Procesamiento del lenguaje Natural, nº 41 (2008), pp. 259-266 recibido 7-05-2008; aceptado 16-06-2008

Comparación y combinación de los sistemas de traducción automática basados en n-gramas y en sintaxis

Comparison and system combination of n-gram-based and syntax-based machine translation systems

Maxim Khalilov y **José A. R. Fonollosa**
Centre de Recerca TALP
Universitat Politècnica de Catalunya
Campus Nord, C. Jordi Girona, 1-3
Barcelona, Spain
(khalilov,adrian)@gps.tsc.upc.edu

Resumen: En este artículo se comparan dos sistemas basados en dos aproximaciones diferentes de traducción automática: El denominado sistema de la Traducción Automática Aumentado con Sintaxis (SAMT / TAAS), basado en una sintaxis subyacente al modelo basado en frases, y el sistema de traducción automática estadística (TAE) basado en n-gramas en el cual el proceso de traducción está basado en el modelado estocástico del contexto bilingüe. Se realiza una comparación de la arquitectura de los dos sistemas paso a paso y se comparan también los resultados en base a las medidas automáticas de evaluación de la calidad de traducción y los recursos computacionales para una pequeña tarea árabe-inglés que pertenece al dominio de noticias. Finalmente, se combinan las salidas de ambos sistemas para obtener una mejora significativa de la calidad de la traducción.
Palabras clave: Traducción automática estocástica, traducción basada en sintaxis, n-gramas, combinación de sistemas

Abstract: In this paper we shall compare two approaches to machine translation: the Syntax Augmented Machine Translation system (SAMT), which is a syntax-driven translation system, underlain by phrase-based model, and the *n*-gram-based Statistical Machine Translation (SMT), in which a translation process is based on statistical modeling of the bilingual context. We provide a step-by-step comparison of the systems, reporting results in terms of automatic evaluation metrics and required computational resources for a smaller Arabic-to-English translation task from the news domain. Finally, we combine the output of both systems that yield to significant improvement of translation quality.
Keywords: Statistical machine translation, syntax-based translation, n-grams, system combination

1. Introducción

La inclusión de información sintáctica en los sistemas de traducción automática estocásticos (sistemas híbridos sintáctico-estocásticos) es un tema actual de investigación en Traducción Automática (TA). Los denominados modelos clásicos de IBM basados en palabras que aparecieron a principios de la década de 1990 fueron mejorados incluyendo la posibilidad de trabajar a nivel de frases (entendidas como secuencias de palabras) tal como se describe en (Koehn, Och, y Marcu, 2003) o en la implementación más reciente: MOSES MT (*http://www.statmt.org/moses/*).

En paralelo a la aproximación basada en frases[1], ha aparecido la aproximación basada en *n*-gramas (Mariño et al., 2006), derivada de la traducción basada en *Transductores de Estados Finitos* (Casacuberta, Vidal, y Vilar, 2002). Los sistemas basados en *n*-gramas trabajan con unidades bilingües, denominadas tuplas, compuestas por una o más palabras del lenguaje fuente y por una o más palabras del lenguaje destino.

En contraste o complementando a los sis-

[1]En este artículo la palabra "frase"se utiliza como la traducción directa de la palabra inglesa "phrase"

ISSN: 1135-5948

temas tradicionales de TAE, han ganado fuerza los sistemas basados en sintaxis y los modelos basados en la jerarquía de frases. Una muestra representativa de los sistemas de traducción basados en sintaxis incluye los que están basados en gramática bilingüe sincrónica (Melamed, 2004), en los modelos árboles de derivación-a-cadena (parse tree-to-string) y en los mapeos árbol-a-árbol no isomorfos (Charniak, 2003).

Basándose en las probabilidades relativas de las frases, Chiang (2005) introdujo un modelo jerárquico de frases, que puede considerarse como una generalización coherente del modelo clásico basado en frases (este modelo permite crear múltiples generalizaciones dentro de cada frase). El sistema TAAS (SAMT en inglés) (Zollmann y Venugopal, 2006) es una implementación del sistema de TA, que ofrece una mayor generalización de esa aproximación, en que las categorías sintácticas, extraídas de la parte destino del árbol de derivación, se asignan a las frases estructuradas jerárquicamente.

En este artículo se comparan las diferencias y las similitudes de la traducción estadística basada en n-gramas de las unidades de traducción y el sistema TAAS, que opera con las categorías no terminales y una Gramática Sincrónica Libre de Contexto (GSincLC). La comparación se ha realizado en una pequeña tarea de traducción árabe-inglés del dominio de noticias (*News*); el corpus de entrenamiento incluye aproximadamente 1,5M tokens.

El resto de la presentación está organizado de la siguiente manera: En la Sección 2 el sistema TAAS de la CMU-UKA[2] se describe brevemente, en la siguiente Sección se describe el sistema de TAE basado en n-gramas. En la Sección 4 se presentan la metodología, la descripción de los experimentos y los resultados alcanzados, y finalmente en la Sección 5 se discuten los resultados y se presentan las conclusiones.

2. Sistema TAAS

Uno de los mayores comentarios críticos de los modelos basados en frases es la escasez de datos. Este problema es incluso más serio cuando los lenguajes fuente y destino, o ambos son de mucha inflexión y ricos en morfología. Además, los modelos basados en frases tienen dificultades para considerar reordenamientos de larga distancia, porque el modelo de distorsión se basa únicamente en la distancia del movimiento y los recursos computacionales crecen rápidamente con la distancia considerada (Och y Ney, 2004).

Un intento satisfactorio de abordar este problema fue la introducción y discusión del sistema de TA basado en las frases generalizadas y estructuradas jerárquicamente tal como se describe en Chiang (2005). Este sistema opera con sólo dos etiquetas (una categoría de frases sustancial y una etiqueta de unión[3]) y un trabajo reciente (Zollmann y Venugopal, 2006) presenta una mejora importante si las categorías sintácticas completas o parciales (obtenidas de los árboles de derivación del lenguaje destino) están asignadas a las frases.

2.1. Modelado

Un formalismo para la Traducción Aumentada con Sintaxis es la Gramática Probabilística Sincrónica Libre de Contexto (GPSinLC), la cual se define en términos de los conjuntos de terminales de los lenguajes fuente y destino y de un conjunto de no terminales:

$$X \longrightarrow \langle \gamma, \alpha, \sim, \omega \rangle$$

donde X es un elemento no terminal, γ es una secuencia de elementos terminales relativos a la parte fuente y no terminales, α es una secuencia de elementos terminales relativos a la parte destino y no terminales, $\sim$ es el mapeo uno-a-uno del espacio de categorías no terminales en γ al espacio de no terminales en α, y ω es el peso no negativo asignado a la regla.

El conjunto de no terminales se compone de las categorías sintácticas que corresponden al conjunto P*enn Treebank* de la parte destino, un conjunto de las reglas de unión y de una etiqueta especial que representa la categoría por defecto, denominada según las reglas "del estilo Chiang", que no se corresponde con ninguna otra categoría del árbol de derivación. Consecuentemente, todas las reglas puramente lingüísticas están incluidas en la tabla del mapeo de frases.

[2]Carnegie Melon University - University of Karlsruhe

[3]"Glue rule"

2.2. Anotación, generalización y poda de las reglas

La tabla puramente léxica que sostiene el sistema TAAS está identificada como se describe en Koehn et al. (2003) y está basada en el alineado de palabras, generado según el método *grow-diag-final* (Och y Ney, 2003).

La parte destino del corpus de entrenamiento ha sido procesada con el parser de Charniak (Charniak, 2000), y cada frase se ha anotado con el constituyente que cubre la parte destino de las reglas. El conjunto de no terminales se ha extendido por las categorías condicionales y adicionales de acuerdo con la Gramática Combinatoria Categorial (Steedman, 1999). Las reglas se construyen, por ejemplo, como *RB+VB*, representando un constituyente de unión de dos categorías adyacentes, i.e. un adverbio y un verbo, o como *DT\NP*, que indica un grupo nominal incompleto, a el que le falta el determinante inicial.

El procedimiento recursivo de generalización de las reglas coincide con el que se propuso en Chiang (2005), pero violando las restricciones introducidas para una gramática que contenía sólo una categoría (por ejemplo, las reglas que contienen elementos generalizados adyacentes).

Por lo tanto, cada regla existente

$$N \longrightarrow f_1 \dots f_m / e_1 \dots e_n$$

puede ser extendida por la regla existente

$$M \longrightarrow f_i \dots f_u / e_j \dots e_v$$

donde $1 \leq i < u \leq m$ y $1 \leq j < v \leq n$ para obtener una regla nueva

$$N \longrightarrow f_1 \dots f_{i-1} M_k f_{u+1} \dots f_m / \\ e_1 \dots e_{j-1} M_k e_{v+1} \dots e_n$$

donde k es un índice de un no-terminal M que indica la correspondencia uno-a-uno entre los M tokens nuevos en los dos lados.

La figura 1 muestra un ejemplo de extracción de las reglas iniciales. Estas reglas son extendidas posteriormente gracias a la estructura jerárquica del modelo (figura 2).

La poda de reglas es necesaria debido a que el tamaño del conjunto de las reglas generalizadas puede ser enorme y se realiza en base a las frecuencias relativas y la naturaleza de las reglas: las reglas no léxicas que han ocurrido solamente una vez se descartan directamente, las reglas condicionadas por la fuente con una frecuencia de aparición menor a un umbral también son eliminadas, mientras que las reglas que no contienen no-terminales nunca pueden ser podadas.

2.3. Decodificación y las funciones características

El proceso de decodificación se lleva a cabo utilizando un modelo loglineal "top-down"que decodifica una oración fuente enriquecida con GPSinLC de modo que la calidad de traducción sea representada por un conjunto de las funciones para cada regla, i.e.:

- *Las probabilidades condicionales*, dado las categorías fuente, las destino o las categorías por la izquierda;
- *Las funciones de pesos léxicos*, como se ha presentado en Koehn et al. (2003);
- *Los contadores* del número de palabras en la parte destino y del número de aplicaciones de las reglas;
- *Las características binarias* que reflejan el contexto de la regla (si es puramente léxica o puramente abstracta, entre otras);
- *Las penalizaciones* por rareza y desequilibrio de la regla.

El proceso de decodificación se puede representar como una búsqueda (operación $argmax$) en el espacio de probabilidad de los terminales del lenguaje destino, que es similar al parsing monolingüe con una gramática libre de contexto. Los pesos de las funciones características se optimizan en base a la maximización de la medida BLEU (Zollmann y Venugopal, 2006).

3. El sistema de n-gramas

Una descripción detallada del sistema basado en n-gramas se encuentra en Mariño et al. (2006). El problema de la TAE se formula en términos de los lenguajes fuente (f) y destino (e) y se define de acuerdo con la ecuación (1). Se puede reformular como la selección de la traducción con la probabilidad más alta del conjunto de las oraciones destino (2):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} = \quad (1)$$

$$= \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} \quad (2)$$

donde I y J representan el número de palabras en los idiomas fuente y destino respectivamente.

Los sistemas más recientes operan con las unidades bilingües extraídas del corpus paralelo a partir del alineado de palabras. Los logaritmos de las probabilidades asociadas a las funciones características son combinados linealmente (aproximación *loglineal*) para definir una función cuya maximización establece la traducción (Och y Ney, 2002) tal como se muestra en la formula (3):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (3)$$

donde h_m se refiere a las funciones características y λ_m a los pesos que corresponden a cada modelo.

3.1. El sistema de traducción

La aproximación basada en n-gramas se considera como una alternativa a la traducción basada en frases bilingües, donde la secuencia de palabras del idioma fuente es segmentada en frases monolingües que son traducidas individualmente para formar la oración destino (Koehn, Och, y Marcu, 2003).

La traducción basada en n-gramas considera la traducción como un proceso estocástico que maximiza la probabilidad conjunta $p(f, e)$, en base a una descomposición en n-gramas bilingües. La parte principal del sistema así construido es un modelo de traducción (un modelo de lenguaje (ML)), basado en las unidades bilingües denominadas tuplas. Las tuplas se extraen del alineado de palabras de acuerdo con unas condiciones que definen una segmentación única (Mariño et al., 2006).

Mientras que la TAE de frases considera el contexto solamente para reordenar las frases pero no para la traducción, los sistemas basados en n-gramas condicionan las decisiones de traducción en las decisiones previas de traducción.

3.2. Las características adicionales

Al igual que los sistemas basados en frases, los sistemas basado en tuplas más recientes implementan una combinación lineal de los logaritmos de la probabilidad asignada a la traducción por el modelo de traducción y otras características adicionales:

- *el ML* de n-gramas *de palabras* del lenguaje destino;
- *el ML* de n-gramas *de tags* del lenguaje destino (un modelo de n-gramas de las etiquetas gramaticales o Part-Of-Speech tags (POS));
- *el modelo de penalización* para las traducciones más cortas, que compensa la tendencia a la generación de traducciones con un menor número de palabras;
- *los modelos léxicos* en cada dirección (de fuente a destino y viceversa) como se describe en Och y Ney (2004).

3.3. El reordenamiento de palabras extendido

La aproximación basada en tuplas se trata de partida como una traducción monótona ya que el modelo está basado en el orden secuencial de las tuplas durante el entrenamiento, aunque es necesario introducir estrategias de reordenamiento para obtener buenos resultados en algunas tareas de traducción. El modelo de distorsión extendido ha sido implementado tal como se presenta en Crego y Mariño (2006). Basándose en el alineado de palabras, las tuplas se extrajeron siguiendo la técnica denominada *unfolding*, mediante el cual las tuplas se dividen en tuplas más pequeñas y estos trozos se secuencian en el orden de las palabras destino. La estrategia de reordenamiento está apoyada por un ML de 4-gramas de los tags POS del texto fuente reordenado. Un ejemplo de extracción de tuplas, en contraste con la construcción de reglas basada en chunks como se hace en el sistema TAAS, se muestra en la figura 1.

3.4. Decodificación y optimización

La herramienta de decodificación MARIE, distribuida gratuitamente, se ha utilizado como motor de búsqueda del sistema de traducción. Los detalles de su funcionamiento están descritos en Crego et al. (2005). El decodificador implementa un algoritmo de búsqueda en haz (beam-search) con poda basada en histograma. Dado el corpus de desarrollo y las traducciones de referencia, los pesos de la combinación loglineal se ajustan mediante el método de optimización denominado *simplex* (con el objetivo de maximizar la medida BLEU) y un re-ranking de la lista n-best tal como se describe en *http://www.statmt.org/jhuws/*.

4. *Experimentos*

4.1. Evaluación de los sistemas

Los resultados experimentales se obtuvieron utilizando únicamente las primeras 50K líneas del corpus del dominio de noticias (*News*) ofrecido en la evaluación de sistemas de traducción NIST'08. Los datos estadísticos del corpus pueden verse en la tabla 1. Los conjuntos de desarrollo y de test tenían 4 traducciones de referencia y contenían 663 y 500 oraciones respectivamente.

Las medidas de la evaluación automática se obtuvieron ignorando las diferencias entre mayúsculas y minúsculas (case-insensitive). Se consideraron las métricas clásicas BLEU y NIST, junto al mPER, el mWER y el METEOR. El alineado de palabras se obtuvo automáticamente utilizando el programa GIZA++ (Och y Ney, 2004) en las dos direcciones y simetrizando las dos salidas mediante la unión.

	Árabe	Inglés
Oraciones	50 K	50 K
Palabras	1,41 M	1,57 K
Longitud media	28,15	31,22
Vocabulario	51,10 K	31,51 K

Cuadro 1: El material de entrenamiento.

Los experimentos se hicieron en la máquina Pentium IV Dual Intel Xeon Quad Core X5355 2.66 GHz con 24 Gb de RAM. Los resultados estimados del tiempo computacional y del espacio de memoria son aproximados.

4.2. El preprocesado del árabe

El árabe es un idioma VSO (verbo-sujeto-objeto) con una morfología de esquemas vocálicos en el que las palabras se componen de raíces y afijos así como clíticos pegados a las palabras. Para el preprocesado utilizamos una aproximación parecida a la que se muestra en Habash y Sadat (2006), basada en el sistema MADA+TOKAN para eliminar ambigüedades y tokenización. Para la eliminación de ambigüedades en los diacríticos se empleó exclusivamente la estadística de unigramas. Para la tokenización usamos el esquema D3 con opción –TAGBIES. Este esquema parte del siguiente conjunto de clíticos: f+, b+, k+, l+, Al+ y de los clíticos pronominales. La opción –TAGBIES produce los tags POS Bies para todos los tokens.

4.3. Los experimentos con el sistema TAAS

Para realizar los experimentos seguimos la directriz, disponible on-line: *http://www.cs.cmu.edu/∼zollmann/samt/*. El script que forma parte del sistema de TA MOSES se utilizó para crear el alineado $grow - diag - final$ y extraer las frases puramente léxicas, que posteriormente se utilizaron para inducir la gramática del TAAS. La parte destino (el inglés) del corpus de entrenamiento se procesó con el Penn Treebank parser de Charniak (Charniak, 2000).

Los procedimientos de extracción y filtrado de reglas se restringieron en base a la concatenación de los conjuntos de desarrollo y de test, permitiendo solamente las reglas de hasta 12 elementos en la parte destino y con el criterio de ocurrencia mínima de cero para todas las reglas, pero las reglas puramente abstractas (sin expresiones léxicas) se eliminaron.

El número de frases del estilo Moses extraídas con el sistema basado en frases fue $4{,}8M$, mientras que el número de las reglas generalizadas que representaban el modelo jerárquico creció sensiblemente hasta $22{,}9M$ de las cuales $10{,}8M$ fueron podadas en el proceso de filtrado.

El tamaño del vocabulario de elementos elementales del Penn Treebak es 72, mientras que el número de elementos generalizados, que incluye las categorías adicionales y truncadas es de $35{,}7K$.

El decodificador de búsqueda en haz (beam-search) $FastTranslateChart$ se utilizó como motor de entrenamiento MER con objeto de ajustar los pesos de las funciones características y generar las traducciones "N-best" y "1-best", entrecruzando la búsqueda intensiva con un modelo de lenguaje estándar de 3-gramas (Venugopal, Zollmann, y Vogel, 2007). En el proceso de optimización, el número de iteraciones se limitó a 10, la extracción a la lista 1000-best y se utilizó la medida BLEU como criterio de optimización.

La tabla 2 muestra un resumen de los recursos computacionales necesarios en cada paso de la traducción. Los resultados para el sistema TAAS, junto a los resultados para el sistema de n-gramas y la combinación de sistemas (tal y como se explica en la subsección 4.6) se presentan en la tabla 5.

Paso	Tiempo	Memoria
Parsing	1,5h	80Mb
Extracción de reglas	10h	3,5Gb
Poda & fusión	3h	4Gb
Ajuste de pesos	40h	3Gb
Prueba	2h	3,0Gb

Cuadro 2: TAAS: Recursos computacionales.

4.4. Los experimentos con el sistema basado en n-gramas

Como ya se ha mencionado anteriormente, el modelo principal del sistema basado en n-gramas es un ML basado en 4-gramas de las unidades bilingües, que contiene: 184.345 1-gramas[4], 552.838 2-gramas, 179.466 3-gramas y 176.221 4-gramas.

Junto a este modelo, el sistema de tuplas implementa una combinación loglineal de un ML de 5-gramas del idioma destino estimado con la parte inglesa del corpus paralelo, además de los modelos de POS basados en 4-gramas de los lenguajes fuente y destino. Los tags POS *Bies* se utilizaron para la parte del árabe como se muestra en la subsección 4.2, la herramienta TnT se utilizó para extraer los POS para el inglés (Brants, 2000).

El número de las tuplas no únicas, extraídas inicialmente fue $1{,}1M$, que se podaron de acuerdo con el número máximo de opciones de traducción por tupla en el lado fuente (30). Las tuplas con NULO en la parte fuente se adjuntaron a la unidad anterior o próxima (Mariño et al., 2006).

Los pesos de las características adicionales se ajustaron de acuerdo con el criterio del BLEU máximo. Las exigencias de tiempo y de memoria RAM se presentan en la tabla 3.

Paso	Tiempo	Memoria
Estimación de modelos	0,2h	1,9Gb
Reordenamiento	1h	—
Ajuste de pesos	15h	120Mb
Prueba	2h	120Mb

Cuadro 3: TAE: recursos computacionales.

4.5. Significación estadística

Hemos realizado una prueba de la significación estadística basada en el método "boostrap resampling" como se presenta en Koehn (2004). Para un nivel de confianza del 98 % y 1000 re-extracciones, las traducciones generadas por el TAAS y por el sistema basado en n-gramas son estadísticamente diferentes de acuerdo con el BLEU (43,20±1,69 para el TAAS contra $46{,}42 \pm 1{,}61$ para el sistema de n-gramas).

4.6. Combinación de sistemas

Motivados por el hecho de que muchos sistemas de TA generan traducciones muy diferentes pero de calidad similar, incluso si los modelos que participan en el procedimiento de traducción son semejantes, decidimos combinar las salidas del sistema sintáctico y del sistema de traducción automática puramente estadística. Lo hicimos utilizando la lista de las traducciones más probables generadas por los dos sistemas (1000-best).

Para ello se utilizó el algoritmo del Riesgo Mínimo de Bayes tal como se introdujo en Kumar y Byrne (2004). La tabla 5 demuestra los resultados de la combinación de sistemas en el conjunto de test, contrastada con la traducción *oráculo* hecha como una selección de las traducciones con el BLEU más alto de la unión de las dos listas, la generada por el TAAS y por el sistema de n-gramas.

Además analizamos el porcentaje de contribución de cada sistema a la combinación de sistemas: 55-60 % de las mejores traducciones vienen de la lista 1000-best generada por el sistema de n-gramas en ambos casos (combinación de sistemas y "*oráculo*").

Experimentos	TAAS	N-gramas
Combinación de sistemas	39 %	61 %
Oráculo	44 %	56 %

Cuadro 4: El porcentaje de las oraciones generadas por cada sistema.

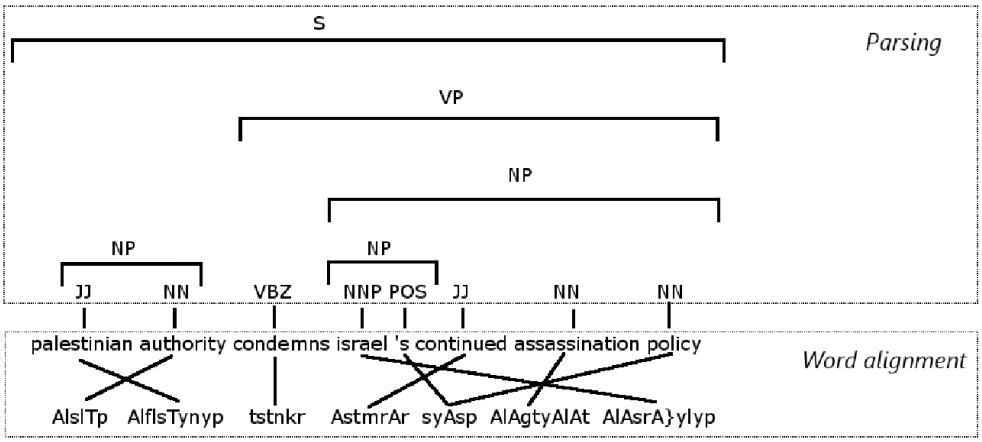
5. Discusión y conclusiones

En esta presentación hemos comparado dos sistemas de traducción automática: un sistema estadístico basado en n-gramas (TAE) y el denominado sistema de traducción automática aumentado con sintaxis (TAAS). La aproximación basada en n-gramas proporciona mejores prestaciones en la tarea analizada. La comparación se ha realizado utilizando el mismo material de entre-

[4]Este número también corresponde al tamaño del vocabulario del modelo bilingüe

	BLEU	NIST	mPER	mWER	METEOR
TAAS	43,20	9,26	36,89	49,45	58,50
TAE basada en n-gramas	46,39	10,06	32,98	48,47	62,36
Combinación de sistemas	48,00	10,15	33,20	47,54	62,27
Oráculo	61,90	11,41	28,84	41,52	66,19

Cuadro 5: Evaluación de la traducción árabe-inglés.



TUPLES (unfolded):
palestinian# AlflsTynyp
authority#AlsTp
condemns#tstnkr
's continued assassination policy#AstmrAr syAsp AlAgtyAlAt
israel# AlAsrA}ylyp

SAMT (example of initial rules):
JJ -> palestinian, AlflsTynyp
NN -> authority, AlslTp
VBZ -> condemns, tstnkr
...
NP -> palestinian authority, AlslTp AlflsTynyp
VBZ+NP -> condemns israel 's continued assassination policy, tsnkr AstmrAr syAsp AlAgtyAlAt AlAsrA}ylyp
VBZ/VP -> israel 's continued assassination policy, AstmrAr syAsp AlAgtyAlAt AlAsrA}ylyp

Figura 1: Ejemplo de la extracción de unidades primitivas en caso de TAAS y de sistema de n-gramas.

SAMT generalized rules (example)

VBZ+NP -> VBZ Israel 's continued assassination policy, VBZ AstmAr syAsp AlAgtyAlAt AlAsrA\}ylyp
VBZ+NP -> VBZ NNP 's continued assassination policy, VBZ AstmAr syAsp AlAgtyAlAt NNP

Figura 2: Ejemplo de las reglas generalizadas (TAAS).

namiento y las mismas herramientas para el preprocesado y el alineado palabra-a-palabra.

Respecto al tamaño de la memoria ocupada y el coste computacional, el sistema de n-gramas ha obtenido también resultados significativamente mejores que los del sistema TAAS, principalmente por el tamaño claramente inferior del espacio de búsqueda.

Se han obtenido resultados comparativos de interés al respecto de las medidas PER y WER: de acuerdo con el PER, el sistema TAE supera a su rival en un 10 % relativo, mientras que la mejora de WER apenas alcanza un 2 %. Esto puede ser explicado en base a que el sistema TAE, comparado con el TAAS, traduce mejor el contexto, pero pro-

duce más errores de reordenamiento. Dado que los idiomas árabe e inglés son lenguas con mucha disparidad en el orden de palabras y debido a la tendencia de producir las unidades cortas, el sistema de n-gramas trata peor los reordenamientos de larga distancia. Sin embargo, al introducir el contexto de las palabras en el modelo de traducción, captura de forma más eficiente las dependencias bilingües de corta distancia.

Finalmente, se ha conseguido una mejora muy significativa mediante la combinación de las salidas de los dos sistemas basados en principios de traducción diferentes.

Como trabajo futuro se va a aplicar la misma metodología de comparación y combinación de sistemas a otras tareas de traducción.

Bibliografía

Brants, T. 2000. TnT – a statistical part-of-speech tagger. En *Proceedings of ANLP-2000*.

Casacuberta, F., E. Vidal, y J. M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. En *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, páginas 39–44.

Charniak, E. 2000. A maximum entropy-inspired parser. En *Proceedings of NAACL 2000*, páginas 132–139.

Charniak, J. 2003. Learning non-isomorphic tree mappings for machine translation. En *Proceedings of ACL 2003 (companion volume)*, páginas 205–208.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. En *Proceedings of ACL 2005*, páginas 263–270.

Crego, J. M., J. Mariño, y A. de Gispert. 2005. An Ngram-based Statistical Machine Translation Decoder. En *Proceedings of INTERSPEECH05*, páginas 3185–3188.

Crego, J. M. y J. B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Habash, N. y F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. En *Proceedings of the Human Language Technology Conference of the NAACL*, páginas 49–52.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. En *Proceedings of EMNLP 2004*, páginas 388–395.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based machine translation. En *Proceedings of HLT-NAACL 2003*, páginas 48–54.

Kumar, S. y W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. En *Proceedings of HLT/NAACL 2004*.

Mariño, J. B., R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, y M. R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.

Melamed, I.D. 2004. Statistical machine translation by parsing. En *Proceedings of ACL 2004*, páginas 111–114.

Och, F. y H.Ñey. 2003. A systematic comparison of various statistical alignment models. En *Computational Linguistics*, volumen 29(1), páginas 19–52.

Och, F. y H.Ñey. 2004. The alignment template approach to statistical machine translation. En *Computational Linguistics*.

Och, F. J. y H.Ñey. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. En *Proceedings of ACL 2002*, páginas 295–302.

Steedman, M. 1999. Alternative quantifier scope in ccg. En *Proceedings of ACL 1999*, páginas 301–308.

Venugopal, A., A. Zollmann, y S. Vogel. 2007. An Efficient Two-Pass Approach to Synchronous-CFG Driven Statistical MT. En *Proceedings of HLT/NAACL 2007*, páginas 500–507.

Zollmann, A. y A. Venugopal. 2006. Syntax augmented machine translation via chart parsing. En *Proceedings of NAACL 2006*.