UNA EVALUACIÓN EXHAUSTIVA DE SISHITRA, UN PARADIGMA HÍBRIDO EN TRADUCCIÓN AUTOMÁTICA

Jesús González[1], *A. Giménez*[1], *Jorge González*[1], *A. L. Lagarda*[1], *J. R. Navarro*[2], *L. Eliodoro*[2], *V. Félix*[2], *P. Peris*[1], *F. Casacuberta*[1]

[1]Departamento de Sistemas Informáticos y Computación
Universitad Politécnica de Valencia
{jegonzalez, agimenez, jgonzalez, alagarda, pperis, fcn}@dsic.upv.es

[2]Instituto Tecnológico de Informática
Universitad Politécnica de Valencia
{jonacer, leliodoro, victorf}@iti.upv.es

RESUMEN

SisHiTra representa un marco de computación en el que se combinan técnicas deductivas e inductivas para el desarrollo de sistemas de traducción automática de dominio abierto entre idiomas parejos. El uso de máquinas de estados finitos para modelar todas las fuentes de conocimiento ha permitido el diseño de un software muy eficiente. El rendimiento de esta aproximación se demuestra sobre el par de lenguas español y catalán, consiguiendo un compromiso muy interesante entre precisión y eficiencia computacional.

1. INTRODUCCIÓN

El objetivo de la *Traducción Automática* (TA) es desarrollar aplicaciones que permitan la traducción de textos sin intervención humana. Sin embargo, las tecnologías actuales de traducción no son capaces de cubrir las demandas de una traducción de alta calidad entre cualquier par de idiomas.

Las aproximaciones basadas en el conocimiento, en las que el conocimiento lingüístico experto se formaliza de un modo computable han tenido cierto éxito al abordar una traducción sin ningún tipo de restricciones. Sin embargo, las técnicas basadas en corpus, en las que los modelos estadísticos se infieren automáticamente a partir de ejemplos de textos, han alcanzado resultados muy competitivos en tareas semánticamente restringidas.

Existen modos diferentes de representar el conocimiento lingüístico, y, en particular, los transductores de estados finitos [1, 2, 3] han demostrado ser muy eficientes tanto en la implementación de sistemas deductivos como inductivos. Asimismo, las técnicas basadas en modelos de estados finitos han permitido el desarrollo de herramientas útiles para el procesamiento del lenguaje natural[4, 5, 6, 7, 8].

SisHiTra representa un marco de computación en el que, a través de un conjunto de modelos de estados finitos, se combinan técnicas deductivas e inductivas para el desarrollo de sistemas de TA de dominio abierto entre idiomas parejos [9].

El principal objetivo de SisHiTra es el de proporcionar traducciones de alta calidad con propósitos diseminativos. Por supuesto, es el objetivo de cualquier sistema de TA; sin embargo, en nuestro caso, es un aspecto especialmente importante que se ha tenido en cuenta en el diseño de cada etapa. Por lo tanto, consideramos que una traducción óptima es aquella que no parece el resultado de una traducción, sino que parece, más bien, fruto de un proceso de generación directa en el idioma destino. Esto no suele ser un problema para un traductor profesional, aunque es crucial en un sistema automático. Por ejemplo, un ser humano resuelve ambigüedades semánticas con cierta facilidad, mientras que en TA suelen representar un serio problema. Como consecuencia, la evaluación de la calidad de SisHiTra se hace en términos de la distancia entre las hipótesis de traducción y sus respectivas referencias, cuya optimalidad lingüística ha sido determinada a priori por una serie de expertos.

Por otro lado, el uso de máquinas de estados finitos para modelar todas las fuentes de conocimiento ha permitido un diseño muy eficiente del sistema. Así, la evaluación de la productividad de SisHiTra se ha realizado a través del tiempo de respuesta que proporciona al usuario.

Por último, el rendimiento de este paradigma híbrido se demuestra sobre el par de lenguas español y catalán, consiguiendo un compromiso muy interesante entre precisión y eficiencia computacional.

Este trabajo ha sido parcialmente respaldado por la "Agència Valenciana de Ciència y Tecnologia" bajo la subvención GRUPOS03/031 y por el proyecto español ITEFTE (TIC2003-08681-C02-02).

2. MARCO ESTADÍSTICO

Sea $\mathcal{W} = \{w_1, w_2, \ldots, w_N\}$ el vocabulario de entrada y sea $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ el de salida. El problema de la TA puede formularse estadísticamente como sigue: dada una frase $\mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_L \in W^*$, se debe buscar la secuencia de elementos de salida $\hat{\mathbf{t}} \in C^*$ que maximice la probabilidad a posteriori[1]:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{s}) \; . \tag{1}$$

Empleando la regla de Bayes y dado que el proceso de maximización es independiente de la frase de entrada **s**, la Ecuación (1) se puede reescribir como:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \Pr(\mathbf{t}) \Pr(\mathbf{s}|\mathbf{t}) \; . \tag{2}$$

En esta ecuación, las probabilidades contextuales (o modelo de lenguaje), $\Pr(\mathbf{t})$, representan todas las secuencias posibles de elementos de la salida, mientras que las probabilidades de emisión (o modelo de transferencia), $\Pr(\mathbf{s}|\mathbf{t})$, establecen la relación entre los vocabularios de entrada y salida.

Esta ecuación puede resolverse adecuadamente bajo ciertas condiciones aceptables: en primer lugar, se asume una traducción monótona elemento a elemento, proporcionando hipótesis de la misma longitud que la frase de entrada. Este hecho implica la utilización de un modelo de fertilidad 1. Por lo tanto, se pueden incorporar ciertas asunciones propias de los modelos de Markov para simplificar el problema.

Por una parte, se asume que las probabilidades contextuales para un elemento determinado dependen exclusivamente de los *n* elementos inmediatamente anteriores. Por otra parte, podemos asumir que las probabilidades de emisión únicamente dependen del símbolo de salida correspondiente. Así, desde un punto de vista generativo, las hipótesis se producirían monótonamente de izquierda a derecha, obteniéndose exactamente un símbolo de salida por cada elemento de la entrada.

Para los modelos de Markov de primer orden (bigramas), el problema se reduce a resolver la siguiente ecuación:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}_1,\ldots,\mathbf{t}_L} \left(\prod_{1\ldots L} P(\mathbf{t}_i|\mathbf{t}_{i-1}) P(\mathbf{s}_i|\mathbf{t}_i) \right) \tag{3}$$

Los parámetros de esta ecuación se pueden representar mediante un *Modelo Oculto de Markov* (en inglés *Hidden Markov Model*) en el que los estados y los símbolos de salida están unívocamente asociados. Las probabilidades contextuales, $P(\mathbf{t}_i|\mathbf{t}_{i-1})$, son las probabilidades de transición entre estados y las probabilidades del modelo de transferencia, $P(\mathbf{s}_i|\mathbf{t}_i)$, se pueden ver como distribuciones de probabilidad entrada-salida. El problema de maximización que se describe en la Ecuación (3) se resuelve por medio del algoritmo de Viterbi[10].

[1]Por simplicidad, $\Pr(X = x)$ y $\Pr(X = x \mid Y = y)$ se denotan como $\Pr(x)$ y $\Pr(x \mid y)$

2.1. Desambiguación probabilística

El objetivo de la desambiguación sintáctica es el de decidir la categoría léxica a la que pertenece un elemento dentro de un contexto determinado. Esta categorización no sólo es útil para reducir la ambigüedad sobre la traducción, sino también para otros propósitos lingüísticos, como, por ejemplo, la detección de sintagmas, indispensable para llevar a cabo las concordancias pertinentes.

La desambiguación sintáctica se realiza a través de la aproximación estadística descrita en la Ecuación 3, en la que el vocabulario de salida se refiere al conjunto de categorías léxicas.

Además de la desambiguación sintáctica, la TA precisa desambiguar semánticamente los elementos de la entrada antes de convertirlos a la lengua destino. Los métodos de desambiguación semántica tratan de determinar el significado implícito de un elemento dentro de un contexto. De nuevo, SisHiTra emplea modelos estadísticos para esta tarea.

Del mismo modo que en el caso sintáctico, la desambiguación semántica se lleva a cabo mediante la aproximación estadística descrita en la Ecuación 3.

Los modelos estadísticos están ganando popularidad por diversos motivos, especialmente por su facilidad de generación frente a los métodos deductivos. Las técnicas probabilísticas aprenden automáticamente a partir de corpus, evitándose el proceso de producir conocimiento lingüístico, con el ahorro temporal y económico que ello conlleva. Sin embargo, la obtención de los corpus para el entrenamiento de los modelos tampoco es trivial.

Por una parte, los modelos empleados en la desambiguación sintáctica necesitan un corpus monolingüe y segmentado. A cada segmento se le asigna una categoría. Por otra parte, los modelos para la desambiguación semántica en SisHiTra requieren corpus paralelos, es decir, corpus en los que el texto se encuentre emparejado biunívocamente entre ambas lenguas. A tal fin, se recolectaron diversos corpus a partir de algunas publicaciones electrónicas bilingües (periódicos, textos oficiales, etc.), y se paralelizaron mediante diversos algoritmos de alineamiento.

3. ARQUITECTURA DEL SISTEMA

SisHiTra es un paradigma de computación para el desarrollo de sistemas de traducción automática de dominio abierto entre idiomas parejos. Una versión previa de la arquitectura de SisHiTra se puede encontrar en [11].

Se han utilizado una serie de metodologías innovadoras, basadas en modelos de estados finitos, para representar las diferentes fuentes de conocimiento. Se han empleado transductores estocásticos como estructuras de datos para los accesos al diccionario, así como en la comunicación entre módulos. En los procesos de desambiguación, se han aplicado los conocidos Modelos de Markov[12].

SisHiTra sigue una arquitectura modular basada en máquinas de estados finitos, por lo que proporciona un

entorno de trabajo homogéneo y eficiente. El proceso de traducción de un idioma fuente a un idioma destino se realiza en 4 fases, combinando elementos tanto de naturaleza lingüística como estadística (véase la Figura 1).

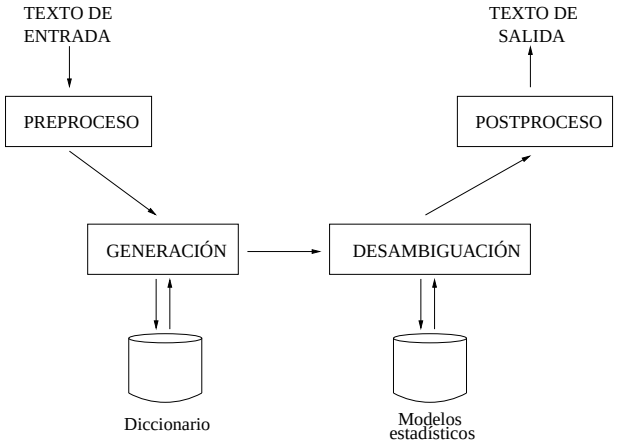

Figura 1. Arquitectura del sistema

La información parcial durante las etapas de traducción también se representa por medio de modelos de estados finitos. La arquitectura de SisHiTra se estructura del siguiente modo:

Preproceso: se encarga de dividir el texto de entrada en frases, y éstas, en elementos indivisibles (signos de puntuación, números, abreviaturas, nombres propios, palabras generales, etc.).

Generación: a través de una serie de accesos al diccionario, se produce un grafo sintáctico que representa todos los posibles análisis de la frase de entrada junto con todas sus posibles traducciones. En la Figura 2 se observa un diseño muy simple de la base de datos que constituye el diccionario, que para el par de lenguas que nos ocupa, dispone de un total de 80544 entradas.

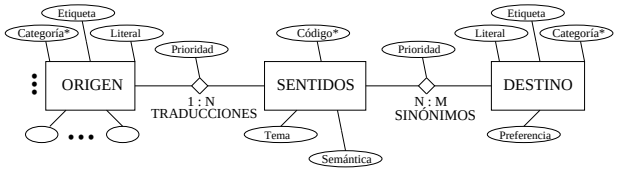

Figura 2. Diseño del diccionario

El uso de una aproximación basada en modelos de estados finitos resulta fundamental para una respuesta en tiempo real del sistema.

Desambiguación: este módulo realiza un proceso completo de desambiguación usando modelos estadísticos. En primer lugar, se desambigua morfosintácticamente a través de un proceso de etiquetado PoS, seleccionando la ruta sintáctica más probable dentro del grafo de análisis. Esto implica una segmentación de la frase de entrada, seleccionando una categoría léxica para cada segmento. A continuación, se realiza una desambiguación semántica, también por medio de métodos basados en el contexto.

Postproceso: aplica varias transformaciones basadas en reglas para transformar las frases de salida (no expresadas todavía en lenguaje natural) en frases correctas del idioma destino. En primer lugar, se realiza una concordancia de género y número a nivel de sintagma nominal. Seguidamente, se flexionan las palabras que admiten flexión (verbos, adjetivos o sustantivos), teniendo en cuenta algunos de sus rasgos lingüísticos, tales como género, número o persona. En tercer lugar, se realizan contracciones de palabras, elisiones de vocales, y se implementa correctamente el uso de pronombres. Por último, se realiza una conversión a texto plano en un lenguaje natural legible y sensible a mayúsculas.

Los módulos de SisHiTra se han desarrollado siguiendo dos esquemas diferentes: por un lado, los módulos basados en reglas (preproceso y postproceso) se han implementado en *Flex*, una herramienta muy útil para generar programas que realizan búsquedas de expresiones regulares en textos; por otro lado, los módulos de análisis (generación y desambiguación) se han escrito en *C*, siguiendo una estrategia de decodificación basada en el algoritmo de Viterbi.

4. MOTIVACIONES LINGÜÍSTICAS

Aquí nos limitaremos a mostrar un ejemplo de cómo se ha manejado la información lingüística para obtener una traducción de alta calidad. Para ilustrarlo, hemos seleccionado un aspecto lingüístico puntual que requiere una solución diferenciada en cada lengua. En concreto, nos referimos al comportamiento del complemento directo animado en español y catalán. A tal fin, reseñaremos la solución ofrecida por nuestro sistema.

Tanto en español como en catalán, el complemento directo se inserta directamente (de ahí el nombre) sin mediación de ningún tipo de partícula. Así, decimos:

Compré los tomates para el gazpacho.
Vaig comprar els tomàquets per al gaspatxo. (4)

y tanto en una lengua como en la otra, nuestros *tomates* aparecen ligados directamente al verbo.

Pero, cuando este complemento se refiere a un ente animado (por ejemplo, una persona), en español se introduce mediante la preposición *a*. Por contra, en catalán es indiferente que sea o no animado, ya que continúa conectándose al verbo (haciendo algunas salvedades) sin ningún tipo de elemento.

Ayer vimos a tu madre.
Ahir vam veure la teua mare. (5)

La solución que ofrece SisHiTra para esta frase es exactamente la que aparece en el ejemplo (5).

La elaboración de la regla lingüística que permite traducir correctamente estos casos es relativamente fácil. Tan sólo hay que etiquetar automáticamente aquellos verbos que son transitivos (es decir, que requieren un complemento directo) y de generar una regla que suprima la preposición *a* en catalán, cuando coincidan en la misma secuencia. Ahora bien, puede suceder que se den casos como:

Vimos a las tres de la tarde a tu madre.

Vam veure a les tres de la vesprada la teua mare. (6)

donde concurren un verbo transitivo y dos complementos introducidos por la preposición *a*. ¿Cómo puede decidir el traductor automático qué preposición debe suprimir? De hecho, la traducción resultante tiende a eliminar la preposición del primer complemento:

Vam veure les tres de la vesprada la teua mare. (7)

La verdadera dificultad reside, justamente, en identificar el complemento animado. También debemos contemplar que éste, en un uso vivo de la lengua, puede aparecer en un orden relativamente libre. Aún más, puede tratarse del complemento indirecto de un verbo transitivo (que a su vez posea un complemento directo introducido, por ejemplo, por la partícula que). En definitiva, se trata de considerar en abstracto todos los posibles contextos de concurrencia.

Si profundizamos un poco más en este fenómeno, observaremos que en él hay implícitas dos vertientes lingüísticas que tradicionalmente aparecen separadas. Por un lado, desde un punto de vista morfosintáctico, la presencia de un verbo transitivo que necesita completar su significación por medio de un complemento. Y por otro, semánticamente, que dicho complemento se refiera a una entidad animada. En cuanto al primero, la identificación automática de las categorías, su etiquetado, etc. o la inclusión de información de este tipo (y de muchos otros) en los diccionarios electrónicos, no suponen ninguna novedad. Sin embargo, en relación al segundo, la dificultad reside justamente en ser capaces de identificar la significación (animado vs no animado) del complemento.

Por nuestra parte, desde una perspectiva cognitivista, consideramos que ambos aspectos forman parte de un mismo todo y, por eso mismo, no pueden separarse. En ese sentido, los elementos que integran nuestro diccionario, se definen tanto por sus rasgos morfosintácticos como semánticos. Por ejemplo, el complemento directo que constituían nuestros *tomates* del ejemplo (4), poseen una etiqueta que señala su categoría gramatical (y, por tanto, su comportamiento sintáctico) y la subclase a la que pertenece dentro de ésta.

Así, decimos que se trata de un nombre[13] (N–), dado que pertenece a una clase léxica abierta, flexiva, etc., que prototípicamente es núcleo de un sintagma que, a su vez, puede ejercer las funciones de sujeto, de complemento, etc. Además decimos que es un nombre común (NC-), ya que presenta los rasgos de [+extensión], [+intensión] y [+referencialidad inherente]. En este caso, como en otros, hemos considerado que no nos es necesaria más información para nuestro propósito. Si ahora pasamos a la mencionada *madre*, comprobaremos que también aparece definida como un nombre común, pero ahora animado (NCA), porque para el par de lenguas con las que trabajamos este rasgo semántico sí que es pertinente para resolver dicha particularidad sintáctica.

Lo que tratamos de decir con todo ello es que la información semántica no sólo puede formalizarse (aunque sea muy someramente) sino que, además, puede resultar muy útil para tratar de resolver aspectos sintácticos. Es evidente que previamente hay que decidir y seleccionar qué rasgos semánticos pueden contribuir a una mejor resolución en la traducción entre un par de lenguas determinado. Creemos que esta digresión es necesaria para hacer ver que la voluntad de obtener traducciones automáticas de alta calidad precisa continuas reflexiones y revisiones sobre el trabajo realizado.

5. EVALUACIÓN

El rendimiento de toda aplicación de traducción comprende dos aspectos bien diferenciados y hasta cierto punto incompatibles: precisión y tiempo de respuesta. Ambos aspectos son importantes, ya que una traducción perfecta pero que requiera un tiempo excesivo de ejecución nos sirve de tan poco como una traducción instantánea pero completamente errónea. Debe llegarse a un equilibrio entre ambas aumentando la precisión lo máximo posible sin sacrificar la velocidad. Éste ha sido uno de los objetivos del nuevo diseño modular de SisHiTra y en los siguientes apartados veremos los resultados obtenidos.

Para evaluar el rendimiento de nuestra aproximación, disponemos de un corpus que consta de 240 pares de frases bilingües, extraídos de diversas fuentes, con una media de $18{,}3$ palabras por frase, contabilizando un total de 4387 palabras.

5.1. Otros sistemas

interNOSTRUM[14] es un sistema clásico de traducción indirecta por transferencia morfológica avanzada, semejante a la empleada en los sistemas comerciales para PC. Tiene una estructura modular, muy similar en muchos aspectos a la de SisHiTra, donde el análisis morfológico, el *PoS-tagging*, los diccionarios bilingües, y la información de formato cooperan todos juntos para proporcionar traducciones aproximadas del español al catalán en tiempo real. Se puede procesar texto en cualquiera de los siguientes formatos: ANSI, HTML, y RTF.

SALT[2] es un sistema de TA completamente basado en el conocimiento que posee un método interactivo que minimiza errores, añadiendo naturalidad a las traducciones

[2] http://www.cult.gva.es/salt/salt_programes_salt2.htm

entre español y catalán. Además, es un corrector ortográfico muy potente que detecta barbarismos, perífrasis o locuciones incorrectas, combinaciones incorrectas de pronombres, errores de concordancia, etc. Finalmente, es un programa de autoaprendizaje que enseña catalán a partir de los propios errores del usuario, con cientos de consejos sobre léxico, gramática, el uso apropiado de mayúsculas o minúsculas, signos de puntuación, etc.

AutomaticTrans[3] es un servidor de traducción fácilmente integrable en una red TCP/IP para procesar una gran cantidad de datos (miles de ficheros diarios) con excelente calidad. Es una plataforma de traducción que permite la selección explícita de un estilo linguístico, consiguiendo así una comunicación homogénea por la que es bien conocido. Con AutomaticTrans, que es más barato que las aproximaciones tradicionales a la traducción (pero no de peor calidad), los documentos se traducen simultáneamente a varios idiomas.

Comprendium[4] también es un sistema de TA comercial, desarrollado por la empresa Translendium SL, que se encuentra en Barcelona y forma parte del grupo europeo Braintribe, líder en Gestión de Contenidos de Empresa multilingües. La tecnología de Comprendium es líder mundial en TA tanto por el número de pares de idiomas así como por la funcionalidad ofrecida por el sistema. Éste consiste en un motor de traducción, con una estructura modular de gramáticas y léxicos, el cual realiza un análisis morfosintáctico del texto de entrada para luego traducirlo al idioma destino. El motor puede conectarse a una serie de módulos de memorias de traducción así como a un editor profesional de diccionarios. El usuario puede acceder al mismo a través de un servidor multiusuario, bien desde un cliente web, o bien desde una aplicación profesional monousuario. Se pueden crear distintas configuraciones del producto según las necesidades del usuario. Comprendium es el sucesor de INCYTA, sistema pionero en TA entre español y catalán.

5.2. Precisión

El WER es una medida objetiva de la calidad de la traducción que calcula la distancia de edición a nivel de palabra entre una hipótesis de traducción y una traducción de referencia predefinida. La distancia de edición calcula el número de sustituciones, inserciones y borrados que son necesarios para trasformar la hipótesis de traducción en la traducción de referencia. El número acumulado de errores para todas las frases de test se divide entre el total de palabras del mismo, y el porcentaje resultante nos muestra el número medio de palabras incorrectas. Los resultados de WER para SisHiTra son similares a los obtenidos por el resto de sistemas analizados, tal y como vemos en la Tabla 1.

Una desventaja del WER es que sólo compara la hipótesis de traducción con una única traducción de referencia.

[3]http://www.automatictrans.es

[4]http://www.translendium.com

Sistema	WER
interNOSTRUM	12.6
SALT 3.0	12.2
AutomaticTrans	**10.3**
Comprendium	11.5
SisHiTra	11.8

Tabla 1. WER para varios sistemas de traducción

Esto no ofrece ningún margen a traducciones correctas con un estilo de escritura diferente. Por lo tanto, para evitar este problema, evaluamos nuestro sistema mediante el MWER, que considera varias traducciones de referencia para una misma frase de test, calcula la distancia de edición de la hipótesis con todas ellas, y retorna el mínimo WER obtenido. El MWER ofrece una medida más realista que el WER ya que permite una mayor variabilidad en el estilo de traducción.

Utilizaremos un total de 3 referencias, incorporando una serie de traducciones generadas por lingüistas a partir de la referencia original.

Sistema	MWER
interNOSTRUM	6.5
SALT 3.0	6.1
AutomaticTrans	5.2
Comprendium	5.7
SisHiTra	**3.3**

Tabla 2. MWER para varios sistemas de traducción

Los resultados de MWER para SisHiTra son los mejores de entre los cinco sistemas analizados, tal y como vemos en la Tabla 2.

5.3. Tiempo de respuesta

El tiempo de respuesta es el tiempo transcurrido desde que el usuario lanza una petición hasta que el sistema produce el resultado correspondiente.

Para estimar el tiempo de respuesta de SisHiTra, hemos medido el tiempo empleado por cada módulo para procesar el corpus de evaluación. Así, sumando los tiempos empleados por módulo, obtenemos una estimación sobre el tiempo de respuesta de nuestro sistema, considerando despreciables los tiempos internos de comunicación entre módulos.

Los experimentos se realizaron en un computador con un procesador Pentium 4 a 3.2 GHz y 2 GB de memoria. Hemos repetido cada prueba 20 veces con el fin de que los resultados estén dotados de la máxima fiabilidad posible, calculando la media y la desviación típica de los tiempos obtenidos (véase la Tabla 3).

De este modo, dividiendo el total de palabras del corpus de evaluación, 4387, por el tiempo medio observado para su procesamiento, $0{,}8745$ segundos, podremos obtener una estimación de la velocidad de nuestro sis-

	Tiempo	
Módulo	Media	Desv. Típica
Preproceso	0.0180	0.0041
Generación	0.2605	0.0147
Desambiguación	0.1815	0.0067
Postproceso	0.4145	0.0051
Total	0.8745	0.0188

Tabla 3. Tiempos de respuesta en segundos

tema, $\frac{4387}{0{,}8745} \approx 5017$ palabras/segundo, considerado como tiempo real. Tan sólo interNOSTRUM reporta una velocidad comparable de 5000 palabras por segundo. El resto de sistemas presentan una productividad inferior.

6. CONCLUSIONES

En este artículo, hemos realizado una evaluación exhaustiva de SisHiTra, un paradigma de computación en el que se combinan técnicas deductivas e inductivas para el desarrollo de sistemas de traducción automática de dominio abierto entre idiomas parejos.

El rendimiento de esta aproximación se ha demostrado sobre el par de lenguas español y catalán, comparando los resultados obtenidos frente a otros sistemas de traducción automática. El uso de máquinas de estados finitos para modelar todas las fuentes de conocimiento ha permitido un diseño muy eficiente del sistema, logrando un compromiso muy interesante entre precisión y eficiencia computacional.

Como trabajo futuro, pretendemos desarrollar un prototipo inverso de traducción, es decir, catalán⇒español, a partir del diccionario actual. Por otro lado, y dentro del marco lingüístico, hemos iniciado manualmente el marcaje y la revisión de las entradas del diccionario. A la vez, estamos perfilando una serie de reglas lingüísticas considerando todos los contextos posibles, para ajustar los resultados. Además, los principales aspectos a mejorar son:

- Tiempo de respuesta, optimizando el módulo de postproceso cuya carga es excesiva porcentualmente.
- El desarrollo de una serie de aplicaciones de usuario que permitan la explotación masiva de las características de SisHiTra.
- Emplear SisHiTra sobre otras lenguas.

7. BIBLIOGRAFÍA

[1] L. Karttunen, "Citation of unpublished documents," Tech. Rep., XEROX Palo Alto Research Center, 1993.

[2] Emmanuel Roche y Yves Schabes, Eds., *Finite-State Language Processing*, Bradford Book. MIT Press, Cambridge, Massachusetts, USA, 1997.

[3] Emmanuel Roche, "Finite state transducers: parsing free and frozen sentences," pp. 108–120, 1999.

[4] M. Mehryar, "Finite-state transducers in language and speech processing," 1997.

[5] Mehryar Mohri, Fernando Pereira, y Michael Riley, "The design principles of a weighted finite-state transducer library," *Theoretical Computer Science*, vol. 231, no. 1, pp. 17–32, 2000.

[6] Kemal Oflazer, "Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction.," *Computational Linguistics*, vol. 22, no. 1, pp. 73–89, 1996.

[7] Emmanuel Roche y Yves Schabes, "Deterministic part-of-speech tagging with finite-state transducers," *Computational Linguistics*, vol. 21, no. 2, pp. 227–253, 1995.

[8] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, C. Martinez D. Llorens, S. Molau, F. Nevado, M. Pastor, D. Pico, y A. Sanchis., "Some approaches to statistical and finite-state speech-to-speech translation," *Computer Speech and Language*, vol. 18, pp. 25–47, 2004.

[9] J. González, A. L. Lagarda, J. R. Navarro, L. Eliodoro, A. Giménez, F. Casacuberta, J. M de Val, y F. Fabregat, "SisHiTra: a Spanish-to-Catalan hybrid machine translation system.," in *5th SALTMIL Workshop on Minority Languages*, Genoa, Italy, 23 May 2006, pp. 69–73.

[10] A. Viterbi, "Error bounds for convolutional codes and an asymtotically optimal decoding algorithm," *Annals of the New York Academy of Sciences*, vol. IT-13, pp. 260–269, 1967.

[11] José R. Navarro, Jorge González, David Picó, Francisco Casacuberta, Joan M. de Val, Ferran Fabregat, Ferran Pla, y Jesús Tomás, "Sishitra: a hybrid machine translation system from spanish to catalan.," in *EsTAL*, 2004, pp. 349–359.

[12] A. Sanchis, D. Picó, J.M. del Val, F. Fabregat, J. Tomás, F. Casacuberta, y E. Vidal, "A morphological analyser for machine translation based on finite-state transducers," in *MT Summit VIII*, September 2001, pp. 305–309.

[13] M.J. Cuenca, *Sintaxi fonamental: les categories gramaticals*, Empúries, 1996.

[14] M. Forcada, A. Garrido, R. Canals, A. Iturraspe, S. Montserrat-Buendia, A. Esteve, S. Ortiz Rojas, H. Pastor, y P.M. Pérez, "The spanish-catalan machine translation system internostrum," *0922-6567 - Machine Translation*, vol. VIII, pp. 73–76, 2001.