INFORME FINAL DE LA EVALUACIÓN PARA TRADUCCIÓN AUTOMÁTICA DE TEXTO EN CASTELLANO A LENGUA DE SIGNOS ESPAÑOLA

1.- Introducción

El presente Plan de Evaluación para la Traducción Automática de Texto (en castellano) a Lengua de Signos Española (LSE) trata de promover la investigación y avance del estado del arte en Traducción Automática mediante:

- La definición de una tarea novedosa
- La creación y gestión de evaluaciones formales de las prestaciones de las tareas
- La provisión de herramientas y utilidades de evaluación

Las evaluaciones proporcionan una importante contribución a los esfuerzos de investigación y la comprobación de las capacidades técnicas. Su intención es ser de provecho para todos los investigadores que trabajan en el problema de la traducción entre lenguajes humanos. Para ello, las evaluaciones se han diseñado para resultar simples, ceñidas a temas esenciales y accesibles.

La evaluación consiste en la traducción de texto en un lenguaje origen, Castellano, a un lenguaje destino, Lengua de Signos Española. Los textos consistirán en frases de uso administrativo.

Los participantes se comprometen a la presentación de los resultados de la evaluación en una sesión especial que tendrá lugar durante las IV Jornadas en Tecnología del Habla. La participación se realiza a modo individual o equipo formado por un máximo de 3 investigadores donde al menos uno de ellos debe ser estudiante.

2.- Medición de prestaciones

Las prestaciones se medirán usando una técnica automática de puntuación basada en co-ocurrencia de N-gramas llamada BLEU-4.

Esta técnica de puntuación evalúa las traducciones de segmento en segmento. En este caso los segmentos serán frases de texto delimitadas en el fichero origen, su organización debe ser preservada por la traducción. En este contexto un N-grama es una secuencia de N símbolos, donde las palabras y los signos de puntuación se cuentan por separado. La técnica de co-ocurrencia de N-gramas puntúa la traducción según el número de N-gramas que ésta comparte con una o más traducciones de referencia de alta calidad.

BLEU-4 fue desarrollado por IBM (1) y proporciona estimaciones estables de las prestaciones del sistema que encajan bien con las valoraciones humanas de la calidad de la traducción. Detalles del uso de la técnica para medir la calidad de las traducciones se pueden encontrar en la página web del NIST (2).

(1) *Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu* (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation". http://domino.watson.ibm.com/library/CyberDig.nsf/home. (clave = RC22176)

(2) http://www.nist.gov/speech/tests/mt/doc/ngram-study.v26.pdf

3.- Condiciones de evaluación

Se proporcionará una base de datos de entrenamiento formada por 200 frases de texto y su traducción a Lengua de Signos Española. Se ofrecerá una única traducción por frase de origen. La evaluación está orientada a un dominio concreto de aplicación: frases pronunciadas por un funcionario que atiende a las personas que desean sacar o renovar el DNI o el pasaporte.

4.- Formato de los archivos

Se utilizará el mismo formato que el utilizado pro NIST: se dispone de un conjunto de etiquetas SGML para dar formato a los archivos de prueba, traducidos y de referencia para el proceso de evaluación. Los sistemas de traducción deben ser capaces de recibir los documentos origen en este formato y producir la salida traducida de la misma forma.

4.1.- Archivo origen

Cada archivo origen para la evaluación comienza con la etiqueta (**srcset**), seguida de un documento (**DOC**). Cada uno de éstos estará formado por uno o varios segmentos (**seg**). A su vez, cada segmento dispondrá de un identificador. Toda etiqueta de apertura dispondrá de su correspondiente etiqueta de cierre. Un ejemplo de archivo origen es el siguiente:

```
<srcset setid="mt-cast-v0" srclang="Castilian">
<DOC docid="NYT-doc1" genre="text">
<seg id="1"> TEXTO EN CASTELLANO </seg>
<seg id="2"> TEXTO EN CASTELLANO </seg>
…
<seg id="n"> TEXTO EN CASTELLANO </seg>
</DOC>
</srcset>
```

La etiqueta (**srcset**) tiene dos atributos requeridos, **setid** y **srclang**, y uno implícito, **trglang**. El primero contiene el nombre del conjunto de prueba y será distinto para cada uno de ellos. El segundo la lengua original, castellano. El último corresponde a la lengua de las traducciones del sistema y normalmente no se utiliza.

La etiqueta (**DOC**) tiene dos atributos requeridos, **docid** y **genre**, y uno implícito, **sysid**. Los primeros identifican el documento y el tipo dentro del conjunto, y el último no se utiliza normalmente.

La etiqueta (**seg**) tiene un atributo, **id**, que identifica el segmento.

4.2.- Archivo de prueba de texto traducido

Cada archivo de prueba traducido comienza con la etiqueta (**tstset**), seguida de un documento (**DOC**) que contiene la traducción con el sistema de traducción identificado (**sysid**). Cada una de éstas estará formada por uno o varios segmentos (**seg**). A su vez, cada segmento dispondrá de un identificador. Toda etiqueta de apertura dispondrá de su correspondiente etiqueta de cierre. Un ejemplo de archivo de prueba traducido es el siguiente:

```
<tstset setid="mt-cast-v0" srclang="Castilian" trglang="LSE">
<DOC docid="NYT-doc1" genre="text" sysid="JTH_castellano_grande">
<seg id="1"> TEXTO TRADUCIDO A LSE </seg>
<seg id="2"> TEXTO TRADUCIDO A LSE </seg>
…
<seg id="n"> TEXTO TRADUCIDO A LSE </seg>
</DOC>
</tstset>
```

La etiqueta (**tstset**) tiene dos atributos requeridos, **setid** y **srclang**, y uno implícito, **trglang**. El primero contiene el nombre del conjunto de prueba que debe coincidir con el del conjunto origen. El segundo la lengua de origen, castellano. El último corresponde a la lengua de las traducciones que debe ser LSE.

La etiqueta (**DOC**) tiene dos atributos requeridos, **docid** y **genre**, y uno implícito, **sysid**. Los primeros identifican el documento y el tipo dentro del conjunto, y el último se utiliza para identificar el sistema de traducción.

4.3.- Archivo de referencia

Los archivos de referencia tendrán la misma estructura que los archivos de prueba traducidos, excepto en el uso de la etiqueta (**refset**) en lugar de (**tstset**).

5.- Datos de evaluación

Para la evaluación se proporcionará un conjunto de prueba compuesto por 81 frases en castellano, sobre las que se ejecutará la traducción automática. Para la evaluación se debe generar una única traducción por frase.

6.- Procedimiento para la evaluación

El procedimiento con las fechas para la evaluación es el siguiente:

- El 21 de Julio de 2006 se dispondrá de los planes de evaluación y se abre el periodo de inscripción.
- La fecha límite de inscripción será el 15 de Septiembre de 2006.
- A partir del 16 de Agosto de 2006 se podrá disponer del material de entrenamiento y desarrollo para las distintas evaluaciones. Es necesario estar inscrito en la evaluación para recibir el material.
- El 18 de Octubre de 2006 se liberarán las bases de datos para la evaluación.
- El 27 de Octubre de 2006 a las 24:00 es la fecha límite para recibir los resultados en el formato y método indicados.
- El 3 de Noviembre de 2006 se enviarán los resultados de la evaluación.

7.- Envío de traducciones

Las traducciones se enviarán por correo electrónico a la organización. Deberán ser completas, conteniendo por tanto todo el conjunto de datos de evaluación.

Las traducciones incluidas en un fichero con el formato indicado en el apartado 4.2 y con la extensión "**.sgm**" deberán enviarse a: **Rubén San-Segundo Hernández** (lapiz@die.upm.es) . Se debe enviar un fichero por cada uno de los sistemas que se desee evaluar.

Los resultados estarán disponibles una vez se hayan enviado dichos resultados a los participantes. Esto permitirá realizar análisis previos a la celebración de las IV Jornadas de Tecnologías del Habla.

Cada participante deberá remitir una descripción del sistema enviado a la evaluación, que deberá incluir:

- Identidad elegida para el sistema (**`sysid`**)
- Condiciones de evaluación (base de datos de entrenamiento)
- Descripción de la aproximación algorítmica

Esta descripción se enviará en formato de texto ASCII o PDF. Las descripciones recibidas se distribuirán como parte del material de análisis de la evaluación.

8.- Evaluación de sistemas

Para la evaluación de los diferentes sistemas se computaron las siguientes medidas:

- **WER (Word Error Rate):** porcentaje de signos erróneos en la frase resultado de la traducción.
- **PIWER (Position Indepent WER):** porcentaje de signos erróneos en la frase resultado independiente de la posición del signo.

- **BLEU (Bilingual evaluation understudy)**
- **NIST:** Similar al BLEU pero con ponderación de la longitud de las frases.

8.1 Sistema GRAH-DEE por UPV-EHU

Contacto: Alicia Pérez, email: webperaa@lg.ehu.es,
Tel: 946015364

El sistema **sysid = upv-ehu sfst** consiste en un transductor estocástico de estados finitos con suavizado por back-off. El suavizado permite modelar los n-gramas no vistos. Adicionalmente, incluye un tratamiento para palabras desconocidas que se desarrolla en la fase de búsqueda. Este mecanismo considera que la palabra desconocida es equivalente a alguna de las palabras vistas, pero se queda sin definir hasta analizar el análisis de la cadena de entrada.

8.2 Sistema DSIC-DSI por UPV

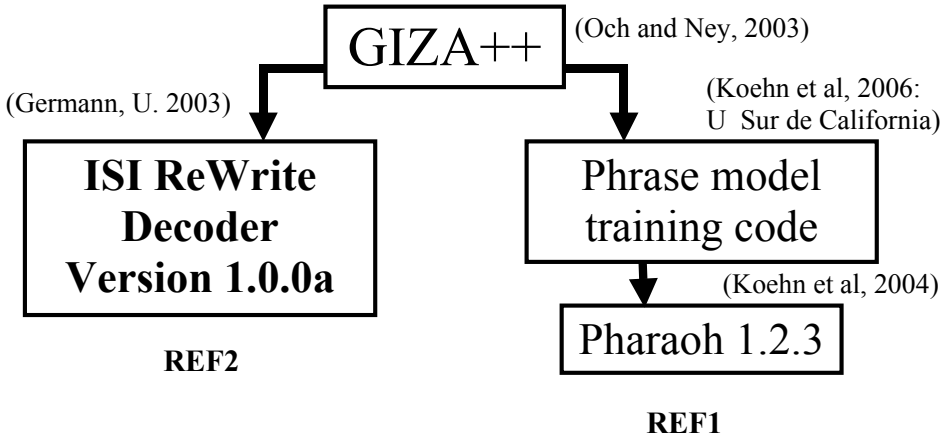
Contacto: Daniel Ortiz, email: dortiz@iti.upv.es,
Tel: 963877069

El sistema de traducción upv_phrase utilizado en la evaluación de la tarea de traducción de texto a lengua de signos, consiste en un traductor basado en el enfoque algorítmico de ramificación y poda, y está apoyado en modelos de traducción estadísticos. Específicamente, el sistema utiliza un algoritmo de pila que lleva a cabo una maximización en base a modelos estadísticos de traducción y lenguaje. Como algoritmo de traducción se emplearon modelos estadísticos de secuencias, y como modelos de lenguaje se emplearon los bien conocidos modelos estadísticos de n-gramas.

8.3 Sistemas REF1 y REF2

Contacto: Rubén San-Segundo, email: lapiz@die.upm.es,
Tel: 915495700 ext 4228.

Con el objetivo de incrementar el interés de la evaluación en esta actividad se realizaron experimentos con dos sistemas de traducción basada en modelos estadísticos que se muestran en la siguiente figura.



El modelo de traducción estadística se obtiene utilizando el programa GIZA++ para los dos sistemas. En el caso del REF1 se entrena un modelo de frase que se aplica en traducción utilizando el programa Pharaoh versión 1.2.3. En el caso de REF2 se utiliza el ISI ReWrite Decoder versión 1.0.0 que utiliza un modelo de traducción de palabras.

8.4 Resultados finales

Los principales resultados de los sistemas comentados son los siguientes:

SISTEMA	WER	PIWER	BLEU	NIST
UPV-EHU	**28.90**	**20.93**	**0.62**	**7.06**
DSIC-DSI	35.33	23.79	0.54	6.48
REF 2	40.13	27.97	0.50	6.12
REF 1	53.59	37.67	0.21	5.23

El mejor sistema ha sido el propuesto por UPV-EHU donde cabe resaltar la estrategia de tratamiento para palabras desconocidas en el espacio de búsqueda. Esta estrategia ha conseguido marcar diferencias importantes en un tarea con un elevado número de palabras fuera de vocabulario.