Dan Maxwell / Klaus Schubert / A. P. M. Witkam (eds.)
New Directions in Machine Translation Conference Proceedings, Budapest 18/19- 8-1988
Budapest: John von Neumann Society for Computing Sciences / Dordrecht/Providence: Foris Publishers
ATAMIRI - Interlingual MT Using the Aymara Language

ATAMIRI
Sistema de traducción interlingüe utilizando el lenguaje Aymara

Iván Guzmán de Rojas
Traducción del trabajo presentado en Budapest, Agosto de 1988

1. Introducción

El Sistema ATAMIRI ha sido desarrollado por el autor en La Paz, de 1980 a 1985, a manera de *hobby*, desde 1985 a 1987 como una actividad semi-professional. Desde 1987, este proyecto de traducción por máquina a pequeña escala ha continuado con la colaboración de un equipo lexicográfico consistente en tres personas y una computadora con el apoyo de sus primeros usuarios.

ATAMIRI se encuentra operacional en la Oficina de Traducciónes de la Comisión del Canal de Panamá, ayudando en la traducción de documentos técnicos y legales, del Inglés al Español. El Centro Internacional de Traducciones (ITC) de Wang en Panamá usa también este sistema. Este año, ATAMIRI está siendo probado en los ITCs europeos de Wang en una operación piloto para traducir manuales técnicos escritos en Inglés simultáneamente al Alemán, Holandés y Francés. Tenemos planes para expandir el número del lenguajes meta para traducir al Italiano y Sueco a fines de 1989. A nivel de prueba, el sistema sirve también para trabajar partiendo del Español como idioma fuente y el Aymara, Portugués y Hungaro como lenguajes meta.

La versión actual de ATAMIRI opera en una computadora Wang VS. La mayoría de las subrutinas del sistema fueron escritas inicialmente en BASIC y ahora están siendo convertidas a PL/I. El sistema requiere de una instalación multiusuaria, con por lo menos 2 MB de memoria y un disco dedicado de por lo menos 60 MB. Se requiere de un software de procesamiento de textos adecuado para manejar los textos de entrada y salida.

2. Representación Matricial del Lenguaje

El proyecto ATAMIRI, a pesar de contar con recursos limitados, ya ha sido capaz de mostrar resultados prácticos trabajando en ambiente multilingüe. Este hecho primicial tiene que ver con el concepto de interlingua subyacente en el diseño del sistema.

La representación de lenguaje utilizada se define formalmente como sigue:

- |M> llamado CONCEPTOR, es una entidad invariante en un espacio tensorial de tercer rango de N dimensiones.

- <i,j,k;L| llamado UNITOR is un tensor de tres subindices en el espacio de referencia L que subtiende el espacio tensorial.

- <i,j,k;L|M> llamado EXPRESOR es la matriz obtenida por el producto escalar de la dos entidades indicadas anteriormente, de esta manera conforma la correspondiente componente de |M > en el marco de referencia L.

El número de dimensiones N es siempre finito, pero puede ser muy grande dependiendo del tamaño de las sintagmatecas de cada idioma en el universo de textos considerado. L varía desde 0 hasta el número de lenguajes implementados en el sistema. L=0 es el interlingua.

La unión ordenada de todos los expresores para un cierto conceptor |M >, dada por:

$$U < i,j,k,l \mid M >$$

representa la oración que expresa el mensaje M en lenguaje L.

En estas formas de representación matemática no nos interesan en resultados cuantitativos, pero si queremos más bien asegurarnos que sus transformaciones tensoriales reflejarán las traducciones de un lenguaje a otro. Por esta razón estamos obligados a restringir la clase de grupo de transformaciones permitidas. Hemos adoptados dos tipos de restricciones:

TOPOLOGICAS: Restricciones para asegurar una correspondencia de uno a uno de las componentes en cualquier marco de referencia para un par de lenguajes. Así se evita la adición de componentes.

METRICAS: Restricciones para evitar multiplicaciones por otros factores que no sean unos o ceros. Por ejemplo, no se admiten cambios en el "tamaño" de las componentes. Las transformaciones se reducen a permutaciones.

En todo lenguaje natural, la representación correcta nos obliga a manejar transformaciones donde la "distorción", "creación" y "destrucción" de componentes sean permitidas. En efecto, la versión actual de ATAMIRI está basada en una representación extendida del lenguaje tensorial utilizando tales operadores no-lineales. Aquí solamente discutiremos componentes matriciales con el objeto de ilustrar la manera en que ATAMIRI usa una representación sintáctica formal con el Aymara como interlingua.

Tomando en cuenta las anteriores restricciones, es ahora válido visualizar una traducción de oraciones del idioma fuente L_1 al lenguaje meta L_2 como una transformación de la matriz:

$$< i_2,j_2,k_2;L_2 \mid M > = \sum (i_2 ,j_2 ,k_2;L_2 \mid i_1 ,j_1 ,k_1;L_1) < i_1,j_1,k_1;L_1 \mid M >$$

Aquí la suma corre sobre ambos conjuntos de índices i,j,k para cada par de lenguajes L_1 y L_2. Para simplicidad, escribimos este transformación como sigue:

$$< 2 \mid M > = (L_2 \mid L_1) < 1 \mid M >$$

La correspondiente transformación inversa se escribe:

$$< 1 \mid M > = [L_1 \mid L_2] < 2 \mid M >$$

Estas fórmulas proporcionan la representación de las dos oraciones, en lenguajes L_1 y L_2, expresando el mismo mensaje M. Los coeficientes de transformación conforman la matriz $(L_2|L_1)$ y la matriz inversa $[L_1|L_2]$ que representa la correspondiente regla de reordenamiento sintáctico de esa clase de oración para efectuar la traducción en ese par de lenguajes.

Las componentes del tensor < L | M > representan cada de los elementos de la oración del mensaje M expresado en un lenguaje dado L. Su "valor" es el símbolo adoptado para codificar las categorías sintácticas utilizadas por el parcelador de tres niveles operando bajo un modelo de representación específico.

Obviamente se requiere que el conjunto de categorías sintácticas adoptados en un modelo de representación dada, tiene que proporcionar descripciones consistentes de todas las operaciones de parcelación efectuadas con oraciones expresadas en cualquiera de los lenguajes incluidos en el modelo. Llamamos lenguaje *canónico* al lenguaje natural o el lenguaje formal cuyo categorías sintácticas son adoptadas como base de codificación.

El sistema ATAMIRI es capaz de operar bajo cualquier lenguaje canónico definido por el usuario. La versión actual opera bajo un modelo de representación en el que el idioma Aymara actúa como el lenguaje canónico. Este lenguaje estructurado sufijalmente, gracias a sus categorías sintácticas definidas posicionalmente, es sumamente adecuado para este propósito, ya que las posiciones de sufijos se pueden vincular directamente a los subindices de las matrices. Hay también otras propiedades algorítmicas en el Aymara, que no necesitan ser discutidas aquí, que hacen de esta antígua herramienta Andina para la comunicación humana un lenguaje canónico ideal.

La representación matricial de lenguaje aquí explicada muy sucintamente, ofrece tres enormes ventajas:

- LA SINTAXIS EXTERNA definida por las matrices de la transformación, cuyos valores de los coeficientes pueden ser almacenados en una base de datos sintáctica, de modo externo a la subrutinas del parcelador. De esta manera es posible enriquecer los conocimientos sintacticos del sistema sirviéndonos de experiencias previas de traducción

de textos sin tener que tocar el programa. Esto le da a ATAMIRI una poderosa capacidad de "aprender" sintaxis.

- Capacidad de traducción simultánea MULTILINGÜE, ya que ahora el usuario puede escoger el conjunto de coeficientes de transformación y las direcciones de traducción, es decir, los idiomas que utilizará como lenguaje fuente y como lenguajes meta. El usuario puede también decidir si va cargar en memoria una o más de los conjuntos correspondiente a varios lenguajes meta para acelerar el procesamiento de las traducciones.

- INTERLINGUA: El puente de transformación diseñado para minimizar costos en el desarrollo de la base de datos sintáctica multilingüe.

El concepto de interlingua como se aplica en el sistema ATAMIRI necesita todavía una mayor explicación. Llamamos interlingua a uno de los lenguajes en el modelo de la representación, el cual sirve siempre como el marco de referencia común para las transformaciones en que intervienen todos los otros lenguajes.

Por ejemplo, si con nuestro modelo de representación tenemos que cubrir N lenguajes, necesitamos almacenar solamente N-1 conjuntos de matrices de transformación, una conjunto para cada lenguaje transformado relativo al interlingua. Las transformaciones entre cualquier par de lenguajes pueden ser definidas inmediatamente por un producto de matrices, como se explica a continuación:

- A es el lenguaje interlingua.

- X y Y son dos lenguajes cualquiera.

- (A | X) es la matriz de transformación de X a A.

- (A | Y) es la matriz de la transformación de Y a A.

- [Y | A] es la matriz de transformación inversa de A a Y.

- (Y | X) = [Y | A] (A | X) es la transformación resultante de X a Y.

De otro modo habríamos necesitado desarrollar $(N-1)^2$ conjuntos de matrices de transformación, y trabajar con ellos! Por ejemplo para N=21, se requieren 400 conjuntos.

La primera implementación de ATAMIRI usa una modalidad de representación en la que el AYMARA es tanto el lenguaje canónico como el interlingua. Sin embargo el sistema permite al usuario una elección irrestricta del lenguaje para ser utilizado como interlingua, siempre y cuando sea un lenguaje bien definido en el modelo de la representación; es decir, posee un conjunto consistente y completo de categorías sintácticas con respecto a los demás lenguajes en el modelo (el interlingua no puede ser un lenguaje subconjunto). De esta

explicación vemos que ATAMIRI no utiliza un interlingua semántico. El léxico Aymara no es usado para nada, a menos que el usuario requiera una traducción al Aymara.

3. Confrontación con la praxis

La subrutina del analizador sintáctico reserva un área de memoria máxima de medio MB para almacenar los coeficientes de transformación requeridos para traducción simultánea del idioma fuente a cinco lenguajes meta. Después de tres años de experiencia, tanto en corridas de prueba corre como en operación normal en ambiente de producción, el número de conjuntos de transformación almacenados todavía no nos ha obligado a sobrepasar ese valor máximo. Este hecho muestra cómo sólo unas pocas estructuras elementales son necesarias para describir oraciones de la *vida real,* cuando se usan descriptores simbólicos de tres niveles para las categorías sintácticas. Estamos hablando de menos de mil fórmulas sintagmaticas por lenguaje.

En el trabajo hemos encontrado que ocurren pocos casos frecuentes, donde el lenguaje canónico tuvimos que extenderlo más allá de la gramática Aymara, introduciendo categorías vacías para este lenguaje. Por ejemplo, el uso de verbos auxiliares para el tiempo futuro en algunos lenguajes, como el Inglés, no tiene su equivalente en Aymara. Tales categorías, que son vacías en algún lenguaje, nos obliga introducir algoritmos deviantes utilizando operadores de "creación" y "destrucción". Este desvío de transformaciones lineales causaron bastante incremento en código para los algoritmos de manejo de sintáxis. Sin embargo, el incremento en tiempo adicional de ejecución no fue mayor al 5%.

Otra dificultad encontrada en la praxis, al traducir de Inglés al Alemán o al Holandés, fue la necesidad de almacenar demasiadas matrices de transformación redundantes para el manejo de "Gliedsätze". Por ejemplo, la oración compuesta:

La oración condicional en Inglés: **If K_I, I_1.**

Tiene su equivalente en Alemán: **Wenn K_2, I_2.**

La dificultad radica en que I_2 no es la traducción "normal" de I_1, cuando es tomada como una oración sola. I_2 es más bien una forma "distorsionada". Por lo tanto el sistema tendría que olvidar que los coeficientes de la transformación para la oración compuesta pueden ser construídos a partir de los correspondientes coeficientes para la oración componente, y tendría que pedir una nueva matriz (redundante en Inglés y en los otro lenguajes donde tal "distorción" no se presenta) para toda la oración compuesta.

Aquí otra vez, con la ayuda de operadores deviantes de "distorción", aún a costo de desarrollar un código lógico altamente intrincado, el tamaño de la base de datos sintáctica puede mantenerse reducido, haciendo mucho más fácil la tarea de "enseñanza" del lenguaje.

Puesto que los errores sintácticos en el *draft* tienen una fuerte incidencia negativa en el trabajo de afinado del texto traducido, el rendimiento del sistema para el Alemán y el

Holandés ha sido mejorado significativamente al extender el modelo de representación como para incluir la capacidad de manejo de transformaciones con "distorción".

La experiencia de trabajo con el sistema ATAMIRI muestra que la representación matricial de lenguaje, con el Aymara como interlingua, cuando se complementa con operadores no lineales deviantes, puede convertirse en una herramienta muy poderosa y económica para el manejo de cualquier sintáxis en ambiente multilingüe.

Anexo: Ejemplo ilustrativo

Al final de la presentación de este trabajo en la conferencia de Budapest, se me acercaron algunos participantes para comentarme que encontraron mi *paper* demasiado abstracto, expresando su deseo de ver algunos ejemplos ilustrativos para aclarar conceptos. Efectivamente, el cortísimo tiempo que se tiene a disposición para exponer un trabajo en este tipo de conferencias internacionales, me obligó a producir un artículo demasiado sintético, abusando, en cierta manera, con el manejo de las herramientas de una formulación matemática. A continuación presento de modo complementario al artículo, algunos de los ejemplos ilustrativos con que expliqué a dichos participantes aspectos centrales del concepto de interlingua tal como se lo entiende en el diseño de ATAMIRI.

Consideremos un mensaje M dado por la oración en Inglés (L=1):

M = ‘Your sister wants to buy a yellow dress.’

Este es el texto fuente que deseamos traducir al Español (L=2) y al Alemán (L=3). Acordemos que el lenguaje interlingua es el Aymara (L=0), que también es el lenguaje canónico de nuestro sistema traductor.

El analizador morfológico y lexical de ATAMIRI determina que el mensaje M en Inglés se encuentra segmentado de la siguiente manera:

$$|M|^{E} = | \text{ your } | \text{ sister } | \text{ wants } | \text{ to buy } | \text{ a } | \text{ yellow } | \text{ dress } | \text{ . } |$$

Igualmente determina que el vector sintáctico de M en Inglés se compone de las siguientes categorias de nivel k=1 así ordenadas:

$$(|M|^{E}) = (> \text{ s } \text{ v } \text{ w } \text{ d a s . })$$

Al mismo tiempo, ya que para este efecto el sistema consultó a su base de datos lexicográfica, obtiene las siguientes traducciones asintácticas (Ta) del mensaje M al Aymara, al Castellano y al Alemán:

$$Ta|M|^{A} = | \text{ -ma } | \text{ kullaka } | \text{ muni } | \text{ alasiña } |\text{mä}| \text{ q’ellu } | \text{ isi } | \text{ . } |$$
$$Ta|M|^{C} = | \text{ tu } | \text{ hermana } | \text{ quiere } | \text{ comprar } | \text{ un } | \text{ amarillo } | \text{ vestido } | \text{ . } |$$
$$Ta|M|^{D} = | \text{ deine } | \text{ Schwester } | \text{ will } | \text{ kaufen } |\text{ein } | \text{ gelbes } | \text{ Kleid}| \text{ . } |$$

Obviamente ninguna de estas traducciones asintácticas constituyen oraciones bien formadas, ni en Aymara ni en Castellano ni en Alemán. De lo que se trata es conseguir que el ordenador ponga estas oraciones de acuerdo al orden sintáctico correcto que corresponde en cada uno de estos idiomas.

Una operación reiterativa del parcelador, esta vez sobre el vector sintáctico ($|M|^E$), proporciona sus componentes de nivel k=2:

$$|(|M|^E)| = | > s | v | w | d a s | . |$$

Las correspondientes categorías sintácticas, ahora de nivel k=2, se representan y ordenan así:

$$(|(|M|^E)|) = (P V W O .)$$

De modo recursivo aplicamos nuevamente el parcelador y obtenemos que:

$$|(|(|M|^E)|)| = | P | V W O | . |$$

Así las correspondientes categorías sintácticas, ahora de nivel k=3, son:

$$(|(|(|M|^E)|)|) = (X Y .)$$

Este es el nivel sintáctico mas profundo al que llega el parcelador de ATAMIRI. Ahora la representación matricial de los elementos de parcelación en los diferentes niveles deben ser "traducidos", es decir se deben obtener los reordenamientos adecuados para la sintáxis de cada uno de los lenguajes meta. Lo que ha logrado el parcelador es obtener los *expresores*:

<1,1,1,1|M> = > <2,1,1,1| M> = s
<1,2,1,1|M> = v
<1,3,1,1|M> = w
<1,4,1,1|M> = d <2,4,1,1| M> = a <3,4,1,1| M> = s

<1,1,2,1|M> = P
<1,2,2,1|M> = V
<1,3,2,1|M> = W
<1,4,2,1|M> = O

<1,1,3,1|M> = X
<1,2,3,1|M> = V <2,2,3,1|M> = W <3,2,3,1|M> = O

La sintagmateca de ATAMIRI permite conocer los coeficientes de transformación para llevar la configuración de un expresor a su correspondiente en el lenguaje interlingua (L=0), y viceversa. Por ejemplo la última línea de expresores, nivel k=1, se transforma así:

<1,1,1,0|M> = s <2,1,1,0| M> = > Es decir: P_0 = s >

como resultado de la operación de traducción sintáctica:

$$<i,1,1,0|M> = \sum_{i} (0\ 1;\ 1\ 0) <í,1,1,1|M> \quad \text{donde el subíndice } \mathbf{í} \text{ corre de 1 a 2.}$$

Estas operaciones de traducción sintáctica permiten reordenar los elementos de los expresores en cada nivel. Comenzando de los niveles más profundos hasta la superficie se efectúan las operaciones de modo recursivo. El resultado se puede resumir en los esquemas:

$$|(|M|^{E})|^{A} = | s > |) | (| d\ a\ s | w | (| v | . |$$

$$|(|M|^{E})|^{C} = | > s | v | w | d\ s\ a | . |$$

$$|(|M|^{E})|^{D} = | > s | v | d\ a\ s | w | . |$$

Ahora podemos escribir las traducciones **Ts** sintácticamente correctas:

$$Ts|M|^{A} = | \text{kullaka} | \text{-ma} | \text{-xa} | / | \text{mä} | \text{q'ellu} | \text{isi} | \text{alasiña} | / | \text{muni} | . |$$

kullakamax mä q'ellu isi alasiñ muni.

$$Ts|M|^{C} = | \text{tu} | \text{hermana} | \text{quiere} | \text{comprar} | \text{un} | \text{vestido} | \text{amarillo} | . |$$

$$Ts|M|^{D} = | \text{deine} | \text{Schwester} | \text{will} | \text{ein} | \text{gelbes} | \text{Kleid} | \text{kaufen} | . |$$

Es de hacer nota que en el caso de la traducción al Aymara se han introducido los dos elementos sintácticos **-xa** y **/** que no tienen expresores equivalentes en el texto original en Inglés; el sistema los ha introducido por la acción de operadores de *creación* que son activados por los parámetros almacenados en la sintagmateca para el Aymara.

En la praxis con textos de la vida real, la traducción no se puede reducir a reordenamientos sintácticos, ya que los problemas de desambiguización lexical interfieren en la correcta asignación de categorías sintácticas a las palabras que aparecen en el texto original. Una parte importante de estos problemas pueden ser atacados con técnicas basadas en las restricciones sintácticas típicas de cada idioma. Por ejemplo, en Inglés, a un artículo no le puede seguir un verbo modal, the modo que una expresión que contenga *the can* permite reconocer que *can* es el sustantivo (*envase de lata*) y no el verbo modal (*puede*). Este tipo de ambiguedades pueden ser manejadas bastante bien por ATAMIRI; sin embargo, el problema es mas complicado en situaciones en que intervienen aspectos semánticos. El diseño para manejar ambiguedades semánticas con los métodos de representación matricial está contemplado en nuestros planes para las próximas versiones del sistema ATAMIRI.