Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 101-108 recibido 03-04-2016 revisado 21-04-2016 aceptado 03-05-2016

Traducción Automática usando conocimiento semántico en un dominio restringido

Automatic translation using semantic knowledge in a restricted domain

Lluís-F. Hurtado, Iván Costa, Encarna Segarra, Fernando García-Granada, Emilio Sanchis
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Camino de Vera s/n
lhurtado@dsic.upv.es

Resumen: El propósito que sigue este trabajo es incorporar conocimiento semántico a la traducción automática con el objetivo de mejorar la calidad de ésta en dominios restringidos. Nos centraremos en la traducción entre inglés, francés y español en el contexto de consultas telefónicas a un servicio de información ferroviaria. Se han desarrollado varias estrategias para la incorporación de la semántica en el proceso de traducción. Algunas de estas aproximaciones incorporan la semántica directamente en los elementos al ser traducidos, mientras que otras utilizan una interlingua o lengua pivote que representa la semántica. Todas estas aproximaciones han sido comparadas experimentalmente con una traducción automática basada en segmentos léxicos que no incorpora conocimiento semántico.
Palabras clave: traducción automática, semántica, dominios restringidos

Abstract: The purpose of this work is to add semantic knowledge to a machine translation process in order to improve its quality for restricted domains. We will focus on the translation between English, French and Spanish, in a task of telephonic query to a railway information service. Many strategies have been developed for the incorporation of semantics into the translation process. Some of these approaches directly incorporate semantics into the elements to be translated and some others use an interlingua or pivot language that represents the semantics. All of these approaches have been experimentally compared to an automatic translation based on lexical segments that do not incorporate semantic knowledge.
Keywords: automatic translation, semantics, restricted domains

1 *Introducción*

En traducción automática, así como en la mayor parte de aplicaciones del procesamiento automático del lenguaje natural, existen dos tipos de tareas diferentes según su ámbito de aplicación. El primero, el ámbito de propósito general, en el que el sistema ha sido entrenado de manera genérica, de forma que se pueden procesar frases de cualquier contexto. Por ejemplo, Europarl (Koehn, 2005) es un corpus que recoge sesiones del parlamento europeo en las que se ha hablado de distintos temas. El segundo ámbito representa un dominio restringido, referente a un sistema que ha sido entrenado para procesar frases dentro de un contexto concreto. Este es el caso del corpus DIHANA (Benedí et al., 2006), que contiene diálogos referentes a consultas sobre información ferroviaria.

Entre las distintas aproximaciones a la traducción automática, en este trabajo nos centraremos en la traducción estadística, también referida como SMT (Statistical Machine Translation). El objetivo de esta aproximación es obtener buenos traductores a partir de estadísticas obtenidas de un corpus. Para entrenar un traductor de un idioma origen (o fuente) a otro destino mediante un sistema de traducción automática estadística es necesario disponer de un corpus paralelo. La calidad de un sistema de traducción automática estadística depende de las características del corpus paralelo de entrenamiento disponible. Hay dos factores importantes: el tamaño del corpus paralelo de

* Este trabajo ha sido financiado por el MINECO y fondos FEDER en el proyecto TIN2014-54288-C4-3-R

ISSN 1135-5948

entrenamiento y el dominio. Un conjunto de datos de entrenamiento pequeño conduce a modelos de traducción pobremente estimados y, en consecuencia, a una mala calidad en la traducción. Por esta razón, la calidad de la traducción, empeora cuando no tenemos suficientes datos de entrenamiento para el dominio específico objetivo.

Los métodos de adaptación al dominio se pueden dividir en dos amplias categorías. La adaptación al dominio puede hacerse en el corpus, por ejemplo, mediante la ponderación y la selección de los datos de entrenamiento (Gao et al., 2002). La adaptación también puede hacerse mediante la adaptación de los modelos de traducción. Algunos trabajos apuestan por la incorporación de conocimiento semántico para mejorar los resultados de la traducción, por ejemplo en (Banchs y Costa-jussà, 2011) se propone el uso de ciertas características semánticas para la SMT basada en el uso de Latent Semantic Indexing.

También se obtienen malos resultados cuando se trabaja con idiomas para los que no se dispone de recursos suficientes. Para estos casos se ha propuesto el uso de una lengua pivote o interlingua como paso intermedio entre la traducción entre dos idiomas. En (Habash y Hu, 2009) se propone el uso del inglés como lengua pivote entre árabe y chino, en (Babych, Hartley, y Sharoff, 2007) se propone el uso de una lengua pivote cercana a la lengua fuente y se reporta una comparación con la traducción directa.

El objetivo de este trabajo es incorporar conocimiento semántico al proceso de traducción automática para mejorar su calidad en dominios restringidos. La incorporación de este conocimiento se lleva a cabo siguiendo dos estrategias: o bién añadimos etiquetas semánticas a las palabras o secuencias de palabras en el lenguaje fuente o destino, o bién utilizamos la representación semántica como lengua pivote para la traducción. En este trabajo se exploran diferentes etiquetados semánticos y se compararan los resultados con los correspondientes a la traducción automática sin conocimiento semántico usando el corpus DIHANA, que contiene frases procedentes de diálogos referentes a consultas sobre información ferroviaria.

Las diferentes estrategias empleadas para incorporar este conocimiento semántico son: uso de la representación semántica mediante etiquetas semánticas asociadas a cada palabra y el uso de una lengua pivote basada en la representación semántica.

Para estimar el traductor automático estadístico se ha usado el conjunto de software MOSES (Koehn y et al., 2007). En este trabajo se han seleccionado IRSTLM (Federico, Bertoldi, y Cettolo, 2008) para la generación de modelos de lenguaje, y GIZA++ (Och y Ney, 2003) para la generación de alineamientos entre frases en distintos idiomas como herramientas externas. Como resultado de la fase de entrenamiento se han estimado diferentes modelos de traducción, dependiendo de la calidad de los alineamientos entre frases que se obtuvieron mediante el software alineador. Posteriormente estos modelos se han mejorado haciendo uso de un conjunto de Desarrollo.

Para la evaluación de los resultados utilizaremos la métrica BLEU. El BLEU se calcula dependiendo de la precisión de n-gramas indicada, habitualmente se emplea una precisión de 4, y a tal métrica se la llama BLEU-4 (Koehn, 2010). Es importante saber que BLEU no tiene en cuenta la relevancia de las palabras que son sinónimas, por lo que sólo cuenta un acierto cuando las palabras que compara son iguales.

2 Descripción del corpus

El corpus utilizado en este trabajo ha sido el corpus DIHANA. Se compone de 900 diálogos en español en el ámbito de llamadas telefónicas realizadas a un servicio de información de trenes. Todos estos diálogos fueron adquiridos mediante la técnica del Mago de Oz. Para disponer de un corpus paralelo se han traducido esos mismos diálogos a francés y a inglés. El corpus se ha dividido en tres conjuntos: Entrenamiento, Desarrollo (para ajuste de parámetros) y Prueba. El conjunto de Entrenamiento se ha traducido a francés e inglés mediante traductores web de propósito general sin supervisión. Se optó por emplear cuatro traductores diferentes para realizar cada una de las traducciones, de manera que se mitigaran los errores de traducción que podía haber producido cada uno de ellos individualmente. Estos traductores web son: Google Translate, Bing, Lucy y OpenTrad. En el caso de las traducciones de las frases que conforman los corpus de Desarrollo y Prueba para inglés y para francés estas traducciones fueron hechas por hablantes nativos de ambas lenguas.

El corpus en español está etiquetado desde el punto de vista semántico. Todas las frases del corpus están segmentadas y cada segmento tienen asociado una etiqueta semántica asignada de forma manual, además toda la frase tiene asociada una representación semántica en términos de Frames (Tabla 1).

Frase
hola quiero saber el horario de ida de Palencia a Oviedo el viernes dieciocho de junio
Conceptos Semánticos
hola:cortesia quiero saber:consulta el horario de:<hora> ida:tipo_viaje de Palencia:ciudad_origen a Oviedo:ciudad_destino el viernes dieciocho de junio:fecha
Frame
TIPO-VIAJE:ida (HORA) CIUDAD-ORIGEN:Palencia CIUDAD-DESTINO:Oviedo FECHA:[viernes-18-06-??,viernes-18-06-??]

Tabla 1: Ejemplo frase corpus DIHANA

El conjunto original en castellano contiene 6226 frases, un total de 47186 palabras y un vocabulario compuesto de 719 palabras diferentes. A partir del corpus original se han definido los conjuntos de Entrenamiento, Desarrollo y Prueba. Para los corpus en inglés y francés se han utilizado las cuatro traducciones diferentes proporcionadas por los respectivos traductores web. Esto hace que el corpus original con las frases en castellano tenga 4887 frases para entrenamiento mientras que los corpus en Francés e Inglés contienen 19548. Para equiparar el número de frases se han replicado las frases en castellano, de manera que el corpus de Entrenamiento en castellano contiene también 19548 frases.

Las Tablas 2, 3 y 4 describen los conjuntos de Entrenamiento, Desarrollo y Prueba.

	ES	EN	FR
frases	19548	19548	19548
palabras	147720	152368	142642
vocabulario	638	1277	1927

Tabla 2: Conjuntos de Entrenamiento

Puede verse que el vocabulario tanto para inglés como para francés es mucho mayor que el original en español. Esto se debe a que las traducciones proporcionadas por los traductores web introducen nuevas palabras que además son diferentes entre los diferentes traductores.

	ES para			
	Inglés	Francés	EN	FR
frases	340	277	340	277
palabras	2619	2013	2744	2175
vocabulario	263	235	287	253

Tabla 3: Conjuntos de Desarrollo

El número de frases de los conjuntos de Desarrollo para francés e inglés son diferentes porque una vez adquirido el corpus multilingüe se hizo un control de calidad que descartó algunas frases, mayoritariamente en francés. Se debe tener en cuenta que para el ajuste de parámetros se necesita que el conjunto de frases para el lenguaje origen y para el lenguaje destino contengan el mismo número de frases, ya que uno se traducirá y se comparará con el otro, que será la referencia para ajustar los pesos de cada modelo de traducción. En consecuencia, en español hay dos conjuntos de Desarrollo distintos, uno para cuando el otro idioma es el inglés y otro para cuando es el francés.

	ES	EN	FR
frases	1000	1000	1000
palabras	7637	7701	7650
vocabulario	412	604	586

Tabla 4: Conjuntos de Prueba

A destacar que tanto los conjuntos de Desarrollo como los conjuntos de Prueba son muy pequeños, pero que sin embargo han sido traducidos por humanos.

3 Representación de la semántica y su uso en el modelo de traducción

Inicialmente disponemos de unas etiquetas semánticas y un Frame asociados a cada una de las frases del corpus en castellano. Para incluir esta información en las frases del conjunto de Entrenamiento hemos seguido diferentes aproximaciones. Recordemos que sólo el corpus en español está segmentado y etiquetado semánticamente.

3.1 Etiquetas semánticas asociadas a cada palabra

Esta primera aproximación consiste en añadir a cada palabra en cada frase del corpus de Entrenamiento en castellano la etiqueta

semántica que tiene asociado el segmento al que pertenece la palabra. Al usar esta representación en nuestros modelos conseguimos que las mismas palabras asociadas a distintos conceptos se cuenten como distintas y las probabilidades de aparición de cada palabra, tanto sola como en bigramas o trigramas, sea diferente. Aumentamos el vocabulario de la tarea. Obtenemos los nuevos conjuntos para Entrenamiento, Desarrollo y Prueba, ahora con información semántica.

La Tabla 5 muestra un ejemplo obtenido con este tipo de representación.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Representación
quiero#consulta ir#consulta a#ciudad_destino Segovia#ciudad_destino desde#ciudad_origen Valencia#ciudad_origen

Tabla 5: Representación mediante etiquetas semánticas asociadas a cada palabra.

3.2 Semántica en destino y en origen

Dado que los etiquetados semánticos sólo están disponibles en castellano, es posible usar éstos de dos formas distintas: donde el castellano es la lengua origen (semántica en origen), y cuando es la lengua destino (semántica en destino). En el caso de la semántica en destino, para entrenar el sistema de traducción usamos la aproximación del apartado anterior para modelar la lengua destino. Así obtenemos un modelo de lenguaje y traducción que traducirá desde un idioma fuente (con sólo palabras) a castellano con información semántica.

En el caso de la semántica en origen, el idioma fuente será el castellano enriquecido mediante información semántica, y se traducirá al francés o al inglés. Como resultado obtendremos una traducción al inglés o al francés en la que solamente habrá palabras.

3.3 Uso de interlingua basada en pares atributo-valor

La aproximación mediante interlingua consiste en definir un lenguaje pivote en el proceso de traducción. Una representación con interlingua implica entrenar sistemas que traduzcan desde un lenguaje origen a interlingua, más otros sistemas que traduzcan desde interlingua a un lenguaje destino.

Esta aproximación trata de traducir una frase en un idioma fuente por otra en un idioma destino que tenga un contenido semántico similar, en lugar de traducir palabra por palabra. De esta manera, se intenta conservar el significado de la frase original más que su literalidad. Esto brinda un abanico más grande de posibilidades, ya que tendremos más opciones a la hora de realizar una traducción. Sin embargo, también cabe destacar que la métrica BLEU no será tan representativa con esta aproximación. Una frase en el lenguaje destino, con significado equivalente pero distintas palabras en la frase de referencia obtendría con BLEU una puntuación baja, pese a poder ser una traducción válida.

En esta aproximación hemos empleado el Frame contenido en las frases del corpus DIHANA como lenguaje pivote. Cada frase de entrenamiento se traducirá en la secuencia de pares atributo-valor del frame correspondiente. Un ejemplo de este etiquetado se muestra en la Tabla 6. Nótese que cada par atributo-valor se considera un símbolo en el lenguaje pivote.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
TIPO-VIAJE-nil () CIUDAD-ORIGEN-Valencia CIUDAD-DESTINO-Segovia

Tabla 6: Representación mediante una interlingua basada en pares atributo-valor

Con un corpus de Entrenamiento para la interlingua que contenga este tipo de frases podremos entrenar sistemas de traducción que traduzcan de castellano, inglés o francés a interlingua, y después podremos entrenar otros sistemas de traducción que traduzcan de interlingua a castellano, inglés o francés.

3.4 Uso de interlingua basada en secuencias atributo-valor

Esta representación está basada en la anterior, pero considerando ahora el atributo y su valor como dos símbolos distintos. Un ejemplo de este etiquetado aparece en la Tabla 7.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
TIPO-VIAJE nil () CIUDAD-ORIGEN Valencia CIUDAD-DESTINO Segovia

Tabla 7: Representación mediante una interlingua basada en secuencias atributo-valor.

3.5 Uso de interlingua basada en secuencias de conceptos y valores

La tercera representación con interlingua está basada en todas las representaciones anteriores. En esta aproximación representaremos una frase mediante la secuencia de etiquetas semánticas de la segmentación de esa frase en el corpus DIHANA. Además, se tienen en cuenta las apariciones de valores en la representación de Frame, pero sustituyendo los valores concretos por valores genéricos para aumentar la cobertura del modelo. Esta representación persigue conseguir mejores resultados equiparando los tokens del lenguaje origen con una secuencia de atributos que normalmente es mayor a la secuencia de pares atributo-valor que hay en el Frame. En la Tabla 8 vemos un ejemplo de este tipo de representación.

4 Experimentos y Resultados

Se ha llevado a cabo una evaluación experimental para comprobar la idoneidad de las representaciones semánticas propuestas en el proceso de traducción automática estadística. Como punto de partida, experimento baseline, realizaremos experimentos en los cuales no se ha utilizado ninguna información semántica.

Frase
quiero ir a Segovia desde Valencia
Conceptos semánticos
quiero ir:consulta a Segovia:ciudad_destino desde Valencia:ciudad_origen
Frame
TIPO-VIAJE:nil () CIUDAD-ORIGEN:Valencia CIUDAD-DESTINO:Segovia
Representación
consulta ciudad_destino#*ciudad*1 ciudad_origen#*ciudad*2

Tabla 8: Representación mediante una interlingua basada en secuencias de conceptos y valores.

4.1 Experimentos con etiquetas asociadas a cada palabra

Se han hechos experimentos tanto para cuando la representación semántica (el etiquetado semántico sólo está disponible para castellano en el corpus DIHANA) está en el destino como cuando está en el origen.

En el experimento con etiquetas asociadas a cada palabra en destino se han llevado a cabo a su vez otros dos tipos de experimentos. En el primer tipo hemos empleado en la fase de desarrollo un conjunto de datos sin etiquetas asociadas de manera que se maximiza el BLEU para traducir de palabras en un idioma origen a palabras en castellano. Sin embargo, para el segundo tipo de experimentos usamos un conjunto de datos con etiquetas asociadas para la fase de desarrollo, de manera que tenemos un sistema que maximiza el BLEU para traducir de palabras en un lenguaje origen a castellano con etiquetas asociadas.

En la Tabla 9 están los resultados para el baseline y para los experimentos con etiquetas asociadas a cada palabra en destino, haciendo uso en un caso de un conjunto de Desarrollo con etiquetas semánticas asociadas y otro conjunto de Desarrollo sin etiquetas semánticas asociadas. En el caso de traducir desde el inglés obtenemos mejores resultados ajustando los pesos mediante el uso de un conjunto de datos sin etiquetas asociadas y conseguimos mejorar al baseline. Sin embargo, al traducir desde el francés obtenemos mejores resultados usando un conjunto de datos para desarrollo con etiquetas asociadas y se mejora también el baseline.

	EN-ES	FR-ES
Baseline	39.06	37.91
Desarrollo=Etiquetas Sem	39.01	**39.38**
Desarrollo=Palabras	**39.49**	39.03

Tabla 9: BLEU de la representación semántica con etiquetas asociadas a cada palabra.

A continuación se muestra un ejemplo de traducción EN-ES y FR-ES:

(1) Hello, good morning. I'd like times for trains to Cuenca → hola buenos días quisiera horarios para trenes a Cuenca

(2) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → hola quisiera saber horarios de trenes para ir a Barcelona

En la Tabla 10 se han vuelto a realizar los dos experimentos para semántica en destino, pero esta vez sin eliminar las etiquetas asociadas a las palabras en la salida de traducción y empleando un conjunto de Prueba con etiquetas asociadas para que sirva de referencia. De este modo estamos midiendo no solo el acierto del traductor a nivel de palabras sino además el acierto considerando los conceptos semánticos asociados a ellas. Ésta es una tarea más compleja que la anterior, y justifica el hecho de que se obtenga un BLEU menor, pues el rango de palabras que tiene para acertar al traducir al castellano sin etiquetas asociadas es de 719 palabras, mientras que al castellano con etiquetas asociadas es de 1746 palabras.

	EN-ES	FR-ES
Desarrollo= Etiquetas Sem	33.44	**34.03**
Desarrollo=Palabras	**33.99**	33.61

Tabla 10: BLEU de la representación semántica con etiquetas asociadas a cada palabra, evaluando las etiquetas semánticas.

A continuación se muestra un ejemplo de traducción EN-ES y FR-ES:

(3) Hello, good morning. I'd like times for trains to Cuenca → hola#cortesia buenos#cortesia días#cortesia quisiera#consulta horarios# <hora> para# <hora> trenes# <hora> a#ciudad_destino Cuenca#ciudad_destino

(4) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → hola#cortesia quisiera#consulta saber#consulta horarios# <hora> de# <hora> trenes# <hora> para# <hora> ir# <hora> a#ciudad_destino Barcelona#ciudad_destino

Por último en la Tabla 11 se muestran los experimentos usando etiquetas asociadas a cada palabra en origen. En este caso el lenguaje origen usa siempre las etiquetas asociadas a cada palabra del castellano, que se traduce a francés o inglés.

Los resultados del baseline obtienen menos BLEU que cuando el castellano es la lengua destino. La razón puede estar en que las traducciones de inglés y francés han sido obtenidas por cuatro traductores web diferentes, generando un vocabulario mayor y que hace más difícil que se den aciertos.

	ES - EN	ES - FR
Baseline	**22.01**	26.91
Etiquetas Sem	21.72	**27.01**

Tabla 11: BLEU de la semántica en origen

A continuación se muestra un ejemplo de traducción ES-FR:

(5) sí# <afirmacion> me#consulta podrías#consulta decir#consulta el# <precio> precio# <precio> del#numero_relativo_orden primero#numero_relativo_orden del#numero_relativo_orden último#numero_relativo_orden → oui vous me dire le prix du premier du dernier

4.2 Experimentos usando una interlingua basada en pares atributo-valor

La Tabla 12 muestra los resultados obtenidos usando una interlingua basada en pares atributo-valor. Se hace uso de los pares atributo-valor del *Frame* de cada frase.

Se ha hecho un experimento para cada combinación de traducciones posible, cabe destacar los experimentos hechos para traducir de un idioma a interlingua y después otra vez a ese mismo idioma. En este caso lo que se hace es traducir por palabras o segmentos que expresen los mismos atributos semánticos. Dicho de otro modo, lo que hacemos al traducir al interlingua es "comprender" el significado de la frase, y al traducir desde un interlingua traducimos a una frase

con ese significado. En la traducción usando una interlingua, aunque generemos una frase con el mismo significado y las palabras a la salida del proceso de traducción sean sinónimos de las palabras en la referencia, la métrica BLEU las contará como un error.

	BLEU
ES-interlingua-ES	18.77
EN-interlingua-ES	8.06
FR-interlingua-ES	8.55
ES-interlingua-EN	6.34
EN-interlingua-EN	8.54
FR-interlingua-EN	4.60
ES-interlingua-FR	10.45
EN-interlingua-FR	7.86
FR-interlingua-FR	11.78

Tabla 12: BLEU de una representación mediante Interlingua de pares atributo-valor.

A continuación se muestra un ejemplo de traducción ES-IL e IL-ES mediante una interlingua basada en pares atributo-valor:

(6) hola buenos días mira quería saber horarios de trenes para ir de Castellón a Barcelona → TIPO-VIAJE-ida TIPO-VIAJE-nil HORA CIUDAD-ORIGEN-Castellón CIUDAD-DESTINO-Barcelona → me gustaría saber el horario de un viaje de ida para ir de Castellón a Barcelona

En el ejemplo se observa una traducción correcta que sin embargo proporciona un BLEU bajo.

4.3 Experimentos usando una interlingua basada en secuencias atributo-valor

Con esta representación se espera aumentar la cobertura del modelo dado que no hará falta que atributo y valor aparezcan juntos para poder ser traducidos. En la Tabla 13 se muestran los resultados obtenidos para esta aproximación realizando un experimento por cada par posible de idiomas. Como se esperaba, si lo comparamos con la tabla 12 los resultados en BLEU han aumentado. El vocabulario en esta aproximación es menor (370 tokens) que en la representación de interlingua anterior, debido a que no hay combinaciones concepto-valor sino que tenemos los conceptos por un lado y los valores por el otro, dando un número menor de tokens.

A continuación se muestra un ejemplo de traducción FR-IL e IL-EN mediante una interlingua basada en secuencias atributo-valor:

	BLEU
ES-interlingua-ES	19.96
EN-interlingua-ES	10.10
FR-interlingua-ES	11.75
ES-interlingua-EN	10.10
EN-interlingua-EN	11.33
FR-interlingua-EN	8.58
ES-interlingua-FR	10.87
EN-interlingua-FR	7.98
FR-interlingua-FR	10.25

Tabla 13: BLEU de la representación mediante Interlingua de secuencias atributo-valor.

(7) bonjour je voudrais connaître les horaires des trains pour aller à Barcelona → TIPO-VIAJE nil HORA CIUDAD-DESTINO barcelona → i'd like to know the schedules of trains to barcelona

4.4 Experimentos usando una interlingua basada en secuencias de conceptos y valores

Para esta última representación de interlingua, y al igual que para las anteriores se ha llevado a cabo un experimento por cada combinación de idiomas (ver Tabla 14). Los resultados en BLEU son generalmente mayores que las obtenidas por las representaciones de interlingua anteriores. Esto se debe a que al hacer uso de los conceptos en lugar de sólo los atributos del *Frame* podemos enriquecer la representación semántica, tal es el caso de la etiqueta *cortesía*. Además, las otras representaciones interlingua que hemos usado en este trabajo sólo usan el *Frame*, y este tiene los pares atributo-valor ordenados canónicamente, con lo que la tarea de reordenamiento en el proceso de traducción era también más difícil.

A continuación se muestra un ejemplo de traducción EN-IL e IL-ES mediante una interlingua basada en secuencias conceptos y valores:

(8) Well, I want leave on the thirtieth of July → afirmacion consulta m_salida fecha#30X07 tipo_viaje#vuelta july → sí me gustaría salir el próximo día treinta de vuelta july

	BLEU
ES-interlingua-ES	32.67
EN-interlingua-ES	17.44
FR-interlingua-ES	17.86
ES-interlingua-EN	8.62
EN-interlingua-EN	9.18
FR-interlingua-EN	7.60
ES-interlingua-FR	12.71
EN-interlingua-FR	8.32
FR-interlingua-FR	11.87

Tabla 14: BLEU de la representación mediante una interlingua de secuencias de conceptos y valores.

5 Conclusiones

En este trabajo hemos presentado diferentes alternativas para añadir conocimiento semántico a los sistemas de traducción automática en el ámbito de una tarea de dominio restringido. Estas alternativas se pueden agrupar en dos clases. En una de ellas se usan los conceptos semánticos que tiene el corpus DIHANA para enriquecer el proceso de traducción. En la otra se usa una lengua pivote como paso intermedio en el proceso de traducción entre idiomas.

Con la primera de las aproximaciones hemos obtenido resultados superiores a los del baseline, es decir, a la traducción cuando no se usa información semántica. Se puede inferir de todo esto que si trabajamos con un corpus de dominio cerrado, tal como DIHANA, y podemos asignar etiquetas semánticas a las palabras entonces es posible mejorar los resultados de la traducción con estas representaciones siempre y cuando el corpus sea lo suficientemente grande.

En cuanto al uso de una lengua pivote, si bien la puntuación en BLEU es menor que en la otra aproximación, la propia métrica BLEU tampoco refleja bien su calidad. Al traducir de una lengua origen a una interlingua se está haciendo una interpretación del significado de las palabras de la lengua origen, y al traducir de este interlingua a la lengua destino se está generando una frase que expresa el significado deseado. Es frecuente pues que en la traducción aparezcan palabras diferentes pero que sin embargo tienen el mismo signficado.

Los resultados obtenidos indican que es alentador continuar en esta dirección. Esto implicaría, entre otras medidas, desarrollar una nueva métrica para cuantificar la calidad de una traducción, de forma que se tengan en cuenta las similitudes semánticas además de las léxicas.

Bibliografía

Babych, B., A. Hartley, y S. Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. En *Proceedings of the MT Summit XI*, páginas 412–418.

Banchs, R. E. y M. R. Costa-jussà. 2011. A semantic feature for statistical machine translation. En *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, páginas 126–134. ACL.

Benedí, J.-M., E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, y A. Miguel. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. En *Proceedings of LREC 2006*, páginas 1636–1639.

Federico, M., N. Bertoldi, y M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. En *INTERSPEECH*, páginas 1618–1621.

Gao, J., J. Goodman, M. Li, y K.-F. Lee. 2002. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing*, 1(1):3–33.

Habash, N. y J. Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language. En *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, páginas 173–181. ACL.

Koehn, P. y et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. En *Proc. of ACL demonstration session*, páginas 177–180.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Procs. of Machine Translation Summit X*, páginas 79–86.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edición.

Och, F. J. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.