Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 335-342 recibido 26-04-2005; aceptado 01-06-2005

Clasificación y generalización de formas verbales en sistemas de traducción estocástica

Adrià de Gispert **José B. Mariño** **Josep M. Crego**

Centre de Recerca TALP
Universitat Politècnica de Catalunya (UPC)
Campus Nord UPC. 08034-Barcelona
{agispert,canton,jmcrego}@gps.tsc.upc.es

Resumen: En esta comunicación se propone un método para incorporar conocimiento lingüístico relativo a las formas verbales en sistemas estocásticos de traducción. Por medio de una clasificación basada en conocimiento de dichas formas, y de su sustitución por el lema del verbo principal durante la fase de entrenamiento, se consigue un mejor alineado en palabras, cuya consecuencia es una mejor estimación del modelo de traducción. Además, a partir de las formas verbales observadas en el entrenamiento es posible generalizar con éxito y proporcionar traducciones a nuevas formas no vistas anteriormente. El método propuesto es evaluado en una tarea de traducción del inglés al español de dominio restringido, donde se alcanza una mejora significativa.
Palabras clave: traducción estocástica, conocimiento lingüístico, formas verbales, morfología

Abstract: This paper introduces a method to incorporate linguistic knowledge regarding verb forms into an stochastic machine translation model. By means of a rule-based classification of these forms, and by substituting them by the base form of the head verb during the training stage, we achieve a better statistical word alignment, which leads to a better estimate of the translation model. Furthermore, a successful generalization strategy can be devised to produce a new translation for unseen verb forms from the translations of seen verb forms. An evaluation of this method in an English to Spanish limited-domain translation task is presented, producing a significant performance improvement.
Keywords: stochastic machine translation, linguistic knowledge, verb forms, morphology

1. Introducción

Actualmente, la investigación en el campo de los sistemas estocásticos de traducción goza de una creciente popularidad entre la comunidad científica. Los buenos resultados alcanzados por este planteamiento en múltiples evaluaciones en tareas de dominio limitado e ilimitado justifican este gran interés.

Sin embargo, la mayoría de sistemas estocásticos del estado del arte parten del nivel superficial de la palabra como única fuente de conocimiento, ignorando así cualquier información morfológica, sintáctica o, en general, lingüísticamente más informada. Si bien esto no representa una limitación importante para lenguas con poca flexión de formas (y por lo tanto con una talla del vocabulario reducida) como el inglés, sí supone una limitación importante al trabajar con lenguas altamente flexivas como el español. Los errores mostrados en el siguiente ejemplo son fruto de esta falta de información morfológica, que impide al sistema relacionar las diversas formas en que se expresa un nombre, adjetivo o, sobre todo, un verbo.

i was told that the service in this hotel is very good $\Longrightarrow$ **yo estaba dicho** `que el` **servicios** `en este hotel está muy bien`

En esta comunicación se presenta una solución híbrida que incorpora cierto conocimiento lingüístico dentro del enfoque estocástico a la traducción, y en concreto, relativo a las formas verbales. Para ello, se presenta una clasificación basada en reglas de todas las formas verbales, que permite conside-

ISSN: 1135-5948

rar separadamente, a efectos de aprendizaje del sistema de traducción, pronombres, verbos auxiliares y sufijos derivados de la flexión verbal por un lado, y lema del verbo principal por otro. De esta forma se mejora el modelo de traducción al concentrar las distintas formas de un mismo verbo en una única unidad de traducción (sección 3).

Por otro lado, el uso de esta clasificación permite el diseño de estrategias de generalización a formas verbales no vistas en el material de entrenamiento a partir de las formas vistas (sección 4).

Para realizar experimentos se ha trabajado con el par de lenguas inglés – español, y se presentan resultados obtenidos tanto en el alineado en palabras del entrenamiento, como en una tarea de traducción de dominio limitado del inglés al español (sección 5). Por último, en la sección 6 se presentan conclusiones, junto a ideas para investigaciones futuras.

2. Trabajos previos

En la línea de esta comunicación podemos encontrar algunos trabajos recientes. En (Ueffing y Ney, 2003) también se muestra un posible enfoque para el tratamiento de las formas verbales en el caso inglés – español. Sin embargo, los autores optan por unir los pronombres personales ingleses a la forma del verbo con el fin de generar un vocabulario inglés más amplio que pueda corresponderse con el español. Por el contrario, nuestra propuesta va en la dirección opuesta al reducir la talla del vocabulario e incrementar así la frecuencia de aparición de las unidades de traducción.

También para el caso del español (y del serbio), otra posibilidad radica en descomponer las formas flexivas en morfema y afijos, considerando cada uno de ellos como palabras independientes en el modelo de traducción, como se presenta en (Popovic y Ney, 2004). Sin embargo, los autores no proporcionan resultados de traducción al español (sólo del español al inglés), ya que ello les obligaría a incorporar una estrategia de generación de la forma flexiva a partir de morfema y afijos.

Por último, cabe mencionar los trabajos de (Lee, 2004) o (Nießen y Ney, 2004) relacionados con la introducción de transformaciones morfológicas en el material de entrenamiento de sistemas estocásticos de traducción, en especial para el caso del árabe – inglés y del alemán – inglés, respectivamente.

3. Planteamiento de la traducción estocástica

Para realizar la traducción de la oración f de una lengua fuente en la oración d de una lengua destino, a partir del modelado de máxima entropía (Och y Ney, 2002) se suele usar una combinación log-lineal de funciones de características que pueden gobernar la traducción, como se expresa en la siguiente ecuación:

$$\hat{d}_1^I = \arg max_{d_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(d_1^I, f_1^J) \right\} \quad (1)$$

Las funciones de características elementales, derivadas del modelo del canal ruidoso introducido en (Brown et al., 1993), son:

- un modelo de traducción $Pr(d|f)$ basado en cadenas de palabras
- un modelo del lenguaje destino $Pr(d)$

aunque típicamente se combinan con modelos de distorsión del orden de las palabras, modelos de probabilidad léxica, penalizaciones a las traducciones cortas para compensar la preferencia del modelo de lenguaje por las traducciones cortas, etc.

Sin embargo, este planteamiento no considera clases de unidades de traducción, y por lo tanto, trata todas las formas verbales de un verbo, o todos las formas singular y plural de un sustantivo, como unidades completamente distintas sin ninguna relación. A continuación se propone un modelo que intenta abordar esta problemática por medio de una clasificación basada en conocimiento lingüístico. En concreto, se clasifican, para cada idioma, las formas verbales (incluyendo pronombre personal, verbo principal y auxilares) al lema del verbo principal. Como se comenta en la sección 5.2, esta detección se realiza de forma determinista mediante autómatas que implementan simples reglas basadas en información de las palabras, su etiqueta morfológica y su lema.

3.1. Modelo de traducción con clases

Si definimos $\tilde{f}_j$ como un cadena de palabras consecutivas de la frase fuente y $\tilde{d}_i$ como una cadena de la frase destino, cuyas clases

a las que pertenecen son $\tilde{F}_j$ y $\tilde{D}_i$ respectivamente, y que $T = (\tilde{F}_j, \tilde{D}_i)$ es el par de clases fuente y destino utilizado como unidad de traducción (que llamamos tupla), podemos expresar el modelo de traducción como:

$$Pr(\tilde{d}_i|\tilde{f}_j) = \sum_T Pr(\tilde{d}_i, T|\tilde{f}_j) =$$

$$= \sum_T Pr(\tilde{d}_i|T, \tilde{f}_j) Pr(\tilde{D}_i, \tilde{F}_j|\tilde{f}_j) =$$

$$= \sum_T Pr(\tilde{d}_i|T, \tilde{f}_j) Pr(\tilde{D}_i|\tilde{F}_j, \tilde{f}_j) Pr(\tilde{F}_j|\tilde{f}_j) \quad (2)$$

Como se verá en la sección 5, nuestra implementación actual considera una clasificación de grupos verbales de carácter determinista, es decir, sin admitir ambigüedad. Esto implica que no hay dependencia entre $\tilde{F}_j$ y $\tilde{f}_j$, puesto que una implica la otra, con lo cual $Pr(\tilde{F}_j|\tilde{f}_j) = 1$, y la probabilidad se puede simplificar en:

$$Pr(\tilde{d}_i|\tilde{f}_j) = Pr(\tilde{D}_i|\tilde{F}_j) Pr(\tilde{d}_i|T, \tilde{f}_j) \quad (3)$$

En la ecuación 3 intervienen un modelo estándar de traducción basado en cadenas de palabras, pero entrenado sobre un material clasificado; y un modelo de instancia que distribuye la probabilidad entre las distintas formas verbales destino dada la forma verbal fuente y la tupla de traducción de clases utilizada.

De esta forma, mientras la tupla T se centra en la probabilidad de traducción del lema del verbo, el modelo de instancia se concentra en elegir la forma destino (pronombres, verbos y partículas auxiliares, tiempo verbal, etc.) de ese lema dada la forma fuente. De todas formas, la decisión sobre la traducción se toma, en última instancia, en la combinación log-lineal de características.

3.2. Modelo de instancia

Con el fin de estimar el modelo de instancia $Pr(\tilde{d}_i|T, \tilde{f}_j)$ se propone un enfoque directo basado en la frecuencia relativa de cada instancia, de entre todas las tuplas que contienen dicha instancia en su parte fuente, independientemente de su traducción. Matemáticamente:

$$Pr(\tilde{d}_i|T, \tilde{f}_j) = \frac{N(T, \tilde{d}_i, \tilde{f}_j)}{N(T, \tilde{f}_j)} \quad (4)$$

Esta probabilidad, sin embargo, no está definida para todas aquellas formas verbales que, a pesar de clasificarse a un lema conocido (visto en el entrenamiento), no han ocurrido en el material de entrenamiento. Para esos casos, se ha desarrollado una estrategia de generalización descrita en la siguiente sección.

4. Generalización de formas verbales no vistas

Para generar una instancia (o grupo verbal) $\tilde{d}_i$ dada la tupla T y la instancia fuente $\tilde{f}_j$, es posible utilizar la información de las formas verbales vistas. En concreto, se propone realizar una búsqueda entre las instancias vistas en dicha tupla T de formas verbales idénticas a la que se desea traducir, excepto en los rasgos referentes a la persona (sean pronombres personales o sufijos del verbo). En el caso de encontrarse alguna, se genera una nueva instancia destino con la misma forma verbal, excepto en la persona, que se sustituye por la persona de la instancia fuente $\tilde{f}_j$.

A título de ejemplo, supongamos que se desea traducir la frase '*we would have payed* it' del inglés al español, y que en el entrenamiento aparecen las tuplas T_1=(V[pay],V[pagar]), T_2=T(V[pay],V[hacer] el pago) and T_3=T(V[pay] it, lo V[pagar]) que traducen la clase V[pay] presente en la frase a traducir. Sin embargo, nunca se ha observado la forma verbal 'we would have payed' entre las instancias de dichas tuplas. En ese caso, para cada tupla se procede a examinar todas sus formas vistas en busca de instancias idénticas (en palabras, etiquetas morfológicas y lemas) a la que se desea traducir salvo en la información de persona, como se muestra en el cuadro 1, donde no se ha encontrado ninguna instancia útil para la tupla T_2.

Para cada una de estas instancias, se genera una nueva forma verbal en español, por medio de la sustitución de la información relativa a la persona de la forma vista (*habría pagado*, 1a o 3a del singular) por la persona de la forma a traducir (*we*, 1a del plural). Además, cada nueva traducción recibe un peso de acuerdo con el número de apariciones de la forma vista en el entrenamiento (última columna del cuadro 1). Este peso actúa de probabilidad de instancia para estas nuevas formas. Así pues, en el ejemplo se generarían

T_1 = (V[pay] , V[pagar])		
I would have payed	habría pagado	3
you would have payed	habrías pagado	1
you would have payed	pagarías	1
T_2 = (V[pay] , V[hacer]	el pago)	
* would have payed	—	0
T_3 = (V[pay] it , lo V[pagar])		
I would have payed it	lo habría pagado	1

Cuadro 1: Instancias vistas en las tuplas que traducen V[pay] que son útiles para generalizar la forma 'we would have payed'.

las siguientes nuevas formas, con probabilidad:

V_{ins}= we would have payed			
T_1	V_{ins}	habríamos pagado	4/6
T_1	V_{ins}	pagaríamos	1/6
T_3	V_{ins} it	lo habríamos pagado	1/6

En el caso de existir ambigüedad sobre la persona que debe generarse (por ejemplo, al traducir 'you', que puede ser 2a persona de singular o plural en español), se generan todas las posibles formas, puesto que la decisión final sobre cuál es más adecuada en el contexto de la frase se tomará en la combinación log-lineal de características que rige la traducción. De hecho, se espera que el modelo de lenguaje destino contribuya favorablemente a discernir qué casos deben ser descartados dada la composición del material de entrenamiento.

5. *Resultados obtenidos*

En esta sección se evalúa la aproximación presentada con el corpus paralelo inglés – español desarrollado en el marco del proyecto LC-STAR. Primero se describe el corpus, posteriormente se muestran estadísticas relacionas con la detección de formas verbales en dicho corpus, y por último se presentan los resultados alcanzados en alineado automático de palabras y en una tarea de traducción del inglés al español.

5.1. Corpus y preprocesado

El corpus desarrollado en el marco del proyecto LC-STAR consiste en transcripciones de diálogos de habla espontánea en la tarea de la información turística, la planificación de viajes y la concertación de citas. Así pues, por su espontaneidad es un corpus rico en variedad de formas y expresiones, y a menudo contiene frases sin una correcta estructura sintáctica. Este corpus ha sido tratado de la siguiente forma:

- Normalización de contracciones para el inglés (por ejemplo, wouldn't = would not, we've = we have)
- Etiquetado morfológico del inglés por medio de la herramienta de libre distribución *TnT* tagger (Brants, 2000), y extracción de lemas por medio de *wnmorph*, aplicación perteneciente al paquete WordNet (Miller et al., 1991).
- Y etiquetado morfológico del español por medio de la herramienta de análisis *FreeLing* (Carreras et al., 2004), que también proporciona el lema de cada palabra analizada.

Se han separado 350 frases como conjunto de desarrollo (para optimización de parámetros), y 500 frases como conjunto de test. Para ambos conjuntos se dispone de tres traducciones alternativas a efectos de evaluación. En el cuadro 2 se proporcionan las principales estadísticas de los corpus de entrenamiento, desarrollo y test: número de oraciones (**orcn**), número total de palabras (**plbr**), talla de los correspondientes vocabularios (**vcblr**) y longitud media en palabras de las oraciones (**media**).

Lng	**Orcn**	**Plbr**	**Vcblr**	**Media**
Entrenamiento				
in	29998	419113	5940	14.0
es		388788	9791	13.0
Desarrollo				
in	350	6645	841	19.0
Test				
in	500	7412	963	14.8

Cuadro 2: Estadísticas del corpus paralelo inglés - español utilizado.

El número de palabras inglesas desconocidas, es decir, no vistas en el entrenamiento, es de 20 para el conjunto de desarrollo (el 0.3 % del total de palabras) y de 48 para el conjunto de test (el 0.7 % del total de palabras).

5.2. Detección y clasificación de verbos

Para la detección de formas verbales se ha utilizado una estrategia basada en conocimiento como la descrita en (de Gis-

pert, 2005). Mediante autómatas deterministas que implementan reglas de detección que combinan información de las palabras, su etiqueta morfológica y su lema, se realiza una clasificación no ambigua para el inglés y el español separadamente.

Lng	**Verbos**	**desc**	**Lemas**	**desc**
Entrenamiento				
in	56419		768	
es	54460		911	
Desarrollo				
in	856	3 %	120	0 %
Test				
in	1076	5.2 %	146	4.7 %

Cuadro 3: Formas verbales detectadas en el corpus.

En el cuadro 3 se muestra el número de formas verbales detectadas mediante esta técnica (**verbos**), así como el número de lemas a los que se clasifican (**lemas**). Además, para los conjuntos de desarrollo y test, se presenta el porcentaje de formas y lemas desconocidas (**desc**), es decir, que no aparecen en el conjunto de entrenamiento. Es de destacar el hecho de que el número de formas verbales detectadas es diferente, aunque muy parecido, en cada idioma, debido básicamente a dos factores. Por un lado, el mero hecho de que ciertas formas verbales no se traducen con una forma verbal en otro idioma, y por otro, el hecho de que los etiquetadores morfológicos utilizados son estadísticos y, como tales, tienen cierto grado de error que puede llevar a estas diferencias.

5.3. Resultados de alineado de palabras

En este apartado se desea evaluar los efectos que tiene la clasificación basada en conocimiento sobre el aprendizaje del sistema estocástico de traducción. Para ello, se han alineado manualmente 350 frases seleccionadas aleatoriamente del conjunto de entrenamiento, introduciendo enlaces seguros y probables para calcular el Recall, la Precisión, y la Tasa de Error de Alineado, o *Alignment Error Rate* (AER). Estas medidas fueron introducidas para la evaluación de alineados en (Och y Ney, 2000) y han sido usadas ampliamente para esta tarea en evaluaciones previas (Mihalcea y Pedersen, 2003). Matemáticamente, se pueden expresar como:

$$recall = \frac{|A \cap S|}{|S|}, \;\; precision = \frac{|A \cap P|}{|A|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

donde A es el alineamiento hipotetizado, S es el conjunto de enlaces seguros existentes en la referencia manual y P es el conjunto de todos los enlaces de la referencia (seguros y probables).

Mediante la aplicación GIZA++ (Och, 2003) se ha realizado el alineamiento del corpus original de inglés a español y viceversa (ejecutándose 5 iteraciones de los modelos IBM1 y HMM, y 3 iteraciones de los modelos IBM3 e IBM4), y se han obtenido los alineamientos unión e intersección de ambas direcciones. Posteriormente, se ha repetido el proceso con el corpus clasificado (mediante la detección de formas verbales), y después del alineado se han reintroducido las formas verbales con todos los enlaces que obtuvo su clase.

Las formas verbales inglesas que contienen un adverbio entre el pronombre personal y el verbo (como 'I *only* want') han sido tratadas de forma especial: el adverbio ha sido aislado de la clase antes del alineado, y una vez realizado éste, se ha reintroducido a la forma, pero conservando sus enlaces separadamente a los de las palabras que conforman la forma verbal.

El cuadro 4 compara el resultado del alineado unión[1] para ambos casos, observándose una reducción importante del error de alineado en el caso con clasificación (**clas**), lo que claramente demuestra que la riqueza de formas verbales limita las posibilidades de aprendizaje del sistema estocástico de traducción.

	Recall	Precision	AER
básico	74.14	86.31	20.07
clas	76.45	89.06	**17.37**

Cuadro 4: Resultados en alineado de palabras.

[1]La unión del alineado en ambas direcciones proporciona sistemáticamente un menor AER que la intersección.

5.4. Resultados de traducción

Para evaluar los efectos del método de clasificación y generalización propuesto, se ha integrado en un sistema estocástico de traducción cuya combinación log-lineal tiene en cuenta las siguientes características (Crego, Mariño, y de Gispert, 2005):

- un modelo de traducción basado en un N-grama de tuplas como el presentado en (de Gispert y Mariño, 2004)
- un modelo de lenguaje destino basado en N-gramas utilizando la aplicación SRILM (Stolcke, 2002)
- una penalización para compensar la preferencia del modelo anterior por las traducciones cortas

El peso λ_m de cada función de características ha sido optimizado sobre el conjunto de desarrollo utilizando la medida de calidad BLEU (Papineni et al., 2002).

Se han realizado tres experimentos de traducción del inglés al español. En primer lugar, un experimento sin clasificación de verbos y utilizando únicamente las tres características mencionadas (**básico**). Posteriormente, un experimento con clasificación (y por lo tanto, añadiendo el modelo de instancia a la combinación log-lineal), pero sin generalización de las formas desconocidas (**clas**), que se han dejado sin traducción. El último experimento añade esta estrategia al caso anterior (**clas+gen**).

En el cuadro 5 se muestran los resultados obtenidos tanto en el conjunto de desarrollo como en el de test, y para dos medidas ampliamente utilizadas, como son el porcentaje de error en palabras (mWER) y el BLEU.

	desarrollo		test	
	mWER	BLEU	mWER	BLEU
básico	21.32	0.698	23.16	0.671
clas	19.37	0.728	22.22	0.686
clas+gen	19.27	0.727	21.65	0.692

Cuadro 5: Comparativa de resultados de traducción de inglés a español

Como se desprende de los resultados, la clasificación propuesta produce efectivamente una mejora significativa de la calidad de la traducción, incluso en el caso en que no se generalizan las formas no vistas, debido a la menor perplejidad del modelo de traducción obtenido. Por otra parte, el uso de la generalización proporciona una mejora adicional al abordar la traducción de aquellas formas verbales que difícilmente el modelo estocástico original podría tratar.

Resulta interesante estudiar el comportamiento distinto para el caso del conjunto de desarrollo y el conjunto de test. Si bien en el desarrollo la clasificación consigue una reducción de mWER y un aumento de BLEU más importante que en el test, la generalización tiene un efecto opuesto. Esto se justifica por dos razones: por un lado, el mayor porcentaje de formas no vistas en el conjunto de test y, por otro, el hecho de que el 4.7 % de los lemas de los verbos del test son desconocidos (ver cuadro 3), de manera que jamás se podrá traducir dicho lema sin el uso de algún diccionario u otra información extra. Este efecto no está presente en el caso del conjunto de desarrollo.

En la figura 1 se pueden observar dos ejemplos de frases traducidas por el sistema básico, y por las aproximaciones de clasificación con y sin generalización.

6. Discusión y conclusiones

En esta comunicación se ha presentado un método de clasificación basada en conocimiento de las formas verbales para mejorar las prestaciones de un sistema estocástico de traducción. Por medio de la sustitución de las formas por el lema del verbo principal durante la fase de entrenamiento, se consigue un mejor alineado a nivel de palabras, lo que repercute en una mejor estimación del modelo de traducción. Por otro lado, el enfoque de generalización a formas no vistas propuesto proporciona una vía de enriquecimiento lingüístico de los modelos de traducción estrictamente estocásticos.

Sin embargo, a pesar de la mejora obtenida en mWER y BLEU, del estudio de las frases traducidas se desprende que todavía hay un camino a recorrer hacia la gramaticalidad y naturalidad en la traducción estocástica. Aparte del posible tratamiento de los lemas no vistos con la incorporación de un diccionario bilingüe, es necesario un estudio más profundo del modelo de instancia propuesto; el modelo actual reduce en cierta medida la información contextual del modelo de traducción, priorizando la información limitada de la forma verbal fuente para generar la forma destino al contexto de la frase en el que

fuente:	I WAS TOLD that the service IS very good
básico:	yo estaba dicho que el servicio está muy bien
clas:	me habían dicho que el servicio está muy bien
fuente:	In two days' time , if YOU HAVE NOT CALLED me I WILL CANCEL the reservation
básico:	pasado mañana fuera tiempo , si no hemos llamado anular la reserva
clas:	en dos días tiempo , si UNSEEN UNSEEN la reserva
clas+gen:	en dos días tiempo , si no ha llamado la anularé la reserva

Figura 1: Ejemplos de frases traducidas, donde los verbos ingleses detectados se indican en mayúsculas.

se traduce dicha forma. Si bien el modelo de lenguaje debería compensar esta limitación, en algunos casos no es así.

Por otro lado, otra limitación actual del sistema que requiere un estudio más profundo está relacionada con los pronombres clíticos españoles, que al no estar considerados por la herramienta de etiquetado morfológico, no son asociados con su correspondiente en inglés (que se expresa como palabra independiente del verbo) y se ignoran sistemáticamente en la traducción.

6.1. Trabajo futuro

Como trabajo futuro, se van a realizar experimentos de clasificación y generalización de formas verbales con un corpus paralelo de grandes dimensiones, y para distintos tamaños del conjunto de entrenamiento. En concreto, se trabajará con el corpus español – inglés de los debates del Parlamento Europeo, que alcanza los 30 millones de palabras por idioma. A pesar de estas dimensiones, el estudio informal de las traducciones arrojadas por un sistema estocástico de traducción del estado del arte revela que los problemas relativos a formas verbales siguen afectando negativamente las prestaciones.

Más a medio plazo, se prevén hacer otras clasificaciones con motivación lingüística, como la de los sintagmas nominales al lema del sustantivo, independientemente de determinantes e incluso preposiciones que lo precedan. El objetivo es considerar las expresiones 'del hotel', 'algún hotel', o 'nuestros hoteles' como instancias de una misma clase 'N[hotel]', de forma equivalente a la estrategia presentada para formas verbales, mejorando el entrenamiento y abordando los problemas de concordancias entre sustantivos, determinantes y adjetivos presentes en las traducciones actuales.

7. Agradecimientos

Este trabajo ha sido financiado parcialmente por la CICYT a través del proyecto TIC2002-04447-C02 (ALIADO), la Unión Europea mediante el proyecto FP6-506738 (TC-STAR), y el "Departament de Universitats, Recerca i Societat de la Informació" de la Generalitat de Catalunya.

Bibliografía

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Brown, P., S. Della Pietra, V. Della Pietra, y R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, May.

Crego, J.M., J. Mariño, y A. de Gispert. 2005. An Ngram-based statistical machine translation decoder. *Submitted to Interspeech 2005*, April.

de Gispert, A. 2005. Phrase linguistic classification for improving statistical machine translation. *Accepted for Publication at the ACL 2005 Students Workshop*, June.

de Gispert, A. y J. Mariño. 2004. Talp: Xgram-based spoken language translation system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, páginas 85–90, October.

Lee, Y.S. 2004. Morphological analysis for statistical machine translation. En Daniel Marcu Susan Dumais y Salim Rou-

kos, editores, *HLT-NAACL 2004: Short Papers*, páginas 57–60, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Mihalcea, Rada y Ted Pedersen. 2003. An evaluation exercise for word alignment. En Rada Mihalcea y Ted Pedersen, editores, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, páginas 1–10, Edmonton, Alberta, Canada, May. Association for Computational Linguistics.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, K. Miller, y R. Tengi. 1991. Five papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4):235–312.

Nießen, S. y H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, June.

Och, F.J. 2003. Giza++ software. http://www-i6.informatik.rwth-aachen.de/~och/ software/giza++.html.

Och, F.J. y H. Ney. 2000. Improved statistical alignment models. *38th Annual Meeting of the Association for Computational Linguistics*, páginas 440–447, October.

Och, F.J. y H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, páginas 295–302, July.

Papineni, K.A., S. Roukos, R.T. Ward, y W-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics*, páginas 311–318, July.

Popovic, M. y H. Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. *4th Int. Conf. on Language Resources and Evaluation, LREC'04*, páginas 1585–1588, May.

Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

Ueffing, N. y H. Ney. 2003. Using pos information for smt into morphologically rich languages. *10th Conf. of the European Chapter of the Association for Computational Linguistics*, páginas 347–354, April.