TRABAJO FINAL DE GRADO EN TRADUCCIÓN E INTERPRETACIÓN

TREBALL FINAL DE GRAU EN TRADUCCIÓ I INTERPRETACIÓ

Departament de Traducció i Comunicació

TÍTULO / *TÍTOL*

Competencias profesionales asociadas a la traducción automática estadística

Autor/a: Joan Dolz Sánchez

Tutor/a: Amparo Alcina Caudet

Fecha de lectura / Data de lectura: juny 2016

Resumen / Resum:

Los programas de traducción automática son una parte cada vez más importante del juego de herramientas del traductor profesional y la empresa de traducción. Sin embargo, la falta de formación en su uso presenta un obstáculo para muchos traductores y traductoras a la hora de integrarlas en su actividad profesional. Analizamos el uso de estas herramientas con un caso representativo: el programa de traducción automática estadística Moses. A partir de un ejemplo de configuración y uso de este programa, determinamos cuáles son las competencias tecnológicas y lingüísticas asociadas.

Palabras clave/ Paraules clau: (5)

Traducción automática, competencias de la traducción, tecnologías aplicadas a la traducción, traducción de gran volumen, Moses, MosesCore, EuroMatrix.

Índice

1. Introducción

Los primeros intentos de automatizar la traducción tuvieron lugar en la década de 1930. En aquel momento comenzó una vía de investigación que culminaría en 1954 con la primera demostración pública de un sistema de traducción automática del ruso al inglés.

A pesar del optimismo inicial entorno la traducción automática, pronto se hizo evidente que las traducciones que producían los ordenadores no estaban a la altura de las necesidades de la industria. Lo que sí consiguieron automatizar fue la producción en la lengua meta de un texto inteligible (en ocasiones suficiente) que sirviera de punto de partida para el traductor o corrector para producir una traducción final de calidad.

Todavía hoy, casi todos los sistemas de traducción automática requieren la intervención de un ser humano, ya sea preparando el texto de entrada, interaccionando con el sistema mientras tiene lugar la traducción, o revisando el texto producido por el sistema. No obstante, como veremos a lo largo de este trabajo, esta intervención tiene a menudo un carácter marcadamente interdisciplinar. Para el traductor o la traductora que trabaja con un sistema de traducción automática, esto requiere poseer unos determinados conocimientos técnicos. A menudo las entidades que necesitan utilizar traducción automática recurren a ingenieros o lingüistas computacionales para llevar a cabo esta labor.

2. Objetivos

El objetivo de este trabajo es exponer cuáles son los conocimientos necesarios para trabajar con software de traducción automática, qué competencias requiere el traductor, y cuáles son los retos a los que se enfrenta un profesional de la traducción a la hora de trabajar con software de traducción automática. Asimismo, a partir de este estudio, analizaremos algunas de las soluciones propuestas a estos problemas.

3. Metodología

Para realizar este estudio empezaremos por describir el estado de la cuestión a grandes rasgos para así entender la extensión del uso de la traducción automática en el ámbito empresarial e institucional a gran escala. Describiremos las principales estrategias de traducción automática y cuál es el flujo de trabajo habitual con estas herramientas.

A continuación, pasaremos a hacer un análisis más específico utilizando como base

la *suite* de traducción automática Moses para hacer un recorrido por las diferentes fases que conlleva preparar un sistema de traducción automática de estas características, así como el texto que se desea traducir y finalmente revisar la traducción generada por el sistema. Documentaremos también el uso de uno de los sistemas diseñados para acortar las distancias entre el traductor no especializado y este sistema de traducción automática.

Parte teórica

Introducción a la traducción automática
y la traducción automática estadística

4. Introducción a la traducción automática

La idea de automatizar la totalidad o parte de la traducción existe por lo menos desde el siglo XVII. Descartes y Leibniz ya sugirieron formas de abordar el problema, pero fue en 1933 cuando empezaron a idearse (y a patentarse) sistemas completos que pudieran resolverlo de una manera holística.

En 1949, Warren Weaver, miembro de la Fundación Rockefeller, sugirió apicar técnicas de análisis estadístico y de la teoría de la información desarrolladas durante la Segunda Guerra Mundial para construir el primer sistema de traducción automática y poco a poco se empezó a realizar investigación en el campo de la traducción automática por todos los Estados Unidos. Cinco años después se presentó públicamente el primer sistema práctico de traducción automática. Se desarrolló en la Universidad de Georgetown en colaboración con IBM y contaba con un vocabulario de apenas 250 entradas y 6 reglas gramaticales. A pesar de su limitada naturaleza, despertó un gran optimismo que acompañaría el desarrollo de los sistemas de TA durante toda su primera generación. Por primera vez se consideró factible la creación de sistemas capaces de producir traducciones, indistinguibles de aquellas hechas por seres humanos, de manera completamente autónoma (Hutchins y Sommers, 1992). A partir de los años 60, los investigadores empezaron a darse cuenta de lo alejadas que estaban las necesidades reales de la industria de la realidad de lo que eran capaces los sistemas existentes en aquella época (Austermühl, 2001). El futuro de un mundo de la traducción sin traductores humanos resultó estar mucho más lejos de lo que se pensaba.

A lo largo de los años, la investigación en lingüística computacional, el aumento en potencia de cálculo y los distintos avances en las tecnologías y disciplinas relacionadas con la traducción automática han ido introduciendo grandes mejoras, pero casi todos los sistemas siguen dependiendo de la post-edición para producir traducciones de calidad. En algunos casos, también es necesario pre-editar o preparar el texto de entrada.

5. La traducción automática en el ámbito empresarial e institucional

5.1. La aparición de los sistemas para uso comercial

Durante la década de los 80 aparecieron los primeros sistemas de traducción automática para uso comercial. Estos sistemas, que estaban diseñados para ejecutarse en los primeros ordenadores personales (las llamadas “microcomputadoras”, el Amstrad CPC

siendo tal vez la más conocida en España), brindaban una calidad discutible desde el punto de vista lingüístico, pero ofrecían una solución asequible, y resultaban especialmente atractivas para el uso personal y profesional de menor envergadura. Además de estos sistemas, se popularizaron entre las grandes corporaciones, los desarrollados por encargo. La empresa Systran, por ejemplo, desarrollo sistemas a medidas para Aérospatiale, Dornier, la OTAN y General Motors (Hutchins y Sommers, 1992).

Systran (desarrollado por la empresa con el mismo nombre) es uno de los sistemas de traducción automática más antiguos en la industria, y ha sido utilizado por entidades de gran impacto socioeconómico tanto públicas como privadas, entre las cuales se encuentran las Fuerzas Aéreas estadounidenses, o empresas como Xerox o General Motors. La implantación de estos sistemas a finales del siglo XX supuso un aumento notable de la eficiencia en términos económicos y de tiempo (Pigott, 1993).

No obstante, tras la adopción de los sistemas de traducción automática en la Comisión Europea se observó que los traductores necesitaban un tiempo de adaptación antes de poder aprovechar todo el potencial que brindaban. Parte de este proceso de adaptación consistía en aprender a preparar los documentos de entrada y a manejarlos durante las distintas etapas del proceso. Por otra parte, la post-edición de las traducciones difería mucho de los métodos utilizados tradicionalmente, aplicados sobre textos producidos por traductores humanos, y requirió también una adaptación por parte del personal encargado de revisar los textos (Pigott, 1993). Cualquier traductora o traductor que carezca de una formación específica en la utilización de sistemas de traducción automática se encontrará con que, al tratar de utilizar uno de estos sistemas, se ve obligada a sumar a su metodología habitual un conjunto de procesos y actividades pertenecientes a otras disciplinas. Se trata de una competencia instrumental con la que, por su carácter específico, muchos traductores no cuentan y que no necesitan para desarrollar su actividad profesional fuera de estos ámbitos específicos. Existen cursos realizados por empresas cercanas al ámbito del traductor *freelance*, como SDL,[1] que los enfocan como un complemento formativo específico. No obstante, debido a este carácter específico, algunas empresas optarán por recurrir a una plantilla multidisciplinar y aprovecharán una demanda laboral mucho más amplia y rentable

1 http://www.translationzone.com/learning/training/post-editing-machine-translation/

antes de invertir en la formación de sus traductores, en los casos en que éstos se encuentren en la plantilla de la empresa. Por este motivo, la tarea del traductor poco a poco queda relegada a la posedición lingüística (Pym, 2013).

En años recientes, la popularidad de los proyectos de código abierto ha alcanzado también la traducción automática. Especialmente relevantes fueron los proyectos EuroMatrix en 2006 y más adelante EuroMatrixPlus en 2009, cuyos principales objetivos, de acuerdo con la información publicada en su página web,[2] fueron la creación de sistemas de traducción automática para todos los pares de lenguas europeas, la evaluación periódica de la traducción automática en relación con las necesidades económicas y sociales de Europa, así como crear y distribuir tecnologías de traducción automática de código abierto. Estos proyectos culminaron en 2012 con MosesCore, una iniciativa para mantener e impulsar el proyecto Moses, un sistema de traducción automática estadística del cual hablaremos en detalle más adelante.

5.2. Los usos de la traducción automática en el ámbito institucional

Existen estudios que evalúan el uso de la traducción automática en organizaciones multilingües y multiculturales de gran tamaño que revelan la existencia de dos tipos de usuarios. De acuerdo con la clasificación realizada por Yuste Rodrigo (2005), en primer lugar están aquellos que utilizan la traducción producida por el sistema automático como base para desarrollar su propio trabajo traductológico. Por otra parte, están los usuarios que utilizan la traducción automática como una herramienta para acceder a información escrita en un idioma que desconocen y realizar una actividad que pueda estar relacionada directamente con el texto o con la traducción. Este segundo tipo de usuario se encuentra cada vez más a menudo fuera del ámbito institucional.

Hutchins (2003) distingue tres tipos de demanda de traducción automática según el uso que se pretende hacer de ella. Estos tres tipos son "difusión", "absorción" e "intercambio".[3] El uso de la traducción automática para la difusión de información es la que tiene unos requerimientos de calidad mayores en lo que respecta al texto producido, y requieren casi invariablemente la intervención humana. En este sentido, Hutchins

2 http://euromatrix.net

3 En inglés: *dissemination*, *assimilation* e *interchange*, respectivamente. La difusión, como su propio nombre indica, consiste en la propagación de la información traducida, mientras que la asimilación consiste en el consumo directo de dicha traducción con fines informativos y el intercambio en su uso instrumental con intenciones comunicativas, como es el caso de la interpretación.

considera que el sistema de traducción está, en realidad, produciendo un borrador, ya que la revisión final estará siempre a cargo de un humano. Este uso de las tecnologías de traducción automática está estrechamente relacionado con el primer tipo de usuario del que hablábamos (el que utiliza la TA como base para su labor traductológica), y es el que más nos interesa aquí.

Existen numerosas aplicaciones de traducción automática que se ajustan a una gran diversidad de usuarios. La mayor parte de traductores *freelance* y empresas y agencias de traducción utilizan a menudo herramientas cuya existencia se deriva de la investigación en traducción automática de gran volumen,[4] pero que además de ofrecer traducción automática en mayor o menor medida, proporcionan la posibilidad de realizar numerosas tareas de asistencia al traductor, como por ejemplo la creación de memorias de traducción o bases de datos terminológicas. Algunas de ellas son Trados, Transit, Déjà Vu, MultiTrans o Wordfast. Aquí centramos nuestro interés en las herramientas especializadas, enfocadas de manera exclusiva a la traducción automática estadística de gran volumen de texto para su difusión, entre cuyos usuarios encontramos organizaciones gubernamentales, grandes multinacionales y entidades supranacionales como la Comisión Europea (Hutchins, 2010).

5.3. La traducción automática en el ámbito empresarial español

El proyecto de investigación ProjecTA 2015 publicó un estudio (Torres-Hostench, Presas y Cid-Leal, 2016) realizado sobre empresas del ámbito de la traducción en toda la península con el objetivo de determinar el grado de implantación de la traducción automática en el ámbito empresarial español. El estudio reveló que el 52,7% de las empresas encuestadas no utilizan la traducción automática; en un 35,6% esto se debe a que la propia empresa no confía en los resultados que pueda brindar, y en un 20% porque los traductores no lo aceptan como herramienta de trabajo. Entre los motivos que se citan para rechazar la traducción automática están el alto coste que supone formar a los traductores para aprender a usarlos, así como su falta de formación en materias de traducción automática y posedición, y su reticencia a aceptar "cambios en la manera tradicional de trabajar del traductor".

4 En inglés *bulk translation*, se entiende como la traducción sistematizada de grandes cantidades de texto. Normalmente se estructura o contenido se encuentran estandarizados, como es el caso de facturas, informes o formularios.

De la lectura del estudio se desprende que la implantación de la traducción automática en España se ve frenada por dos factores: la falta de confianza y la falta de formación (y los consiguientes gastos en concepto formativo por parte de la empresa). Conforme la implantación de sistemas de traducción automática aumente a nivel internacional, esta resistencia dentro de la industria y el sistema educativo español a invertir en la formación interdisciplinar de traductores supondrá un grave impedimento a su competitividad.

6. Tipología de la traducción automática

Veronica Lawson define la traducción automática como aquella traducción hecha por un ordenador con o sin asistencia humana. Los conceptos de traducción automática y traducción asistida son diferentes, puesto que la diferencia esencial entre ellas, es que la primera ofrece por sí misma una traducción completa. Lawson (1993) define cinco tipos de traducción automática según el proceso:

- De un solo corpus
- De entrada limitada
- Interactivos
- Actualizables por lotes
- *Try-anything* ("todo vale")

Los sistemas de un solo corpus están configurados para trabajar con un conjunto de textos determinado, a menudo muy específicos o especializados. Suelen utilizarse en la investigación y a menudo no son efectivos para traducir textos ajenos al corpus asociado a ellos.

Los sistemas de entrada limitada parten de textos simplificados, pre-editados o directamente escritos en un lenguaje restringido, diseñado específicamente para funcionar con estos sistemas de traducción. La adaptación puede consistir en procesar de manera sistemática el texto, eliminando ambigüedad, simplificando su estructura temática u homogeneizando su vocabulario. Otra manera del material de entrada adaptado es incrementar la cantidad de información de los símbolos de un texto añadiendo marcadores sintácticos y gramaticales explícitos (Koehn, 2007), o adaptar el texto manualmente de acuerdo con cualquiera de los dos enfoques. El objetivo es que el texto sea claro, simple y directo (Austermühl, 2001).

Existen también sistemas de traducción automática que requieren información

adicional que una persona debe ir proporcionando al programa conforme éste se encuentra con problemas de análisis estadístico que es incapaz de resolver por sí mismo. Estos son los llamados sistemas "interactivos", y la persona encargada de trabajar con ellos debe preparar el texto y revisar la traducción final, además de supervisar el proceso de traducción en tiempo real.

La cuarta categoría es la de sistemas actualizables. Estos sistemas permiten añadir vocabulario nuevo al traductor si durante el proceso de revisión se detectan carencias al respecto.

Finalmente, los sistemas denominados *try-anything*, como su nombre indica, están diseñados para lidiar con todo tipo de textos, cuentan con una mayor tolerancia a errores que otros sistemas y también son más propensos a producirlos.

La clasificación de Lawson de los sistemas de traducción automática de acuerdo con su funcionamiento nos permite organizarlos desde un punto de vista operativo. Sin embargo, en una disciplina tan cambiante y que ha visto tantos avances en un período histórico tan relativamente corto, una clasificación de acuerdo con los métodos subyacentes puede resultar más útil. Bonnie J. Dorr, Pamela W. Jordan y John W. Benoit (1998) realizaron un estudio sobre las distintas arquitecturas y principios en las que se basan los sistemas de traducción automática. Establecen dos grupos (con una extensísima clasificación en subgrupos): los sistemas simbólicos, cuyos métodos se basan en teorías lingüísticas o en propiedades lingüísticas del lenguaje, y los sistemas estadísticos, cuyos métodos utilizan técnicas de análisis estadístico sobre corpus de textos paralelos para producir traducciones. Este último grupo, que es el que nos ocupa, se ha convertido, en parte gracias al aumento en la disponibilidad de la potencia computacional (Dorr et al, 1998), en el paradigma dominante en el ámbito de la investigación y la traducción automática de gran volumen (Koehn, 2007).

7. La traducción automática estadística

Tal y como explican Dorr, Jordan y Benoit (1998) en su descripción de los sistemas de traducción automática basados en la estadística, se puede calcular la probabilidad de que una serie de palabras en la lengua meta sea la traducción de una serie de palabras en la lengua de origen a partir de las relaciones estadísticas entre estas palabras, siendo dichas relaciones extrapoladas del corpus alineado en ambas lenguas.

Expresado formalmente este enunciado tiene la forma:

$$P(T \mid S) \sim P(T) * P(S \mid T)$$

Es decir, la probabilidad de que la cadena de palabras T en la lengua meta sea la traducción de la cadena S en la lengua de partida es proporcional al producto de las probabilidades de que T sea una cadena correcta en la lengua meta y que S en la lengua de partida sea una traducción de T en la lengua meta. Esencialmente, lo que hace un sistema de traducción automática estadística es buscar la combinación de palabras T cuya probabilidad sea lo más alta posible de acuerdo con esta ecuación.

En algunos casos se utiliza un enfoque híbrido y se utiliza información lingüística de apoyo. En otros, esta apreciación se realiza independientemente de las propiedades lingüísticas de los textos y es puramente estadística.

8. El programa de traducción Moses

En 2005, Hieu Hoang, un estudiante de la Universidad de Edimburgo bajo la supervisión de Philipp Koehn, empezó a trabajar en un conjunto de herramientas para la traducción automática llamado Moses. El proyecto estaba concebido como el sucesor de un proyecto similar creado anteriormente por Koehn como parte de su doctorado.[5] Un año después, en 2006, Moses pasó a estar subvencionado por la Unión Europea dentro del proyecto EuroMatrix (y más tarde EuroMatrixPlus). Finalmente, en 2012, la Comisión Europea creó el proyecto MosesCore con el objetivo de aunar los intereses del mundo académico y comercial en la traducción automática de código abierto.[6] Philipp Koehn se encuentra actualmente al cargo del mantenimiento de Moses.

Aunque Moses todavía no existía cuando Lawson elaboró su clasificación, si hubiésemos de buscarle un lugar en ella, podríamos considerarlo un sistema de traducción de entrada limitada, aunque no es el texto lo que se prepara para poder trabajar con el sistema, sino que es el sistema el que se "entrena" para poder trabajar con un determinado tipo de textos. Este entrenamiento consiste en analizar un corpus de

5 Este proyecto se llama *Pharaoh* y su funcionamiento es similar al de Moses, ya que utiliza modelos de lenguaje y de traducción generados a partir de un corpus para "decodificar" el texto de partida.

6 Esta información se encuentra publicada en la página web del proyecto MosesCore, actualizada el 17 de diciembre de 2014. Puede encontrarse más información sobre la historia de Moses, EuroMatrix y EuroMatrix en sus respectivas páginas web, detalladas en la bibliografía de este trabajo.

textos paralelos y extraer información estadística de las correspondencias entre palabras. El paradigma con el que trabaja es pues estadístico.

Si, por ejemplo, vamos a utilizar Moses para traducir textos médicos especializados del inglés al francés, deberemos disponer de un corpus de textos paralelos de estas características y seguir el proceso de preparación antes iniciar la traducción. Dado que la información utilizada para la traducción se extrae principalmente del análisis de los textos paralelos, la fase preparativa juega un papel muy importante en la efectividad del traductor automático. Dicha preparación, dependiendo de los textos y del equipo informático utilizado, puede tardar entre horas y días en completarse.

Podemos decir que la traducción con Moses consta de dos fases diferenciadas: una preparatoria, llamada "de entrenamiento", en la que se configura el sistema para traducir un par de idiomas a partir de textos paralelos y una fase de traducción propiamente dicha. La primera fase sólo deberá realizarse una primera vez para cada par de idiomas con los que se quiera trabajar, ya que los archivos de configuración que se generan pueden reutilizarse tantas veces como se quiera para proyectos de características similares.

8.1. Estructura de Moses

Los componentes de Moses pueden clasificarse de la siguiente manera:

COMPONENTES PRINCIPALES	SUBCOMPONENTES
Herramientas de entrenamiento Se trata de un conjunto de pequeñas utilidades escritas en el lenguaje de programación Perl.	**Herramientas de alineación de corpus y extracción de datos estadísticos del mismo** **Constructor del modelo lingüístico** Este subcomponente se encarga de analizar el corpus de la lengua meta para inferir su sintaxis. Pueden utilizarse diferentes programas externos. Algunos de ellos son: IRSTLM, RandLM, KenLM, OxLM y NPLM[7] **Calibrador y optimizador de los modelos estadísticos** Este subcomponente se encarga de comparar la traducción generada con textos en la lengua meta y la ajusta para lograr una traducción lo más natural posible.
Decodificador Se trata de un sólo programa escrito en el lenguaje de programación C++, con la eficiencia del uso de recursos en mente.	El decodificador utiliza el modelo de lenguaje y el modelo estadístico creados en la fase de entrenamiento para producir traducciones.

7 Puede encontrarse una lista completa en http://www.statmt.org/moses/?n=FactoredTraining.BuildingLanguageModel.

Hemos visto que la traducción automática estadística se basa en el análisis de textos tanto en la lengua de partida como en la lengua meta, a partir de los cuales el programa genera un modelo explicativo del lenguaje que después utiliza para generar las traducciones. A continuación, en la parte empírica de este trabajo, documentamos el proceso de configuración y uso del programa Moses como ejemplo representativo de un programa de traducción automática estadística.

Parte empírica

Análisis del sistema de traducción automática Moses
y las competencias asociadas

9. Descripción del proceso de configuración y uso de un sistema de TA

Vamos a recorrer los distintos pasos del proceso de configuración del sistema y generación de traducciones. Debe tenerse en cuenta que Moses permite ajustar su funcionamiento de muchas maneras y ofrece una funcionalidad muy amplia. El objetivo de este trabajo es elaborar una documentación introductoria al funcionamiento del software, por lo que se describirá solamente su uso básico para preparar un modelo de traducción y utilizar el decodificador.[8]

9.1. Recursos necesarios

Los ejemplos e instrucciones que utilizaremos corresponden a un sistema de tipo UNIX (sea Mac OS X o Linux) en el que las herramientas ya hayan sido instaladas.[9]

Los archivos que utilizaremos para entrenar el sistema de traducción se generarán a partir de un corpus alineado, que consistirá en una serie de frases alineadas en dos pares de archivos diferentes (cada par consiste en un archivo para cada idioma. En nuestro, caso español e inglés). El primer par se utilizará para entrenar el sistema de traducción, y el segundo para optimizarlo. En los ejemplos que siguen, nos referiremos a estos archivos como `frases.en`, `frases.es`, `frases2.en`, `frases2.es`. Estos archivos se encuentran dentro de un subdirectorio `corpus`.

9.2. Preparación del corpus paralelo

El primer paso consiste en separar las frases de manera que cada palabra y signo de puntuación quede rodeado de espacios (este paso se conoce en inglés como *tokenisation*). A continuación eliminamos las frases para las cuales falte la correspondencia en alguno de los dos idiomas, así como las líneas en blanco y las líneas de más de 80 caracteres de longitud.

Los dos comandos necesarios para la separación en unidades (*tokenisation*) son:

```
mosesdecoder/scripts/tokenizer/tokenizer.perl -l en \
< corpus/frases.en > corpus/frases.tok.en
```

Y

```
mosesdecoder/scripts/tokenizer/tokenizer.perl -l es \
```

8 Una documentación exhaustiva puede encontrarse en la página web de Moses (http://statmt.org) o directamente en el manual de uso de Moses (http://www.statmt.org/moses/manual/manual.pdf).

9 Se asumirá que se han seguido las instrucciones de instalación que pueden encontrarse en el manual de Moses y que por lo tanto las herramientas se encuentran en sus ubicaciones por defecto.

```
< corpus/frases.es > corpus/frases.tok.es
```

Y para limpiar las frases (de nuevo, teniendo en cuenta las rutas a los archivos):

```
mosesdecoder/scripts/training/clean-corpus-n.perl \
corpus/frases.tok en es corpus/frases.clean 1 80
```

Este último comando procesa los dos archivos, `frases.tok.es` y `frases.tok.en`, y genera dos archivos limpios: `frases.clean.es` y `frases.clean.en`.

Todo este proceso se repitió con los archivos `frases2.es` y `frases2.en` para obtener también `frases2.clean.es` y `frases2.clean.en`.

9.3. Creación de un modelo de lenguaje

Para producir texto idiomático, Moses utiliza un modelo de lenguaje para la lengua meta (español en este caso) que extrae a partir del mismo corpus. Esto tuvo lugar dentro de un directorio nuevo usando los comandos *mkdir* (para creación de directorios) y *cd* (para movernos dentro del directorio creado):

```
mkdir lm && cd lm
```

Y utilizando estos dos comandos lo generaremos:

```
../mosesdecoder/bin/lmplz -o 3 \
<../../corpus/frases.clean.es > frases.arpa.es
```

Y

```
../mosesdecoder/bin/build_binary frases.arpa.es frases.blm.es
```

Finalmente volvemos a nuestro directorio de trabajo antes de seguir con el siguiente paso:

```
cd ..
```

9.4. Entrenamiento del sistema de traducción

Para el proceso de entrenamiento del sistema crearemos un directorio donde se almacenarán todos los archivos intermedios y la configuración final que utilizaremos para la realizar traducciones. Utilizando el siguiente comando creamos y entramos en dicho directorio:

```
mkdir working && cd working
```

Dentro de este directorio ejecutaremos el comando que iniciará el proceso de entrenamiento. Este comando realiza dos funciones. Primero convierte el corpus alineado en un formato adecuado para Moses, y a continuación lo analiza y extrae la información estadística necesaria para realizar traducciones. En nuestro caso, el proceso

llevó cuatro horas. Dependiendo del equipo que se utilice y del tamaño del corpus, puede llevar entre horas y días. El comando es el siguiente:

```
nohup nice ../mosesdecoder/scripts/training/train-model.perl \
--root-dir train -corpus ../corpus/frases.clean \
-f en -e es -alignment grow-diag-final-and \
-reordering msd-bidirectional-fe \
-lm 0:3:$(readlink -f ../lm/frases.blm.es):8 \
-external-bin-dir ../mosesdecoder/tools >& training.out &
```

Nótese que se utiliza *nohup* para que el proceso continúe aunque cerremos la ventana del terminal y *nice* para hacer explícita la prioridad que el sistema operativo ha de dar a nuestro comando. Estos dos comandos pueden omitirse o ajustarse dependiendo del entorno en el que estemos trabajando. En el caso de este trabajo, se utilizaron los valores por defecto, de manera que todos los recursos del ordenador quedaron copados. Con el parámetro *-n* del comando *nice* puede reducirse este uso de recursos. Esto tendría el efecto de alargar el tiempo necesario para completar el entrenamiento de Moses, pero nos permitiría utilizar el ordenador mientras tanto o, en el caso de que estuviésemos trabajando en un servidor compartido, no alterar el trabajo de otros usuarios.

9.5. Calibración y optimización de los archivos de entrenamiento

El último paso consiste en calibrar y optimizar el sistema de traducción utilizando el resto del corpus (los archivos `frases2.en` y `frases2.en`):

```
nohup nice ../mosesdecoder/scripts/training/mert-moses.pl \
../corpus/frases2.en ../corpus/frases2.es \
../mosesdecoder/bin/moses train/model/moses.ini \
–mertdir ../mosesdecoder/bin/ &> mert.out &
```

Cuando este proceso termina, encontramos el archivo de configuración para Moses, `moses.ini`, en el directorio `working/mert-work`. Este archivo hace referencia a gran parte de los archivos intermedios generados durante todo el proceso de preparación, por lo que es importante mantenerlos en sus respectivas ubicaciones para que Moses pueda encontrarlos.

9.6. Traducción con Moses de un texto de muestra

Una vez ha sido completado el entrenamiento, la traducción en sí solo requiere usar el siguiente comando. El texto que queremos traducir se encuentra en un archivo llamado

`texto-partida.txt` y queremos guardar la traducción en un archivo `traduccion.txt`:

```
mosesdecoder/bin/moses -f working/mert-work/moses.ini \
-i texto-partida.txt > traduccion.txt
```

9.6.1. Análisis de la traducción generada

Se ha utilizado la configuración de Moses generada en este ejemplo para traducir el siguiente párrafo:

> In the context of a capital increase, Drillisch placed 17.4 million in new shares, with a total value of 106.4 million euros. The two companies have founded the joint venture MSP in order to buy competitor Freenet.

El texto producido automáticamente por Moses es el siguiente:

> En el contexto de una capital increase, Drillisch colocó 17.4 millones en nuevos shares, con un valor total de 106.4 millones de euros. Las dos empresas han fundó la empresa conjunta MSP para comprar competidor Freenet.

Como puede verse, el texto producido necesita una corrección. Se encuentran problemas de falta de correspondencias léxicas (*capital increase* y *shares*) así como de correspondencia gramatical (*han fundó* en vez de *han fundado*) y malas equivalencias (*placed 17..4 million in new shares* donde "placed" está traducido como "colocó", cuando una mejor opción hubiese sido "invirtió") También falta un artículo (*to buy competitor Freenet* traducido como *para comprar competidor Freenet*) que fue omitido en el original en inglés respondiendo a su estilo periodístico. Además, la traducción utiliza el punto como separador decimal en vez de la coma.

9.7. Análisis de competencias necesarias

Como hemos visto, el proceso de preparación del corpus y de entrenamiento del sistema se componen de distintas partes interdependientes. Debido a su complejidad, es un proceso que debe entenderse, ya que dependiendo de las circunstancias pueden aparecer problemas inesperados en cualquier fase.

De acuerdo con el esquema que acabamos de presentar, distinguimos las siguientes competencias que no pertenecen al ámbito de la traducción:

9.7.1. Competencias cognitivas

- Conocimientos de codificación de archivos de texto y conversión entre distintos sistemas de codificación.
- Conocimientos sobre el paradigma de traducción basado en estadística.
- Conocimientos sobre el funcionamiento interno de los distintos componentes de Moses y cómo interaccionan.

9.7.2. Competencias procedimentales

- Uso de la línea de comandos UNIX o Windows para navegar y modificar el sistema de ficheros. Especialmente *mkdir* para la creación de directorios y *cd* para navegar el sistema de ficheros. También *rm* y *rmdir* para eliminar respectivamente archivos y directorios.
- Uso de la línea de comandos para ejecutar programas en Perl así como ejecutables binarios. Es especialmente importante conocer *nohup*, *nice* y *readlink*.
- Debido a los altos requisitos de memoria y capacidad de cálculo del programa, saber lidiar con los potenciales problemas de falta de memoria y espacio en disco que puedan surgir, ya sea a través de la línea de comandos o un interfaz gráfico.

9.7.3. Competencias actitudinales

Debido a la naturaleza modular de Moses y su complejidad, es necesario tener una actitud abierta a la experimentación en la resolución de problemas. Como se vio en la parte teórica, la falta de dicha actitud aqueja a muchos profesionales de la traducción en España y presenta el primer obstáculo en la obtención de las demás competencias.

9.7.4. Competencias relacionadas con la traducción

También es evidente, a partir del breve fragmento que hemos traducido, que los textos producidos necesitan de la revisión de una persona con competencias lingüísticas, si no de traducción, al menos de redacción en la lengua meta.

10. Discusión

La traducción automática estadística se encuentra en proceso de expansión debido a su relevancia en el contexto institucional y empresarial. Las herramientas destinadas a satisfacer esta creciente demanda tienen una estructura muy modular y a menudo toman una forma a medio camino entre la aplicación especializada de uso corporativo y la herramienta del investigador. Están diseñadas de forma que puedan ser desensambladas, analizadas y modificadas en cualquier momento para, por una parte, poder adaptarse a nuevos propósitos, y por otra, sacar provecho de nuevas técnicas descubiertas en el laboratorio.

Sin embargo, para el traductor o la traductora que se incorpora a un proyecto de investigación o a un puesto en un organismo o empresa que hace uso de estas herramientas, se presentan dos tipos de obstáculos. Por una parte, es muy probable que durante su formación haya tratado muy poco o nada en absoluto las herramientas de traducción automática,[10] por otro, estas herramientas no están diseñadas para que sea fácil aprender a usarlas. Esto, como se ha visto, frena su implementación en el ámbito empresarial, lo cual dificulta todavía más que el traductor o la traductora obtenga la experiencia necesaria para suplir su falta de formación.

Como hemos visto, es imprescindible que las personas encargadas de supervisar el proceso de traducción automática sean competentes tanto tecnológica como lingüísticamente. Muchas empresas optan por la contratación de dos profesionales cuyas competencias combinadas cubran las necesidades de la herramienta. Esto supone en algunos casos un gasto adicional para la empresa (Torres-Hostench, Presas, Cid-Leal, 2016) y en otros un obstáculo en el acceso al empleo para el traductor.

Una posible solución que permitiría que un único operario cubriese todas las necesidades del proceso de traducción automática, sería la creación de interfaces gráficos para cruzar la laguna formativa del traductor, eliminando parte de las competencias necesarias. Volviendo a Moses, vemos que existen soluciones de este tipo, como MosesGUI,[11] pero ninguna de ellas puede aspirar a cubrir todo el abanico de funcionalidad del programa. Esto sin duda presenta buenas oportunidades de negocio para las empresas de desarrollo de software que se encuentren en posición de ofrecer

10 Cabe diferenciar las herramientas de traducción automática, como Moses, de las de traducción asistida, como Trados, más comunes en los programas formativos universitarios y de empresa.

11 MosesGUI es la interfaz oficial, cuyo código fuente está disponible en: https://github.com/moses-smt/mosesdecoder/tree/master/mingw/MosesGUI.

soluciones personalizadas a empresas e instituciones. Al mismo tiempo, esto significa que no existe una vía predeterminada a la que puedan acogerse los directores de proyecto a la hora de decidir cómo abordar este problema.

Otra solución, que aborda el problema desde el punto de vista, no de eliminar, sino de satisfacer las competencias necesarias, es formar a los traductores y traductoras, ya sea desde el ámbito empresarial o desde el ámbito académico. Han habido propuestas de reforma educativa en este sentido. El proyecto LETRAC,[12] por ejemplo, propuso una serie de elementos de carácter tecnológico para su inclusión en los programas de formación de traductores en universidades europeas con el fin de mejorar el acceso de los estudiantes al mercado laboral (Alcina, 2008).

11. Conclusión

Para el uso de herramientas de traducción automática estadística es imprescindible contar con profesionales competentes en el ámbito lingüístico y en el uso de las tecnologías pertinentes.

La necesidad de una competencia lingüística es inherente a la función de estas herramientas, que en última instancia es traducir. Por otra parte, las competencias tecnológicas necesarias van más allá de las adquiridas por un traductor o una traductora en su formación, pero solo son necesarias como consecuencia del diseño de las herramientas. Es decir, no son inherentes a su función. Es por esto que la potencial carencia del profesional de la traducción en este ámbito puede suplirse ofreciendo formación complementaria, o bien evitarse completamente adaptando las herramientas para permitir su uso a personal no especializado.

12. Valoración personal

Hemos visto que un profesional de la traducción debe aspirar a ser más que traductor. En ocasiones lingüista, o ingeniero. Se trata de una profesión estrechamente ligada a la dimensión humana del lenguaje y la comunicación, que al mismo tiempo está recibiendo cada vez más apoyo tecnológico en distintas áreas de su actividad. El traductor debe ver este avance de la tecnología no como una invasión de su territorio, sino como una oportunidad para ampliar su capacidad como profesional. Debe adaptarse

12 *Language Engineering for Translators Curricula*.

si quiere integrarse en nuevos ámbitos profesionales tecnológicos que, por su naturaleza interdisciplinar, emergen estrechamente ligados a la traducción.

13. Bibliografía

ALCINA, A. (2002). Estrategias y recursos en la enseñanza de la Informática aplicada a la traducción. *Actes del primer Simposi sobre l'Ensenyament a distància i semipresencial de la Tradumàtica*, Bellaterra.

ALCINA, A. (2008). Translation technologies. Scope, tools and resources. *Target 20:1*. John Benjamins Publishing Company.

AUSTERMÜHL, F. (2001). *Electronic Tools for Translators* (pp. 153-176). Manchester: St. Jerome Publishing.

DORR, B. J., JORDAN, P. W., y BENOIT, J. W. (1998). *A Survey of Current Paradigms in Machine* Translation. (Technical Report: LAMP-TR-027, UMIACS-TR-98-72, CS-TR-3961). College Park: Universidad de Maryland.

HUTCHINS, J. (2003). Commercial systems: the state of the art. En Somers, H. (Ed), *Computers and translation: a translator's guide* (pp. 161-174). Ohio: John Benjamins Publishing Company, Ohio.

HUTCHINS, J. (2010). Outline of Machine Translation Developments in Europe and America. *JAPIO 2011 Yearbook*.

HUTCHINS, J y SOMMERS, H. (1992). *Introduction to machine translation*. Londres: Academic Press.

KOEHN, P. Moses. Statistical Machine Translation System. Recuperado el 20 de febrero de 2016 de. http://www.statmt.org/moses/?n=Moses.Overview

KOEHN, P. Pharaoh. A beam search decoder for phrase-based statistical machine translation models. Recuperado el 20 de febrero de 2016 de http://www.isi.edu/licensed-sw/pharaoh/

KOEHN, P et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 177-180). Praga: Association of Computational Linguistics.

LAWSON, V. (1993). A Translator's Map of Machine Translation. En Vasconcellos, M. (Ed), *Technology as Translation Strategy*. Ohio: John Benjamins Publishing Company.

PIGOTT, I. M. (1993). MT in Large Organizations: Systran at the Comission of the European Communities. En Vasconcellos, M. (Ed), *Technology as Translation Strategy*. Ohio: John Benjamins Publishing Company.

PYM, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta : journal des traducteurs* (vol. 58, n. 3), 487-503. Montreal: Les Presses de l'Université de Montréal.

TORRES-HOSTENCH, O., PRESAS, M., CID-LEAL, P. (coords.) (2016). El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: Informe de investigación ProjecTA 2015. Bellaterra.

YUSTE RODRIGO, E. (2005). La traducción automática, de su evolución y papel actual en la industria de la localización. En Reineke, D (director y coordinador), *Traducción y localización. Mercado, gestión y tecnologías* (pp. 161-186). Las Palmas de Gran Canaria: Anroart Ediciones.