UNIVERSIDAD POLITÉCNICA DE PUEBLA

PROGRAMA ACADÉMICO DE
INGENIERÍA EN INFORMÁTICA

Módulo de Traducción Aplicado a la Búsqueda de Respuestas Multilingüe

Patricia Lorena Irigoyen Sánchez

Reporte Técnico PII-06-08-09

COMITÉ EVALUADOR

Dra. Rita Marina Aceves Pérez (*Asesor*)
M.C. Rebeca Rodríguez Huesca (*Sinodal*)
Dra. María Auxilio Medina Nieto (*Sinodal*)

PROFESOR(A) DE PROYECTO DE INVESTIGACIÓN II

Dra. María Auxilio Medina Nieto

Juan C. Bonilla, Puebla
Agosto 2009

ÍNDICE

CAPÍTULO 1. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 INTRODUCCIÓN

Cuando un usuario requiere realizar búsqueda de información, la entrada a cualquier sistema es una 'petición' o 'consulta'. Esta consulta es de vital importancia para obtener la información que el usuario necesita. En ocasiones, la búsqueda de la información se realiza en varios idiomas, lo que significa que se requiere de su traducción. Si se considera a un grupo de idiomas, en este caso 'Español', 'Francés' e 'Italiano', cada uno maneja su propia estructura. En este trabajo no se tomará en cuenta la sintaxis o la estructura léxica de los idiomas, sino que se aprovecharán los resultados de traductores o máquinas de traducción con la finalidad de generar una traducción funcional para la búsqueda de respuestas.

Los métodos que se implementarán tienen la función de seleccionar o generar una traducción adecuada para encontrar la información que un usuario desee. En la actualidad, en Internet se cuenta con una vasta cantidad de información examinada por millones de personas, esto se resume a un sin número de consultas que pueden o no ser eficientes para el objetivo de los consultores. Para garantizar que la consulta tenga un resultado correcto se desarrollan métodos que la evalúan. "Dada la inmensa información presente en la web y en colecciones privadas de documentos, surge la necesidad de técnicas que permitan extraer información relevante [Juárez G. A., 2007]". [Aceves-Pérez R., 2007] desarrolló 2 métodos: "Selección de la mejor traducción" y "Reformulación de la pregunta", que en este trabajo se retomarán y se implementarán con la finalidad de optimizar la traducción desde una consulta hecha por máquinas de traducción para que ésta sea lo suficientemente funcional y así poder obtener resultados prometedores. Las características de cada máquina de traducción se especifican en la sección 2.3.

1.2 OBJETIVO GENERAL

Implementar un módulo de traducción de una pregunta en Español, Francés e Italiano combinando la salida de máquinas de traducción para generar una 'consulta' que permita obtener resultados satisfactorios en un sistema de búsqueda de respuestas multilingüe.

1.3 OBJETIVOS ESPECÍFICOS

1. Utilizar los resultados generados a partir del software de traducción Systran, InterTiran, TraduceGratis para adquirir consultas o entradas a sistemas de búsqueda de respuestas multilingüe.
2. Implementar el algoritmo para traducción "Selección de la mejor traducción"

3. Implementar una interfaz para donde el usuario pueda manejar dicho algoritmo

1.4 JUSTIFICACIÓN

Se sabe que para obtener información relevante, es decir, información útil al usuario interesado en ella, es necesario hacer consultas apropiadas al sistema que se utilizará. Por esta razón; surge la necesidad de implementar un módulo de traducción basado sólo en la traducción de la pregunta con base en la hipótesis siguiente: “Si la pregunta es eficaz, entonces la información encontrada también lo será”.

El proyecto consistirá en desarrollar un ‘prototipo’ de traducción de la pregunta para obtener resultados que le permitan a los sistemas de búsqueda de respuestas multilingüe encontrar mayor calidad y cantidad de respuestas que no podría encontrar con una sola máquina de traducción.

1.5 CRONOGRAMA DE ACTIVIDADES

Actividades	*Septiembre-Diciembre 2008*												
	Semanas												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Investigación de literatura	☑												
Comprensión y planteamiento del problema.		☑											
Entrega de propuesta del proyecto de investigación (Capítulo I).			☑										
Exploración de método "selección de la mejor traducción".				☑									
Búsqueda del método "Reconstrucción de la pregunta".					☑								
Estudio de resultados arrojados por Systran.						☑							
Estudio de resultados arrojados por InterTiran.							☑						
Estudio de resultados arrojados por Traduce Gratis.								☑					
Entrega de avances del protocolo de proyecto de investigación (Capítulo II).									☑	☑			
Elaboración del capítulo III del proyecto de investigación											☑	☑	
Entrega de la propuesta del proyecto de investigación													☑

Actividades	*Mayo-Agosto 2009*												
	Semanas												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Análisis de lenguajes de programación a utilizar.	☑	☑											
Conexión con páginas de los traductores			☑	☑									
Instalación CMU-Cam_Toolkit_v2					☑	☑							
Tratamiento de las colecciones del EFE94 y EFE95							☑						
Estudio y diseño de la interfaz de usuario para el módulo de traducción								☑					
Desarrollo de la interfaz de usuario									☑	☑			
Obtener la traducción y el valor de perplejidad de las preguntas										☑	☑		
Pruebas												☑	
Desarrollo de documentación													☑

1.6 ALCANCES Y LIMITACIONES

Alcances

El proyecto manejará 160 preguntas, se utilizarán 3 máquinas de traducción: Systran, InterTiran y Traduce Gratis. El acceso a estas máquinas es en línea, cada una ofrece traducciones a Español, Francés, Italiano, entre otros idiomas. El proyecto tiene futuro ya que si se logra obtener una traducción con mayor funcionalidad, es decir, menor perplejidad, se tendrá la oportunidad de desarrollar un proyecto más grande, por ejemplo, utilizando grabaciones de voz. Así mismo, al inicio sólo se consideran tres idiomas, pero con el desarrollo y el avance del proyecto, se podrán incorporar más, esto para incrementar el número de traducciones y obtener una mejor respuesta. Actualmente se consideran dos algoritmos para elegir o reconstruir una traducción, sin embargo, pueden generarse otros.

Limitaciones

A causa de los caracteres especiales usados en algunos idiomas como el árabe, japonés y coreano, este proyecto no podrá hacer traducciones de estos idiomas. Otra limitación es que solamente se podrán utilizar máquinas de traducción en línea gratuitas.

1.7 RECURSOS DE HARDWARE Y SOFTWARE

El sistema de cómputo para el desarrollo del proyecto tiene las siguientes características:

Hardware	Software
Procesador Celeron 3.0 Ghz	Systran version 5.0
Disco duro 80 Gb	InterTiran version 4.5
Memoria RAM 1Gb	Traduce Gratis versión 4.3
Unidad DVD	Motor de búsqueda Google versión 6.0
Teclado	Lenguaje de programación Java
Ratón	Sistema operativo Windows XP o Windows Vista
	HTML Parser versión 1.6
	JCreator LE versión 3.50.010 Copyright

	2000-2005 Xinox Software
	Eclipse versión 3.2.0 (c) Copyright Eclipse contributors and others 2000-2006.

CAPÍTULO 2. MARCO TEÓRICO

2.1 INTRODUCCIÓN A LOS SISTEMAS BR

Los Sistemas de Búsqueda de Respuestas (BR) se desarrollan debido a la gran cantidad de información que los usuarios obtienen desde Internet a sus computadoras [6]. Se sabe que no toda la información que viaja en el Internet a cualquier parte del mundo es funcional para las personas que requieren de ella. Por lo tanto, un Sistema BR es aquel que dada una consulta cualquiera, en un idioma específico tal como el Español, obtiene respuestas desde documentos verificados [2].

La ventaja de los Sistemas BR es la transparencia que tienen los métodos para realizar las traducciones y las búsquedas necesarias para dar al usuario información relevante; todos los métodos se inician en el sistema de búsqueda de respuestas de forma automática, facilitándole al usuario la obtención de los resultados.

Sin embargo; los métodos aplicados a la traducción de la pregunta en los sistemas desarrollados hasta el momento, aún no tienen los resultados esperados para satisfacer a los usuarios ya que la respuesta no es concreta a la pregunta insertada. Con la investigación presentada en este documento se espera mejorar los resultados para construir un módulo de traducción eficiente.

Un Sistema BR consiste de tres bloques, los cuales se explican en la siguiente sección.

2.1.1 FUNCIONAMIENTO DE UN SISTEMA BR

Según [3], existen tres bloques en un Sistema BR, cada uno posee una tarea específica, éstos son:

1. *Procesamiento de la pregunta*. Este bloque trata de identificar sobre qué se está preguntando, por ejemplo, una persona, un lugar o algún acontecimiento.
2. Análisis de respuestas candidatas. En este bloque, como el nombre lo indica, se analizan las respuestas con mayor número de aparición en los documentos consultados. Estas respuestas se convertirán en respuestas candidatas.
3. Extracción de respuestas. En este bloque se eligen las mejores respuestas candidatas para ser mostradas al usuario de acuerdo a algún criterio numérico.

2.1.2 PREGUNTAS EN SISTEMAS BR MULTILINGÜE

Existen sistemas de búsqueda de respuestas monolingüe, es decir, sólo manejan un idioma. Un sistema de respuestas multilingüe es aquel que cuenta con máquinas de traducción (traductores) que manejan una consulta que puede estar en diferente idioma para obtener respuestas que se encuentran en documentos escritos en otros idiomas. Esto se debe a la necesidad que tiene el usuario de apoderarse de información [2]. El funcionamiento de estos sistemas calcula por medio de ciertos métodos la mejor traducción de las consultas como de las respuestas candidatas. Un sistema de búsqueda de respuestas multilingüe está formado por los tres bloques anteriormente descritos con la adición de máquinas de traducción, en este caso traductores en línea.

2.2 ALGORITMO PARA OBTENER TRADUCCIONES DE UNA PREGUNTA

En la actualidad los métodos aplicados en algunos sistemas ofrecen un nivel de error de aproximadamente 20% [1]. Este proyecto pretende reducir este nivel, para ello se propone implementar un módulo de traducción utilizando el siguiente método: selección de la mejor traducción. Este método se propuso en [1]. En la sección siguiente se describe el algoritmo mencionado anteriormente.

2.2.1 SELECCIÓN DE LA MEJOR TRADUCCIÓN

El método selección de la mejor traducción consta de tres pasos [1]:

1. Traduce la pregunta en cierto número de máquinas de traducción, para este proyecto se utilizan Systran, InterTiran y Traduce Gratis.
2. Selecciona la mejor de las traducciones con base en la medida de perplejidad.
3. Envía la traducción seleccionada a un sistema de búsqueda de respuestas y de esta manera obtiene la respuesta correcta.

La Figura 1 muestra el diagrama del método.

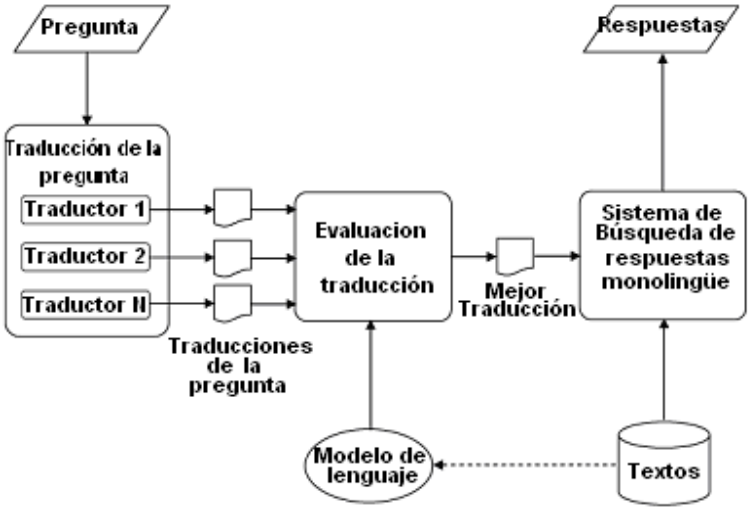


Figura 1. Método selección de la mejor traducción [1]

En el método se aplicará la siguiente fórmula que calcula la pertinencia de una traducción:

$$H = -\frac{1}{Q}\sum_{i=1}^{Q} \log P(w_i | w_{i-1}, w_{i-2}, \ldots, w_{i-N+1})$$

donde W_i es una palabra en la secuencia de palabras

$P(w_i)$ indica la probabilidad de W_i

Q es el número de palabras del examen de datos

N es el orden de modelo de secuencia.

El puntaje final es expresado mediante la perplejidad, su fórmula es: $B=2^H$, donde

B es el puntaje de perplejidad.

H es la pertinencia de la traducción.

Entre menor sea el valor de la perplejidad, será mejor la traducción.

2.3 TRADUCTORES EN LINEA

Los traductores que se emplean en el proyecto se acceden en línea; se podrá realizar un enlace a la interfaz que se desarrollará. A continuación se presentan algunas características de éstos.

2.3.1 INTERTIRAN

El traductor en línea InterTiran cuenta con una colección de idiomas de entre los cuales están los idiomas necesarios para este proyecto. Permite conexiones a aplicaciones y por ende la extracción de la información, en este caso, la traducción de las preguntas. [9] Permite una traducción entre 1600 pares de lenguajes además de soluciones a preguntas frecuentes por el usuario.

2.3.2 SYSTRAN

Las características del traductor Systran se listan en la Tabla 1, la cual incluye las actividades que se pueden realizar y la versión del software que las implementa [8], [10].

Tabla 1. Descripción de actividades en Systran

Actividad	*Versión de Systran*
Explorar el Web en su idioma materno o navegar y traducir páginas Web	Plug-in de SYSTRAN para Internet Explorer
Traducir documentos de Microsoft Word	Plug-in de SYSTRAN para Microsoft Word
Traducir hojas de cálculo de Microsoft Excel	Plug-in de SYSTRAN para Microsoft Excel
Traducir presentaciones en Microsoft PowerPoint	Plug-in de SYSTRAN para Microsoft PowerPoint
Traducir correo electrónico en Microsoft Outlook	Plug-in de SYSTRAN para Microsoft Outlook
Traducir cualquier texto	SYSTRAN Clipboard Taskbar (SCT)
Traducir archivos de PDF	SYSTRAN Translation Project Manager (STPM), SYSTRAN MultiTranslate Utility (SMTU), plug-in de SYSTRAN para Microsoft Word, plug-in de SYSTRAN para Internet Explorer.
Alinear páginas Web y archivos DOC, RTF, TXT, HTML, XHTML y PDF; realizar correcciones y edición posterior	SYSTRAN Translation Project Manager (STPM)
Definir tareas de traducción para un conjunto de documentos y ejecutar trabajos programados de traducción en lotes (modo “batch”)	SYSTRAN MultiTranslate Utility (SMTU)
Desarrollar diccionarios personalizados basados en la terminología de su empresa o industria.	SYSTRAN Dictionary Manager (SDM)
Añadir su terminología específica a las traducciones	SYSTRAN Dictionary Manager (SDM)
Crear y utilizar *Translation Memories*	SYSTRAN Dictionary Manager (SDM)
Crear *Normalization Dictionaries*	SYSTRAN Dictionary Manager (SDM)
Extraer terminología para crear diccionarios personalizados	SYSTRAN Translation Project Manager (STPM), SYSTRAN MultiTranslate Utility (SMTU), plug-in de SYSTRAN para Microsoft Office, plug-in de SYSTRAN para Internet Explorer

2.3.3 TRADUCE GRATIS

Según [11], las características principales del traductor Traduce Gratis son:

- Añade una barra de traducción al navegador Internet Explorer
- Traduce la página Web activa preservando los enlaces y el formato
- Incluye los siguientes idiomas: inglés, español, italiano, francés y alemán
- Traduce Gratis propone más de una traducción
- Busca palabras y frases en diccionarios, conjuga verbos en diferentes idiomas
- Permite seleccionar diccionarios especializados antes de traducir.

Traduce Gratis utiliza diccionarios especializados (Economía y Negocios, Informática, Medicina, Jurídico, Seguridad Social); la traducción se realiza entre las siguientes parejas de idiomas:

- Español -> Inglés e Inglés -> Español
- Español -> Italiano e Italiano -> Español
- Español -> Francés y Francés -> Español
- Inglés -> Italiano e Italiano -> Inglés
- Inglés -> Francés y Francés -> Inglés
- Italiano -> Francés y Francés -> Italiano
- Italiano -> Alemán y Alemán -> Italiano

2.4 TRABAJO RELACIONADO

2.4.1 EXTRACCIÓN DE RESPUESTAS MEDIANTE APRENDIZAJE AUTOMÁTICO UTILIZANDO ATRIBUTOS LÉXICOS

En el trabajo extracción de respuestas mediante aprendizaje automático utilizando atributos léxicos [3] se documenta que los esfuerzos realizados en los sistemas BR son insuficientes para tratar preguntas de tipo factual, en especial en el idioma Español, ya que una palabra tiene muchos significados o se puede usar para diferentes objetivos. [3] menciona que el resultado de los sistemas actuales se debe a los métodos implementados en el bloque de traducción de la pregunta, por la complejidad de la combinación de formas léxicas. [3]

propone un enfoque basado en aprendizaje automático utilizando 17 características léxicas del idioma Español.

2.4.2 TÉCNICAS LINGÜÍSTICAS APLICADAS A LA BÚSQUEDA TEXTUAL MULTILINGÜE

En el trabajo realizado por [6] se documenta que aún no existe un sistema de búsqueda de respuestas ideal ya que existen problemas que son característicos propios del lenguaje. Ese trabajo considera problemas tales como la ambigüedad léxica, la variación de terminología y el translingüismo en el acceso a la información. [6] comenta que la investigación de técnicas lingüísticas propone realizar experimentos con sistemas como WordNet y el sistema Website Term Browser (WTB).

En comparación con el presente trabajo: "Módulo de traducción aplicado a la búsqueda de respuestas multilingüe", se desarrolla un módulo donde se pretende disminuir los problemas de lenguaje mediante la aplicación de métodos que evalúen la perplejidad de las traducciones hechas para obtener una respuesta óptima. Toma en cuenta la perplejidad que posee cada traducción. Evalúa el sistema con el CLEF. Las colecciones empleadas para el proyecto son las del EFE 94 y EFE 95.

***CLEF: Cross Language Evaluation Forum Workshop**

CAPÍTULO 3. DISEÑO DE INVESTIGACIÓN

3.1 INTRODUCCION

En el diseño de la investigación se dará a conocer los datos necesarios para la realización del proyecto. En este caso, los datos principales son preguntas que utilizando alguna máquina de traducción se traducirán de acuerdo a algún método elegido.

3.2 CASOS DE USO

La Tabla 2 describe uno de los requerimientos principales del sistema. Los restantes se consideran casos de uso, éstos se muestran en la Figura 3 y se describen en las tablas 4-6.

Tabla 2. Requerimiento "traducción de la pregunta"

Identificador:	TR
Nombre de requerimiento:	Traducción de la Pregunta
Descripción corta:	Representa pasar a un idioma diferente una pregunta en lenguaje natural
Descripción detallada:	1. El usuario inserta una pregunta en lenguaje natural 2. El usuario hace una petición para muestra de resultados (traducción y perplejidad).

Tabla 3. Descripción de casos de uso

	Caso de Uso 1	Caso de Uso 2	Caso de Uso 3
Nombre	CU-Inserción de pregunta	CU-Alcanzar la mejor traducción	CU-Presentación de la información
Descripción	Se propone una pregunta en lenguaje natural	Ejecución de método “Selección de la mejor traducción”	Vista de resultados
Datos de entrada	Pregunta	Método “Selección de la mejor traducción”	Solicitud de presentación de resultados
Datos de salida	Mensaje de confirmación: “Pregunta en proceso…”	Mensaje de confirmación: “Método procesando pregunta…”	Mensaje de confirmación: “Resultado del método…”

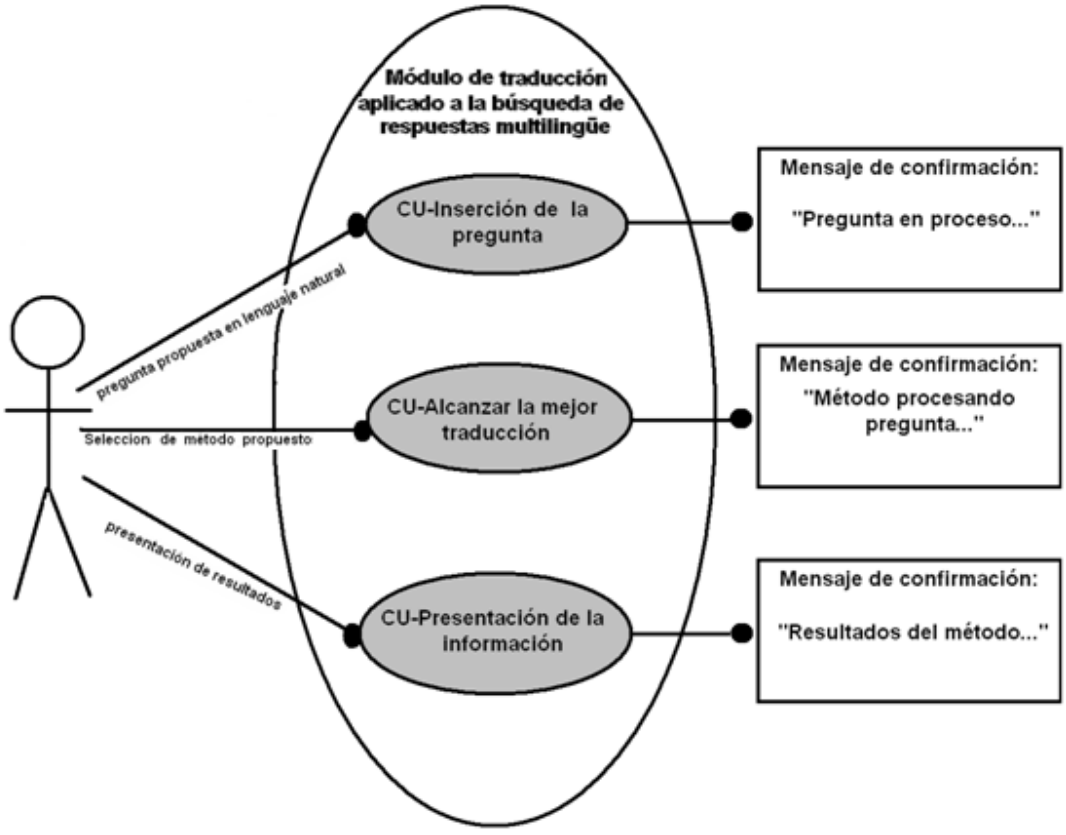


Figura 3. Ejemplificación de casos de uso

Tabla 4. Caso de uso inserción de pregunta

Descripción:	Se propone una pregunta en lenguaje natural
Actor principal:	Administrador
Personal involucrado e interés:	Asesor de proyecto: desea revisar la inserción de la pregunta
Precondiciones:	- Autentificación del usuario - Uso correcto de signos de interrogación
Garantías de éxito:	La traducción de la pregunta será fiable
Escenario principal de éxito o flujo básico:	El usuario insertará la pregunta en el sistema

Tabla 5. Caso de uso alcanzar la mejor traducción

Descripción:	Selección de uno del método propuesto
Actor principal:	Administrador
Personal involucrado e interés:	Asesor de proyecto: desea revisar los métodos de traducción de la pregunta
Precondiciones:	El administrador debe de haber insertado una pregunta en lenguaje natural
Garantías de éxito:	Si el usuario insertó la pregunta correctamente, la traducción será funcional
Escenario principal de éxito o flujo básico:	El usuario elegirá el método en la pantalla del sistema.

Tabla 6. Caso de uso presentación de la información

Descripción:	Vista de resultados
Actor principal:	Administrador
Personal involucrado e interés:	Asesor de proyecto: desea revisar la traducción de la pregunta

Precondiciones:	El usuario debió insertar una pregunta en lenguaje natural y posteriormente debió hacer la petición de muestra de resultados con el botón “aceptar”.
Garantías de éxito:	Si el usuario insertó la pregunta correctamente, la traducción será funcional
Escenario principal de éxito o flujo básico:	El sistema mostrará en pantalla los resultados.

3.3 CASOS DE PRUEBA

La Tabla 7 contiene la descripción de casos de prueba para los casos de uso de la Tabla 3.

Tabla 7. Descripción de casos de prueba

	Caso de prueba 1	**Caso de prueba 2**	**Caso de prueba 3**
Nombre	CP-Inserción de pregunta para el CU- Inserción de pregunta	CP-Alcanzar la mejor traducción para CU-Alcanzar la mejor traducción	CP-Presentación de la información para CU-Presentación de la información
Datos de entrada	¿Quién es el presidente de Canadá?	Método	Solicitud de presentación de resultados
Datos de salida	Mensaje de confirmación: “Who is the president from Canada?	Mensaje de confirmación: “Método [Selección de la mejor traducción] procesando [Who is the president from Canada?]	Mensaje de confirmación: “Resultado del método [Selección de la mejor traducción] = ¿Quién es el presidente de Canadá?

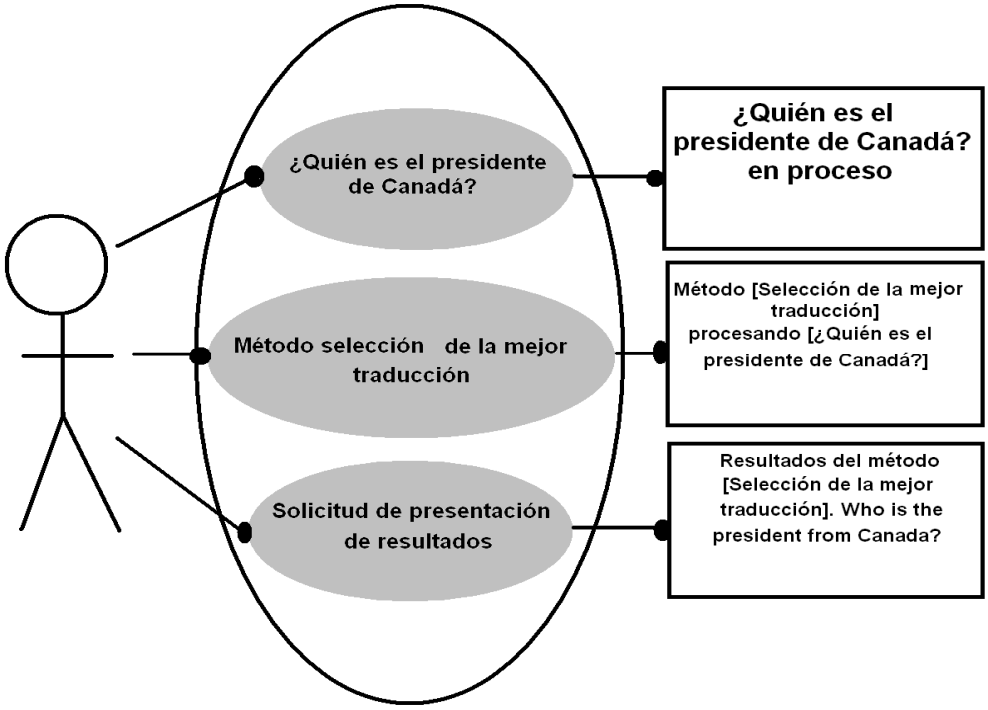


Figura 4. Ejemplificación de casos de prueba

3.4 Manejo de datos

El manejo de los datos y las herramientas que se utilizarán son:

- **Idiomas:** Español, Francés e Italiano.
- **Colecciones de búsqueda**. Documentos del Cross Language Evaluation Forum Workshop (CLEF 2004). Se utilizarán 160 preguntas relacionadas con el trabajo. Las colecciones del tienen 300 preguntas y de ellas se seleccionaron 160
- **Máquinas de traducción:** Systran, InterTiran y TraduceGratis.

Los datos que se utilizarán en este proyecto son sólo preguntas relacionadas con la consulta que se le pasará a un sistema de búsqueda de respuestas multilingüe. Esto es porque cualquier usuario puede hacer cualquier pregunta y así hacer uso del módulo de traducción. Los datos en este proyecto se organizarán por los siguientes tipo de pregunta: ¿Dónde?, ¿Cómo?, ¿Cuándo? y ¿Por qué?.

Los documentos y preguntas del CLEF 2004 son públicos, por lo tanto, sin costo para el proyecto. Además del programador del proyecto, cualquier usuario de un sistema de búsqueda de respuestas podrá tener acceso. El administrador será el responsable de dar mantenimiento al sistema. El mantenimiento será anual y las mejores traducciones de la pregunta serán presentados en formato *.txt. La Tabla 8 describe las características de los actores del sistema.

Tabla 8. Especificación del actor

Nombre:	Administrador
Descripción:	Persona que programa y gestiona la inserción y traducción de la pregunta
Características:	Persona con conocimientos en programación en java
Relaciones:	Actualiza el sistema Realiza la catalogación de las preguntas realizadas

La Figura 5 muestra el diagrama conceptual del sistema, el cual indica que desde la página principal se mostrará un menú de ayuda o se iniciará una sesión

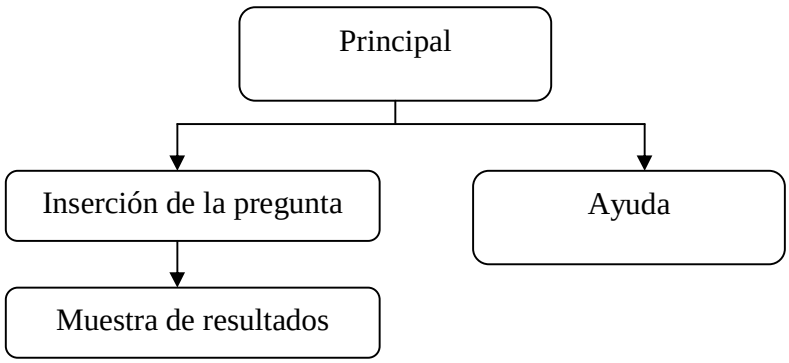


Figura 5. Ejemplificación de casos de prueba

Los datos de entrada, datos de salida y la descripción de algunos módulos de la Figura 5 se muestran en las Tabla 9-11.

Tabla 9. Descripción del módulo inserción de la pregunta

Descripción:	El usuario insertará una pregunta en lenguaje natural y se ejecutará el método “Selección de la mejor traducción”
Datos de entrada:	El usuario insertará una pregunta en lenguaje natural
Datos de salida:	Mensaje de confirmación

Tabla 10. Descripción del módulo muestra de resultados

Descripción:	El sistema mostrará los resultados al usuario
Datos de entrada:	Pregunta insertada por el usuario
Datos de salida:	El sistema mostrará la pregunta en lenguaje natural insertada por el usuario, la traducción obtenida y el método que se usó para su traducción

Tabla 11. Descripción del módulo de ayuda

Descripción:	Es una guía para el usuario
Datos de entrada:	NULA
Datos de salida:	El sistema mostrará al usuario la información necesaria para manejarlo

La implementación de los módulos de la Figura 5 se describe en el capítulo siguiente.

CAPÍTULO 4. IMPLEMENTACIÓN

4.1 Introducción

Este capítulo se divide en 3 secciones: 1) conexión con las páginas de los traductores en línea, 2) generación del modelo de lenguaje y 3) implementación del método: selección de la mejor traducción. La división anterior es conforme a las etapas de un sistema de búsqueda de respuestas (ver la Figura 1).

La Figura 6 ilustra el escenario de uso del sistema, el rectángulo de la izquierda representa el módulo de traducción, debajo están escritos los usuarios. El rectángulo de la derecha muestra la secuencia de tareas principales.

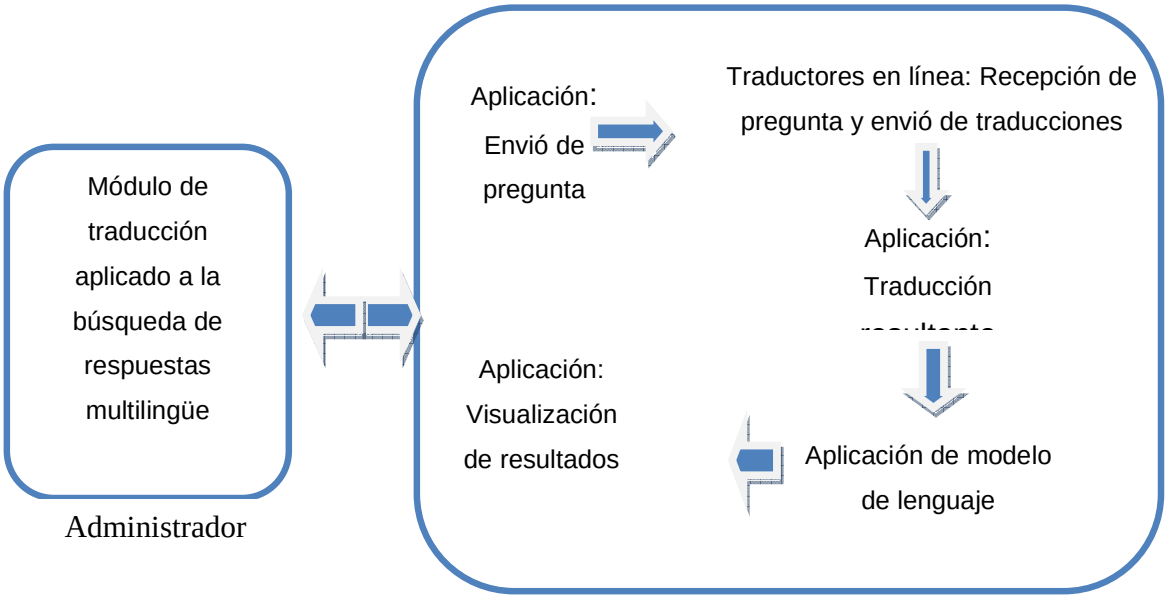


Figura 5. Ejemplificación del módulo de traducción

4.2 Conexión con traductores

La conexión con los traductores en línea: Systran, InterTiran y Traduce Gratis se realiza por medio de la herramienta llamada HTML Parser que analiza código HTML.

Método ***conexión ()***

```
private void Conexion_All() {
        try {
            parserSys =new Parser(ptsys,Parser.DEVNULL);
            parserInterTiran =new Parser(urlInterTiran,Parser.DEVNULL);
            parserTraduceGratis =new Parser(urlTraduceGratis,Parser.DEVNULL);
        }
        catch (ParserException e) {
            e.printStackTrace();
        }
        extraccion_Sys();
        extraccion_Intertiran();
        extraccion_TraduceGratis();
}
```

La URL de la página de cada traducción se pasa a través de una variable String, en este caso: ptsys (para systran), urlInterTiran (para InterTiran) y urlTraduceGratis (para TraduceGratis).

Método ***extracción ()***

Esta función para la extracción para systran, Intertiran y Traduce Gratis. La función es la misma para Intertiran y Traduce Gratis, solo se cambia la URL en la primera línea del Try. Esta función extrae todo el contenido de las páginas web en forma de texto.

```
protected void extraccion_Sys() {
    try {
        parser = new Parser (urlSys, Parser.DEVNULL);
        NodeIterator nd=parser.elements();
            while(nd.hasMoreNodes()){
                NodeIterator cp=nd;
                NodeList b=nd.nextNode().getChildren();
                SimpleNodeIterator sn=null;
                try{
                    sn=b.elements();
                    while(sn.hasMoreNodes()){
                        txtInfoCopia.setText(txtInfoCopia.getText()+sn.nextNode().toHtml().intern());
                    }
                }
                catch(Exception en){
                en.printStackTrace();
                txtInfoCopia.setText(txtInfoCopia.getText()+cp.nextNode().toHtml().intern());
                }

            }
    }
    catch (Exception e) {
    e.printStackTrace();
    }
txtInfoCopia.setText(txtInfoCopia.getText());

}
```

Anteriormente se ha extraído solo la línea (líneaIT, líneaTG, líneaS) donde se encuentra la traducción identificada por un textarea. Para extraer solo la traducción se utilizan las siguientes líneas de código. La traducción se queda en la variable String llamada “ins”

```
boolean bandera=false;
for(int i=0;i<=lineaIT.length();i++){
          if(lineaIT.charAt(i)=='>'){
                bandera=true;
                continue;
          }
```

```
else {
      if(lineaIT.charAt(i)=='<'){
            bandera=false;
            continue;
      }
}
if(bandera){
      ins+=lineaIT.charAt(i);
   }
```

4.3 Generación de modelo de lenguaje

Un modelo de lenguaje es la colección del vocabulario usado del propio lenguaje. Un modelo de lenguaje está compuesto por:

- LEXICO: vocabulario
- GRAMATICA: estructura y vocabulario

El modelo de lenguaje se realizara con ayuda de las colecciones del EFE 94 y EFE 95 proporcionadas por el CLEF [12].

La generación del modelo de lenguaje se realizó por medio de la herramienta CMU-Cam_Toolkit_v2. Se siguieron los pasos:

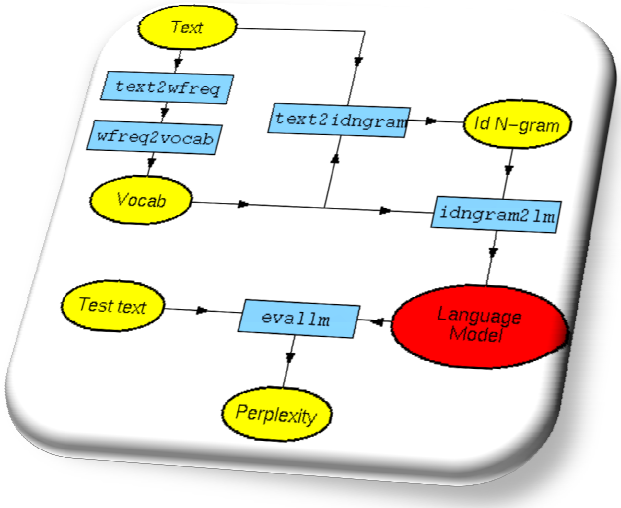


Figura 5. Secuencia de modelo de lenguaje

1. Se bajó la herramienta para la generación del modelo de lenguaje anteriormente mencionada.
2. Se modificó el archivo toolkit.h de la carpeta src de la siguiente forma: #define STD_MEM 100 a #define STD_MEM 200. Esto para aumentar la memoria que se utilizará para manejar el modelo de lenguaje.
3. Se abre una terminal en la cual se ingresó en la carpeta src: patricia@Patricia:-/Escritorio/CMU-Cam_Toolkit_v2/src$
4. Se ejecuta el comando Make Install
5. Las instrucciones se guardan en la carpeta **bin**
6. Posteriormente se trataron los documentos de la colección de EFE94 y EFE95: se quitaron las etiquetas y se unieron en un sólo archivo de texto.
7. Con este archivo al que se le dio el nombre de EFE.text se genera el modelo de lenguaje.
8. cat EFE.text | ./text2wfreq | ./wfreq2vocab -top 20000 > EFE.vocab. En esta instrucción se concatenan los resultados de la frecuencia de las palabras y el vocabulario resultante en un archivo llamado EFE.vocab.
9. cat EFE.text | ./text2idngram -vocab EFE.vocab | \ ./idngram2lm -vocab EFE.vocab -idngram - \ -binary a.binlm -spec_num 5000000 15000000 . Esta instrucción crea el modelo de lenguaje a partir del último archivo EFE.vocab.
10. echo "perplexity -text b.text" | ./evallm -binary EVALLPerplexity.binlm . Con esta instruccion se obtiene la perplejidad de una pregunta contenida en un archivo, en este caso b.text.

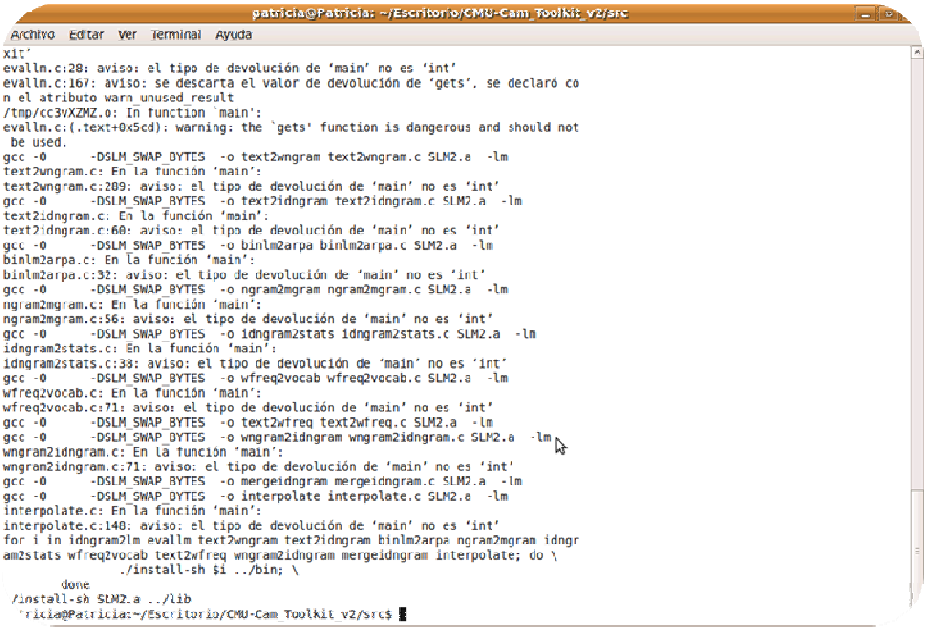


Figura 6. Instalación de la herramienta CMU-Cam_Toolkit_v2

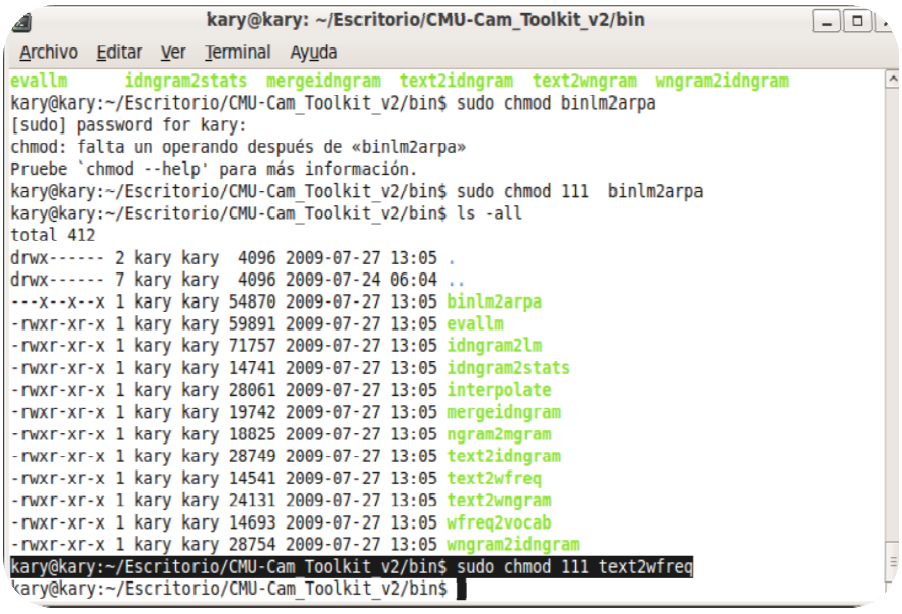


Figura 7. Instrucciones generadas en la carpeta Bin

En la Figura 6 se muestra como debe de quedar la instalación de la herramienta de modelo de lenguaje y en la figura 7 se muestra las 12 instrucciones ejecutables que se generaron en la carpeta Bin.

4.4 Implementación del método “selección de la mejor traducción”

Las traducciones generadas en archivos txt a partir de la conexión y la extracción de las páginas web son dadas como entradas al modelo de lenguaje que calcula la perplejidad con la siguiente instrucción, donde b.txt es una de las la traducciones:

echo "perplexity -text b.text" | ./evallm -binary EVALLPerplexity.binlm

Para guardar el cálculo de la perplejidad en un archivo, se utilizó la siguiente instrucción:

echo "perplexity -text b.text" | ./evallm -binary EFE.binlm > EFE.text

Cuando se obtuvieron ya las perplejidades de cada traducción, entonces se realizó la siguiente función que calcula la menor perplejidad a partir de la lectura de 3 archivos

```
public void CalculaPerplejidad(){
        File       a = null;
        File       d = null;
        File       g = null;
        FileReader         b = null;
        FileReader e = null;
        FileReader                  h = null;
        BufferedReader   c = null;
        BufferedReader             f = null;
        BufferedReader             i = null;

        boolean j=false;
   try {

        //para Intertitran
    a = new File ("PTInterTiran.txt");
    b = new FileReader (a);
    c = new BufferedReader(b);
    //para Systran
    d = new File ("PTSystran.txt");
    e = new FileReader (d);
    f = new BufferedReader(e);
    //para TraduceGratis
    g = new File ("PTTraduceGratis.txt");
    h = new FileReader (g);
    i = new BufferedReader(h);
    String linea,linea2,linea3;
```

```
//traduccion systran
 while((linea2=f.readLine())!=null){
        if(j=false){
        PTS_T="Traducción Sys:"+linea2; System.out.println(PTS_T);
        j=true;
        }
        else
                PTS=linea2;
 }//fin while
 j=false;
 //    traduccion TG
 while((linea2=i.readLine())!=null){
        if(j=false){
        PTTG_T="Traducción TG:"+linea2; System.out.println(PTTG_T);
        j=true;
        }
        else
                PTTG=linea2;
 }//fin while
 j=false;
 //    traduccion IT
 while((linea2=c.readLine())!=null){
        if(j=false){
        PTIT_T="Traducción IT:"+linea2;System.out.println(PTIT_T);
        j=true;
        }
        else
                PTIT=linea2;
 }//fin while
 j=false;
  PerplejidadS=Double.valueOf(PTS).doubleValue();
  PerplejidadIT=Double.valueOf(PTIT).doubleValue();
  PerplejidadTG=Double.valueOf(PTTG).doubleValue();
  ResultPerplejidad=0.0;
 ResultPerplejidad=Math.min(PerplejidadS,Math.min(PerplejidadIT,PerplejidadTG));
 System.out.println(PTS_T+"\n"+PTTG_T+"\n"+PTIT_T);
}     catch(Exception p){
            p.printStackTrace();
          }
 }
```

4.5 Interfaz

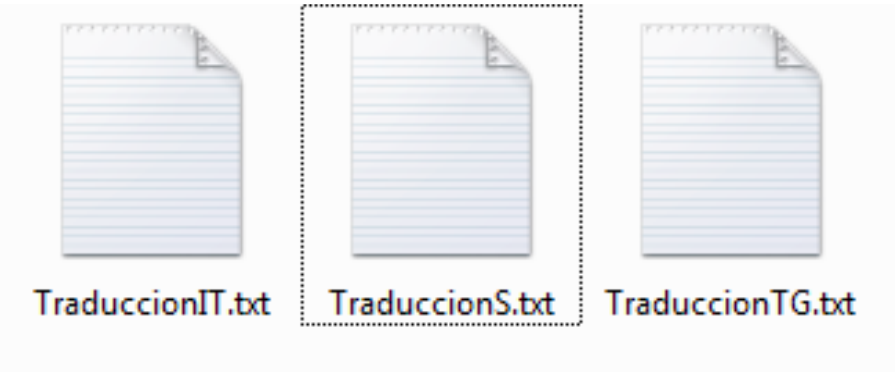
La interfaz del módulo de traducción tiene una sola pantalla, la cual está divida en tres partes: la parte superior contiene los datos descriptivos del módulo; en la parte izquierda se introducen los datos de entrada y en la parte derecha se muestran los resultados.

Figura 6. Interfaz del sistema

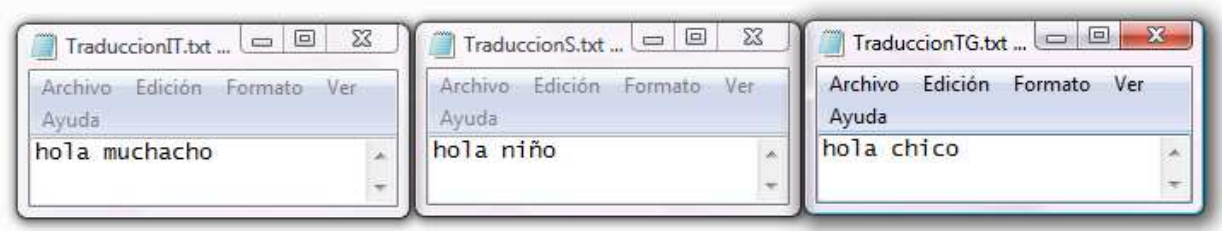
Según la Figura 6, la función principal del sistema es regresar a la vista del usuario la traducción correcta para mandarla a un sistema de búsqueda de respuestas. Esto lo realiza por medio del método programado antes mencionado. En la primera parte se inserta la pregunta en lenguaje natural, se elige el par de idiomas (origen-destino) y se presiona el botón "Aceptar". El botón llevará a cabo las tareas de conectar la aplicación con los traductores en línea ya antes descritos, también se encarga de mandar los parámetros a los traductores, es decir, la pregunta y el par de idiomas. A cada una de las traducciones se les aplicará un modelado de lenguaje, el cual usa y evalúa su perplejidad. La traducción con menor perplejidad se mostrará como resultado.

CAPÍTULO 5. RESULTADOS

Al ejecutar el programa se generan los siguientes archivos, que contienen las traducciones extraídas de las páginas web del traductor asociado (Systran, InterTiran y TraduceGratis).

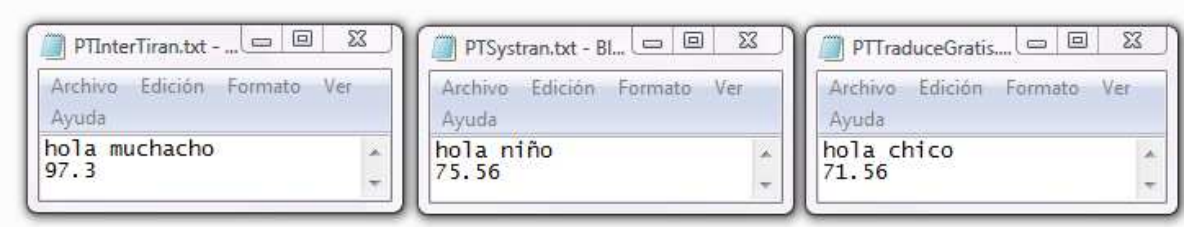


Cada archivo contiene:



Estos archivos se mandan al modelo de lenguaje para obtener la perplejidad de cada traducción. Esto genera 3 nuevos archivos:

Que contienen:



Estos archivos son leídos por el programa y calcula cual es la menor perplejidad de las tres traducciones

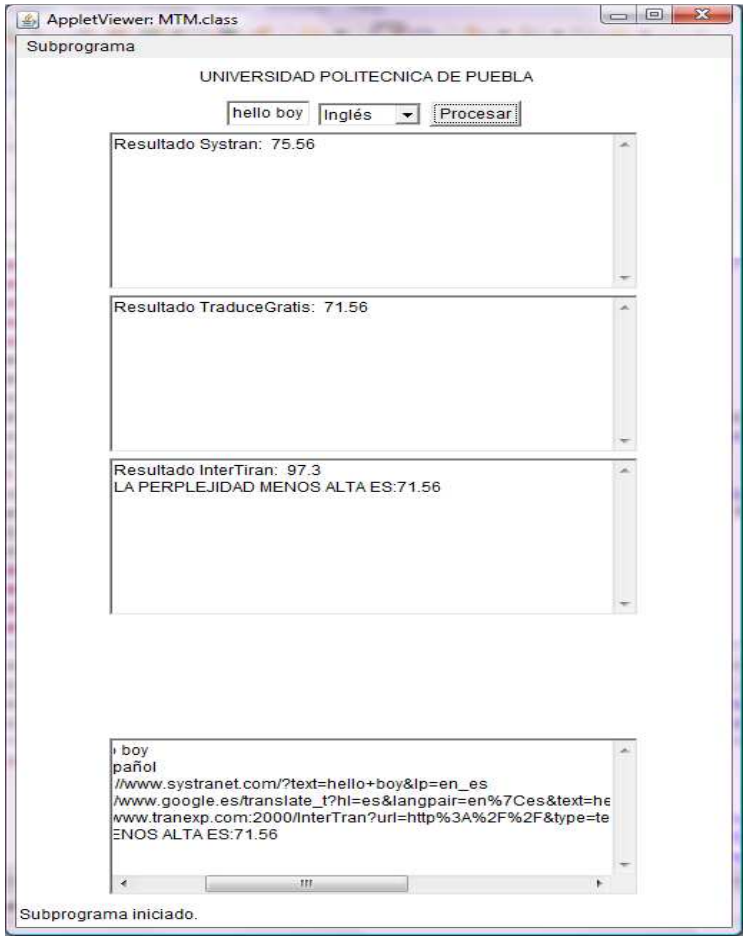


Figura 7. Ejemplo de la traducción “hello boy” de Ingles a Español mostrando su perplejidad

CAPÍTULO 6. CONCLUSIONES

Este proyecto es la primera fase de un sistema de Búsqueda de Respuestas Multilingüe. Esta fase constituye la traducción de la pregunta.

El método que se programó para seleccionar la mejor traducción depende de obtener diferentes traducciones con diferentes traductores automáticos, por lo tanto, es necesario encontrar páginas web de traductores que presten su servicio de forma gratuita, lo cual no siempre es posible. Ya que los traductores utilizados no tienen restricciones.

Se realizó un modelo de lenguaje con colecciones de periódicos del EFE, de los años 94 y 95. Para ello se utilizó una herramienta creada para la plataforma Unix: CMU Toolkit.

El trabajo futuro es realizar la segunda fase del proyecto que es la fusión de respuestas candidatas.

El procesamiento de las colecciones del EFE 94 y EFE 95 se generaron en un solo archivo.

REFERENCIAS

[1] Aceves P. R., Montes G. M., Villaseñor P. L., 2007. Enhancing Cross-Language Question Answering by Combining Multiple Question Translations. Lecture Notes in Computer Science, Vol. 4394, Springer 2007.

[2] Fernández A., Ureña A., Díaz V. 2002. "Construcción de un sistema de recuperación multilingüe en la web". Cross Language Evaluation Forum Workshop (CLEF). 2002.

[3] Juárez G. A., Extracción de Respuestas mediante Aprendizaje Automático utilizando Atributos Léxicos. Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Tesis de Maestría, 2007.

[4] Gatius M., González M., 2001. Un sistema de diálogo multilingüe dirigido por la semántica". Tesis de Doctorado. Software Deparment.2001.

[5] *Montes G., M., Villaseñor P, L.,Pérez C, M., Gómez S, J. M.,Sanchis-Arnal, E. & Rosso, P*. Joint Participation in Cross Language Evaluation Forum Workshop (CLEF) 2005: Experiments in monolingual question answering. In working notes of cross language evaluation forum workshop. 2005, Vienna, Austria. Instituto nacional de astrofísica, óptica y electrónica (INAOE-UPV) 2005.

[6] Peñas P. A. 2003, Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe.

[7] Strzalkowski, T. 1999, Natural Language Processing Information Retrieval. Kluwer, Boston,MA.

[8] Systran 5.0 Guía del usuario. Systran Software, inc.

Páginas Web

[9] Traductor InterTiran. © 1998-2004 v 8.0
http://www.tranexp.com:2000/Translate/result.shtml

[10] Traductor Systran Copyright 2009 Systran v 6.0
http://www.systransoft.com/

[11] Traductor Traduce Gratis: © 200-2009 v 8.0
http://www.traducegratis.com/

[12] Multilingual Question Answering, Answer Fusion, Cross Language Evaluation Forum Workshop (CLEF):

http://www.clef-campaign.org/