SISTEMA DE TRADUCCIÓN DE TEXTO A LENGUA DE SIGNOS UPV_PHRASE

Daniel Ortiz-Martínez[1] *, Ismael García-Varea*[2] *, Francisco Casacuberta*[1]

[1] Universidad Politécnica de Valencia

[2] Universidad de Castilla-la Mancha

RESUMEN

En este artículo se describen brevemente las características del sistema de traducción etiquetado como upv_phrase que participó en la evaluación de la tarea de traducción de texto a lengua de signos organizada con motivo de las IV Jornadas en Tecnología del Habla celebradas en Zaragoza. Adicionalmente también se describen la principales características del conjunto de datos de entrenamiento disponible para la citada tarea.

1. CORPUS DE ENTRENAMIENTO

Una de las tareas a evaluar propuestas en las IV Jornadas en Tecnología del Habla celebradas en Zaragoza consistió en la traducción automática de texto (en castellano) a Lengua de Signos Española (LSE).

La Tabla 1 muestra las principales estadísticas del conjunto de datos de entrenamiento disponible para la tarea.

	Texto	**LSE**
Pares de frases	200	
Pares de frases distintos	190	
Palabras	2240	2075
Vocabulario	443	260
Longitud media frases	11.2	10.4

Tabla 1. Principales cifras del corpus de entrenamiento para la tarea de traducción de texto a lengua de signos.

Como puede apreciarse, no se dispone de una cantidad muy elevada de pares de frases. Por otro lado, los tamaños de vocabulario son relativamente grandes teniendo en cuenta el reducido tamaño del corpus.

El conjunto de frases en castellano a traducir estaba compuesto por un total de 81 frases, de las cuales un total de 39 (esto es, un $48\,\%$) contenían palabras no vistas en el vocabulario.

Pese a no disponer de un corpus de validación con el que hacer experimentos de traducción, podemos afirmar, a la vista de los datos presentados, que la tarea de traducción propuesta posee una dificultad cuando menos moderada.

Este trabajo ha sido parcialmente subvencionado por el proyecto CICYT TIC2003-08681-C02-02, la *Agencia Valenciana de Ciencia y Tecnología* dentro del contrato GRUPOS03/031, la *Generalitat Valenciana*, y el proyecto HERMES II (Vicerrectorado de Investigación - UCLM-06/07)

2. DESCRIPCIÓN DEL SISTEMA DE TRADUCCIÓN

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.

En los siguientes apartados se aporta, en primer lugar, una breve descripción de los fundamentos de la traducción automática estadística, en segundo lugar se introducen las técnicas de estimación de modelos estadísticos de secuencias de palabras empleados; por último, se describen brevemente los algoritmos de pila utilizados para hacer la búsqueda apoyada en los modelos estadísticos de traducción y lenguaje.

2.1. Traducción automática estadística

El proceso de traducción puede formularse desde un punto de vista estadístico de la siguiente forma: Sea $f_1^J = f_1 \dots f_J$ la frase origen que queremos traducir en su equivalente en el lenguaje destino $e_1^I = e_1 \dots e_I$. Se considera que cualquier posible frase de la lengua destino es traducción de la frase origen con una probabilidad a posteriori determinada $Pr(e_1^I|f_1^J)$. Según la regla de Bayes, la frase destino que buscamos $\hat{e}_1^I$ será aquella que maximiza[1] el producto del modelo del lenguaje destino $Pr(e_1^I)$ y el modelo de traducción $Pr(f_1^J|e_1^I)$. La ecuación que modela el proceso es la siguiente:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \qquad (1)$$

[1] Obsérvese que la expresión debe ser maximizada también para I; sin embargo, para simplificar se asume que I es conocida.

En función de cómo se concibe la relación entre las palabras origen y destino, se han propuesto diferentes *modelos de traducción*; que intentan explicar la manera en que se genera la frase destino a partir de la frase origen. Esta relación se describe a través del concepto de *alineamiento*; dando lugar a diferentes *modelos estadísticos de alineamiento*. Los modelos de alineamiento IBM fueron propuestos en [1]. Estos modelos pertenecen a la categoría de los *modelos basados en palabras*, ya que asumen que en el proceso de traducción se establecen relaciones entre palabras individuales de las frases origen y destino. En los últimos tiempos, se ha demostrado que estos modelos no capturan adecuadamente la información de contexto a la hora de traducir, por lo que se han propuesto modelos que trabajan con grupos de palabras en lugar de palabras, los cuales han constituído una mejora con respecto a los modelos basados en palabras originales descritos en [1].

2.2. Estimación de modelos estadísticos de secuencias de palabras

Los modelos de secuencias de palabras surgen como alternativa a los modelos de palabras (o más comúnmente conocidos como modelos de IBM [1]) para superar las limitaciones que presentan, de modo que en lugar de trabajar con diccionarios estadísticos de palabras trabajan con diccionarios estadísticos de secuencias.

La traducción, usando modelos de secuencias de palabras, de una frase de entrada f_1^J en la frase destino equivalente e_1^I consiste, desde un punto de vista generativo, en escoger la forma en que dicha frase de entrada es segmentada en K secuencias $f_1^J = \tilde{f}_1^K$, seleccionar las secuencias en el lenguaje destino que traducen las secuencias origen y, por último, reordenar; con lo que terminamos obteniendo $e_1^I = \tilde{e}_1^K$. Podemos asumir que las relaciones entre las palabras origen y destino se sumarizan mediante una variable oculta $\tilde{\mathbf{a}} = \tilde{a}_1^K$, que contiene todas las decisiones que se hacen durante la historia generativa.

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}, \tilde{f}_1^J|\tilde{e}_1^I) \\ &= \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}|\tilde{e}_1^I) Pr(\tilde{f}_1^J|\tilde{\mathbf{a}}, \tilde{e}_1^I) \end{aligned} \quad (2)$$

Se pueden hacer diferentes asunciones a partir de la ecuación anterior, aunque lo normal es que los modelos terminen convirtiéndose en diccionarios estadísticos de segmentos. Por ejemplo, en [2] se propone el siguiente modelo:

$$p_\theta(f_1^J, e_1^I) = \alpha(e_1^I) \sum_{\tilde{\mathbf{a}}} \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) \quad (3)$$

donde $\tilde{a}_k$ denota el índice de la secuencia origen $\tilde{e}$ que se alinea con la k'ésima secuencia destino $\tilde{f}_k$ y se asume que todas la posibles segmentaciones tienen la misma probabilidad.

Estimación de parámetros del modelo

Una posible técnica de estimación de los parámetros de los modelos de secuencias de palabras pasa por utilizar matrices de alineamiento a nivel de palabra [3]. Dichas matrices se obtendrán automáticamente como subproducto de la estimación de modelos IBM. Concretamente, dado un par de frases y su correspondiente matriz de alineamiento A, se extraerán aquellos pares de secuencias que sean *consistentes* con la matriz de alineamiento. La condición de consistencia viene dada por la ecuación (4) [4]. La Figura 1 muestra un ejemplo en el que se da un par de frases con su matriz de alineamiento, y el conjunto de todos los pares bilingües consistentes que se pueden extraer.

$$\begin{aligned} \mathcal{BP}(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n} : \forall (i', j') \in A : \\ j \leq j' \leq j+m \iff i \leq i' \leq i+n\} \end{aligned} \quad (4)$$

secuencia origen	secuencia destino
La	the
casa	house
verde	green
casa verde	green house
La casa verde	the green house
.	.
casa verde .	green house .
La casa verde .	the green house .

Figura 1. Conjunto de pares bilingües consistentes (derecha) para una matriz de alineamiento dada (izquierda).

Una vez que se han extraído los pares de secuencias, las probabilidades del modelo se calculan a través de las frecuencias relativas de los mismos.

$$p(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} count(\tilde{f}, \tilde{e})} \quad (5)$$

Un importante inconveniente que poseen los modelos de secuencias es la gran cantidad de espacio en memoria requerida por sus parámetros, por lo que serán necesarias técnicas especiales para su estimación y manejo cuando se trabaja con corpus muy grandes.

Para obtener los modelos de secuencias utilizados por el sistema upv_phrase se utilizó la herramienta Thot de estimación de modelos estadísticos de secuencias. Dicha herramienta es una herramienta de código abierto que puede descargarse en *Sourceforge*: `http://thot.sourceforge.net/`

2.3. Búsqueda con algoritmos de pila

Los algoritmos de pila fueron introducido inicialmente en el dominio del reconocimiento del habla por Jelinek en 1969[5], su funcionamiento consiste en el desarrollo incremental de hipótesis parciales, cada una de las cuales tiene asignada una *puntuación* también parcial. Dichas hipótesis son almacenadas en una pila o cola de prioridades,

de manera que quedan ordenadas por su puntuación dentro de la pila.

Los algoritmos de pila son algoritmos que construyen su solución iterativamente, extrayendo de la cima de la pila la hipótesis más prometedora (aquella con mayor puntuación) en cada iteración y sometiéndola a un proceso de *expansión*. El concepto de *expansión* de una hipótesis puede definirse, usando la terminología básica de los problemas de búsqueda, como la aplicación de todos los operadores posibles sobre esa hipótesis.

Lo expresado anteriormente puede resumirse con el diagrama de flujo mostrado en la Figura2

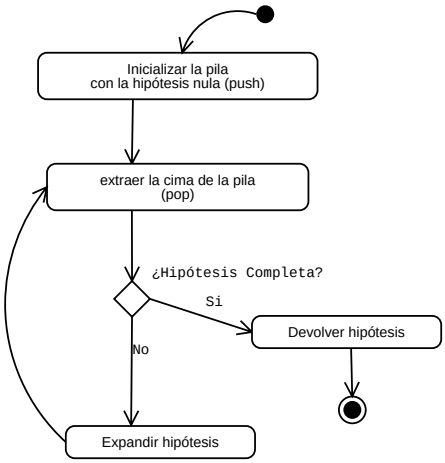

Figura 2. Diagrama de flujo del algoritmo *stack decoding*

3. BIBLIOGRAFÍA

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, y R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] J. Tomás y F. Casacuberta, "Monotone statistical translation using word groups," in *Procs. of the Machine Translation Summit VIII*, Santiago de Compostela, Spain, 2001, pp. 357–361.

[3] P. Koehn, F. J. Och, y D. Marcu, "Statistical phrase-based translation," in *Proceedings of the HLT/NAACL*, Edmonton, Canada, May 2003.

[4] Franz Joseph Och, *Statistical Machine Translation: From Single-Word Models to Alignment Templates*, Ph.D. thesis, Computer Science Department, RWTH Aachen, Germany, October 2002.

[5] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM Journal of Research and Development*, vol. 13, pp. 675–685, 1969.