INFORME FINAL DE LA EVALUACIÓN DE TRADUCCIÓN AUTOMÁTICA DE CASTELLANO A EUSKARA

Nerea Ezeiza

n.ezeiza@ehu.es

1.- Introducción

El propósito de esta evaluación es el fomentar la investigación en sistemas de traducción automática entre castellano y euskara e impulsar la colaboración entre grupos teniendo en cuenta que existen menos recursos tanto bilingües como monolingües relacionados con el euskara.

Con dicho propósito se plantea un plan de evaluación que consiste en traducir artículos divulgativos de consumo de castellano a euskara. Para ello se plantean dos opciones. La primera consiste en desarrollar un traductor utilizando todos los grupos los mismos recursos para poder comparar diferentes aproximaciones bajo las mismas condiciones. La segunda opción permitirá presentar sistemas que utilicen otros recursos disponibles, como pueden ser corpus adicionales, diferentes procesadores de lenguaje o diccionarios.

Los participantes se comprometen a la presentación de los resultados de la evaluación en una sesión especial que tendrá lugar durante las V Jornadas en Tecnología del Habla. La participación se realiza a modo individual o equipo formado donde el representante del mismo deberá ser estudiante. Cada equipo podrá presentar uno o más sistemas.

2.- Medición de prestaciones

Las prestaciones se medirán usando varias técnicas de puntuación automática. Estas medidas comparan la salida de la traducción automática con las traducciones manuales de referencia. Se usarán las medidas habituales BLEU, NIST, WER y PER. BLEU (Papineni *et al.* 2002) y NIST (Doddington 2002) medidas basadas en la similitud de subsecuencias (N-gramas) de la traducción automática y la de referencia. La tasa de error por palabra WER mide las inserciones, supresiones y sustituciones entre la traducción automática y la de referencia. La tasa de error por palabra independiente de la posición PER calcula la distancia entre el conjunto de palabras que contiene la traducción automática y el de la traducción de referencia. Para la evaluación se utilizarán las herramientas de TC-STAR (www.tc-star.org).

3.- Condiciones de evaluación

Se evaluarán dos tipos de sistemas: los sistemas basados en los recursos proporcionados y los sistemas de recursos sin restricciones. Para los sistemas basados en recursos limitados se proporcionarán 58.000 frases para entrenamiento, aproximadamente 1.500 para ajustar el sistema y otras 1.500 para test. Se ofrecerá una única traducción por frase de origen.

El corpus se proporcionará en formato texto, tokenizado y alineado a nivel de frase y se entregará por separado la parte correspondiente a cada idioma. Alternativamente, se proporcionará el texto lematizado y etiquetado automáticamente. Cada línea corresponderá a una frase tanto en el fichero de origen como en el de destino. Todos los procesos aplicados son automáticos, por lo que el corpus contendrá errores. El test final consistirá en traducir varios artículos de divulgación del mismo

dominio no incluidos en el material inicial y estará procesado de forma análoga al resto de material.

4.- Procedimiento para la evaluación

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

- El 1 de Mayo de 2008 se dispondrá de los planes de evaluación y se abre el periodo de inscripción.
- La fecha límite de inscripción será el 31 de Mayo de 2008.
- A partir del 16 de Junio de 2008 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 15 de Septiembre de 2008 se proporcionarán los datos para la evaluación.
- El 30 de Septiembre de 2008 a las 24:00 es la fecha límite para recibir los resultados.
- El 31 de Octubre de 2008 se publicarán los resultados de la evaluación entre los participantes.

5.- 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. Deberán enviarse a: Nerea Ezeiza (nerea.ezeiza@gmail.com). Se debe enviar un fichero por artículo y por sistema 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 V 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 (datos de entrenamiento)
- En el caso de los sistemas sin restricciones, deberán indicarse las características de los recursos utilizados para el desarrollo del mismo:
  - En el caso del corpus debe indicarse:
    - si se ha utilizado el corpus proporcionado o no
    - si es un corpus alternativo, una breve descripción del contenido, tamaño aproximado, si es monolingüe o bilingüe y si está públicamente accesible
  - En el caso de los diccionarios, tamaño aproximado, si son monolingües o bilingües y si está públicamente accesible
  - En el caso de procesadores del lenguaje
    - tipo de procesador (analizador morfológico / POS / sintáctico superficial / semántico /...)
    - si está públicamente accesible
    - idiomas a los que se les ha aplicado
- Descripción de la aproximación algorítmica

Esta descripción se enviará con la misma plantilla utilizada para las comunicaciones regulares de las V Jornadas en Tecnología del Habla. Las descripciones recibidas se distribuirán como parte del

material de análisis de la evaluación.

6.- Evaluación de sistemas

De los cinco equipos que se inscribieron en la tarea, únicamente tres de ellos han finalizado la tarea. Los equipo presentados son Avivavoz, IXA y UPV-PRHLT. El test consta de 1003 oraciones en castellano procedentes de artículos de diversos dominios de la revista Eroski Consumer (http://www.consumer.es). Estos artículos han sido alineados manualmente y las traducciones al euskara de las oraciones han sido supervisadas para evitar grandes diferencias en longitud y contenido (semántico) de las mismas. No se han hecho cambios sustanciales, por lo que las oraciones de referencia siguen bastante fielmente el original.

Cada uno de los equipos ha enviado la traducción del test correspondiente a cuatro configuraciones diferentes, por lo que la evaluación final consta de 12 sistemas.

6.1.- Métricas utilizadas

Para la evaluación de los diferentes sistemas se han computado las siguientes medidas: BLEU, NIST, WER y PER. Para la evaluación se han utilizado las herramientas de TC-STAR (www.tc-star.org).

Para comparar entre sí los sistemas y valorar todas esas medidas en conjunto, se han escalado los resultados al rango de 0 a 1, lo que nos dará como resultado un ranking relativo y un peso por cada medida. El peso máximo (1) corresponde al sistema con mejor resultado y el mínimo (0) al peor. El valor final dado a cada sistema corresponderá a la suma de sus cuatro pesos.
En concreto, para las medidas BLEU y NIST se calcula de la siguiente manera:

$$D_{bleu} = \frac{BLEUi - MIN}{MAX - MIN} \qquad D_{nist} = \frac{NISTi - MIN}{MAX - MIN}$$

donde BLEUi/NISTi corresponde a la medida del sistema *i,* y MAX y MIN corresponden a los resultados mejor y peor respectivamente.

En el caso de WER y PER, dado que los resultados mejores corresponden a los valores menores, las diferencias se calculan según las siguiente fórmulas:

$$D_{WER} = 1 - \frac{WERi - MIN}{MAX - MIN} \qquad D_{PER} = 1 - \frac{PERi - MIN}{MAX - MIN}$$

Todas las medidas han sido valoradas por igual, aunque se podría haber planteado alternativas a esta combinación de métricas.

6.2.- Descripción de los sistemas

A continuación de presenta la descripción de los sistemas propuestos. El corpus empleado consta de 58.000 frases para entrenamiento, aproximadamente 1.500 para desarrollo y otras 1.500 para test procedentes de la revista Eroski Consumer (Alcázar 2005). Se facilitaron los textos etiquetados y lematizados. Algunos de los sistemas han utilizado esta información y otros se han desarrollado basándose en palabras.

Todos los equipos han realizado el aprendizaje con el texto en minúsculas, aunque el único equipo que ha restaurado las mayúsculas en la salida ha sido Avivavoz.

6.2.1.- Avivavoz

Nombre de contacto: José B. Mariño (canton@gps.tsc.upc.edu)

Avivavoz-System 1

Avivavoz-System1 es un sistema basado en frases que utiliza los parámetros por defecto de Moses. No se han utilizado ni las etiquetas POS ni los lemas proporcionados para construir el sistema. En ambas lenguas se han convertido las mayúsculas en minúsculas y se ha tokenizado. Al finalizar la traducción se han restaurado las mayúsculas utilizando la herramienta *disambig* de SRILM. El sistema aplica una simetrización *grow-diagonal-final* en el alineado, un modelo de lenguaje de 5-gramas y reordenamiento lexicalizado.

Avivavoz-System 2

Avivavoz-System2 también es un sistema basado en frases construido con Moses. En este caso, las etiquetas POS de euskara han sido utilizadas para crear un modelo de lenguaje POS. El alineado y el reordenamiento se realizan como en el sistema anterior. Finalmente, además de utilizar el modelo de lenguaje de 5-gramas de palabras, incluye un modelo de lenguaje de 7-gramas de POS.

Avivavoz-System 3

Avivavoz-System3 incluye las características de Avivavoz-System2, y difiere de éste en que se ha realizado un preproceso adicional previo a la fase de alineado. El corpus en euskara ha sido segmentado antes del alineamiento y se ha alineado con el corpus de castellano. Después de alinear ambos corpus, se ha desegmentado la parte de euskara para recuperar las palabras originales. El proceso de desegmentación modifica los enlaces de manera que mantiene la misma relación entre palabras. Por ejemplo, si el segmento en euskara 'a' está vinculado a la palabra en castellano 'x' y el segmento euskara 'b' a la palabra 'z', entonces el desegmentados enlazará la palabra en euskara 'ab' con las palabras en castellano 'x' y 'z'.

Avivavoz-Final

Avivavoz-Final incluye todas las características de Avivavoz-System3 y añade un modelo de lenguaje de 7-gramas de lemas.

6.2.2.- UPV-PRHLT

Nombre de contacto: Germán Sanchís (gsanchis@dsic.upv.es)

-Sysid: UPV-PRHLT
- Datos de entrenamiento: los proporcionados por la organización, en lowercase, exclusivamente.
- Procesadores del lenguaje empleados: FreeLing 2.0
  - Tipo de procesador: parentizado
  - Publicamente accesible: sí, en http://garraf.epsevg.upc.es/freeling/
  - Aplicado a: Castellano

- Descripción de la aproximación algorítmica:
  - Extracción de segmentos mediante SITGs, inicializadas heurísticamente y con dos iteraciones de reestimación de las probabilidades.
  - Componentes del modelo log-lineal: modelos directo e inverso de conteo, modelos directo e

inverso léxicos, modelos directo e inverso sintácticos.
- Búsqueda monótona.

6.2.3.- IXA[1]

Nombre de contacto: Gorka Labaka (gorka.labaka@ehu.es)

Sistema Baseline

Sistema basado en Moses, entrenado sobre el corpus facilitado por la organización. En este sistema no hemos llevado a cabo ningún procesado en el texto original ni hemos incluido ningún tipo de información morfológica adicional.

Los modelos usados en el traductor son los que se utilizan en la configuración por defecto de Moses. Un modelo del lenguaje basado en 5-gramas, cinco modelos de traducción (probabilidad léxica en ambas direcciones, probabilidad traducción directa e inversa, penalización por *phrase*) y un modelo de reordenamiento lexicalizado[2] además del reordenamiento basado en distancia.

Finalmente, se ha usado Minimum Error Rate Training (MERT) para optimizar los pesos de cada modelo sobre la métrica BLEU.

Sistema Matrex

Este sistema se basa en el baseline, por lo que utiliza los mismos modelos y procesos de optimización. La diferencia entre ambos sistemas se encuentra en la extracción de frases. Mientras el sistema Baseline extrae las frases mediante el algoritmo presentado en (Koehn *et al*. 2003), el sistema Matrex añade a lo anterior unas frases extraídas mediante métodos de traducción automática basada en ejemplos (Stroppa and Way 2006), donde las frases coinciden con los sintagmas marcados por un parser.

Sistema Seg

Este sistema también está basado en el sistema Baseline por lo que utiliza los mismos modelos que este. La diferencia en este caso consiste en el preproceso del texto en euskara, donde cada palabra se ha dividido en diferentes tokens, separando por un lado los prefijos (si los hay), por otro lado el lema y por último los sufijos (si los hay). Para separar los distintos elementos de cada palabra nos hemos basado en el análisis morfológico conseguido por *EusTagger*. Este cambio conlleva que el modelo de lenguaje utilizado en el traductor esté basado en estos nuevos tokens en vez de en palabras.
Tras el proceso de traducción necesitamos una fase de generación donde basados en la salida de Moses (formada por los tokens sobre los que ha sido entrenado) generamos la forma final de las palabras en euskara. Para poder utilizar un modelo de lenguaje basado en palabras después del proceso de generación, hemos usado una lista n-best donde el traductor devuelve las 200 traducciones con mayor probabilidad y tras la generación se reordena esta lista incorporando un modelo de lenguaje basado 4-gramas.

---

1 Este trabajo ha sido subvencionado por Gobierno Vasco, mediante la ayuda predoctoral concedida a Gorka Labaka (código BFI05.326)

2 http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc1

Sistema Matrex+Seg

El sistema final, es una combinación de los sistemas Matrex y Seg. Además de entrenar el sistema sobre el texto segmentado, con todo lo que ello conlleva (postproceso de generación, reordenamiento de lista 200-best basado en el modelo de lenguaje), se incorporan las frases extraídas usando técnicas de traducción automática basada en ejemplos. Este sistema es el que mejores resultados ofrece y el que presentamos a la competición como sistema final.

6.3.- Resultados finales

Estos son los resultados de todos los sistemas evaluados (sin tener en cuenta la tipografía):

SISTEMA	BLEU	NIST	PER	WER	Dbleu	Dnist	Dper	Dwer	Total
avivavoz-system1	0.0801	3.8702	64.1728	81.2158	0.898	0.697	0.449	0.438	2.481
avivavoz-system2	0.0780	3.8120	64.7789	82.0359	0.732	0.540	0.275	0.254	1.801
avivavoz-system3	0.0791	3.7936	65.6881	82.9867	0.819	0.490	0.014	0.041	1.364
avivavoz-final	0.0812	3.8985	64.2203	81.6021	0.984	0.773	0.435	0.351	2.544
UPV-PRHLT.n1.g2.VII	0.0687	3.6145	65.6584	83.1709	0.000	0.007	0.022	0.000	0.029
UPV-PRHLT.n3.g2.VII	0.0695	3.6542	65.0582	82.5113	0.063	0.114	0.195	0.148	0.520
UPV-PRHLT.n5.g2.Vlex	0.0696	3.6118	65.7357	82.7193	0.071	0.000	0.000	0.101	0.172
UPV-PRHLT.n5.g2.VII	0.0711	3.6522	65.5574	82.6420	0.189	0.109	0.051	0.118	0.468
IXA.baseline	0.0731	3.7382	64.9929	82.1072	0.346	0.341	0.213	0.238	1.139
IXA.seg	0.0814	3.9281	63.4359	80.2650	1.000	0.853	0.660	0.650	3.164
IXA.matrex	0.0772	3.8004	64.7136	81.7982	0.669	0.509	0.294	0.307	1.779
IXA.matrex+seg	0.0810	3.9825	62.2534	78.7022	0.969	1.000	1.000	1.000	3.969

Resultados de los sistemas evaluados

Dado que en un principio cada equipo propone un sistema final, se han evaluado éstos por separado, y los resultados son los siguientes:

SISTEMA	BLEU	NIST	PER	WER	Dbleu	Dnist	Dper	Dwer	Total
Avivavoz-final	0.0812	3.8985	64.2203	81.6021	1.000	0.746	0.405	0.264	2.414
PRHLT.n5.g2.VII	0.0711	3.6522	65.5574	82.6420	0.000	0.000	0.000	0.000	0.000
IXA.matrex+seg	0.0810	3.9825	62.2534	78.7022	0.980	1.000	1.000	1.000	3.980

Resultados de los sistemas finales

El mejor sistema es el propuesto por IXA, en el que se realiza la traducción basada en segmentación morfológica, lo que obliga a aplicar generación morfológica para obtener las palabras finales. Además realiza procesado sintáctico para extraer pares de sintagmas que se incorporan en las tablas de traducción.

El sistema Avivavoz ha obtenido resultados similares al de IXA, aunque los datos de WER y PER son peores.

Cabe destacar que el sistema UPV-PRHLT no ha utilizado información de segmentación morfológica y la traducción ha sido monótona, lo cual ha podido influir en el menor rendimiento de su propuesta.

En cuanto a los resultados tras la restauración de mayúsculas, la caída en el rendimiento de los sistemas es evidente, como se puede observar en la siguiente tabla:

SISTEMA	BLEU	NIST	PER	WER
UPC-system1	0.0681	3.4881	67.0430	83.0818
UPC-system2	0.0668	3.4310	67.6789	83.9375
UPC-system3	0.0675	3.4320	68.5168	84.8110
UPC-final	0.0689	3.5079	67.1738	83.4977

Resultados con restauración de mayúsculas

Referencias

Alcázar A. 2005. "Corpus Consumer: Towards linguistically searchable text". Invited Speaker. Bilbao-Deusto International Student Conference in Linguistics.

Doddington, G. 2002 "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". In Proc. ARPA Workshop on Human Language Technology, San Diego, California, March 2002.

Koehn P., Och F.J., and Marcu D. 2003. "Statistical Phrase-Based Translation". Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), May 27-June 1, Edmonton, Canada.

Papineni K., Roukos S., Ward T., and Zhu W.-J. 2002 "Bleu: a method for automatic evaluation of machine translation". In Proc. of the 40th Annual Meeting of the ACL, Philadelphia, PA, July 2002, pp. 311–318.

Stroppa, N. and Way A. 2006. "MaTrEx: DCU Machine Translation System for IWSLT 2006". In Proceedings of the International Workshop on Spoken Language Translation, pages 31-36, Kyoto, Japan.