Desambiguación lingüística guiada por dominios de conocimiento distribuidos para traducción de lengua española a LESCO

Johan Serrato-Romero
TEC Digital
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
jserrato@itcr.ac.cr

Mario Chacón-Rivas
Escuela de Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
machacon@itcr.ac.cr

***Resumen*—La necesidad de un conjunto grande de ejemplos de frases para las metodologías actuales de traducción y la creciente investigación sobre las jóvenes y dinámicas lenguas de señas, implican la búsqueda de nuevas estrategias para flexibilizar la inclusión de nuevo vocabulario y producirlo correctamente de acuerdo al contexto del discurso. En este artículo se presenta el diseño e implementación de una arquitectura distribuida de hilos semánticos para un sistema de traducción de lengua española a la Lengua de Señas Costarricense (LESCO). Es mostrada la efectividad de un algoritmo de reconocimiento contextual propuesto para esta arquitectura a través de pruebas de textos divulgativos y el uso de corpus lingüísticamente validados.**

***Palabras clave*—desambiguación por contexto, lengua de señas, traducción automática, LESCO, accesibilidad**

I. Introducción

La Lengua de Señas Costarricense (LESCO) es la lengua materna de la comunidad sorda costarricense, que como otras lenguas de señas en el mundo, posee una gramática y léxico muy diferentes a las de una lengua oral. Su origen y establecimiento son lingüísticamente recientes, y eso implica la poca cantidad de recursos tales como un vocabulario estandarizado, detallados corpus ejemplificando su uso o alguna clase de representación textual que la denoten completamente.

Los investigadores en traducción automática hacia una lengua de señas aprovechan distintos esfuerzos realizados para la interpretación estructurada del texto de entrada, para después centrarse en encontrar la mejor forma de transferir esa interpretación inicial, llevarla a un medio de reproducción como un avatar tridimensional y hacerlo eficientemente, con el fin de ofrecer prototipos funcionales capaces de modificar y aumentar su vocabulario.

El mayor problema actualmente es el acoplamiento entre el estado del arte en traducción automática y las necesidades de información para hacer funcional un traductor de señas, específicamente la definición de ejemplos de frases en un corpus de lengua de señas computacionalmente asequible, que aún debe ser generado y validado por la comunidad sorda, por lo que se debe considerar otros enfoques para generar resultados prácticos a corto plazo.

La propuesta de esta investigación consiste en elaborar una arquitectura inspirada en un modelo de procesamiento lingüístico mental concurrente, donde varios *gestores de contexto* (ver sección III–B) reconocen patrones en la entrada y asignan un valor de pertenencia para su dominio lingüístico/temático, para generar una representación correcta si el valor de pertenencia es alto. Entre los principales objetivos está abordar el problema de desambiguación contextual con el enfoque de distribuir su análisis entre estos módulos especializados y evaluar el grado de facilidad en la integración de nuevo vocabulario guiado por contextos temáticos en forma incremental.

El resto de este documento se encuentra conformado de la siguiente manera: la sección II menciona el trabajo más reelevante relacionado a traducción automática a señas. Sigue la sección III describe la metodología propuesta, para luego en la sección IV explicar los experimentos y principales resultados obtenidos. La sección V muestra las conclusiones y la sección VI termina describiendo los trabajos futuros.

II. Trabajo relacionado

Existen varios trabajos relacionados con la traducción de español hacia la Lengua de Señas Española (LSE) con un enfoque general de la arquitectura basada en reglas y el proceso secuencial de los análisis morfológico - sintáctico - semántico ([1], [2]). El trabajo de [3] representa un traductor basado en reglas basado en un estudio de contraste de la gramática de esa lengua de señas con el castellano, logrando mejoras respecto a una versión estadística de la herramienta. Para la generación de las señas, [4] presenta un enfoque basado en un modelo fonológico flexible para facilitar ajustes manuales en la definición de las señas de su vocabulario.

Una propuesta sobresale por parte de [5], que plantea un nuevo modelo de traductor basado en *triggering rules*, que serían definiciones de reglas de generación de elementos lingüísticos para una lengua de señas. Estas reglas serían ejecutadas por módulos intermedios de reconocimiento de patrones en el texto de entrada, de modo que ante la detección de un patrón específico, se hace la relación con la regla de

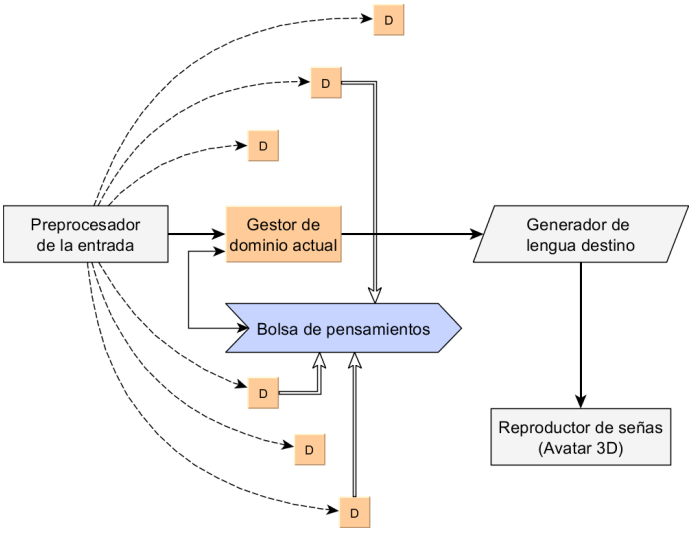

Figura 1. Arquitectura propuesta. El procesador envía a todos los gestores de dominio los *tokens* del texto a traducir, luego las flechas blancas indican los resultados de los dominios más significativos que van a terminar en la bolsa de pensamientos. El gestor de dominio actual interactúa con la bolsa para elegir el mejor contexto, y brinda el resultado al generador de la lengua destino, que se encarga de construir la descripción física de las señas que reproducirá el avatar.

generación correspondiente, minimizando con ello problemas relacionados con la complejidad morfológica de la lengua de señas y la falta de un corpus paralelo representativo para el desarrollo de la investigación. Del trabajo publicado, aún queda resolver el problema de construir una traducción coherente a partir de los resultados disparados por las reglas generadoras.

III. Arquitectura implementada

La arquitectura para traducción automática de lengua española a LESCO guiada por dominios de conocimiento distribuidos, es ilustrada en la figura 1.

La hipótesis es que la distribución del problema de traducción a contextos temáticos, con reconocimiento de patrones lingüísticos, aumente la calidad de traducción final al especializarse la generación de representaciones para la LESCO. Lo anterior corresponde con las observaciones de trabajos recientes en el campo de la traducción a lengua de señas tales como [5] y [3], así como su relación con el estudio de contextos definitorios [6].

Este procesamiento del lenguaje es naturalmente paralelizable y modularizable, de modo que permite establecer jerarquías de redes, donde un módulo principal es el que toma la decisión final de la iterpretación de la entrada. Otra ventaja de esta modularización es la posibilidad de mezclar varias técnicas de enfoques de traducción especializados en contextos en los que funcionen muy bien, simplificando su implementación y mantenimiento.

La arquitectura propuesta se implementa en C++ como un módulo de servicio para la desambiguación, adaptado a la estructura del prototipo del Traductor LESCO. Esta se compone de un servidor FreeLing 4.0 para el análisis lingüístico de la entrada en español [7], una base de datos *PostgreSQL 9.4* para la gestión de datos de vocabulario en español y LESCO, un *back-end PHP 5.6* con lógica de traducción y un *front-end* JavaScript incluyendo un avatar tridimensional programado en *Unity 5.4*. La figura 2 muestra la forma en que se comunican estos componentes.

A continuación se describe la forma en que cada parte de la arquitectura de gestores de domino fue construida para su utilización en el Traductor LESCO.

A. *Preprocesador*

El preprocesador de la entrada realiza dos tareas principales:

1. Analizar sintácticamente el texto de entrada con una herramienta de parseo morfosintáctico. Se utiliza FreeLing para esta tarea.
2. Publicar de forma secuencial cada frase, construyendo una estructura que es compartida por los hilos de ejecución que representan a los demonios. Cada frase incluye un análisis de dependencia con estructura de árbol sintáctico.

B. *Gestores de contexto (demonios)*

La implementación de los demonios se basa en un sistema concurrente con un hilo que toma el rol del gestor actual para centralizar los resultados de los demás gestores. Se han programado gestores de tipo *contextual*, encargados de detectar las frases más representativas de un tema específico, tan general o específico como se requiera dependiendo del tamaño del corpus que lo alimente. Al terminar el procesamiento de cada frase, la sincronización se realiza con barreras de la biblioteca *pthreads*, donde cada uno de los hilos creados guarda su resultado en la estructura de la bolsa de pensamientos y se prepara para la siguiente frase.

Cada gestor de contexto ejecuta el mismo procedimiento genérico para el análisis de la frase, la diferencia está en los datos del modelo del contexto:

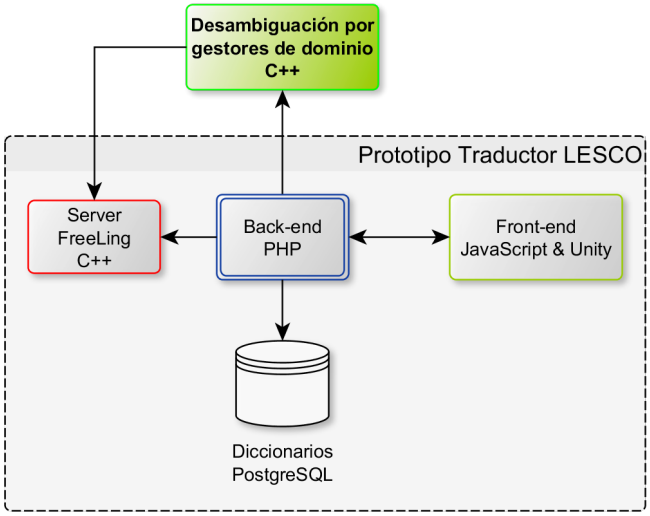

Figura 2. Integración de módulo desambiguador con el prototipo actual del Traductor LESCO.

1. En la inicialización del programa, existe una carpeta `models` que contiene los archivos de modelo de contexto. Por cada uno de ellos se creará un hilo de gestor contextual.
2. Al terminar el análisis de la entrada, se procede a analizar el árbol de dependencia correspondiente a una frase por parte de los gestores y se busca un patrón de entrada en el modelo descrito por el archivo. Este recorrido otorga un *porcentaje de pertenencia* de dicha frase al contexto temático del gestor.

C. Bolsa de pensamientos

La bolsa de pensamientos, ilustrada en la figura 3, consiste en un vector de referencia hacia los resultados de los gestores, donde existe el valor semántico de su interpretación por contexto. Cada posición del vector esta asociada al identificador entero del demonio, en una relación uno a uno.

Además, el *marcador de contextos* es otra estructura que lleva el valor acumulado de los índices de cada gestor contextual. La finalidad de este marcador es que sea una forma de medir qué contexto tiene mayor incidencia en los patrones del texto en un momento determinado, independientemente del contexto con mayor puntaje en una frase determinada. Esto puede ayudar a seleccionar el contexto ante casos como marcadores muy parecidos entre contextos.

D. Comportamiento del gestor actual

El gestor actual ejecuta una sección del algoritmo de análisis donde se escoge un contexto y se puntúa la frase en la bolsa de pensamientos. La implementación de la selección para el gestor actual es así:

- Al analizar la frase inicial, el demonio consciente es el primero en una lista indexada de hilos. Cada vez que se desconoce el contexto de una frase, es decir, todos los índices de pertenencia en 0, este hilo se selecciona como el gestor actual.
- Al finalizar el análisis de una frase, el demonio consciente revisa la bolsa de pensamientos para escoger el contexto con el índice de pertenencia más alto. Cada puntuación de frase se suma al marcador de contexto, que determina el valor semántico del documento analizado en los contextos que vayan surgiendo.

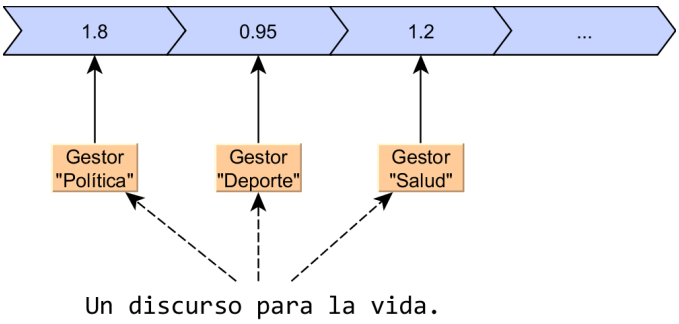

Figura 3. Ejemplo de una bolsa de pensamientos con tres interpretaciones de tres demonios con su correspondiente valor semántico.

- El demonio consciente limpia la bolsa de pensamientos para proceder al análisis de la siguiente frase.

E. Interfaz con generación de LESCO

Para la generación se acoplan los resultados del proceso de desambiguación con el módulo de síntesis ya programado en el Traductor LESCO. Para este proceso, el módulo correspondiente recibe los datos en formato JSON con la secuencia de frases analizadas por el módulo desambiguador. A continuación, pasará por varias descripciones particulares para su reproducción en un avatar, incluyendo gestos manuales, velocidades, formas de mano y deletreos en caso de conceptos sin seña en la base de datos del generador.

IV. Resultados experimentales

Los experimentos están orientados al análisis de textos de carácter informativo para el público general, usando corpus especializados y validados lingüísticamente, donde se espera que estas variables se manifiesten en el análisis de los resultados y así ayudar a identificar criterios para la elección de un contexto temático, como es determinar índices numéricos de aceptación de una frase. Para comprobar esto con la arquitectura implementada, se realizan dos clases de pruebas: una con textos largos relacionados a una contexto temático y otra con análisis de tres grupos de oraciones individuales.

Para las pruebas de textos largos, se analizan los índices de pertenencia generados por el algoritmo de puntuación de cada contexto, sacando sus relaciones y determinando las condiciones para determinar el comportamiento de los porcentajes de pertenencia adecuados para aceptar un contexto.

Los datos a probar para el primer grupo experimental están compuestos de cinco contextos temáticos específicos: economía, ambiente, derecho, informática y medicina. Los recursos textuales usados son de dos tipos:

- Para construir los modelos de contexto, se utilizaron los archivos en español del corpus temático del Instituto Universitario de Lingüística Aplicada (IULA)[1], una organización de postgrado especializada en estudiar el comportamiento de varias lenguas a partir de esta fuente de datos [8].
- Para las entradas de texto, se han escogido notas informativas sobre cada uno de los cinco temas que maneja el corpus de la IULA, además de una nota adicional sobre un tema desconocido (música).

Como primer resultado, se han determinado variables de la configuración del módulo de análisis contextual que modifican los resultados del proceso, apreciables en la tabla I.

Otro resultado derivado de los experimentos realizados con corpus lingüísticamente validados, es que los índices de pertenencia de frases y varios picos de detección corresponden con el índice promedio de oraciones propias del contexto, estando alrededor del 0.05, lo cuál evidencia una cota arrojada por el algoritmo de desambiguación propuesto. La figura 4

[1]El IULA es parte de la Universitat Pompeu Fabra (UPF), Barcelona, España.

Tabla I
VARIABLES DEL PROCESO PARA ANÁLISIS CONTEXTUAL

Variable	Efectos esperados
Tamaño de corpus	Grande: menor porcentaje de pertenencia por nodo. Pequeño: mayor porcentaje de pertenencia por nodo.
Cantidad de contextos	Muchos: más gestores reaccionan ante oraciones desconocidas. Pocos: casi ninguna opción en contextos desconocidos.
Extensión de la entrada	Una entrada grande daría un resultado más exacto; un entrada muy pequeña mantendría un nivel de ambigüedad significativo.
Ambigüedad temática de la entrada	Un texto que trata de varios temas conocidos por el traductor puede contener patrones compartidos entre contextos y producir índices de pertenecia similares.
Estilo de redacción	Tanto para corpus como para la fuente a analizar, este criterio puede hacer la detección de patrones sea más o menos efectiva para textos generales. Por ejemplo: glosarios, diálogos, definiciones, texto técnico descriptivo, narraciones.

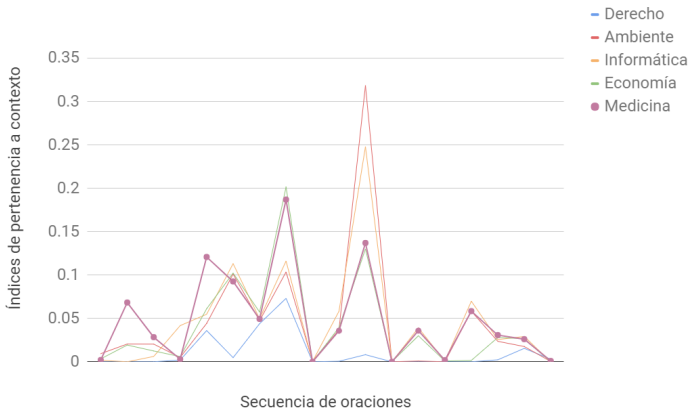

Figura 4. Comportamiento de gestores de contexto sobre el ejemplo de *Medicina*. Se destacan con un punto los índices de cada frase obtenidos por el gestor de contexto "Medicina".

ilustra este comportamiento de los gestores de dominio para la nota sobre el contexto *Medicina*, donde existen picos de detección del gestor "Medicina" que superan este valor de forma más regular respecto a los otros gestores de contexto.

V. CONCLUSIONES

La implementación de la arquitectura funciona y ofrece resultados esperados con corpus especializados validados lingüísticamente, ofreciendo la flexibilidad buscada para la generación de señas guiada por contextos y casos especiales de generación de la lengua de señas.

En el algoritmo de detección de contexto propuesto, existe un umbral de acción alrededor del 0.05 de índice de pertenencia, que es superado por la mayor parte de las frases correctamente clasificadas. Este comportamiento ayuda a definir una cota mínima para la aceptación del resultado de un gestor de contexto.

Para el algoritmo de detección propuesto, existen factores que permiten cambiar la calidad de las respuestas del sistema: tamaño de corpus, cantidad de contextos, extensión del texto a traducir y grado de ambigüedad temática del texto a traducir conocidos.

VI. INVESTIGACIÓN FUTURA

A partir de la investigación realizada se identificaron nuevas líneas de trabajo que ofrecen oportunidades de mejorar en las diversas áreas de conocimiento involucradas en este proceso:

- **Gestores de patrones lingüísticos además de contexto**, que detectarían estructuras lingüísticas importantes por su regularidad de generación en LESCO, por ejemplo: listas de términos, números, fechas y deletreo.
- **Nodos genéricos en el grafo contextual**, para aislar conceptos de tipo número, fecha, magnitud y otros conceptos cuya generalización aporta significado en el contexto.
- **Gestor de contexto como red neuronal**, de modo que pequeñas redes neuronales especializadas detecten patrones de texto directamente de forma más controlada y se puedan evitar problemas de aprendizaje como el *overfitting*.
- **Jerarquía contextual de demonios**, pudiendo tener demonios generales que utilizan el resultado de detección de otros más específicos, pudiéndose construir ontologías de conocimiento.
- **Detección de similitudes lingüísticas entre distintos contextos**, dado que es posible que dos o más gestores de contexto puedan detectar valores similares de índice de pertenencia de una misma frase, se podría inferir que la semántica de esta puede ser la misma en los contextos involucrados.

REFERENCIAS

[1] S. Baldassarri and F. Royo-Santas, *An Automatic Rule-Based Translation System to Spanish Sign Language (LSE)*. London: Springer London, 2009, ch. 1, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1007/978-1-84882-352-5_1

[2] V. López-Ludeña, R. San-Segundo, J. M. Montero, R. Córdoba, J. Ferreiros, and J. M. Pardo, "Automatic categorization for improving spanish into spanish sign language machine translation," *Computer Speech and Language*, vol. 26, no. 3, pp. 149–167, june 2012. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2011.09.003

[3] J. Porta, F. López-Colino, J. Tejedor, and J. Colás, "A rule-based translation from written spanish to spanish sign language glosses," *Comput. Speech Lang.*, vol. 28, no. 3, pp. 788–811, May 2014. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2013.10.003

[4] F. López-Colino and J. Colás, "Hybrid paradigm for spanish sign language synthesis," *Universal Access in the Information Society*, vol. 11, no. 2, pp. 151–168, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10209-011-0245-9

[5] M. Filhol, M. N. Hadjadj, and B. Testu, "A rule triggering system for automatic text-to-sign translation," *Universal Access in the Information Society*, pp. 1–12, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10209-015-0413-4

[6] G. Sierra, "Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos," *Linguamática*, vol. 1, no. 2, pp. 13–37, December 2009. [Online]. Available: http://linguamatica.com/index.php/linguamatica/article/view/38

[7] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA, May 2012. [Online]. Available: http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

[8] Proyecto corpus. corpus textual especializado plurilingüe. https://www.iula.upf.edu/corpus/corpuses.htm. Fecha de acceso: Mayo 2017.