

Descubrimiento Automático de Información Semántica en Documentos Escritos en la Lengua Indígena Náhuatl

Integrantes:

Dra. Carmen Carlota Martínez Gil. Universidad de la Cañada. (Directora del Proyecto)
Lic. Alejandro Zepolteca Pérez. Universidad de la Cañada.
M. en C. Venustiano Soancatl Aguilar. Universidad del Istmo.
M. en C. Rosa María Ortega Mendoza. Instituto Tecnológico Superior del Oriente del Estado de Hidalgo.

Objetivo general:

Descubrir información semántica de forma automática de documentos escritos en náhuatl.

Introducción

El náhuatl es la lengua indígena más hablada en el territorio mexicano, alrededor de un millón y medio, de acuerdo a los reportes del INEGI. Además, es y ha sido un idioma valioso por su importancia histórica, lingüística, literaria y nacionalista.

El náhuatl es uno de los lenguajes de América más estudiado y documentado, se pueden encontrar una gran cantidad de documentos escritos en náhuatl de los cuales podemos extraer información importante y valiosa para la presente y futuras generaciones.

Nuestro interés por desarrollar recursos y herramientas para el náhuatl se centra principalmente en:

- a) Debido al desconocimiento y mal manejo de los términos de ésta lengua, estamos perdiendo gran parte de nuestra cultura. Por ello, es importante conocer y extraer la información contenida en documentos escritos en náhuatl;
- b) Preservar y difundir un idioma con raíces históricas, para que ésta no desaparezca ya que si la perdemos, parte de nuestra esencia e identidad como mexicanos se perderá;
- c) Para poder comprender y entender las culturas que usan ésta lengua y poder comunicarnos con las personas que sólo hablan náhuatl.

En este proyecto planteamos el desarrollo de un sistema para extraer información semántica de documentos escritos en náhuatl, con el objetivo a futuro de desarrollar herramientas más sofisticadas tales como: traductor náhuatl-español, sistemas de extracción de información, clasificación de textos, sistema de búsqueda de respuestas en documentos escritos en náhuatl, entre otros.

Áreas relacionadas con el proyecto

Para poder implementar estas herramientas más complejas es necesario primero desarrollar los recursos lingüísticos (conjunto de datos del lenguaje en formato legible por la computadora y que son usados en la construcción, mejoramiento y evaluación de los sistemas del lenguaje natural), los cuales se clasifican en:

- **Corpus** (colección de textos en lenguaje natural, elegido para caracterizar un estado o variedad de un lenguaje y actúa como repositorio de información de cual se puede manipular su contenido para extraer conocimiento).

- **Herramientas** (ayudan a analizar los textos, entre las más comunes están: etiquetadores de partes de la oración, analizadores morfológicos y analizadores sintácticos).
- **Recursos léxicos** (contienen un conjunto de palabras válidas en el lenguaje, así como también pueden contener propiedades lingüísticas, el significado de las palabras y relaciones entre las palabras o grupo de palabras. Algunos ejemplos de recursos léxicos son: listas de palabras, diccionarios, tesauros, ontologías, glosarios, entre otros).

Las dos principales áreas de investigación en Ciencias de la Computación relacionadas con el desarrollo de este proyecto son:

- **Procesamiento de Lenguaje Natural:** sub-disciplina de la Informática y la Lingüística, que se encarga de producir sistemas informáticos que posibiliten la comunicación hombre-hombre u hombre-máquina mediante el lenguaje natural. El objetivo de PLN es estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural.
- **Aprendizaje Automático:** (definición según Tom Mitchell, El objetivo del Aprendizaje Automático es desarrollar técnicas que permitan a las computadoras *aprender*. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos).

Resultados preliminares

Este proyecto se inició en Septiembre de 2009, financiado por el Programa de Mejoramiento del Profesorado (PROMEP) de la Subsecretaría de Educación Superior. Los resultados y avances obtenidos hasta el momento son:

1. Elaboración de un *Corpus* de documentos escritos en náhuatl-español. Dichos textos primero fueron investigados, después digitalizados y por último estandarizados y organizados en cuatro categorías: Poesía, Religión, Cuentos y Varios. En Varios tenemos textos relacionados con leyendas, narraciones, crónicas, reflexiones, testamentos, entre otros.
2. Desarrollo de recursos léxicos para el lenguaje indígena náhuatl. En esta etapa construimos tres diccionarios: uno de términos náhuatl-español, otro con palabras en náhuatl y cuya traducción al español es una frase y por último un diccionario con términos específicos del náhuatl de la región del norte de Oaxaca.
3. Se ha diseñado e implemento un sistema de software para representar oraciones en náhuatl y sus respectivas etiquetas en un árbol sintáctico. Así como también se desarrollo un sistema de software para obtener prefijos y sufijos de las palabras en náhuatl de un texto.
4. Finalmente, se implemento un sistema informático para agregar información semántica a los términos en náhuatl.

Participación en Eventos de Divulgación

A continuación se listan los eventos donde se ha participado presentando avances y resultados del proyecto de investigación:

No.	EVENTO	DESCRIPCIÓN DE LA ACTIVIDAD	INSTANCIA / ORGANIZACIÓN EN QUE SE REALIZÓ	FECHA
1	4to. Día de Sistemas del ITSA	Taller "Estudio de los algoritmos de Aprendizaje Automático usando Textos en Náhuatl en el Sistema WEKA"	Instituto Tecnológico Superior de Atlixco	14-Jun-11
2	Invitación	Ponencia "Análisis y Estudio del Idioma Náhuatl Mediante Algoritmos de Procesamiento de Lenguaje Natural"	Instituto Tecnológico Superior de Ciudad Serdán	20-May-11
3	Invitación	Ponencia "Uso de Técnicas de Inteligencia Artificial para el Estudio y Análisis de la Lengua Indígena Náhuatl"	Instituto Tecnológico Superior de Atlixco	12-Abr-11
4	Seminario Permanente de Investigación y Divulgación en Ciencias Sociales y Administrativas, Humanidades e Informática	Ponencia "Técnicas de Procesamiento de Lenguaje Natural Aplicadas a la lengua Indígena Náhuatl"	Universidad del Istmo	20-May-10
5	1er. Jornada de la Informática	Ponencia "Construyendo un <i>Corpus</i> de Textos para la Lengua Indígena Náhuatl"	Universidad de la Cañada	26-Mar-10

Publicaciones

- [1] G. Cortés-Mendoza, A. Zempoalteca-Pérez, V. Soancatl-Aguilar, R. M. Ortega-Mendoza, C. Martínez-Gil. *Sistema de Software para la Traducción de Términos Náhuatl-Español*. Memoria en extenso en el VIII Encuentro de Participación de la Mujer en la Ciencia. Centro de Investigaciones en Óptica. León Guanajuato, México. 2011. **ISBN: 978-607-95228-2-7**.
- [2] Martínez-Gil C.C., Zempoalteca-Pérez A., Soancatl-Aguilar V., Ortega-Mendoza R.M., *Developing Linguistic Resources for the Nahuatl Indigenous Language*. Special Issue in Advances in Artificial Intelligence and Applications. Research in Computing Science. G. Arroyo-Figueroa (Ed). Vol. 51. 2010, pp 197-201. **ISSN: 1870-4069**.
- [3] Martínez-Gil C.C., Zempoalteca-Pérez A., Ortega-Mendoza R.M., Soancatl-Aguilar V. *Desarrollo de Recursos Lingüísticos para la Lengua Indígena Náhuatl*. 3er. Encuentro de Investigadores UABJO. pp. 131-134. Oaxaca, México. Mayo 2010.

Trabajo Futuro

Como trabajo futuro inmediato continuaremos agregaremos información semántica a los términos del diccionario y continuaremos aumentando la cantidad de textos del *corpus* y los términos de los diccionarios. Como trabajo futuro a largo plazo consideramos construir un etiquetador de partes de la oración para el náhuatl así como también un *lematizador*.