: “y sacó 8 en el examen”, al mismo nodo (situación).

El tema propuesto de la investigación, pretende contextualizar el intercambio de discurso escrito entre una computadora y un humano de modo que la máquina responda lo más apropiadamente posible al intercambio. Para ello se deberá procesar el lenguaje natural comenzando el análisis en niveles más simples y construyendo las representaciones de un nivel superior dado en base a los niveles anteriores. En cada paso existen problemas técnicos y teóricos, algunos ya resueltos en cierto grado y algunos por resolverse en el transcurso y desarrollo de la investigación. Por lo tanto, la tarea propuesta consiste en que el sistema computacional “comprenda” la conversación y además “contextualice” la misma mantenida con un humano, lo anterior corresponde a una parte de la línea de investigación de Procesamiento del Lenguaje Natural (PLN), área que se está investigando a nivel mundial.

Es obvio que para realizar todo el proceso correctamente, el sistema debe entender el texto, es decir, construir las redes semánticas. Esto es algo que los sistemas modernos no saben hacer todavía. Aunque en la época actual existen las aplicaciones mencionadas, normalmente se basan en

heurísticas, en especial si se trata de los niveles semántico o pragmático; es decir, las aplicaciones funcionan y son útiles, pero todavía no alcanzan la calidad que se desea (Gelbukh, A. y Sidorov, G. 2006).

Resultados y Objetivos

Como resultado de la investigación se pretende realizar tareas de Transformación morfológica entre las palabras. Para este paso, se usan los diccionarios morfológicos y los métodos matemáticos de autómatas de estados finitos como un scanner. Posteriormente, se efectúa la Transformación sintáctica entre la representación morfológica y la representación sintáctica. Aquí, se usan los diccionarios sintácticos y los métodos matemáticos de gramáticas libres de contexto.

Por último, se persigue la Transformación semántica entre la representación sintáctica (la secuencia de árboles) y la representación semántica (la red semántica). Es decir, se identifican las palabras que refieren a la misma entidad (o situación). Para la transformación semántica se usan los diccionarios semánticos, las reglas de transformación y los métodos de inferencia lógica. Teniendo en cuenta casos que puedan reflejar la ambigüedad.

El objetivo principal de la investigación a desarrollar consiste en “Proponer una arquitectura que posibilite la interacción entre el ser humano y la computadora contextualizando el discurso escrito”.

Formación de Recursos Humanos

Se pretende formar recursos en el área de Teoría de la Computación, esto hace referencia tanto a docentes como alumnos de la Facultad de Ciencias Exactas, Físicas y Naturales, en el Departamento de Informática, cuya área Curricular es relativamente nueva en el Departamento. Para tan fin se promoverá la realización de tesis de posgrado y tesinas de grado.

Referencias

[1] GELBUKH, A. y SIDOROV, G. (2006). Procesamiento Automático del español con enfoque en recursos léxicos grandes.

[2] GELBUKH, A. y SIDOROV, G. (2006). Procesamiento Automático del español con enfoque en recursos léxicos grandes.

[3] MOLINA, M. (2004). Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático. Dpto. de Sistemas Informáticos y computación. Universidad Politécnica de Valencia. España.

[4] GELBUKH, A. y SIDOROV, G. (2006). Procesamiento Automático del español con enfoque en recursos léxicos grandes.

[5] HOPCROFT, J. y ULLMAN, J. (1997). Introducción a la Teoría de Autómatas y Computación.