SISTEMAS EVOLUTIVOS TRADUCTORES

Fernando Galindo Soria
www.fgalindosoria.com fgalindo@ipn.mx www.laredi.com

Creado en la Ciudad de México el 12 de Agosto del 2008
Ultima actualización el 27 de Septiembre del 2011

En este trabajo se muestra en una forma incremental (o sea que primero se hace una versión tipo hola mundo y se va complicando paso a paso) *como construir un traductor evolutivo*.

En la página
http://www.fgalindosoria.com/eac/evolucion/
Se encuentra información y mas artículos sobre sistemas evolutivos.

En el articulo *Sistema Evolutivo Traductor*, de José Rafael Cruz Reyes
www.fgalindosoria.com/eac/evolucion/libro_sistemas_evolutivos/III8-TRD.DOC
Viene una descripción de cómo realizar un traductor evolutivo.

1. Sistemas Evolutivos de Reescritura

En principio el traductor evolutivo mas elemental es simplemente una tabla con dos columnas, en la primer columna que llamaremos *Texto* se pone la información del idioma A y en la segunda columna llamada *Traducción* la del idioma B.

Texto	Traducción
Texto en el idioma A	Traducción en el idioma B
.....	.....
.....	.....

Y con un programa como el siguiente:

Programa()
{
 i=1
 lee $Texto_x$
 mientras (($Texto_x$!= $Texto_i$) y (*no fin de archivo*))i++
 si (*no fin de archivo*) escribe $Traducción_i$
 sino //entra al dialogo
  escribe "desconozco $Texto_x$ dame su traducción"
  lee $Traducción_x$
  almacena en el archivo $Texto_x$, $Traducción_x$
}

Se tiene un traductor básico.

Como se puede ver este programa es muy simple de hacer, y *tiene la ventaja de que no requiere tener almacenado el conocimiento previamente*, ya que si el archivo estuviera vacío y se preguntara por $Texto_X$ el programa pediría $Traducción_X$ y ya tendría la regla $Texto_X$ -> $Traducción_X$ con lo que, *el sistema evolutivo puede empezar ha construir la base de conocimiento y 'aprender' desde cero, en tiempo real y fácilmente.*

Pueden encontrar ejemplos sobre este tema en
Sistemas Evolutivos de Reescritura
www.fgalindosoria.com/eac/evolucion/sistemas_evolutivos_reescritura/sistemas_evolutivos_reescritura.pdf

En este caso es el usuario el que interactúa con el sistema, aunque es muy elemental recomiendo que lo desarrollen como un hola mundo y como base para las siguientes versiones.

Como siguiente paso conviene hacer una primera versión de un sistema no supervisado o semi supervisado del sistema de reescritura (es decir sigue siendo un sistema de reescritura, pero en lugar de que uno le de las reglas de reescritura el sistema las encuentra).

Por ejemplo se le pueden dar a la computadora dos archivos, uno con un trabajo en un idioma y el otro con el mismo trabajo, pero en otro idioma, la idea del sistema es que construya la tabla de reescritura poniendo párrafos completos, es decir el primer párrafo en un idioma lo empalma con el primer párrafo del segundo idioma, el segundo párrafo de un idioma se empalma con el segundo del otro y así sucesivamente, si se quiere tener un sistema semi supervisado, los archivos encontrados se dan como entrada al sistema evolutivo de reescritura presentado anteriormente.

2. Manejo de Elementos

Hasta aquí es relativamente simple y no se requiere mucho conocimiento, para las siguientes etapas se requiere un poco mas de conocimiento pero nuevamente los hola mundos son simples.

Básicamente lo anterior ha sido un manejo de caja negra, pero el siguiente paso es un manejo de caja blanca (o sea abrir la caja negra y encontrar los elementos del sistema y sus relaciones), por lo que las siguientes etapas consisten en *la búsqueda de elementos, la búsqueda de estructuras y la integración de reglas transformacionales*.

La búsqueda de elementos surge porque los párrafos están formados por elementos, como por ejemplo oraciones, palabras y letras.

Aun a nivel de oraciones ya empiezan a tenerse algunos problemas, ya que no necesariamente las oraciones de dos párrafos son equivalentes, es decir que la primera oración no necesariamente corresponde a la primera del otro idioma o la segunda oración corresponde a la segunda.

Aun así, el sistema mas simple que recomiendo realizar, es uno que supone que las oraciones si preservan su posición en el párrafo, en esta versión se toman un texto y su traducción y en lugar de separarlos por párrafos se separa por oraciones, en general el proceso es muy parecido a los vistos anteriormente. Este sistema es ineficiente pero es la base para los siguientes.

2.1 Manejo de Unidades Léxicas

El siguiente problema se presenta porque cotidianamente manejamos una palabra como una cadena de caracteres (letras, números, signos) separados por espacios, por ejemplo en la oración "el perro ladra" tenemos tres palabras: "el", "perro", "ladra", el problema es que estamos manejando elementos semánticos y los elementos semánticos no necesariamente son cadenas separadas por blanco, por ejemplo "Buenos Aires" es un elemento semántico que se refiere a una ciudad, "buenos" es otro elemento y "aires" es otro, por lo que en el traductor es conveniente tratar de traducir elementos semánticos y no solo palabras, *a la unidad básica con significado semántico propio se le llama unidad léxica* (la unidad léxica corresponde en muchos casos a una palabra, pero existen otros muchos casos en los cuales eso no es cierto).

En el documento
Concatenación y Desconcatenación: Operaciones Fundamentales de la Lingüística
www.fgalindosoria.com/eac/evolucion/libro_sistemas_evolutivos/III3-COD.DOC
Se encuentran ideas sueltas sobre el problema que involucra encontrar los elementos de un sistema en general.

Existen muchas formas de atacar este problema pero muchas se basan en lo que se conoce como factorización (o sea en la búsqueda de cadenas que se repiten varias veces en un texto y que se ven como un factor dentro del texto).

Las dos variantes mas generales del método son:
Se buscan las palabra (cadenas separadas por blancos) y a todas las palabras iguales se les asigna la misma etiqueta, luego si se encuentra una combinación de etiquetas repetida varias veces se supone que toda esa combinación también es por si sola una unidad léxica y se le asigna su propia etiqueta (por ejemplo si la cadena de etiquetas X8X3X14 aparece repetida varias veces se le asigna a toda esa cadena su propia etiqueta Xi).

Otro método parte al revez: tomar un texto y buscar la cadena mas grande que se repita en el texto y la sustituye por una etiqueta X1, tomar la siguiente cadena que se repita y la sustituye por una cadena X2 y así sucesivamente, mientras existan cadenas, con este método asumimos que cada Xi es una "unidad léxica",

En el articulo
Algunas propiedades matemáticas de los sistemas lingüísticos
www.fgalindosoria.com/linguisticamatematica/propiedades_matematicas_sistemas_linguisticos/prop_mate_sistemas_linguisticos.pdf
Se presenta la factorización y otras técnicas de inferencia gramatical.

Métodos parecidos se usan para encontrar las partes de una palabra, como se ve en el articulo
Sistema Evolutivo Graficador de Moléculas Orgánicas
Escrito por Jesús Manuel Olivares Ceja
www.fgalindosoria.com/eac/evolucion/libro_sistemas_evolutivos/V3-GRMO.DOC

3. Manejo de Estructuras

Otro de los problemas que se presenta en el desarrollo de un traductor es que *la estructura de los elementos en una oración no necesariamente se preserva entre un idioma y otro*. Por ejemplo en las oraciones en español e ingles "*el perro blanco*" que tiene la estructura *psa* y "*the white dog*" con la estructura pas, la estructura en español *psa* es equivalente a la estructura en ingles *pas*, los elementos no están en el mismo orden, pero las dos oraciones significan lo mismo, por lo que, *p*ara resolver este problema se incluye en el traductor lo que se conoce como *reglas transformacionales*, (en este ejemplo la regla psa -> pas), representadas nuevamente por dos columnas, en la primera se pone la estructura del idioma A y el segunda la del idioma B

psa	pas ....

Entonces para hacer la traducción, se traducen los elementos y se les aplican las reglas transformacionales.

4. Enfoque estadístico

Aunque a nivel de párrafo se pueden realizar buenas traducciones, ya que en general un párrafo es una unidad semántica bastante consistente (en el sentido de que si aparece un párrafo en una posición en un idioma, se preserva en la misma posición en su traducción), no necesariamente es cierto para todos los casos y no necesariamente se preserva la posición de los elementos entre un texto y su traducción.

Uno de los enfoques mas usados para tratar esta situación es un enfoque estadístico, por ejemplo se asume que si un elemento ***X*** tiene la máxima frecuencia en un idioma, su equivalente ***Y*** tiene la máxima frecuencia en el otro.

Por lo que el siguiente paso consiste en desarrollar un sistema que busca todos los elementos del documento (pueden ser todos los párrafos o las oraciones o palabras) en el primer idioma y los ordena por frecuencia de aparición. También se toman todas los elementos del documento en el segundo idioma, se ordenan también por frecuencia, y se asocian las dos tablas de tal manera que el elementos con máxima frecuencia en el primer idioma se asocia con el que tiene máxima frecuencia en el segundo idioma y así sucesivamente.

La eficiencia de este tipo de sistemas depende del tamaño de los textos donde se buscan los elementos (entre mas grandes sean mas probable es que se tengan mejores traductores) y de las técnicas que se usen.

Prácticamente el mismo método que se aplica para traducción de oraciones se puede aplicar para la traducción de unidades léxicas, solo recordando que una unidad léxica puede estar formada por varias palabras separadas por blancos, por lo que, si en el traductor estadístico de oraciones se buscan unidades léxicas y se hace el análisis estadístico de las unidades léxicas el sistema sigue funcionando.

Encontrar las regla transformacionales también se puede hacer en forma automática, para lo cual, cuando se encuentra que dos oraciones son equivalentes, se ve que posición ocupa cada elemento en la primera oración y cual en la segunda y se genera la regla transformacional.

Existe una gran cantidad de literatura sobre el tratamiento estadístico de texto, principalmente en áreas como criptoanálisis y generación de correctores de ortografía (recomiendo que se revise toda la información que se pueda, aunque no toda es sobre traductores, los métodos aunque aparezcan para un área se aplican para las otras).

Una liga sobre correctores ortográficos donde hacen un buen comentario sobre el análisis estadístico es:

How to Write a Spelling Corrector
Peter Norvig, Director of Research Google
http://www.norvig.com/spell-correct.html

Conclusión

La ventaja del enfoque presentado es que desde el primer traductor evolutivo se obtienen resultados y conforme se va avanzando se van afinando, pudiendo pasar de sistemas evolutivos supervisados ha sistemas evolutivos no supervisados, en los cuales la información para hacer que evolucione el sistema se pueden obtener de los usuarios o directamente explorando la red y mediante mecanismos estadísticos prácticamente sin supervisión o con muy poca supervisión.