

SISTEMAS EVOLUTIVOS TRADUCTORES

Borrador

Fernando Galindo Soria

www.fgalindosoria.com fgalindo@ipn.mx www.laredi.com

Creado en la Cd. de México el 12 de Agosto del 2008

Ultima actualización el 12 de Agosto del 2008

La forma que se presenta para desarrollar el traductor es una forma incremental o sea que primero se hace una versión tipo hola mundo y se va complicando paso a paso

De entrada recomiendo el artículo

Sistema Evolutivo Traductor

De José Rafael Cruz Reyes

http://www.fgalindosoria.com/evolucion/Libro_Sistemas_Evolutivos/III8-TRD.DOC

1. Sistemas Evolutivos de Reescritura

En principio el traductor más elemental es simplemente una tabla con dos columnas, en la primera columna *Texto* se pone la información del idioma A y en la segunda columna llamada *Traducción* la del idioma B, y con un programa como el siguiente:

```
Programa()
{
  i=1
  lee Textox
  mientras ((Textox != Textoi) y (no fin de archivo))i++
  si (no fin de archivo) escribe Traduccióni
  sino //entra al dialogo
    escribe "desconozco Textox dame su traducción"
    lee Traducciónx
    almacena en el archivo Textox , Traducciónx
}
```

Se tiene un traductor básico.

Como se podrá 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 a construir la base de conocimiento y 'aprender' desde cero, en tiempo real y fácilmente.*

Pueden encontrar ejemplos sobre este tema en los trabajos sobre Sistemas Evolutivos de Reescritura (Lo encuentran en la pagina de sistemas evolutivos)

<http://www.fgalindosoria.com/evolucion/>

O en la pagina siguiente

http://www.fgalindosoria.com/evolucion/sistemasevolutivosdereescritura/sev_rees.pdf

En este documento es el usuario el que interactúa con el sistema, es muy elemental pero 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.

Como primer ejemplo seria darle 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 primer sistema es que construya la tabla de reescritura poniendo párrafos completos, es decir el primer párrafo en un idioma lo empalmas 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 le 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 la siguiente etapa 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 se pueden describir como la búsqueda de elementos, la búsqueda de estructuras y la integración de reglas transformacionales

La búsqueda de elementos surge porque aunque a nivel de párrafo ya se pueden realizar buenas traducciones, porque en general un párrafo es una unidad semántica bastante consistente, en el sentido de que si aparece un párrafo en un idioma se preserva en su traducción (aunque no necesariamente es cierto para todos los casos), conforme pasamos a niveles mas simples (como oraciones o palabras), cada vez es menos probable que se preserve la equivalencia semántica.

Los párrafos están formados por elementos, como por ejemplo oraciones, palabras y letras y 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 y así sucesivamente, por lo que vamos a empezar a manejar algunas herramientas que se utilizan para tratar el manejo de elementos.

El sistema mas simple que recomiendo realizar es uno que supone que las oraciones si preservan su posición en el párrafo, por lo que 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 Enfoque estadístico

Dado que no necesariamente se preserva la posición de los elementos entre un párrafo y su traducción uno de los enfoques mas usados 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 la siguiente versión consiste en desarrollar un sistema que busca todas las oraciones del documento en el primer idioma y las ordena por frecuencia de aparición, también se toman todas las oraciones del documento en el segundo idioma se ordenan también por frecuencia, y se asocian las dos tablas de tal manera que la oración con máxima frecuencia en el primer idioma se asocia con la que tiene máxima frecuencia en el segundo idioma y así sucesivamente.

La eficiencia de este tipo de sistemas depende del tamaño de los textos, entre mas grande sea mas probable es que se tengan mejores traductores, sin embargo si se quiere hacer mejores versiones es conveniente realizar un mejor tratamiento de los textos.

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

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

How to Write a Spelling Corrector

<http://www.norvig.com/spell-correct.html>

Prácticamente el mismo método que se aplica para traducción de oraciones se puede aplicar para la traducción de palabras, por lo que si en el primer traductor estadístico de oraciones se buscan palabras y se hace el análisis estadístico de palabra el sistema sigue funcionando.

2.2 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 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).

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 verlas como un factor dentro del texto)

Algunas Ligas son

Algunas propiedades matemáticas de los sistemas lingüísticos

<http://www.fgalindosoria.com/linguisticamatematica/mat-ling.PDF>

Concatenación y Desconcatenación: Operaciones Fundamentales de la Lingüística

http://www.fgalindosoria.com/evolucion/Libro_Sistemas_Evolutivos/III3-COD.DOC

Las dos variantes mas generales del método son:

Tomar un texto y buscar la cadena mas grande que se repita en el texto y sustituirla por una etiqueta X1, tomar la siguiente cadena que se repita y sustituirla por una cadena X2 y así sucesivamente mientras existan cadenas, con este método asumimos que cada Xi es una “unidad léxica”,

Otro método parte al revez, 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 (por ejemplo la cadena de etiquetas X8X3X14 aparece repetida varias veces) se supone que toda esa combinación también es por si sola una unidad léxica y se le asigna su propia etiqueta.

Métodos parecidos se usan para encontrar las partes de una palabra como se ve en el articulo

Sistema Evolutivo Graficador de Moléculas Orgánicas

Jesús Manuel Olivares Ceja

http://www.fgalindosoria.com/evolucion/Libro_Sistemas_Evolutivos/V3-GRMO.DOC

Independientemente del método que se use para encontrar las unidades léxicas para hacer el traductor se hace lo mismo a los dos idiomas y se les aplica un tratamiento estadístico para emparejar las cadenas con significados equivalentes

3. Manejo de Estructuras

Otro de los problemas que se presenta en el desarrollo de un traductor es que la estructura de los elementos no necesariamente se preserva entre un idioma y otro por ejemplo en las

oraciones en español e inglés “*el perro blanco*” y “*the white dog*” los elementos no están en el mismo orden por lo que la estructura en español *psa* es equivalente a la estructura en inglés *pas*, para resolver este problema se incluye en el traductor lo que se conoce como reglas transformacionales, representadas nuevamente por dos columnas, en la primera se pone la estructura del idioma A y en la segunda la del idioma B, entonces para hacer la traducción se traducen los elementos y se les aplican las reglas transformacionales.

Encontrar las reglas 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

Conclusión

La ventaja del enfoque presentado es que desde el primer traductor se obtienen resultados y conforme se va avanzando se van afinando, pudiendo pasar desde sistemas supervisados a sistemas 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