

La traducción automática del habla

Dr. Rafael E. Banchs (rbanchs@gps.tsc.upc.es)

Universidad Politécnica de Cataluña

Resumen de la ponencia:

La traducción automática entre dos lenguas constituye un problema que ha llamado la atención de investigadores y estudiosos de diversas disciplinas desde tiempos inmemorables. Tras la aparición de los primeros ordenadores, después de la segunda guerra mundial, la traducción automática comenzó a perfilarse como un problema factible de ser resuelto. Hoy por hoy, este problema constituye un área de investigación aún naciente pero muy activa, y está despertando mucho interés, no sólo desde el punto de vista científico, sino también, desde los puntos de vista económico, político, social y cultural.

El presente trabajo, después de presentar un breve resumen histórico de la disciplina y las distintas aproximaciones existentes para abordar el problema, presenta el estado actual del arte en traducción automática del habla haciendo énfasis en los métodos de traducción estadísticos. Así mismo, se presentan y se discuten dos ejemplos de experiencias prácticas recientes de traducción automática del habla. Finalmente, se discuten los principales retos que actualmente enfrenta la investigación en esta materia, y se plantean las principales líneas de investigación futuras en el área.

Curriculum Vitae:

Rafael E. Banchs es Ingeniero Electrónico egresado de la Universidad Simón Bolívar (Caracas, Venezuela), 1991. Obtuvo el título de Magíster en Ingeniería Electrónica de la misma universidad en 1993, y el título de Ph.D. in Electrical Engineering de la Universidad de Texas en Austin, en 1998. Su área de investigación ha estado orientada al procesamiento de señales, específicamente en problemas de modelado, clasificación y estimación de parámetros. Actualmente se desempeña como investigador, con el Programa Ramón y Cajal, en el área de traducción estocástica del habla en la Universidad Politécnica de Cataluña.

Agradecimientos:

El autor agradece al Dr. José B. Mariño y a Patrik Lambert por sus comentarios sobre el material y su valiosa ayuda para la revisión de este documento.

Un poco de historia

“...entonces dijeron: hagamos una ciudad y una torre tan alta que alcance el cielo y hagamos nuestro nombre famoso (...) pero Dios confundió su lengua de forma que ellos no se entendían los unos a los otros (...) y se esparcieron por todos los lugares y así dejaron de construir la ciudad.” (Génesis 11:4-8)

La historia de la Torre de Babel es una historia sobre la diversidad humana. La diversidad lingüística del ser humano es tan grande como la diversidad geográfica del planeta. Se estima que en el mundo existen alrededor de unas 6000 lenguas diferentes¹, algunas de las cuales se encuentran en vías de extinción. Históricamente, las barreras comunicacionales entre personas, o sociedades, de distinta lengua han constituido un factor determinante en muchas de las actividades del ser humano: el comercio, la política, la religión, el intercambio cultural y la guerra. La organización geopolítica del mundo a lo largo de la historia ha estado íntimamente relacionada con la distribución de los grupos étnicos y lingüísticos en las distintas regiones geográficas. La lengua es, definitivamente, un elemento de identidad colectiva y de diferenciación cultural.

De esta forma, el problema de la comunicación entre personas de distinta lengua es tan antiguo como el origen de las primeras civilizaciones. El origen de la traducción automática ha sido asociado por Hutchins² a los primeros intentos de desarrollar lenguajes universales durante el siglo XVII. No obstante, se tiene conocimiento de esfuerzos para la sistematización de la traducción entre lenguas que datan del siglo IV³.

Para los efectos prácticos, el nacimiento de la traducción automática es una consecuencia de la aparición del ordenador, inmediatamente después de la segunda guerra mundial. Aunado a esto, el nacimiento de la teoría de la información, así como el desarrollo de métodos de codificación y decodificación de mensajes durante la segunda guerra mundial⁴, condujeron a un enfoque “criptográfico” del problema de traducción automática del cual surgió la idea del uso de modelos estadísticos para abordar el problema. Sin embargo, las grandes limitaciones de los primeros ordenadores condujeron a un rápido desencanto y pérdida del entusiasmo en la empresa de la traducción automática, y por tanto, a un subsiguiente abandono del área por parte de muchos investigadores.

En la década de los setenta, con el advenimiento de la inteligencia artificial, el problema de la traducción automática despertó nuevamente el interés de muchos, pero bajo un enfoque totalmente diferente: el uso de un interlingua⁵. La aproximación del interlingua, como se explicará en la próxima sección, se basa en la abstracción de ideas y conceptos extraídos de la lengua de origen y su posterior representación en la lengua destino. De esta manera, el proceso de traducción automática se plantea en términos de la comprensión y entendimiento del lenguaje, en la misma forma en que lo hacen los seres humanos⁶. A pesar del entusiasmo y algunos avances interesantes⁷, esta aproximación al problema de la traducción automática se vio rápidamente inmersa en un callejón sin salida.

Finalmente, en la década de los noventa, la aparición simultánea de dos factores fundamentales hizo posible la puesta en práctica de métodos estadísticos para la traducción automática. Estos factores fueron: primero, el desarrollo de ordenadores con un alto poder de cómputo y capacidades de almacenamiento en memoria prácticamente ilimitadas; y segundo, el desarrollo y disponibilidad de grandes volúmenes de datos bilingües. Se puede considerar, que el trabajo de Peter Brown y sus colegas⁸, en el Watson Research Center de IBM, marcó el inicio de una nueva era para la traducción automática: la era de la traducción automática estocástica. En muy poco tiempo, los métodos estadísticos lograron igualar y superar a los métodos de traducción basados en conocimiento que habían sido objeto de un continuo desarrollo durante algo más de dos décadas.

Simultáneamente, los avances tecnológicos en las áreas de reconocimiento automático del habla y de síntesis de voz⁹ hicieron posible un nuevo sueño: la traducción automática del habla¹⁰. La idea de traducción automática del habla ha redimensionado por completo el problema de la traducción automática, sacándolo de los despachos de los burócratas y las compañías editoriales a la calle y al ciudadano común. La traducción automática del habla encuentra su aplicación en una infinidad de actividades en un mundo globalizado como el de hoy; desde el turismo hasta la política internacional, pasando por el comercio, la educación y el intercambio científico y cultural. Así mismo, el desarrollo de la Internet y de las tecnologías de información y comunicación están aumentando continuamente tanto los recursos como la demanda de los sistemas de traducción automática en general.

En forma cónsona, el interés en el problema de traducción automática, así como la investigación y desarrollo en esta área han aumentado significativamente en los últimos años. En el caso particular de la Unión Europea, donde la diversidad lingüística continúa representando una de las principales barreras al proceso de integración, se han invertido una gran cantidad de recursos en actividades de investigación y desarrollo en esta área. Como ejemplo de este esfuerzo se pueden mencionar, entre otros, los proyectos C-Star, Eutrans, Verbmobil, LC-Star, Nespole!, Fame, y TC-Star.¹¹

No obstante, a pesar de todos los esfuerzos, los avances evidentes y la existencia de una gran cantidad de sistemas de traducción automática disponibles en el mercado, e incluso gratuitamente, el problema de la traducción automática continúa siendo actualmente uno de los retos fundamentales de la inteligencia artificial, del procesamiento del lenguaje natural y de la ingeniería en general¹². Los sistemas de traducción existentes hoy en día tienen grandes limitaciones. Entre las más importantes destacan su aún muy alta tasa de errores en las traducciones obtenidas, su general limitación de vocabulario y restricción a dominios específicos, y su incapacidad para el adecuado manejo de las ambigüedades características de los lenguajes humanos. Adicionalmente, en el caso particular de la traducción automática del habla, el estado actual del arte en reconocimiento automático del habla y las irregularidades gramaticales características del lenguaje hablado, constituyen factores que dificultan aún más la tarea de traducción automática. De esta forma, la traducción automática continúa siendo un área naciente y muy activa, con grandes retos para los investigadores y grandes promesas para los empresarios y el público en general.

En el presente trabajo se describen brevemente las distintas aproximaciones existentes al problema de traducción automática, haciendo un mayor énfasis en los métodos de traducción basados en modelos estadísticos. Posteriormente se describe el estado actual del arte y se presentan algunas experiencias prácticas de investigación y desarrollo orientadas al problema de la traducción automática del habla. Específicamente se describen algunos experimentos y resultados de traducción realizados con las transcripciones de los discursos de las Sesiones Plenarias del Parlamento Europeo¹³. Así mismo, se describe un trabajo de evaluación de una plataforma experimental para la comunicación telefónica entre dos personas de distinta lengua¹⁴. Finalmente, se discuten los retos específicos que la traducción automática del habla impone a los sistemas de traducción actuales, así como algunas de las perspectivas futuras para la investigación y el desarrollo en el área de la traducción automática del habla.

Aproximaciones al problema

En la actualidad, existen diversos enfoques sobre el problema de la traducción automática¹⁵. En términos generales, todos estos enfoques se pueden agrupar en dos grandes categorías. Los métodos basados en el uso de conocimiento, y los métodos basados en el uso de corpus paralelos (grandes volúmenes de datos bilingües). Como su nombre lo indica, los métodos basados en conocimiento, también conocidos como métodos basados en reglas, abordan el

problema de la traducción automática mediante el uso de herramientas de análisis e interpretación que emplean conocimiento morfológico, sintáctico y/o semántico de los idiomas fuente y destino para generar las traducciones correspondientes. Entre los métodos más comunes de esta categoría se pueden mencionar los siguientes: el interlingua, los modelos de transferencia y la traducción directa.

1.- *Interlingua*. Esta aproximación, como se mencionó en la sección anterior, se basa en la interpretación de las ideas y conceptos contenidos en la oración fuente a ser traducida. Estos conceptos, sus significados y las relaciones entre ellos son representados en una forma canónica independiente del lenguaje denominada interlingua. Posteriormente, la traducción es obtenida mediante la generación de una representación en la lengua destino de los conceptos, significados y relaciones codificados en el interlingua. De esta manera, el proceso de traducción automática se plantea en términos de la comprensión y entendimiento de las ideas expresadas en un idioma fuente y la posterior generación del mensaje correspondiente en un idioma destino, de una forma similar a como lo haría un ser humano. En la figura #1, se presenta un ejemplo ilustrativo de traducción mediante el uso de un interlingua.

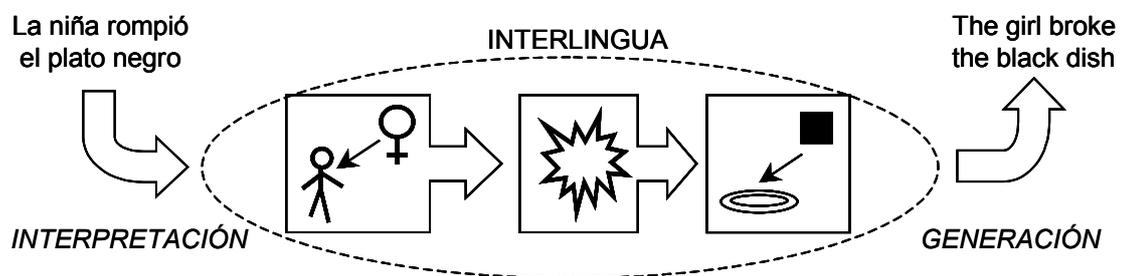


Figura #1: Traducción automática mediante el empleo de un interlingua.

La aproximación del interlingua tiene la gran ventaja de ser multilingüe, en el sentido de que una vez definido el interlingua, la traducción desde un (y hacia un) idioma específico depende única y exclusivamente de las reglas y el conocimiento necesario para la interpretación y la generación de ese idioma.

2.- *Modelos de transferencia (o transfer)*. Esta aproximación se basa en el uso de las diferencias morfológicas, sintácticas y gramaticales entre dos idiomas para transformar la información lingüística de un idioma al otro mediante el uso de un conjunto de reglas de transferencia denominadas modelo de transferencia (o transfer). Este proceso involucra tres etapas específicas: análisis, en la cual la información es extraída del lenguaje fuente; transferencia, en la cual la información es transformada al lenguaje destino; y síntesis, en la cual la forma final en el lenguaje destino es generada a partir de la información transformada. La figura #2 ilustra el funcionamiento de un sistema de traducción basado en un modelo de transfer.

A diferencia del interlingua, que es multilingüe, la aproximación del modelo de transfer es netamente bilingüe, ya que este modelo se basa en las diferencias existentes entre dos idiomas específicos. De esta forma, en un sistema de traducción automática multilingüe, cada par de idiomas necesita el entrenamiento de un modelo de transfer específico. Esto constituye una desventaja importante con respecto al sistema de traducción basado en interlingua. Así por ejemplo, para un sistema de traducción automática multilingüe entre n idiomas, hacen falta un total de $n(n-1)/2$ modelos de transfer; mientras que para un sistema basado en interlingua sólo harían falta n modelos de traducción idioma-interlingua. Sin embargo, como resulta lógico pensar, los sistemas basados en interlingua son mucho más complejos y difíciles de implementar que los modelos de transfer. De hecho, no existe todavía en la práctica ningún sistema de traducción que se base únicamente en el uso de un interlingua.

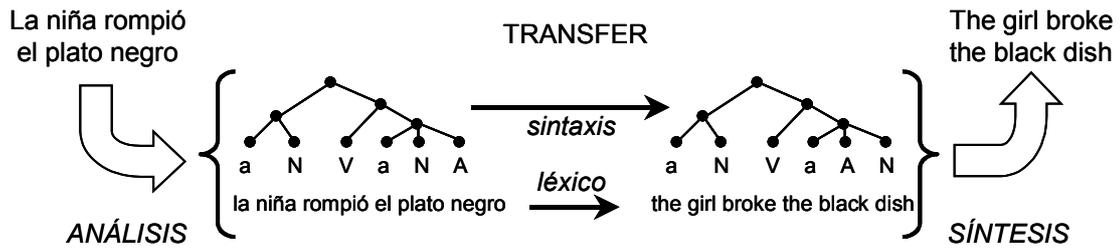


Figura #2: Traducción automática mediante el empleo de un modelo de transferencia.

3.- *Sistemas de traducción directa.* Esta aproximación se basa en la implementación de una secuencia de pasos que permite tratar los distintos aspectos específicos de la traducción de forma independiente. A continuación se presenta un ejemplo que ilustra como sería un proceso muy simple de traducción directa:

Oración fuente: La niña rompió el plato negro.

Paso #1: traducción de las palabras de contenido
Oración intermedia: girl broke dish black

Paso #2: ordenamiento:
Oración intermedia: girl broke black dish

Paso #3: edición final
Oración destino: The girl broke the black dish

Los sistemas de traducción directa están concebidos bajo la premisa de que es preferible una mala traducción a ninguna traducción. De esta forma, estos sistemas tienen como prioridad ser robustos, en el sentido de no depender totalmente de un procedimiento de análisis sofisticado y específico y poder generar siempre una salida. Al igual que los sistemas de traducción basados en modelos de transfer, estos sistemas se diseñan en forma específica para un par de idiomas determinado.

En cuanto a la segunda categoría de métodos de traducción automática previamente mencionados, como su nombre lo indica, se basan en el uso de grandes volúmenes de datos bilingües, o corpus paralelos, como fuente de información para la posterior inferencia de las traducciones requeridas. Entre los métodos más comunes de esta categoría se pueden mencionar los sistemas de traducción basados en ejemplos y los sistemas de traducción estadísticos. Los sistemas de traducción basados en ejemplos utilizan los datos bilingües como una gran base de datos de traducciones. Para generar la traducción de una oración fuente F , simplemente buscan aquella oración que más se le parezca en la base de datos y utilizan su traducción como una aproximación a la traducción de la entrada F . En caso de no hallarse oraciones similares en la base de datos, se procede a realizar una serie de transformaciones a la oración fuente F , tal que sea posible su traducción mediante el empleo de la base de datos. En cuanto a los sistemas de traducción estadísticos, estos serán tratados con lujo de detalles en las siguientes secciones del presente trabajo.

La aproximación estocástica

Tal y como se mencionó en la primera sección, el enfoque estadístico del problema de traducción automática tuvo sus orígenes en los trabajos de codificación y decodificación de mensajes durante la segunda guerra mundial. Sin embargo, la implementación exitosa de estas ideas no pudo concretarse sino hasta principios de la década de los 90, cuando la capacidad

computacional de los ordenadores y la disponibilidad de grandes volúmenes de datos bilingües lo permitieron.

La aproximación estadística al problema de traducción se inspira en el modelo de canal ruidoso, el cual a su vez está fundamentado matemáticamente en el teorema de Bayes. De esta forma, el problema de traducir una oración dada en una lengua fuente –por ejemplo, castellano– a una lengua destino – inglés–, se concibe en términos de la decodificación de una señal original en inglés que fue transformada, al pasar por un canal ruidoso, en la señal observada en castellano.

Desde el punto de vista matemático, lo que se pretende es buscar aquella oración destino D_0 que maximice la probabilidad condicional de una traducción D , dada la oración fuente F que se desea traducir. Así, en el ejemplo de traducción del castellano al inglés, $p(D|F)$ representa la probabilidad de que una oración en inglés D (destino), sea la traducción de una oración dada en castellano F (fuente). De acuerdo con el teorema de Bayes, dicha probabilidad condicional se expresa de la siguiente manera:

$$p(D|F) = \frac{p(F|D) p(D)}{p(F)} \quad \text{Ecuación \#1}$$

El producto $p(F|D) p(D)$ de la ecuación #1 recibe el nombre de modelo de canal ruidoso. En el contexto de la traducción estadística, $p(F|D)$ se denomina el modelo de traducción y representa la probabilidad de que la oración fuente F corresponda a una traducción de D . Similarmente, $p(D)$ se denomina modelo de lenguaje destino y representa la probabilidad de que D constituya una oración válida en el lenguaje destino. De esta forma, el problema de traducción se implementa a través de un problema de búsqueda, en el cual se debe encontrar aquella oración destino D que maximice el producto de las dos probabilidades mencionadas. Debido a que la oración fuente F está dada, el problema de maximizar $p(D|F)$ es totalmente equivalente al problema de maximizar el producto $p(F|D) p(D)$, tal y como se muestra a continuación:

$$\operatorname{argmax}_D \frac{p(F|D) p(D)}{p(F)} = \operatorname{argmax}_D p(F|D) p(D) \quad \text{Ecuación \#2}$$

Este problema de búsqueda, también llamado decodificación, es en realidad un problema de optimización entera, en el cual se debe explorar un espacio que en la práctica es imposible de ser explorado en un tiempo razonable. Por esta razón, los métodos comúnmente usados en la práctica para resolver este problema de búsqueda sólo exploran regiones limitadas en el espacio de las posibles soluciones, de forma que, aunque no garantizan soluciones óptimas, son mucho más eficientes desde el punto de vista computacional¹⁶.

La primera pregunta que surge inmediatamente después del problema planteado en la ecuación #2 tiene que ver con la estimación de los modelos de traducción y de lenguaje destino. Aunque existen muchas formas diferentes de modelar un lenguaje, la manera más común de hacerlo en el contexto estadístico es mediante el uso de un modelo de n-gramas. En teoría, la probabilidad de ocurrencia de una oración D , compuesta por la secuencia de palabras $d_1, d_2, d_3 \dots d_k$, se puede expresar de la siguiente manera usando la regla de la cadena:

$$p(D) = p(d_1) p(d_2 | d_1) p(d_3 | d_2, d_1) \dots p(d_k | d_{k-1} \dots d_2, d_1) \quad \text{Ecuación \#3}$$

En la práctica, las probabilidades del tipo $p(d_j | d_{j-1}, \dots, d_2, d_1)$ que aparecen en la ecuación #3, son aproximadas mediante el uso de probabilidades del tipo $p(d_j | d_{j-1}, \dots, d_{j-n})$, las cuales son mucho más sencillas de calcular cuando se utilizan valores pequeños para n . Estas

aproximaciones se denominan n-gramas, y consecuentemente los modelos de lenguaje que las utilizan se denominan modelos de lenguaje basados en n-gramas¹⁷. La estimación de las probabilidades de los n-gramas se hace de forma muy sencilla mediante el conteo de palabras en un corpus de datos.

En lo que se refiere al modelo de traducción $p(F|D)$, también existen muchas formas diferentes de representarlo. A principios de la década de los 90, Peter Brown y sus colegas propusieron cinco modelos de traducción basados en el alineamiento de palabras, y que hoy son popularmente conocidos como los modelos IBM¹⁸. La existencia de cinco modelos, obedece a una secuencia de complejidad creciente en el entrenamiento de los parámetros. En esta secuencia, cada uno de los modelos es usado como punto de partida para el entrenamiento del modelo siguiente.

El cálculo de los parámetros de los modelos IBM requiere la existencia de un corpus bilingüe alineado, para el cual se deben conocer las probabilidades de todos los posibles alineamientos para cada par bilingüe de oraciones. El concepto de alineamiento se ilustra con un ejemplo concreto en la figura #3, en la cual se presentan cuatro posibles alineamientos diferentes para un par de oraciones de dos palabras cada una.



Figura #3: Cuatro posibles alineamientos para un par de oraciones bilingüe.

Como se puede observar en la figura #3, un alineamiento se compone de un conjunto de enlaces definidos entre palabras de una oración fuente y palabras de su correspondiente oración destino. En principio, cualquier conjunto de enlaces constituye un alineamiento, incluyendo el conjunto vacío y el conjunto de todos los enlaces posibles. En este sentido, el ejemplo mostrado en figura #3 admite un total de 16 alineamientos diferentes.

Debido a que el número de alineamientos puede ser extremadamente grande para oraciones con muchas palabras, en los modelos de IBM se restringe el número de alineamientos posibles a sólo aquellos en los que todas las palabras de la oración fuente están asociadas a un solo enlace. Adicionalmente, los modelos de IBM, incluyen el símbolo *null* al principio de la oración destino como un artefacto que permite modelar la presencia de aquellas palabras en la oración fuente que no se corresponden con ninguna de las palabras presentes en la oración destino. La figura #4 muestra cuatro, del total de nueve, enlaces permitidos por los modelos IBM para el mismo ejemplo representado en la figura #3.

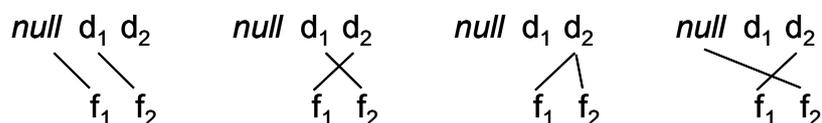


Figura #4: Cuatro de los alineamientos permitidos por los modelos IBM.

Una observación importante que se desprende de la figura #4, es la posibilidad que tienen las palabras del destino a estar asociadas a ningún enlace, un solo enlace o más de un enlace. Esta propiedad se denomina fertilidad, y constituye, conjuntamente con la distorsión, un parámetro esencial para los modelos de traducción $p(F|D)$, a partir del modelo IBM-3 en adelante. La distorsión, por su parte, es un parámetro que modela la probabilidad que tienen las palabras de saltar de una posición a otra.

El concepto de alineamiento es fundamental para la estimación de los parámetros léxicos $t(f|d)$ de los modelos IBM. Estos parámetros modelan la probabilidad de ocurrencia de una palabra fuente f , dada una palabra destino d . Para calcular los parámetros léxicos, es necesario tener una estimación de la probabilidad de los alineamientos para cada par bilingüe de oraciones, ya que la probabilidad entre las palabras fuente y destino se infiere en función de aquellos alineamientos que son más probables. El gran problema de este procedimiento aparece cuando nos damos cuenta de que para poder calcular las probabilidades de los alineamientos es necesario disponer de un estimado de las probabilidades entre las palabras fuente y destino, es decir, ¡ hacen falta los parámetros léxicos !

Este problema, que a primera vista pareciera insalvable, se resuelve mediante el uso de un algoritmo de "bootstrapping" denominado EM¹⁹. Este algoritmo, comienza con un conjunto de valores iniciales de parámetros léxicos $t(f|d)$, a partir de los cuales estima las probabilidades de los alineamientos $p(a|F,D)$, las cuales son a su vez utilizadas para actualizar los parámetros léxicos iniciales. Después de sucesivas iteraciones, el algoritmo converge a un conjunto de valores no necesariamente óptimos, pero generalmente buenos, de parámetros léxicos. Entrar en los detalles de cómo funciona el algoritmo EM no constituye un objetivo de este trabajo. Sin embargo, con el objeto de ilustrar el proceso de entrenamiento de un modelo de traducción, en la siguiente sección se presenta un ejemplo de como es posible inferir los parámetros léxicos para una versión simplificada del modelo IBM-1 a partir de un conjunto de datos²⁰.

Un ejercicio ilustrativo

Para empezar, considérense los datos presentados en la figura #5. Estos datos corresponden a tres pares bilingües de oraciones, donde el lenguaje fuente es el castellano y el lenguaje destino es el inglés. Dos simplificaciones importantes se considerarán en este ejemplo: primero, no se incluirá el símbolo *null* para los alineamientos; y segundo, de todos los alineamientos posibles para cada uno de los pares bilingües de oraciones sólo se considerarán dos. Éstos se denominarán: el alineamiento directo (a_d), representado en la figura con líneas continuas, y el alineamiento cruzado (a_c), representado en la figura con líneas segmentadas.

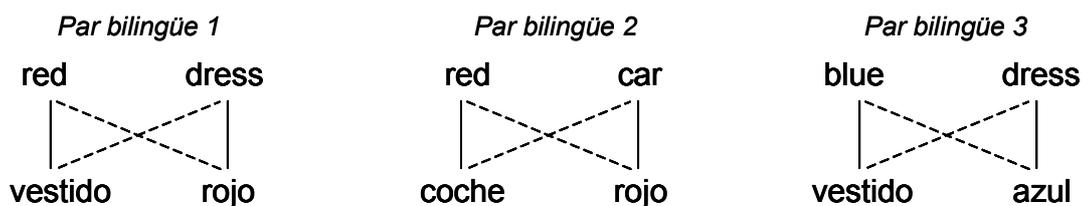


Figura #5: Datos de entrenamiento para el ejemplo a ser considerado.

En la versión simplificada del modelo IBM-1 a ser usada, la probabilidad de ocurrencia de una oración fuente F dada una oración destino D (modelo de traducción) estará dada por la siguiente expresión:

$$p(F|D) = \sum_a p(a,F|D)$$

Ecuación #4

donde $p(a,F|D)$ representa la probabilidad del alineamiento a para la oración fuente F , dada la oración destino D . Nótese que la sumatoria corre sobre todos los alineamientos posibles para un par de oraciones dado, que en este ejemplo son solamente dos: el alineamiento directo y el

alineamiento cruzado. Esta probabilidad de alineamiento se calcula en función de los parámetros léxicos de acuerdo con la siguiente expresión:

$$p(a, F | D) = \prod_k t(f_k | d_{ak})$$

Ecuación #5

donde f_k se refiere a la k -ésima palabra de la oración fuente F , y d_{ak} a la palabra de la oración destino D que está enlazada con f_k en el alineamiento a . Nótese la distinción entre palabras y oraciones en la nomenclatura mediante el uso de minúsculas y mayúsculas, respectivamente. De acuerdo con las ecuaciones #4 y #5, una vez que se conozcan los valores de los parámetros léxicos $t(f|d)$, el modelo de traducción $p(F|D)$ puede ser calculado.

El valor de $t(f|d)$ para un par de palabras f y d , se define en función de una cuenta pesada y normalizada de la cantidad de veces que ambas palabras aparecen enlazadas entre sí en todos los alineamientos posibles de todas las oraciones del corpus bilingüe. El factor de peso que se aplica a cada elemento de la cuenta está definido por la probabilidad del alineamiento dado el par de oraciones bilingües correspondientes $p(a|F, D)$. La expresión matemática para la cuenta pesada $c(f|d)$ es la siguiente:

$$c(f|d) = \sum_n \sum_a p(a | F_n, D_n) \delta(a, f, d)$$

Ecuación #6

donde $\delta(a, f, d)$ vale 1 si las palabras f y d están enlazadas en el alineamiento a de las oraciones F_n y D_n , y vale cero en el caso contrario; n se refiere al n -ésimo par de oraciones del corpus. Nótese que aquí las cuentas se realizan sobre todo el corpus. Los parámetros léxicos $t(f|d)$ se obtienen mediante la subsiguiente normalización de la cuenta pesada $c(f|d)$ obtenida en la ecuación #6. Esta normalización se realiza de acuerdo con la siguiente expresión:

$$t(f|d) = \frac{c(f|d)}{\sum_f c(f|d)}$$

Ecuación #7

El único parámetro que faltaría aún por calcular es el correspondiente a la probabilidad de los alineamientos para un par bilingüe de oraciones $p(a|F_n, D_n)$. Pero esta probabilidad es en realidad una versión normalizada de la probabilidad definida en la ecuación #5, tal y como se muestra a continuación:

$$p(a | F, D) = \frac{P(a, F | D)}{\sum_a P(a, F | D)}$$

Ecuación #8

Las ecuaciones #5, #6, #7, y #8, constituyen la base para implementar una versión simplificada del algoritmo EM. A continuación se ilustrará el proceso de estimación de los parámetros léxicos correspondientes al ejemplo de la figura #5, mediante el uso de dichas ecuaciones en forma iterativa. En las tablas #1 y #2, se presentan los valores calculados para las primeras dos iteraciones de las probabilidades de alineamiento y los parámetros léxicos, respectivamente.

Para iniciar los cálculos se debe asignar un valor inicial a los parámetros léxicos $t(f|d)$. En el ejemplo considerado, los posibles valores para f son cuatro: "vestido", "coche", "rojo" y "azul". Similarmente, los posibles valores para d también son cuatro: "red", "dress", "car" y "blue". De

esta manera, existen en total 16 parámetros léxicos $t(f|d)$ que deben ser estimados. Asumiendo un desconocimiento total de cualquier información relativa a las posibles correspondencias entre los valores de f y d , lo más sensato es inicializar los parámetros léxicos usando una distribución uniforme de probabilidades. Así, los valores iniciales serán $t(f|d)=1/4$.

Con los valores iniciales de los parámetros léxicos y la ecuación #5, se pueden calcular las probabilidades de alineamiento $p(a, F|D)$. Un total de seis probabilidades deben ser calculadas, ya que se dispone de tres pares bilingües de oraciones con dos alineamientos posibles para cada uno. Así por ejemplo:

$$p(a_c, F_1|D_1) = t(f_1|d_2) t(f_2|d_1) = 1/4 \cdot 1/4 = 1/16.$$

En forma similar se calculan las cinco probabilidades de alineamiento restantes. Los valores resultantes se presentan en la columna "Iteración #1: $p(a, F|D)$ " de la tabla #1. Nótese que todos los valores obtenidos son iguales.

En el siguiente paso, se usa la ecuación #8 para normalizar los valores previamente obtenidos. Así por ejemplo:

$$p(a_c|F_1, D_1) = p(a_c, F_1|D_1) / \{ p(a_c, F_1|D_1) + p(a_d, F_1|D_1) \} = (1/16) / (1/16 + 1/16) = 1/2.$$

El resto de los valores se presenta en la columna "Iteración #1: $p(a|F, D)$ " de la tabla #1. Nótese nuevamente que todos los valores obtenidos son iguales.

Tabla #1: Probabilidades de alineamiento para las dos primeras iteraciones.

Par bilingüe	Alineamiento	Iteración #1		Iteración #2	
		$p(a, F D)$	$p(a F, D)$	$p(a, F D)$	$p(a F, D)$
#1	cruzado	1/16	1/2	1/4	4/5
	directo	1/16	1/2	1/16	1/5
#2	cruzado	1/16	1/2	1/4	2/3
	directo	1/16	1/2	1/8	1/3
#3	cruzado	1/16	1/2	1/4	2/3
	directo	1/16	1/2	1/8	1/3

Una vez obtenidas las probabilidades de alineamiento normalizadas $p(a|F, D)$, se procede a calcular las cuentas pesadas $c(f|d)$ definidas en la ecuación #6. Considérese, por ejemplo, la cuenta correspondiente al par de palabras $f=vestido$ y $d=red$. Como estas dos palabras sólo aparecen enlazadas en el alineamiento directo del par bilingüe #1, el parámetro $\delta(a, f, d)$ sólo vale 1 en ese caso y vale cero en todos los demás casos. De esta forma:

$$c(vestido|red) = p(a_d|F_1, D_1) = 1/2.$$

Considérese como un segundo ejemplo las palabras $f=vestido$ y $d=dress$. En este caso, las palabras aparecen enlazadas dos veces en todo el corpus: en el alineamiento cruzado del par bilingüe #1 y en el alineamiento cruzado del par bilingüe #3. De esta forma:

$$c(vestido|dress) = p(a_c|F_1, D_1) + p(a_c|F_3, D_3) = 1/2 + 1/2 = 1.$$

En forma similar, se calculan las cuentas pesadas restantes. Los 16 valores resultantes se presentan en la tabla #2 (sub-tabla correspondiente a las columnas en $c(f|d)$ y las filas en #1).

En el siguiente paso, se calculan los parámetros léxicos $t(f|d)$ mediante la normalización de las cuentas recién obtenidas $c(f|d)$, tal y como se describe en la ecuación #7. Así por ejemplo:

$$t(\text{vestido}|\text{red}) = c(\text{vestido}|\text{red}) / \{c(\text{vestido}|\text{red})+c(\text{rojo}|\text{red})+c(\text{coche}|\text{red})+c(\text{azul}|\text{red})\}$$

$$= (1/2) / (1/2+1+1/2+0) = 1/4.$$

El resto de las normalizaciones se efectúa de forma similar. Los 16 valores resultantes para los parámetros léxicos se presenta en la tabla #2 (sub-tabla correspondiente a las columnas en $t(f|d)$ y las filas en #1).

Tabla #2: Valores de las cuentas y los parámetros léxicos para las dos primeras iteraciones.

Iteración		$c(f d)$				$t(f d)$			
		red	dress	car	blue	red	dress	car	blue
#1	vestido	1/2	1	0	1/2	1/4	1/2	0	1/2
	rojo	1	1/2	1/2	0	1/2	1/4	1/2	0
	coche	1/2	0	1/2	0	1/4	0	1/2	0
	azul	0	1/2	0	1/2	0	1/4	0	1/2
#2	vestido	1/5	22/15	0	1/3	1/10	11/15	0	1/3
	rojo	22/15	1/5	1/3	0	11/15	1/10	1/3	0
	coche	1/3	0	2/3	0	1/6	0	2/3	0
	azul	0	1/3	0	2/3	0	1/6	0	2/3

De esta manera concluye la primera iteración de esta versión simplificada del algoritmo EM. Obsérvese en la tabla #2, como ahora se dispone de unos valores actualizados para los parámetros léxicos mucho mejores que los originales. En este punto, ya se pueden observar parámetros que empiezan a tener algo de sentido, así como por ejemplo $t(\text{coche}|\text{dress})=0$ y $t(\text{rojo}|\text{red})=1/2$. Nótese, sin embargo, que en algunos casos todavía los valores estimados no son buenos.

Para la segunda iteración, se procede nuevamente a calcular las probabilidades de alineamiento $p(a,F|D)$, utilizando la ecuación #5 y los valores actualizados de los parámetros léxicos $t(f|d)$. Luego, los resultados obtenidos son usados para calcular las probabilidades de alineamiento normalizadas $p(a|F,D)$ utilizando la ecuación #8. Los resultados obtenidos se presentan en las columnas correspondientes a la iteración #2 en la tabla #1. Posteriormente, mediante el uso de las ecuaciones #6 y #7, se calculan nuevas versiones actualizadas de las cuentas $c(f|d)$ y los parámetros léxicos $t(f|d)$. Los valores resultantes se presentan en las filas correspondientes a la iteración #2 en la tabla #2.

Si se continúa este proceso iterativo, se encontrará que después de unas 8 o 10 iteraciones todos los parámetros léxicos, con excepción de cuatro de ellos, tendrán un valor muy próximo a cero. Los cuatro parámetros distintos de cero tienden, efectivamente a 1. Ellos son precisamente:

$$t(\text{vestido}|\text{dress}) = t(\text{rojo}|\text{red}) = t(\text{coche}|\text{car}) = t(\text{azul}|\text{blue}) = 1.$$

Hasta este punto, sólo se ha entrenado el modelo de traducción, pues como se mencionó anteriormente, una vez conocidos los parámetros léxicos, el modelo de traducción puede ser estimado usando las ecuaciones #4 y #5.

Para terminar el ejercicio, considérese ahora el problema de decodificación o búsqueda. Se desea hallar la mejor traducción para la oración fuente $F = \text{"coche azul"}$. Nótese que dicha oración no existe en el corpus de entrenamiento que se usó para entrenar el modelo. Para resolver este problema, se deben considerar todas las oraciones destino posibles D y evaluar cual de ellas tiene una mayor probabilidad de ser la traducción correcta de F . Para el reducido

conjunto de alineamientos permitidos en el modelo considerado (directo y cruzado), el espacio de búsqueda se reduce a una docena de oraciones destino posibles de dos palabras cada una.

De esta forma, la probabilidad de traducción queda, de acuerdo con las ecuaciones #4 y #5, definida como:

$$p(\text{"coche azul"}|d_1 d_2) = p(a_d, \text{"coche azul"}|d_1 d_2) + p(a_c, \text{"coche azul"}|d_1 d_2) \\ = t(\text{coche}|d_1) t(\text{azul}|d_2) + t(\text{coche}|d_2) t(\text{azul}|d_1).$$

Para encontrar la traducción más probable, esta expresión debe ser evaluada para todas las oraciones destino posibles $D = "d_1 d_2"$, y los valores resultantes deben ser usados, conjuntamente con un modelo de lenguaje destino $p(D = "d_1 d_2")$, para determinar cuál es la traducción más adecuada. En la tabla #3, se presentan los valores del modelo de traducción $p(F|D)$ entrenado en este ejercicio, así como los valores de un modelo de lenguaje destino hipotético $p(D)$, para todas las oraciones destino posibles D . De la tabla #3 se desprende en forma evidente que la mejor traducción es $D_0 = \text{"blue car"}$.

Tabla #3: Valores del modelo de traducción entrenado y lenguaje destino hipotético para todas las posibles oraciones destino $D = "d_1 d_2"$.

$D = "d_1 d_2"$	$p(F D)$	$p(D)$	$D = "d_1 d_2"$	$p(F D)$	$p(D)$
red dress	0	>0	blue red	0	≈0
red car	0	>0	blue dress	0	>0
red blue	0	≈0	blue car	1	>0
dress red	0	≈0	car red	0	≈0
dress car	0	≈0	car dress	0	≈0
dress blue	0	≈0	car blue	1	≈0

Estado actual del arte

Como se vio en la sección previa a la anterior, los primeros sistemas estadísticos de traducción automática estaban basados en el modelo del canal ruidoso, en el cual la probabilidad de una traducción quedaba expresada en términos del producto de dos modelos: el modelo de lenguaje destino y el modelo de traducción. En los sistemas actuales, el enfoque del canal ruidoso ha sido reemplazado por un enfoque más general que se fundamenta en los principios de máxima entropía²¹. En este nuevo enfoque, la probabilidad de traducción queda expresada en términos de una combinación log-lineal de características. De esta forma, la mejor hipótesis de traducción D , para una oración fuente dada F , estará determinada por la oración destino que maximiza una combinación log-lineal de modelos de probabilidades, tal y como se presenta en la siguiente expresión:

$$\underset{D}{\operatorname{argmax}} p(D|F) \approx \underset{D}{\operatorname{argmax}} \prod_i p_i(F,D)^{\lambda_i}$$

Ecuación #9

donde los λ_i se denominan los pesos o exponentes de la combinación log-lineal, y las funciones $p_i(F,D)$ se denominan características o "features", y corresponden a los distintos modelos estadísticos involucrados.

Nótese, que el modelo del canal ruidoso, puede interpretarse como un caso particular del modelo de combinación log-lineal de características. En este caso particular, las características o "features" utilizados son el modelo de traducción y el modelo de lenguaje destino, y los λ_i correspondientes son iguales a 1 para ambas características.

Para el caso general descrito en la ecuación #9, los exponentes λ_i de la combinación log-lineal deben ser estimados o ajustados de alguna manera. El procedimiento más común consiste en resolver un problema de optimización, mediante el cual se busca un conjunto de valores para dichos exponentes de forma que cierta función objetivo sea minimizada o maximizada. Estas funciones objetivo se definen de acuerdo con alguna medida automática de calidad o error de traducción, y son evaluadas para un conjunto de datos de desarrollo, que en el caso ideal debe ser distinto al conjunto de datos con el que se entrenaron los modelos.

Adicionalmente a este cambio en la concepción del sistema de traducción, los sistemas recientes han reemplazado los modelos de traducción originales que se basaban en palabras por modelos de traducción basados en unidades bilingües compuestas por grupos de palabras. Estas unidades bilingües se conocen con el nombre genérico de frases y, consecuentemente, los sistemas de traducción que las usan se denominan sistemas de traducción basados en frases²². Estas unidades bilingües o frases son extraídas de los alineamientos más probables estimados para un corpus bilingüe y, aunque existen muchas variantes de los métodos de extracción de frases, como regla general se puede establecer que todas las palabras contenidas dentro de una frase sólo pueden estar enlazadas con otras palabras dentro de la misma frase. Nótese que según esta definición, cada par bilingüe de oraciones en un corpus constituye también una frase. Por esta razón, en los métodos de extracción de frases suele limitarse la máxima longitud (en palabras) que pueden tener las frases; de lo contrario, el número de frases puede ser tan elevado que se hace inmanejable, y sus respectivas probabilidades incalculables.

En la figura #6 se presenta un ejemplo ilustrativo de como se extraen frases, de longitud máxima tres, a partir de un par bilingüe de oraciones alineadas.

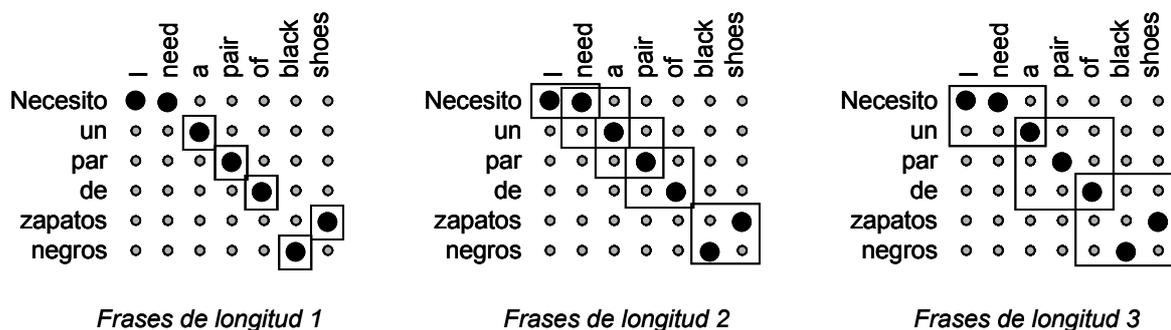


Figura #6: Extracción de frases a partir de un par bilingüe alineado de oraciones.

Una vez extraídas las frases, el modelo de traducción es entrenado mediante el conteo del número de ocurrencias de cada frase en la totalidad del corpus de entrenamiento. Las probabilidades de las frases (modelo de traducción) se calculan normalizando, bien sea en forma absoluta o relativa, los resultados de dicho conteo. En este momento es importante mencionar que las frases más cortas generalmente quedan favorecidas frente a las más largas, por el simple hecho de que aparecen con mayor frecuencia en el material de entrenamiento. Este problema es solventado por la mayor parte de los sistemas de traducción actuales mediante la implementación de alguna –o generalmente ambas– de las siguientes estrategias:

1. Modelos de penalización. Con el objeto de compensar la tendencia a bajar de las probabilidades a medida de que la longitud de las frases aumenta, se suele incluir un modelo de penalización basado en el número de palabras como una característica adicional en la combinación log-lineal. Este modelo, como su nombre lo indica, compensa las probabilidades de las frases penalizando más a las frases más cortas y menos a las frases más largas.

2. Decodificación por listas. Para lograr una justa competencia entre las frases candidatas a una traducción durante el proceso de decodificación, las hipótesis parciales²³ de traducción suelen agruparse en listas según su longitud. De esta manera, las probabilidades de las frases o secuencias de frases que se comparan entre sí en cada paso de la decodificación son, en promedio, del mismo orden de magnitud.

Aunque el paso de los modelos de traducción basados en palabras a los modelos de traducción basados en frases significó una mejora significativamente importante en el estado del arte en traducción automática estadística, los modelos de traducción basados en palabras no han sido abandonados por completo. De hecho, en el nuevo enfoque de combinación log-lineal de características, la mayoría de los sistemas de traducción actuales incluyen modelos léxicos (que no son otra cosa más que modelos de traducción basados en palabras) como características adicionales en la combinación log-lineal²⁴.

Hoy en día, el modelo de combinación log-lineal de características se ha convertido en un verdadero laboratorio de pruebas para la más variada cantidad de características. Los esfuerzos actuales se enfocan principalmente en la búsqueda de características, que basándose en información lingüística y conocimiento humano, sean capaces de incrementar la calidad de las traducciones generadas por los sistemas de traducción basados en frases.

Como un elemento adicional, y frecuente, en muchos de los sistemas de traducción actuales se puede mencionar el uso de técnicas de jerarquizado de hipótesis después de la decodificación. Esta técnica de post-proceso se conoce también como “reranking”, y consiste en realizar una nueva búsqueda sobre el espacio, mucho más reducido, de las mejores hipótesis de traducción identificadas durante la decodificación. Para implementar esta técnica, es entonces necesario que el decodificador sea capaz de generar una lista de las n mejores hipótesis de traducción encontradas, en lugar de una única “mejor” traducción²⁵. La utilización de esta técnica implica en el fondo un proceso de decodificación dividido en dos etapas: en la primera etapa, el decodificador convencional genera una lista de las n mejores hipótesis de traducción; y en la segunda etapa, esta lista es jerarquizada de nuevo mediante el empleo de información adicional o complementaria a la que fue usada en la primera etapa.

En el estado actual del arte en traducción automática estadística existen dos problemas fundamentales que aún no han sido resueltos, estos son: la morfología y el ordenamiento.

1. El problema de la morfología, aunque presente en todos los idiomas, afecta principalmente a aquellos idiomas morfológicamente ricos²⁶. El problema radica en que la riqueza morfológica incide directamente sobre el tamaño del vocabulario de un lenguaje. Cuando los vocabularios son muy grandes, la estimación de los modelos de traducción y de lenguaje se hace mucho más difícil debido a que los datos de entrenamiento se encuentran mucho más esparcidos en el espacio de datos.
2. El problema de ordenamiento es primordialmente un problema de sintaxis y, aunque está presente en todos los idiomas, afecta de forma muy importante la calidad de las traducciones entre lenguas gramaticalmente distantes²⁷. Como ejemplos típicos de esta problemática están los pares inglés-japonés e inglés-árabe. Lo que realmente hace al reordenamiento un problema muy difícil de abordar es que éste constituye un proceso muy costoso desde el punto de vista computacional.

A modo de ilustrar estos dos problemas, en la figura #7 se presenta una gráfica en la que se comparan las calidades de traducción hacia el inglés, para cuatro idiomas fuente diferentes: castellano, francés, alemán y finlandés. Estas comparaciones se hacen en función de los tamaños de vocabulario durante el entrenamiento²⁸. Nótese como las calidades de traducción son significativamente superiores para las traducciones castellano-inglés y francés-inglés, y

como la calidad se deteriora a medida de que aumenta el tamaño del vocabulario de entrenamiento. En el caso particular del finlandés, los problemas de traducción derivados de la morfología son muy importantes; mientras que en el caso del alemán, los problemas de ordenamiento tienen también un alto impacto.

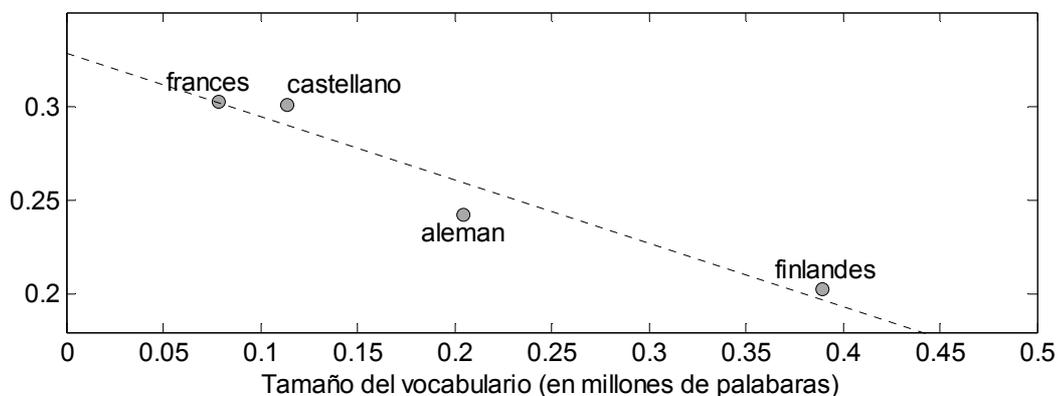


Figura #7: Calidad de traducción en función del tamaño de vocabulario en el entrenamiento.

Un ejemplo de traducción

En esta sección se presentan y discuten experimentos de traducción automática de las transcripciones de las sesiones plenarias del Parlamento Europeo en dos direcciones específicas: inglés-castellano y castellano-inglés. Esta experiencia se enmarca dentro del Consorcio Europeo TC-Star, y constituye la primera experiencia dentro de un proyecto Europeo en la cual se considera la traducción automática del habla utilizando datos de la vida real²⁹. La discusión se centra en los resultados obtenidos con el sistema de traducción basado en n-gramas bilingües de la Universidad Politécnica de Cataluña. Este sistema de traducción, a diferencia de los sistemas convencionales de frases que utilizan frecuencias relativas, utiliza un modelo de n-gramas de un conjunto particular de frases denominadas tuplas³⁰.

Los datos usados para el entrenamiento del sistema de traducción incluyen transcripciones de las sesiones desde abril de 1996 hasta septiembre de 2004. Y los datos de evaluación corresponden a las transcripciones de las sesiones entre el 15 y el 18 de noviembre de 2004. La tabla #4 presenta las estadísticas básicas de los datos de entrenamiento y evaluación utilizados. En la tabla, n° oraciones y n° palabras se refieren al número total de oraciones y palabras en cada conjunto de datos³¹, vocabulario se refiere al número de palabras diferentes, y longitud media al número promedio de palabras por oración.

Tabla #4: Estadísticas básicas de los datos disponibles.

Datos	Lenguaje	n° oraciones	n° palabras	Vocabulario	Longitud media
Entrenamiento	Inglés	$1,22 \times 10^6$	$33,4 \times 10^6$	105×10^3	27,3
	Castellano		$34,8 \times 10^6$	169×10^3	28,4
Evaluación	Inglés	1.094	$26,8 \times 10^3$	$3,9 \times 10^3$	24,5
	Castellano	840	$22,7 \times 10^3$	$4,0 \times 10^3$	27,0

Como se puede observar en la tabla #4, el número total de palabras en los datos de entrenamiento es muy similar para el castellano y el inglés, sin embargo, los vocabularios son muy distintos. De hecho, el vocabulario del castellano es un 60% superior al vocabulario del inglés. Esto encuentra su explicación en la naturaleza mucho más flexiva del castellano, la cual se hace evidente en la diversidad de formas existentes para los sustantivos, adjetivos y verbos,

dependientes del género, número, modo y tiempo. Como se verá más adelante en esta misma sección, esta diferencia entre los tamaños de vocabulario tiene consecuencias importantes en las calidades de traducción obtenidas para la dirección inglés-castellano.

También se puede observar en la tabla #4 que se utilizaron distintos conjuntos de datos para la evaluación; incluyendo el de castellano, un número ligeramente inferior (23%) de oraciones que el de inglés. Sin embargo, los vocabularios son bastante similares en este caso. Para ambos conjuntos de datos de evaluación se disponía de dos referencias de traducción, de forma que las medidas de evaluación fueron calculadas contra dos referencias.

El conjunto de datos de entrenamiento fue preprocesado mediante la aplicación de herramientas estándares para la delimitación de las palabras (“tokens”) y filtrado. En el proceso de filtrado, todos aquellos pares de oraciones con una relación de número de palabras superior a cierto valor fueron removidos de los datos. De la misma forma, todos aquellos pares de oraciones en los cuales al menos una de ellas excedía un número máximo de palabras fueron eliminados también. El objetivo de este preproceso es facilitar el trabajo al algoritmo de alineamiento automático.

El alineamiento automático de palabras fue realizado en ambas direcciones, castellano-inglés e inglés-castellano, mediante el uso de la herramienta GIZA++³². Posteriormente, la intersección y la unión de ambos alineamientos fueron calculados, para la subsiguiente extracción de las unidades bilingües (tuplas) y el entrenamiento de los n-gramas bilingües correspondientes al modelo de traducción utilizado³³.

Para el sistema de traducción utilizado se implementaron cuatro características adicionales al modelo de traducción basado en n-gramas bilingües. Estas características fueron las siguientes:

1. Modelo de lenguaje destino. Esta característica consiste en un modelo de lenguaje basado en 3-gramas de palabras, el cual asigna probabilidades a las hipótesis parciales de traducción dependiendo de la secuencia resultante de palabras destino. Este modelo se entrena utilizando la parte destino de los datos de entrenamiento. De esta forma se entrenó un modelo de lenguaje del castellano para la dirección inglés-castellano y un modelo de lenguaje del inglés para la dirección castellano-inglés.
2. Modelo de penalización de palabras. Este modelo introduce una penalización que depende de la longitud en palabras de la hipótesis de traducción. La idea de este modelo es compensar la preferencia del sistema de traducción por hipótesis de traducción con pocas palabras. Esta penalización se implementa de forma sencilla mediante el uso de un factor que depende directamente del número de palabras contenida en cada hipótesis parcial de traducción.
3. Modelo de léxico fuente a destino. Esta característica consiste en un modelo de traducción basado en palabras –comúnmente denominado también diccionario estadístico– que asigna una probabilidad a cada una de las unidades bilingües (tuplas) que intervienen en las hipótesis de traducción. Las probabilidades correspondientes a cada una de las unidades bilingües (tuplas) se calculan mediante el uso de los parámetros léxicos del modelo IBM-1.³⁴
4. Modelo de léxico destino a fuente. Esta característica es similar a la anterior, salvo que los parámetros léxicos del modelo IBM-1 son extraídos de los alineamientos de destino a fuente en lugar de ser extraídos de los alineamientos de fuente a destino.

Una vez calculados los modelos, los pesos óptimos para la combinación log-lineal fueron estimados para cada dirección de traducción y cada configuración de sistema considerada.

Con el objeto de evaluar las prestaciones del modelo de traducción y de las características utilizadas, cuatro configuraciones diferentes fueron utilizadas:

1. Sistema base. Sólo utiliza el modelo de traducción (3-grama de tuplas).
2. Sistema con refuerzo de destino. Utiliza el modelo de traducción conjuntamente con los modelos de lenguaje destino y de penalización de palabras.
3. Sistema con refuerzo de léxico. Utiliza el modelo de traducción conjuntamente con los dos modelos de léxico.
4. Sistema completo. Utiliza el modelo de traducción conjuntamente con las cuatro características descritas.

La estimación de los pesos se realizó mediante el uso de un algoritmo de optimización basado en el método simplex³⁵, mediante el cual se maximizó el valor del BLEU³⁶ de traducción sobre un conjunto de datos de desarrollo. En todas las optimizaciones realizadas se utilizó el mismo conjunto de datos de desarrollo. Estos datos consistieron en 500 pares bilingües de oraciones, extraídas de las transcripciones de las sesiones parlamentarias entre el 21 y el 28 de octubre de 2004, de forma que este conjunto de datos es independiente de los conjuntos de datos de entrenamiento y de evaluación. Para estos datos se disponía de un total de tres referencias de traducción para cada idioma.

Cada optimización implicó la traducción del conjunto de datos de desarrollo entre unas 20 y 100 veces, dependiendo de la configuración del sistema. Para la traducción se empleó el decodificador basado en n-gramas desarrollado específicamente para el sistema de traducción de la Universidad Politécnica de Cataluña.³⁷

Los valores de los pesos optimizados para cada uno de los sistemas en cada dirección se presentan en la tabla #5. El peso correspondiente al modelo de traducción (3-grama de tuplas) se omite debido a que su valor se mantuvo constante e igual a la unidad para todos los ejemplos presentados. Igualmente se omite la configuración base, ya que esta no incluye ninguna de las cuatro características adicionales.

Tabla #5: Pesos optimizados para las características utilizadas en cada configuración.

Sistema	Dirección	Leng. destino	Penalización	Léxico f-d	Léxico d-f
Destino	castellano-inglés	0,30	0,32	-	-
	inglés-castellano	0,32	0,26	-	-
Léxico	castellano-inglés	-	-	0,48	0,06
	inglés-castellano	-	-	0,17	0,07
Completo	castellano-inglés	0,48	0,28	0,48	0,13
	inglés-castellano	0,80	0,75	0,23	0,18

Finalmente, los conjuntos de datos de evaluación fueron traducidos mediante el uso de los modelos calculados y los respectivos pesos estimados para cada una de las configuraciones consideradas, y para cada una de las direcciones de traducción. Los resultados de todas estas traducciones se presentan en la figura #8, en términos de las medidas de evaluación WER³⁸ y BLEU. Como lo ponen en evidencia los resultados presentados en la figura #8, las traducciones correspondientes a la dirección castellano-inglés son consistente y significativamente mejores que las traducciones en la dirección inglés-castellano. Tal y como se mencionó anteriormente, esto se debe a la mayor talla del vocabulario castellano respecto al inglés en los datos de entrenamiento. Así por ejemplo, la palabra “the” en inglés puede generar cualquiera de las cuatro palabras “el”, “la”, “los” o “las” en castellano, existiendo así una mayor probabilidad de error para estas palabras cuando se traduce de inglés a castellano que cuando se traduce en la dirección contraria. Situaciones similares ocurren con los verbos y los adjetivos para los cuales existen una mayor variedad de formas en castellano que en inglés.

Los resultados de la figura #8 también demuestran que la inclusión de las cuatro características adicionales en el sistema de traducción (sistema completo) produce una importante mejora en la calidad de la traducción en ambas direcciones con respecto al uso del modelo de traducción en forma aislada (sistema base). También se puede comprobar que las características con mayor impacto en la calidad de traducción son los modelos léxicos. Otra observación importante, aunque más evidente para el caso de la dirección inglés-castellano que en la contraria, es que la contribución del modelo de lenguaje destino es más relevante cuando los modelos léxicos están presentes (sistema completo). De hecho, tal y como se puede observar en la tabla #5, el peso optimizado del modelo de lenguaje destino es mayor en la configuración correspondiente al sistema completo que en la configuración correspondiente al sistema con refuerzo de destino.

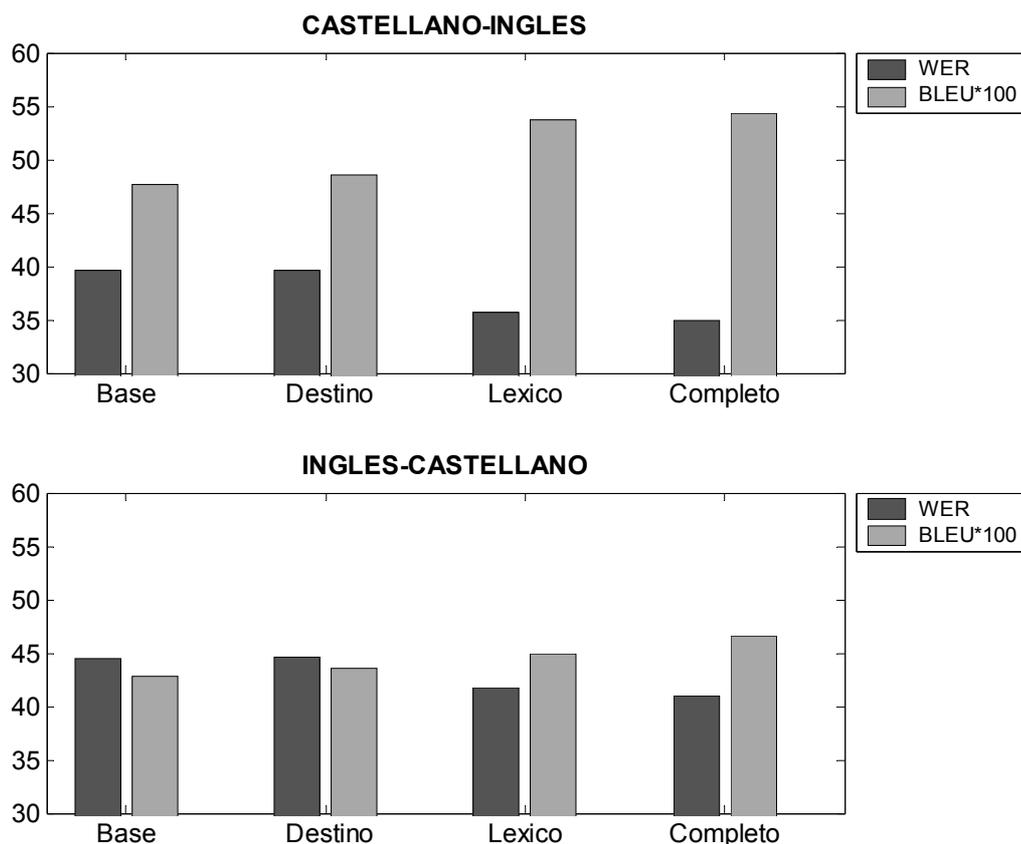


Figura #8: Medidas de evaluación obtenidas para las traducciones de los datos de evaluación.

Nótese de los resultados presentados, que los errores de traducción (WER) obtenidos están por el orden del 35% y el 40% para las direcciones castellano-inglés e inglés-castellano, respectivamente. Aunque las transcripciones del Parlamento Europeo pueden considerarse como un excelente ejemplo de datos reales de dominio amplio, y en consecuencia dichas tasas de error son bastante buenas en lo que respecta al estado actual del arte en traducción, estos resultados están todavía algo lejos de las aspiraciones generales en esta materia. Dichas tasas de error implican que, en promedio, de cada 10 palabras traducidas, entre 3 y 4 de ellas están mal traducidas o colocadas en el lugar incorrecto.

A continuación se presentan, a modo ilustrativo, algunos ejemplos de traducciones obtenidas con las configuraciones correspondientes al sistema completo.

Ejemplo #1:

Fuente: "Do you think that the expenditure on this sitting will be justified in terms of what comes out of it ?"

Destino: “¿ Cree usted que los gastos en esta sesión *null null* justificarse en términos de lo que salga de *null* ?”

Ejemplo #2:

Fuente: “During the last sitting , we adopted a resolution on the problem of human rights in Iran .”

Destino: “Durante la última sesión , hemos adoptado una resolución sobre el problema de los derechos humanos en Irán .”

Ejemplo #3:

Fuente: “They have been held captive ever since .”

Destino: “Se han que permanecen encarcelados desde .”

Ejemplo #4:

Fuente: “Declaro reanudado el período de sesiones del Parlamento Europeo , interrumpido el jueves 28 de octubre de 2004 .”

Destino: “I declare resumed the session of the European Parliament *null* adjourned *null* on Thursday , 28 *null* October *null* 2004 .”

Ejemplo #5:

Fuente: “La caída del muro de Berlín fue el último acontecimiento geopolítico derivado de la primera guerra mundial .”

Destino: “The fall of the Berlin Wall was the last event geopolitical resulting from the first world war .”

Ejemplo #6:

Fuente: “Aunque no lo parezca , están siempre en peligro .”

Destino: “Although I do not *null* appear , are always in danger .”

Evaluación de una plataforma

En esta sección se presentan los resultados de una prueba de aceptación para una plataforma experimental de traducción automática del habla, la cual permite la comunicación telefónica entre dos personas hablando dos lenguas distintas³⁹. Este experimento se enmarca dentro del Consorcio Europeo LC-Star y constituye una experiencia práctica de evaluación de un sistema de traducción automática del habla en tiempo real⁴⁰. En esta prueba, las tareas de reconocimiento y traducción estuvieron restringidas a un dominio turístico y a dos idiomas específicos: castellano y catalán. La prueba se dividió en tres evaluaciones independientes: una orientada a la comunicación, una orientada a una tarea, y una evaluación subjetiva.

1. La evaluación orientada a la comunicación se enfocó en la transmisión de los contenidos de información a lo largo de las etapas del sistema, con el objeto de evaluar hasta que punto el uso del lenguaje hablado ayuda o entorpece la comunicación.
2. La evaluación orientada a una tarea se enfocó en la realización de una tarea específica, con el fin de evaluar si dos usuarios son capaces de completar una tarea dado el estado actual de arte.
3. La última evaluación, consistió en una encuesta subjetiva, mediante la cual se recogieron las impresiones de los usuarios con respecto a la tecnología.

En todos los experimentos se contó con la participación de un total de 12 voluntarios, quienes recibieron un pequeño entrenamiento previo sobre el uso de la plataforma.

Con el objeto de poder ponderar el impacto relativo de cada una de las tres etapas – reconocimiento, traducción y síntesis– dentro del proceso completo de la traducción automática del habla, en la primera evaluación, se compararon los siguientes cuatro sistemas⁴¹:

1. sistema TT: entrada de texto, salida de texto;
2. sistema TV: entrada de texto, salida de voz;
3. sistema VT: entrada de voz, salida de texto; y
4. sistema VV: entrada de voz, salida de voz.

Para la generación de las entradas, se consideraron los siguientes cuatro escenarios:

1. escenario A: cliente haciendo una reserva en un hotel,
2. escenario B: cliente solicitando un servicio del restaurante,
3. escenario C: cliente reservando un billete de avión, y
4. escenario D: operador atendiendo una reserva de hotel.

A cada voluntario se le asignaron dos escenarios diferentes y se le solicitó producir unas 3 ó 5 oraciones relacionadas con cada uno de los escenarios mientras sus voces eran grabadas por el sistema. Inmediatamente después, a cada voluntario se le solicitó escribir 3 ó 5 oraciones para los mismos escenarios y manteniendo el mismo significado de lo que habían grabado. De esta forma se obtuvieron un total de 48 entradas: 24 de voz y 24 de texto, para cada una de las cuales se generaron salidas de voz y de texto. En definitiva, 4 versiones: TT, TV, VT y VV, de cada uno de los 24 casos considerados fueron generadas, i.e. un total de 96 salidas diferentes.

Las salidas generadas fueron escuchadas y/o leídas por los mismos 12 voluntarios, a quienes se les pidió que reportaran por escrito lo que habían entendido en cada caso. A cada voluntario se le asignaron 16 salidas, de forma que cada salida fue evaluada dos veces. Durante la asignación de las salidas que debía escuchar o leer cada voluntario se tuvo cuidado de evitar que algún voluntario evaluara salidas correspondientes a las entradas generadas por él o ella misma, y de evitar que algún voluntario evaluara más de una versión de un mismo caso.

Una vez que los 192 reportes escritos estuvieron listos, estos fueron comparados con sus entradas correspondientes por tres jueces. Cada juez asignó una puntuación a cada reporte en forma independiente y luego dichas puntuaciones fueron promediadas. El criterio de puntuación se basó en la siguiente escala subjetiva de 7 puntos:⁴²

- 6: claramente útil para comunicar la intención
- 5: útil en sentido general
- 4: útil pero menos informativo
- 3: útil pero no adecuado del todo
- 2: casi inútil pero aún algo informativo
- 1: claramente inútil
- 0: nada o basura

En la figura #9 se presentan los resultados de esta evaluación, desglosados por escenario, para cada uno de los cuatro sistemas considerados. De la figura se derivan dos observaciones importantes. En primer lugar, se evidencia que el deterioro más importante del rendimiento de la plataforma de traducción evaluada se debe al sistema de reconocimiento. Nótese como para todos los escenarios considerados, las puntuaciones más bajas se obtuvieron para los dos sistemas con entradas de voz. En segundo lugar, se puede observar que el rendimiento del sistema es prácticamente similar en todos los escenarios, exceptuando el escenario B. La diferencia entre este escenario y los demás, se hace aún más evidente en aquellos casos en los que el sistema de reconocimiento está presente.

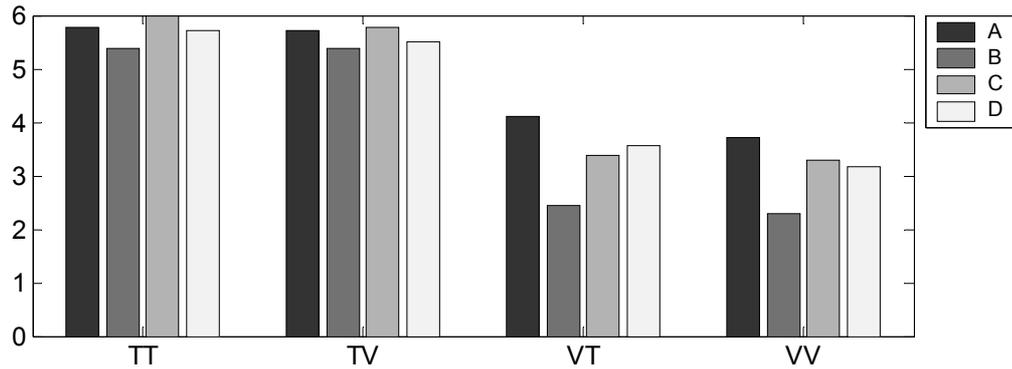


Figura #9: Evaluación subjetiva, por escenario, de los cuatro sistemas considerados.

La segunda evaluación, como ya se mencionó, se enfocó en la realización de una tarea específica, con el fin de evaluar si dos usuarios eran capaces de completar una tarea dado el estado actual de arte. La tarea se definió tomando los escenarios A y D de la primera evaluación, los cuales corresponden a la gestión de una reserva de hotel por parte de un usuario y un operador. La tarea consistió en la consecución de los siguientes ocho objetivos específicos:

1. fecha de llegada
2. número de noches de la estancia
3. tipo de habitación requerida
4. costo por noche del tipo habitación requerida
5. nombre completo del cliente
6. tipo de tarjeta de crédito a ser usada para hacer la reserva
7. número de la tarjeta de crédito, y
8. fecha de caducidad de la tarjeta de crédito.

Los mismos doce voluntarios de la primera evaluación fueron agrupados en parejas, de forma que se realizaron un total de 6 diálogos. A cada participante se le suministraron instrucciones escritas con la información requerida por su rol, y se le solicitó que escribiera durante la conversación lo que fuera logrando en cada uno de los ocho objetivos propuestos. Todas las conversaciones fueron grabadas y transcritas para un posterior análisis de las conversaciones y los logros reportados por escrito. De esta forma, se pudo determinar para cada conversación cuales de los objetivos habían sido realmente logrados y en cuantos intentos de comunicación.

La evaluación de los objetivos en cada dialogo se llevó a cabo utilizando una puntuación ponderada en función de la importancia que los usuarios le asignan a cada objetivo en el momento de la conversación⁴³. De acuerdo con este esquema, la puntuación es una función del éxito en la consecución del objetivo y del número de intentos necesarios para conseguirlo. Así, la puntuación de un objetivo logrado se calcula como el inverso del número de intentos, de forma que un objetivo logrado al primer intento obtiene una puntuación de 1, mientras que un objetivo logrado después de muchos intentos obtiene una puntuación que se aproxima a cero. En forma similar la puntuación de un objetivo no logrado se calcula restando la unidad al inverso del número de intentos. De esta forma un objetivo no logrado y abandonado después del primer intento obtiene una puntuación de cero, mientras que un objetivo no logrado después de muchos intentos obtiene una puntuación de -1 .

Nuevamente, tres jueces evaluaron en forma independiente los ocho objetivos propuestos para cada uno de los seis diálogos realizados. Las puntuaciones totales para cada uno de los objetivos y los diálogos se obtuvieron promediando las puntuaciones de los tres jueces. Estas puntuaciones se presentan en la figura #10.

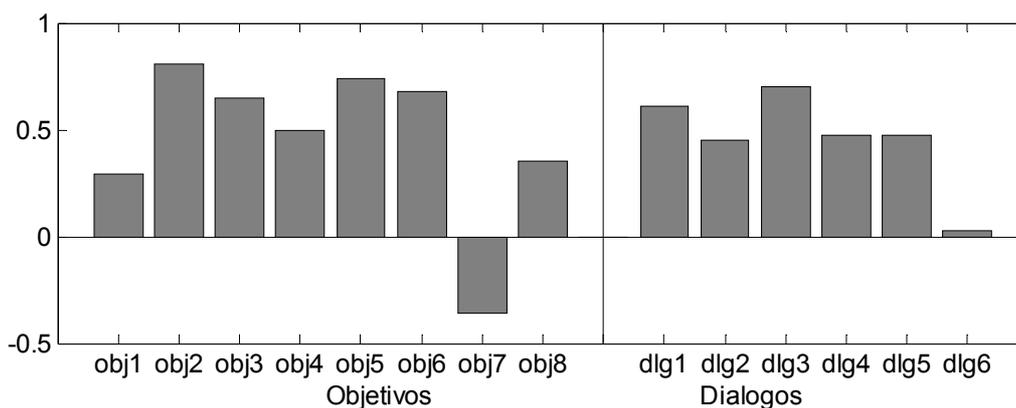


Figura #10: Puntuaciones totales correspondientes a cada objetivo y a cada diálogo.

Como se puede observar en la figura #10, en términos generales, la mayoría de los objetivos fueron logrados en la mayoría de los diálogos. Sin embargo, un análisis más detallado de las puntuaciones de los objetivos dentro de cada uno de los diálogos, revela que en ninguno de los diálogos se logró la totalidad de los objetivos. El objetivo #7 “número de la tarjeta de crédito”, que como se observa en la figura #10 resultó tener la peor puntuación, no pudo ser completado en ninguno de los seis diálogos.

Para el caso particular del objetivo #7, un posterior estudio de las transcripciones permitió identificar un problema básico. En la mayoría de los casos, cuando los clientes decían los 16 dígitos del número de la tarjeta de crédito, hacían pausas entre cada grupo de 4 números, las cuales eran interpretadas por el sistema de reconocimiento del habla como un fin de oración. De esta forma, el operador recibía la información incompleta y solicitaba al cliente que repitiera los números, lo cual, aunado a los tiempos de espera del proceso completo de reconocimiento-traducción-síntesis, generaba una gran confusión. Después de varios intentos, los operadores optaban por anotar lo que habían entendido y proseguir con la siguiente pregunta.

En la figura #10 también se observa que los objetivos consistentemente mejor logrados fueron los correspondientes a “número de noches de la estancia” (2), “nombre completo del cliente” (5), “tipo de tarjeta de crédito a ser usada para hacer la reserva” (6), y “tipo de habitación requerida” (3). En cuanto a los diálogos, con excepción del diálogo #6, siempre se lograron la mayor parte de los objetivos propuestos. Aunque, en términos globales, los resultados de este ejercicio producen una muy buena impresión sobre el estado del arte en traducción automática del habla, desde el punto de vista práctico, es evidente que el estado actual del arte aún no alcanza el nivel de desempeño deseado. Para los efectos reales, ninguna de las seis reservas de hotel intentadas se pudo completar exitosamente.

A pesar de estos resultados, la tercera y última evaluación, mediante la cual se recogieron las impresiones de los usuarios con respecto a la tecnología utilizada, reveló que los usuarios son optimistas en cuanto al uso de los sistemas de traducción automática del habla. Los 12 voluntarios que participaron en las pruebas dieron una puntuación promedio de 3,67 (en una escala de 1: “totalmente en desacuerdo” a 5: “totalmente de acuerdo”) a la pregunta sobre si consideraban que la tecnología era muy útil. Las preguntas sobre si la velocidad del sistema era adecuada y si nunca se habían perdido en el diálogo recibieron puntuaciones de 2,75 y 2,92, respectivamente.

Adicionalmente en la mayoría de escenarios hipotéticos planteados en esta última evaluación, para los cuales se les ofrecía la posibilidad de interactuar con el sistema mediante el uso de la voz, un teclado de ordenador o el teclado de un móvil, la mayoría de los usuarios prefirieron la interacción hablada.

Los retos de la traducción del habla

Tal y como se puede concluir de las experiencias descritas en la sección anterior, el estado actual del arte en traducción automática del habla está todavía algo lejos de las aspiraciones generales en esta materia, y plantea algunos retos fundamentales que deben ser afrontados y resueltos en los próximos años. En particular, traducción automática del habla involucra tres procesos sumamente complicados en sí mismos: el reconocimiento automático del habla, la traducción automática y la síntesis de voz. Los problemas relacionados con cada uno de estos procesos inciden en forma directa sobre el desempeño del sistema en su totalidad. Esta sección describe brevemente algunos de estos problemas.

Como se pudo observar en la evaluación de la plataforma de traducción presentada en la sección anterior, uno de los principales cuellos de botella en el proceso de traducción lo constituye el sistema de reconocimiento del habla. Las tasas de error en reconocimiento son todavía del orden del 10%⁴⁴, lo cual tiene una incidencia bastante importante sobre la calidad de las traducciones obtenidas.

Así mismo, la naturaleza irregular, desde el punto de vista gramatical, característica del lenguaje hablado constituye otro factor de suma importancia que dificulta aún más la tarea de traducción automática. Los problemas conocidos como efectos del habla espontánea, entre los cuales están la repetición de palabras, la producción de palabras entrecortadas, el uso de muletillas y el cambio brusco de ideas al hablar, constituyen retos importantes a ser resueltos tanto desde la perspectiva del reconocimiento como la de la traducción.

La figura #11 presenta una comparación interesante del desempeño de tres sistemas de traducción independientes para tres condiciones diferentes de los datos del Parlamento Europeo⁴⁵: edición final, verbatim y reconocimiento del habla. La condición de edición final se refiere a transcripciones "limpias" en donde los efectos del habla espontánea han sido removidos o corregidos. La condición verbatim se refiere a transcripciones literales. De forma que esta condición permite evaluar los efectos del habla espontánea en la traducción sin incluir los errores de reconocimiento del habla. Finalmente, la condición de reconocimiento del habla se refiere al uso directo de la salida del sistema de reconocimiento como entrada del sistema de traducción. En este caso, tanto los efectos del habla espontánea como los errores de reconocimiento están presentes.

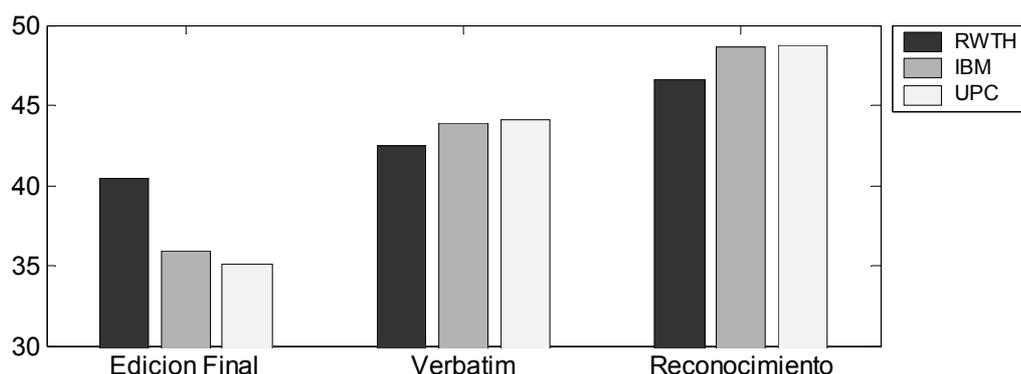


Figura #11: Tasa de error (WER) en traducciones castellano-inglés con tres sistemas de traducción independientes para las tres condiciones de entrada mencionadas.

Como puede observarse en la figura #11, la calidad de la traducción se ve significativamente deteriorada por los efectos del habla espontánea (paso de la condición de edición final a la de verbatim) y por los errores del sistema de reconocimiento del habla (paso de la condición de verbatim a la de reconocimiento).

Otro reto interesante en la traducción automática del habla es la incorporación de información lingüística en el proceso de traducción. En el área de procesamiento del lenguaje natural, se han desarrollado una gran cantidad de herramientas de análisis y recursos que proveen información lingüística que puede ser muy valiosa para el proceso de traducción automática. Sin embargo, la mayor parte de los esfuerzos realizados hasta el momento han sido bastante frustrantes en el sentido de que al tratar de incorporar información lingüística en el proceso de traducción, esta pareciera tener poco o ningún impacto en la calidad de las traducciones obtenidas⁴⁶. Este problema tiene dos aristas. La primera tiene que ver con la manera de incorporar la información lingüística en el proceso de traducción. Esta arista constituye una pregunta abierta y, hoy por hoy, gran cantidad de los esfuerzos de investigación en el área de traducción automática están enfocados en evaluar métodos innovadores para la incorporación efectiva de la información lingüística en la traducción.

La segunda arista se centra en el problema de las medidas de evaluación automática, ya que en la mayoría de los casos estas medidas no son capaces de cuantificar adecuadamente pequeñas mejoras en la calidad de traducción. Las medidas de evaluación automáticas constituyen en realidad una buena medida promedio de la proximidad de un conjunto de traducciones a unas referencias dadas. Sin embargo, como medidas de la calidad de traducción propiamente dicha para oraciones individuales presentan problemas muy graves. Esta problemática se puede resumir en dos puntos específicos:

1. En primer lugar, dada una oración fuente, existen muchas oraciones destino que constituyen traducciones correctas de dicha oración fuente. Si esta información no está contenida en las referencias, es posible que una buena traducción obtenga una mala evaluación. Así por ejemplo, el siguiente ejemplo de traducción extraído de los datos del Parlamento Europeo constituye una traducción totalmente correcta:
Fuente: "Mr President , I want to speak on two matters ."
Destino: "Señor Presidente , quiero *null* hablar sobre dos cuestiones ."
Pero si la referencia fuese "Sr. Presidente yo quisiera referirme a dos asuntos." dicha traducción quedaría evaluada como una traducción regular o pobre.
2. En segundo lugar, no todas las palabras tienen el mismo peso semántico e informativo dentro de una oración. Así por ejemplo, la siguiente traducción, también extraída de los datos del Parlamento Europeo, podría quedar evaluada como una traducción bastante buena siendo realmente una pésima traducción:
Fuente: "La política de la Unión Europea respecto a Cuba no debe cambiar ."
Destino: "The policy of the European Union on *null* Cuba *null* must change ."

Por otra parte, las medidas de evaluación humanas son mucho más confiables que las automáticas, pero a su vez, la evaluación humana es muy costosa, muy lenta y muy subjetiva. En este sentido, uno de los retos, quizá entre los más importantes, de la investigación en traducción automática del habla es el desarrollo de medidas de evaluación confiables, eficientes y robustas que permitan cuantificar la calidad de traducción, tanto en el caso de oraciones aisladas, como en el caso de grandes volúmenes de datos.

Finalmente, otro de los grandes retos de la traducción automática del habla está en el proceso de la síntesis del habla. En este caso, se trata del problema de la prosodia y la expresividad en la voz sintética, lo cual permite una transmisión más adecuada del mensaje mediante la generación de las entonaciones correctas durante el proceso de síntesis⁴⁷. En el contexto de la traducción automática del habla, este problema se ve ampliado al considerarse la adecuada identificación durante el reconocimiento de todos los parámetros para-lingüísticos necesarios para codificar la información de prosodia y expresividad, su posterior transmisión a través del sistema de traducción, y su final empleo en el proceso de generación de la voz sintética para lograr las entonaciones adecuadas.

Conclusiones y perspectivas futuras

Como conclusión de todo lo expuesto a lo largo de las secciones de este trabajo, se puede decir que si bien se han producido avances importantes en el área de la traducción automática del habla a lo largo de los últimos diez años, el estado actual del arte está todavía un poco distante de las aspiraciones generales en la materia. En los ejemplos referidos a lo largo de este trabajo, se evidencian tasas de error en traducción que oscilan entre un 30% y un 50%, según las distintas condiciones, sistemas y direcciones de traducción consideradas. Estas tasas de error son aún, para efectos prácticos, bastante altas.

No obstante, el futuro en la traducción automática del habla es bastante promisorio, y aunque el camino más adecuado para un incremento significativo del rendimiento de los sistemas de traducción actuales no está del todo claro, una gran cantidad de esfuerzos y recursos se están invirtiendo actualmente en investigación y desarrollo en el área de la traducción automática del habla. Entre las principales líneas de investigación que se vislumbran para el futuro inmediato, se pueden mencionar las siguientes:

1. Incorporación de conocimiento lingüístico adecuado, y en una forma adecuada, en el proceso de traducción, con el objeto de solventar problemas específicos y muy frecuentes, tales como: errores de concordancia de género y número, traducciones incorrectas de formas verbales compuestas, y omisiones y/o repeticiones frecuentes de palabras o grupos de palabras.
2. Desarrollo de estrategias eficientes para abordar el problema de ordenamiento de las palabras y/o grupos de palabras en las hipótesis de traducción. El ordenamiento constituye un proceso muy intensivo desde el punto de vista computacional, y de fundamental importancia para la tarea de traducción automática entre algunos pares de idiomas, como por ejemplo: inglés-japonés e inglés-árabe.
3. Desarrollo de nuevos modelos y estrategias que permitan manejar de forma adecuada todos aquellos efectos del habla espontánea que deterioran el desempeño de los sistemas actuales de traducción automática.
4. Integración de los sistemas de reconocimiento del habla y de traducción automática, con el fin de lograr que en conjunto se comporten de una forma más robusta ante los problemas derivados de los errores de reconocimiento.
5. Desarrollo de mecanismos que permitan, al sistema de traducción del habla, identificar y transmitir toda aquella información para-lingüística relevante para una síntesis de voz que se ajuste de la mejor manera posible a los contenidos prosódicos y emocionales del locutor original.
6. Desarrollo de medidas automáticas de evaluación que se aproximen de la mejor manera posible a las medidas de evaluación humanas, en el sentido de ser más confiables, eficientes y robustas, y que permitan cuantificar la calidad de las traducciones, tanto en el caso de oraciones aisladas, como en el caso de grandes volúmenes de datos.

Por lo demás, es tan sólo cuestión de tiempo para que los sistemas de traducción del habla sean una realidad palpable en la vida cotidiana, tal y como lo son ahora el teléfono móvil y la Internet. Para muchos, esto constituirá un gran avance tecnológico en la historia de la humanidad; pero en el fondo sólo se trata de un pequeño salto, tres versículos hacia atrás, en la historia con la que comenzó este documento: "Entonces, el mundo entero tenía un solo idioma y una lengua común..." (Génesis 11:1)

Notas de pie de página

¹ Para una excelente taxonomía de las lenguas y valiosa información adicional sobre los orígenes de los alfabetos, se recomienda visitar el Museo del Alfabeto, disponible en línea a través de la dirección electrónica <http://www.jaars.org/museum/alphabet/index.htm>

² Para un recuento histórico bastante completo del problema de traducción automática, véase el trabajo de Hutchins (1986).

³ Kelley (1979) hace alusión a San Jerónimo como uno de los pioneros en la búsqueda de modelos de traducción. San Jerónimo, quien fue un reconocido políglota que hablaba perfectamente varios idiomas, tradujo del hebreo al latín toda la Santa Biblia.

⁴ Muchas de las ideas que dieron forma a la teoría de la información (Shannon y Weaver, 1949) se inspiraron en trabajos sobre criptografía durante la Segunda Guerra Mundial (Shannon, 1949 y 1951). Estas ideas a su vez inspiraron la visión del problema de traducción como un problema de decodificación (Weaver, 1955).

⁵ Se refiere a una representación semántica, o de significados, en una forma canónica independiente del lenguaje. Para mayor información sobre representaciones semánticas véase el capítulo 14 del libro de texto de Jurafsky y Martín (2000).

⁶ O según se supone que lo hacen, ya que realmente el proceso de traducción humano es extremadamente complejo. Para una revisión de estudios empíricos sobre traducción simultánea, consúltese el trabajo de Gerver (1976); y para un interesante estudio sobre un modelo cognitivo del proceso de interpretación, consúltese el trabajo de Padilla y Bajo (1998).

⁷ Como ejemplos interesantes sobre sistemas de traducción basados en conocimiento, y de las expectativas sobre la aplicación de la inteligencia artificial al problema de traducción automática, se pueden mencionar los trabajos de Carbonell *et. al.* (1981), Tsujii (1986) y Nirenburg *et. al.* (1992), entre otros.

⁸ En este sentido, las dos publicaciones de mayor relevancia son "A statistical approach to machine translation" y "The mathematics of statistical machine translation: parameter estimation" de Brown *et al.* en 1990 y 1993, respectivamente.

⁹ Aunque los temas de reconocimiento y síntesis del habla constituyen elementos intrínsecos al problema de la traducción automática del habla, este trabajo se enfoca únicamente en el aspecto específico de la traducción. Para un tratamiento amplio y detallado sobre el tema de reconocimiento del habla véanse los libros de Rabiner y Juang (1993) y Jelinek (1997); y sobre el tema de síntesis del habla, los libros de Allen *et al.* (1987) y Sproat (1998).

¹⁰ Entre los primeros trabajos que abordan el problema de la traducción automática del lenguaje hablado destaca el de Kay *et al.* (1992).

¹¹ Más información sobre cada uno de estos proyectos puede ser encontrada e sus respectivas direcciones electrónicas:

- *C-Star*: <http://www.c-star.org>
- *Eutrans*: <http://www.cordis.lu/espirt/src/30268.htm>
- *Verbmobil*: <http://verbmobil.dfki.de/verbmobil/overview-us.html>
- *LC-Star*: <http://www.lc-star.com>
- *Nespole!*: <http://nespole.itc.it>
- *Fame*: <http://isl.ira.uka.de/fame/orga.html>
- *TC-Star*: <http://www.tc-star.org>

¹² La dificultad de la traducción automática puede verse desde muchos puntos de vista diferentes. Aunque algunos expertos opinan que sólo la combinación de la fuerza-bruta (métodos estadísticos) y la información lingüística (conocimiento) podrá llevar a los sistemas de traducción automática a un feliz término, la mayoría de los esfuerzos realizados en esta dirección aún no han logrado mejoras

significativamente importantes sobre los sistemas actuales (Knight, 1997) y (Och *et al.* 2003). En cuanto a la complejidad computacional, Knight (1999a) muestra que, aún cuando se consideran los modelos estadísticos más simples, el problema de la traducción estadística es del tipo NP-completo.

¹³ Estos datos están disponibles en la Internet a través de la dirección electrónica del Parlamento Europeo <http://www.europarl.eu.int/>. En el caso particular de los resultados que se presentan en este trabajo, se usó la versión de estos datos preparada por la Universidad de Aachen RWTH en el marco del consorcio europeo TC-STAR (Technology and Corpora for Speech-to-Speech Translation).

¹⁴ Se trata de la plataforma experimental GAIA (Pérez y Bonafonte, 2004) desarrollada en la Universidad Politécnica de Cataluña en el marco del consorcio europeo LC-STAR (Léxica and Corpora for Speech-to-Speech Translation Components).

¹⁵ En esta sección se presentará una muy breve descripción de algunos de los enfoques más comúnmente utilizados para abordar el problema de la traducción automática. Para una descripción más detallada de estos y otros enfoques, se recomienda revisar los capítulos 21 y 13 de los libros de texto de Jurafsky y Martín (2000), y Manning y Schütze (1999), respectivamente.

¹⁶ Los métodos de búsqueda exhaustivos raramente se usan en traducción automática debido a su baja eficiencia computacional. Para una revisión más completa de sobre algunos de los algoritmos de búsqueda más comúnmente usados en el problema de traducción, véanse los trabajos de Wang y Waibel (1997), Tillmann *et al.* (1997), Koehn (2004), y Crego *et al.* (2005).

¹⁷ Para una descripción mucho más completa sobre la estimación de modelos de lenguaje basados en n-gramas, así como de las respectivas técnicas de interpolación y suavizado, se recomienda revisar el capítulo 6 del libro de texto de Jurafsky y Martín (2000), y el capítulo 6 del libro de texto de Manning y Schütze (1999). Para una descripción de la herramienta para entrenar modelos de lenguaje SRILM, véase Stolcke (2002). SRILM es una herramienta de libre distribución y está disponible a través de la página electrónica <http://www.speech.sri.com/projects/srilm>

¹⁸ La descripción matemática formal y completa de estos cinco modelos se encuentra en el trabajo de Brown *et al.* (1993), sin embargo, para un enfoque mucho más digerible sobre estos modelos, se recomienda revisar el tutorial de Knight (1999b).

¹⁹ Las siglas EM se refieren a “Estimation-Maximization”. Como su nombre lo indica, este algoritmo consiste en un proceso iterativo que maximiza las probabilidades entre las palabras fuente y destino, basándose en las co-ocurrencias de palabras existentes en los datos. Este algoritmo fue desarrollado por Baum (1972).

²⁰ El ejemplo aquí presentado se inspira en un ejemplo similar presentado por Knight (1999b).

²¹ Para una interesante introducción al uso de la teoría de la información en el procesado del lenguaje natural, véase el trabajo de Berger *et al.* (1996); y para una descripción más detallada del uso de modelos de máxima entropía en traducción automática específicamente, véase el artículo de Och y Ney (2002).

²² Los sistemas de traducción basados en frases, son considerados actualmente como el estado del arte en traducción automática estadística. Para una descripción detallada de los sistemas basados en frases, véanse Zens *et al.* (2002) y Koehn *et al.* (2003).

²³ Una hipótesis parcial es el resultado de una determinada secuencia de unidades de traducción. Como ya se mencionó, la búsqueda en traducción automática nunca se realiza en forma exhaustiva. El espacio de soluciones se explora mediante la construcción progresiva de traducciones parciales (hipótesis parciales) basándose en la heurística de que buenas traducciones parciales conducen a buenas traducciones completas. Para más detalle sobre la decodificación usando listas, véanse Koehn (2004), y Crego *et al.* (2005).

²⁴ Para más detalles sobre la utilización de modelos léxicos como características en el modelo de combinación log-lineal, véanse Och *et al.* (2003 y 2004) y Mariño *et al.* (2005).

²⁵ Estos métodos de generación de listas de los n mejores candidatos de una búsqueda también son una práctica muy común en los sistemas de reconocimiento del habla. Para mayor información sobre estos algoritmos de búsqueda, véase Eppstein (1994); y para mayor información sobre la aplicación de estos métodos en el problema de traducción automática, véanse *Ueffing et al.* (2002) y Koehn (2004).

²⁶ Para más información sobre las implicaciones de la morfología en el problema de la traducción automática y algunas ideas sobre como tratar el problema de la morfología, véanse los trabajos de Koehn (2002), y Nießen y Ney (2001).

²⁷ Como ejemplos de distintos esquemas de ordenamiento se recomienda revisar los modelos presentados en Brown *et al.* (1993), Och *et al.* (1999), Yamada y Knight (2001), y Zens y Ney (2003).

²⁸ Para una descripción más detallada de estos resultados, véase Banchs *et al.* (2005). Estos resultados fueron obtenidos con la versión de los datos del Parlamento Europeo facilitados para el workshop de la ACL 2005 sobre construcción y uso de textos paralelos. Dichos datos están disponibles a través de la página electrónica del workshop <http://www.statmt.org/wpt05/mt-shared-task/>

²⁹ A diferencia de otros proyectos de traducción automática del habla, en los cuales tradicionalmente se han usado datos generados en los mismos laboratorios de investigación. Para una descripción más completa de estos resultados experimentales, véase el reporte de progreso D5 (Ney *et al.*, 2005) del consorcio TC-STAR (Technology and Corpora for Speech-to-Speech Translation).

³⁰ Para una descripción más detallada de este sistema de traducción, véase el trabajo de Mariño *et al.* (2005).

³¹ Nótese que para los datos de entrenamiento el número de oraciones es el mismo tanto para el inglés como para el castellano. Esto se debe a que el conjunto de datos de entrenamiento está alineado por oraciones, i.e. cada una de las oraciones del conjunto de datos del castellano corresponde a la traducción de una de las oraciones del conjunto de datos del inglés y viceversa.

³² El GIZA++ es uno de los algoritmos más comúnmente usados para el alineamiento automático, palabra a palabra, de corpus bilingües paralelos. Esta herramienta fue desarrollada por Och y Ney (2000), y está disponible en <http://www.fjoch.com/GIZA++.html>

³³ Los conjuntos de tuplas fueron extraídos y los modelos de traducción correspondientes (3-gramas de tuplas) fueron entrenados para cada dirección de traducción, castellano-inglés e inglés castellano, de acuerdo con la metodología descrita en Mariño *et al.* (2005).

³⁴ Los parámetros del conocido modelo IBM-1 fueron definidos por Brown *et al.* (1993) como uno de los elementos fundamentales de los sistemas de traducción basados en palabras. Recientemente, este modelo ha sido utilizado como una característica complementaria en los sistemas de traducción basados en frases y ha demostrado tener un impacto importante en la calidad de las traducciones resultantes (Och *et al.*, 2003 y 2004) y (Mariño *et al.*, 2005).

³⁵ Este método es comúnmente usado en problemas en los cuales no se dispone de información de derivadas. El método simplex utilizado aquí se basa en el algoritmo descrito por Press *et al.* (2002).

³⁶ El BLEU es una de las medidas automáticas de calidad de traducción más frecuentemente utilizadas. Fue definida por Papineni *et al.* (2002), y se basa en el número de n -gramas presentes en la traducción a ser evaluada que coinciden con los n -gramas existentes en las referencias de traducción utilizadas. La escala del BLEU va desde cero, cuando la traducción no coincide con las referencias, hasta 1, para el caso en que la traducción coincide perfectamente con alguna de las referencias. Nótese que esta medida, más que cuantificar la calidad de una traducción como tal, representa una medición del parecido de una traducción a un conjunto de referencias.

³⁷ Este decodificador implementa un método de búsqueda en haz que considera simultáneamente el modelo de traducción y las cuatro funciones características descritas. Esta herramienta, denominada *MARIE*, fue desarrollada por Crego *et al.* (2005).

³⁸ El WER (word error rate), o tasa de error de palabras, es otra medida de calidad de traducción frecuentemente usada. Está basada en el porcentaje de reemplazos, omisiones y/o inserciones de palabras en la traducción obtenida respecto a la referencia para la cual el error sea menor. A diferencia del BLEU, donde 0 representa una pésima traducción y 1 representa la mejor traducción; la escala del WER es porcentual, siendo 0% el valor de WER para la mejor traducción y 100% el valor de WER para la peor.

³⁹ Esta plataforma experimental, desarrollada en el marco del proyecto LC-STAR (Léxica and Corpora for Speech-to-Speech Translation Components), constituye una plataforma distribuida que permite integrar distintos sistemas de reconocimiento del habla, de traducción automática y de síntesis de voz en una sola aplicación. Así mismo, se puede configurar para ser utilizada mediante interfaces de voz (teléfono) ó de texto (consola). Para una descripción detallada de esta plataforma, véase el informe de Pérez y Bonafonte (2004).

⁴⁰ Para una descripción más completa de estos resultados experimentales, véase el reporte de progreso D7.2 (Bonafonte *et al.*, 2005) del consorcio LC-STAR (Léxica and Corpora for Speech-to-Speech Translation Components).

⁴¹ Esta evaluación sigue la metodología presentada por Somers y Sugita (2003) en la evaluación de un sistema comercial de traducción automática del habla entre japonés e inglés.

⁴² Esta escala fue tomada de Somers y Sugita (2003).

⁴³ Esta metodología de evaluación de objetivos ponderados por importancia fue propuesta por Thomas (1999).

⁴⁴ Para una revisión bastante actualizada del estado del arte en reconocimiento automático del habla, se recomienda revisar el reporte de progreso D7 (Gauvain *et al.*, 2005) del consorcio TC-STAR (Technology and Corpora for Speech-to-Speech Translation).

⁴⁵ Estas tres condiciones de entrada son objeto de estudio en el consorcio TC-STAR. Los resultados experimentales para los datos del Parlamento Europeo que fueron presentados en la sección "Un ejemplo de traducción" corresponden a la condición de edición final. Para una descripción más completa de las tres condiciones y sus respectivos resultados experimentales, véase el reporte de progreso D5 (Ney *et al.*, 2005) del consorcio TC-STAR (Technology and Corpora for Speech-to-Speech Translation).

⁴⁶ En este sentido se recomienda revisar el informe final del workshop realizado el verano de 2003 en la Universidad Johns Hopkins (Och *et al.*, 2003). En este workshop se evaluó el impacto en traducción de una gran cantidad de modelos que incorporaban información sintáctica. Los resultados fueron poco alentadores; el mejor modelo resultó ser el popular modelo IBM-1 que había sido definido por Brown *et al.* (1993) diez años antes.

⁴⁷ Para una revisión bastante actualizada del estado del arte en síntesis del habla y con un énfasis especial en el problema de prosodia y expresividad, se recomienda revisar el reporte de progreso D9 (Bonafonte *et al.*, 2005) del consorcio TC-STAR (Technology and Corpora for Speech-to-Speech Translation).

Referencias bibliográficas

- Allen, J., Hunnicut, M., Klatt, D., 1987, *From text to speech: the MITalk system*, Cambridge University Press, Cambridge.
- Banchs, R., Crego, J.M., de Gispert, A., Lambert, P., Mariño, J.B., 2005, "Statistical machine translation of Euparl data by using bilingual n-grams", *Workshop on Data-Driven Machine Translation and Beyond (43th Annual Meeting of the Association for Computational Linguistics)*, Ann Arbor, MI.
- Baum, L.E., 1972, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process", *Inequalities*, 3: 1-8.
- Berger, A., Della Pietra, S., Della Pietra, V., 1996, "A maximum entropy approach to natural language processing", *Computational Linguistics*, 22(1): 39-71.
- Bonafonte, A., Pérez, J., Banchs, R., 2005, "Acceptance testing of the demonstrator", *Deliverable D7.2, LC-STAR*, European Community project ref. no. IST-2001-32216, disponible en: <http://www.lc-star.com/archive.htm>
- Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Sündermann, D., Ziegenhain, U., Adell, J., Agüero, P., Duxans, H., Erro, D., Nurminen, J., Pérez, J., Strecha, G., Umbert, M., Swang, X., 2005, "TTS Progress Report", *Deliverable no. D9, TC-STAR*, European Community project no. FP6-506738, disponible en: http://www.tc-star.org/pages/f_documents.htm
- Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., Roossin, P., 1990, "A statistical approach to machine translation", *Computational Linguistics*, 16(2): 79-85.
- Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R., 1993, "The mathematics of statistical machine translation: parameter estimation", *Computational Linguistics*, 19(2): 263-311.
- Carbonell, J., Cullingford, R., Gershman, A., 1981, "Steps toward knowledge-based machine translation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 3(4): 376-392.
- Crego, J.M., Mariño, J.B., de Gispert, A., 2005, "An n-gram based statistical machine translation decoder", a ser publicado en *Interspeech*, Lisboa, Portugal.
- Eppstein, D., 1994, "Finding the k shortest paths", *IEEE Proceedings of the 35th Symposium on Foundations of Computer Science*, :154-165.
- Gauvain, J.L, Lamel, L., Schwenk, H., Brugnara, F., Schlueter, R., Bisani, M., Stüker, S., Schaaf, T., Mohammed, S.I., Bacchiani, M., Westphal, M., Sivasdas, S., Kiss, I., Giron, F., 2005, "ASR Progress Report", *Deliverable no. D7, TC-STAR*, European Community project no. FP6-506738, disponible en: http://www.tc-star.org/pages/f_documents.htm
- Gerver, D., 1976, "Empirical studies of simultaneous interpretation: a review and a model", en Brislin, R. (editor), *Translation application and research*, Gardner Press, Nueva York.
- Hutchins, W., 1986, *Machine translation: past, present, future*, Ellis Horwood, Chichester, England.
- Jelinek, F., 1997, *Statistical methods for speech recognition*, MIT Press, Cambridge, MA.
- Jurafsky, D., Martin, J., 2000, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall, Upper Saddle River, NJ.
- Kay, M., Gawron, J., Norvig, P., 1992, *Verbmobil: a translation system for face-to-face dialog*, CSLI.
- Kelley, L., 1979, *The true interpreter: a history of translation theory and practice in the west*, St. Martin's Press, Nueva York.

-
- Knight, K., 1997, "Automating knowledge acquisition for machine translation", *AI Magazine*, 18(4).
 - Knight, K., 1999a, "Decoding complexity in word-replacement translation models", *Computational Linguistics*, 25(4): 607-615.
 - Knight, K., 1999b, *A statistical MT tutorial workbook*, disponible en: <http://www.isi.edu/natural-language/mt/wkbk.rtf>
 - Koehn, P., 2002, "Europarl: a multilingual corpus for evaluation of machine translation", artículo no publicado, disponible en: <http://people.csail.mit.edu/people/koehn/publications/europarl/>
 - Koehn, P., Och, F., Marcu, D., 2003, "Statistical phrase-based translation", *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL)*, Edmonton, Alberta.
 - Koehn, P., 2004, "Pharaoh: a beam search decoder for phrase-based SMT", *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.
 - Manning, C., Schütze, H., 1999, *Foundations of statistical natural language processing*, The MIT Press, Cambridge, MA.
 - Mariño, J.B., Banchs, R., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A., Ruiz, M., 2005, "Modelo estocástico de traducción basado en n-gramas de tuplas bilingües y combinación log-lineal de características", a ser publicado en la *Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*.
 - Ney, H., Steinbiss, V., Zens, R., Matusov, E., González, J., Lee, Y., Roukos, S., Federico, M., Kolss, M., Banchs, R., 2005, "SLT Progress Report", *Deliverable no. D5, TC-STAR*, European Community project no. FP6-506738, disponible en: http://www.tc-star.org/pages/f_documents.htm
 - Nießen, S., Ney, H., 2001, "Toward hierarchical models for statistical machine translation of inflected languages", *Workshop on Data-Driven Machine Translation (39th Annual Meeting of the Association for Computational Linguistics)*, :47-54.
 - Nirenburg, S., Carbonell, J., Tomita, M., Goodman, K., 1992, *Machine translation: a knowledge-based approach*, Morgan Kaufmann.
 - Och, F., Tillmann, C., Ney, H., 1999, "Improved alignment models for statistical machine translation", *Workshop on Empirical Methods in Natural Language Processing and Very Large Corpora*, :20-28, University of Maryland, College Park, MD.
 - Och, F., Ney, H., 2000, "Improved statistical alignment models", *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong Kong, China.
 - Och, F., Ney, H., 2002, "Discriminative training and maximum entropy models for statistical machine translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, :295-302, Philadelphia, PA.
 - Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., Radev, D., 2003, "Syntax for statistical machine translation", *Final Report on Johns Hopkins 2003 Summer Workshop*, disponible en: <http://www.clsp.jhu.edu/ws2003/>
 - Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., Radev, D., 2004, "A smorgasbord of features for statistical machine translation", *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (NAACL)*, :161-168, Boston, MA.
 - Padilla, P., Bajo, T., 1998, "Hacia un modelo de memoria y atención en interpretación simultánea", *Quaderns. Revista de traducció*, 2: 107-117.

-
- Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002, "Bleu: a method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA.
 - Pérez, J., Bonafonte, A., 2004, "GAIA: integrated platform for the development of speech translation technologies", *Reporte Interno*, Dpto. de Teoría de la Señal y Comunicaciones, Universitat Politècnica de Catalunya.
 - Press, W.H., Teukolsky, S., Vetterling, W., Flannery, B., 2002, *Numerical Recipes in C++: the Art of Scientific Computing*, Cambridge University Press.
 - Rabiner, L., Juang, B., 1993, *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, NJ.
 - Shannon, C.E., 1949, "Communication theory of secrecy systems", *Bell Sys. Tech. Journal*, 28:656-715.
 - Shannon, C.E., 1951, "Prediction and entropy of printed English", *Bell Sys. Tech. Journal*, 30: 50-64.
 - Shannon, C.E., Weaver, W., 1949, *The mathematical theory of communication*, University of Illinois Press, Urbana, IL.
 - Somers, H., Sugita, Y., 2003, "Evaluating commercial spoken language translation software", *Proceedings of the 9th Machine Translation Summit*, New Orleans, LS.
 - Sproat, R. (editor), 1998, *Multilingual text-to-speech synthesis: the Bell Labs approach*, Kluwer, Dordrecht.
 - Stolcke, A., 2002, "SRILM: an extensible language modeling toolkit", *Proceedings of the International Conference on Spoken Language Processing*, :901-904, Denver, CO.
 - Thomas, K., 1999, "Designing a task-based evaluation methodology for a spoken machine translation system", *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, :569-308, University of Maryland.
 - Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., 1997, "A DP-based search using monotone alignments in statistical translation", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*.
 - Tsujii, J., 1986, "Future directions of machine translation", en *COLING-86*, Bonn.
 - Ueffing, N., Och, F., Ney, H., 2002, "Generation of word graphs in statistical machine translation", *Proceedings of Empirical Methods in Natural Language Processing*, :156-163, Philadelphia, PA.
 - Wang, Y., Waibel, A., 1997, "Decoding algorithm in statistical machine translation", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*.
 - Weaver, W., 1955, "Translation", en Locke, W., y Booth, D., (editores), *Machine translation of languages: fourteen essays*, :15-23, John Wiley & Sons, Nueva York.
 - Yamada, K., Knight, K., 2001, "A syntax-based statistical translation model", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, :6-11, Toulouse, France.
 - Zens, R., Och, F., Ney, H., 2002, "Phrase-based statistical machine translation", *25th German Conference on Artificial Intelligence*, :18-32, Aachen, Springer Verlag.
 - Zens, R., Ney, H., 2003, "A comparative study on reordering constraints in statistical machine translation", *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, :144-151, Sapporo, Japón.