Inteligencia artificial - CT Innovación

Google Translate

Facultat d'Informàtica de Barcelona
Universitat Politècnica de Catalunya

Nil Mamano Grande
Héctor Ramón Jiménez
Isaac Sánchez Barrera

10 de diciembre de 2013
cuatrimestre de otoño
curso 2013–2014

Índice

1. Google Translate

Google Translate es un traductor automático entre múltiples idiomas. Se caracteriza por ser uno de los más **fiables** hoy en día, gracias a las técnicas de inteligencia artificial que veremos. Entre sus características notables, podemos decir que es **gratuito y casi instantáneo**.

Inicialmente, los traductores automáticos, **Google Translate** incluido, usaban sistemas basados en **reglas**. Es decir, intentaban dotar al algoritmo con los conocimientos abstractos lingüísticos que poseemos los humanos, con la intención de hacer un análisis lógico de la estructura de una frase y encontrar una traducción coherente con esos conocimientos.

Sin embrago, esto funcionaba bastante mal debido a que los lenguajes naturales, lejos de los lenguajes formales, son ricos en **excepciones y peculiaridades** difíciles de "codificar".

En 2007, en un gesto de innovación, **Google** desechó el sistema basado en reglas y optó por una estrategia nueva: la **traducción automática estadística**.

Google no inventó la traducción automática estadística, puesto que ya existían nociones desde hacía décadas. Sin embargo, supo aprovecharla para hacer un traductor de mayor **calidad** que sus competidores. Actualmente, la mayoría de traductores automáticos están basados en la traducción automática estadística.

2. Traducción automática estadística

El enunciado formal del problema que aborda la traducción automática es el siguiente: Dado un idioma origen L_1, un idioma objetivo L_2 y una frase f en L_1, dar una traducción t equivalente en L_2.

La **traducción automática estadística** aborda esta situación como un **problema de búsqueda heurística**. De entre todas las frases en L_2, se busca la frase t que maximice el siguiente heurístico:

$$P(f|t) \cdot Q(t)$$

Donde:

- $P(f|t)$ representa **la probabilidad de que alguien entienda** f **al leer** t.
- $Q(t)$ representa **la probabilidad de que alguien use** t. Es inherente a t, independientemente de la frase que se quiera traducir.

Idealmente, el factor $P(f|t)$ procura que la traducción sea **fidedigna**, mientras que el factor $Q(t)$ pretende que el texto generado sea **correcto en el idioma objetivo**.

Así, si una frase t tiene un buen valor heurístico, será correcta sintáctica y gramaticalmente, (por tener un valor de $Q(t)$ alto) y además traducirá correctamente la frase (por tener un valor de $P(f|t)$ alto).

Para ilustrar el rol de estos factores, consideremos a modo de ejemplo que queremos traducir la siguiente frase al inglés:

f : *Luís estudió mucho, mas no pasó el examen de lengua.*

El algoritmo considera las siguientes traducciones:

- t_1 : *Luís studied very, but not passed exam of language.*
- t_2 : *Luís studied a lot, plus he didn't pass tongue's exam.*

En el caso de t_1, tenemos $Q(t_1) = 0.2$ y $P(f|t_1) = 0.8$. Se puede ver que $Q(t_1)$ es bajo, ya que la frase no parece inglés correcto. Sin embargo, sí que parece que alguien que lo lea pudiese entender el significado original. En cambio, para t_2 tenemos $Q(t_2) = 0.8$ y $P(f|t_2) = 0.2$. En este caso,

$Q(t_2)$ es alto ya que parece bien escrito. Sin embargo, es poco probable que el lector entienda el significado original. Por ejemplo, ha capturado erróneamente el significado de "mas" como sinónimo de "suma" en vez de "pero". Esto último puede ser debido a errores ortográficos en los textos usados para crear las reglas de traducción.

La cuestión es como desarrollar unas buenas funciones P y Q, ya que las nociones son bastante abstractas. Aquí es donde entra en juego la **estadística**.

Para encontrar la función $Q(t)$, lo que se hace es analizar enormes cantidades de texto en el lenguaje objetivo. Cuantas más veces aparezca t (total o parcialmente), mayor será $Q(t)$. Así, **expresiones frecuentes** como *"Hola, ¿cómo estás?"* tendrán mayor valor que frases raras como *"Qué gran casa más grande"*.

Para encontrar la función$P(f|t)$, se analizan cantidades lo más grande posibles de texto para los que existe una traducción hecha por humanos entre el idioma origen y el destino. El análisis consiste en ver con qué **frecuencia** una frase de un texto se corresponde (o lo parece) con una frase del otro. Si *"lo siento"* aparece frecuentemente en el mismo sitio que *"I'm sorry"*, es probable que su significado se corresponda.

Debemos notar que para construir las funciones P y Q no hemos tenido que dotar al algoritmo de **ningún conocimiento lingüístico explícito**, todo se infiere a partir del análisis automático. Además, este método **no depende de los lenguajes** en sí. Puede funcionar para cualquier par de lenguajes, siempre y cuando haya suficiente texto disponible. Por contraste, los sistemas basados en reglas requieren del desarrollo manual de las reglas, que son particular de cada lenguaje. Esta propiedad, la gran cantidad de contenido que **Google** tiene indexado y los increíblemente eficientes y precisos algoritmos de búsqueda de **Google Search**, han permitido a **Google Translate** hacer traducciones entre muchos idiomas.

Por otro lado, la calidad dependerá del **tamaño** de las muestras disponibles. Por suerte, para construir Q **sólo** necesitamos texto en el lenguaje objetivo, sin necesidad de tener su traducción, por lo que la disponibilidad suele ser muy alta. Con P no tenemos tanta suerte, ya que **los textos bien traducidos son más limitados**. **Google Translate** ha usado como muestra, entre otras cosas, documentos diplomáticos de la Unión Europea, puesto que estos se **traducen a múltiples idiomas paralelamente**. Para lenguajes de los que no dispone de mucho texto, sus resultados suelen ser peores.

Además de depender del tamaño de la muestra, la calidad también dependerá de la muestra elegida. El estilo literario usado en la muestra quedará reflejado en las traducciones. Por ejemplo, la traducción que **Google Translate** ofrece a *"cooking teacher"* es *"profesora de cocina"*. Esto se debe a que en la muestra de textos usados, el término *profesora* salía al lado de *cocina* **más veces** que *profesor*. Por lo tanto, $Q(\text{"profesora de cocina''}) > Q(\text{"profesor de cocina"})$, indicando erróneamente que la frase está mejor formada. Como veremos más adelante, esto generó cierta **controversia** (ver apartado 4.2).

3. Impacto de *Google Translate* en la empresa

Google ofrece el servicio de traducción de distintas maneras. Por un lado, tiene la traducción de textos no muy largos a través de su página web. También ofrece un servicio gratuito de traducción de páginas web para *webmásters*, además de una API[1] para empresas.

La primera versión de la API estuvo disponible hasta mediados de 2011, cuando Google la declaró obsoleta para su posterior eliminación en diciembre. Esto vino causado *por el abuso continuado*, según sus propias palabras, que se hacía del servicio. Sin embargo, como puede leerse en [2], los responsables de la empresa no quisieron dar más detalles sobre ese abuso continuado.

Más tarde decidieron monetizar el servicio de la API, que la volvieron de pago para la segunda versión y con menos limitaciones que la versión antigua. Pero no es posible ver los resultados

[1]Del inglés *Application programming interface*

económicos y razones reales para el cambio, ya que Google es una gran empresa y en casos como éste es difícil que presenten resultados detallados para el gran público.

Cabe destacar que la propia empresa hace uso de sus servicios de traducción. Un claro ejemplo es el portal de vídeos YouTube, donde utiliza su sistema de traducción para los subtítulos.

Por otro lado, el uso de los servicios de traducción de Google ha ido creciendo con los años. En abril de 2012 tenían más de 200 millones de usuarios mensuales en la página del traductor,[2] y eso sin contar todos los usuarios de otros servicios de la empresa que hacen uso del traductor. También se ha ampliado enormemente el uso a través de dispositivos móviles, cuadruplicándose cada año. Y es notable el hecho de que **el** 92 % **del tráfico** del traductor proviene de fuera de los Estados Unidos.

3.1. Impacto tecnológico

Aunque el servicio actual se lanzó en 2006, los trabajos para implantar la traducción estadística comenzaron en 2003. Los **resultados empezaron a verse en la competición *NIST Machine Translation Evaluation*** de 2005, la principal dentro del ámbito de la traducción automática. **La infraestructura** ya disponible en Google fue **capaz de manejar una gran cantidad de datos**, y fueron los ganadores del concurso.

Sin embargo, **no se podía considerar que los resultados fueran satisfactorios** en la práctica. El sistema tardó más de 40 horas en traducir 1000 oraciones, usando la capacidad de cómputo de 1000 máquinas. Pero el desarrollo de mejores métodos consiguió que un año más tarde, a principios de 2006, se pudieran conseguir traducciones mejores y más rápidas: en traducir una frase se tardaba menos de un segundo. Gracias a eso, sacaron la primera versión del traductor con el chino y el árabe como idiomas (los usados en la competición mencionada).

3.2. Otros posibles impactos derivados

Una de las metas que tiene Google es indexar toda la información producida por el ser humano y hacerla accesible. Para ello, hace unos años, en 2009, compró a la empresa reCAPTCHA, que se encarga de digitalizar libros del dominio público mediante técnicas de reconocimiento óptico de caracteres[3] y de colaboración por parte de los usuarios.

Teniendo en cuenta que **para mejorar las traducciones utilizan textos traducidos por humanos** expertos, pueden aprovechar esta digitalización con libros traducidos para mejorar el sistema de Google Translate. Sin embargo, esto puede suponer un problema causado por la **antigüedad de los textos bajo el dominio público**. Con el paso de los años, los idiomas evolucionan y usar material viejo puede provocar que, aunque correctas, se creen traducciones que el usuario puede considerar malas por su poca cercanía.

Otro provecho que seguramente pueda sacar Google de las traducciones es la publicidad. No en vano, la empresa recibe **la mayor parte de sus ingresos** mediante la **distribución de publicidad** con los programas *AdWords* y *AdSense*. Con un buen sistema de traducción pueden llegar a saber los gustos del usuario independientemente del idioma y conseguir, así, rentabilizar mejor los anuncios.

Un caso donde queda claro que sí que están usando las traducciones para fidelizar al usuario es en las búsquedas. Al realizar una búsqueda en Google, es posible seleccionar el idioma o permitir que busque en idiomas distintos del propio. En estos casos, **traduce la búsqueda** y, además, **propone la traducción de los resultados**. Con esto puede conseguir de manera más fácil que el usuario encuentre lo que busca y, de rebote, que **siga confiando en los servicios**.

Ahora bien, esto son solamente conclusiones a las que se pueden llegar observando a la empresa y conociendo algunos de sus servicios. El tamaño y la complejidad que tiene la empresa,

[2] Disponible en `http://translate.google.com/`

[3] OCR, por sus siglas en inglés

y la poca publicación de resultados, hacen imposible saber el impacto real que tiene el servico Translate en la empresa.

4. Impacto social

4.1. Aplicaciones

Que **Google Translate** ha tenido un impacto importante en la sociedad es indiscutible, sobre todo desde **2007**, cuando la calidad de las traducciones aumentó considerablemente.

El traductor se utiliza de maneras muy diversas y para fines distintos:

Comunicación Como herramienta de comunicación entre personas que entienden diferentes idiomas. Existen programas que permiten la traducción automática durante chats. Además, **Google Translate** es capaz de escuchar las palabras a traducir desde un micrófono y pronunciar la traducción realizada con una voz sintética, facilitando la comunicación más cercana. La **BBC**, por ejemplo, realizó en 2010 un experimento para facilitar la discusión en diferentes lenguajes utilizando **Google Translate** y los resultados fueron bastante satisfactorios (ver [3]).

Traducción automática de sitios web **Google Translate** permite la traducción completa de páginas web. La velocidad del traductor y de los servicios de **Google**, así como su integración con los navegadores hacen que esta tarea resulte increíblemente sencilla.

Fines didácticos No podemos olvidar la cantidad de estudiantes que utilizan el traductor con fines didácticos. Pueden ampliar su vocabulario traduciendo palabras sueltas, aprender gramática observando traducciones completas, cambiar el contexto de las oraciones y observar los resultados... Hasta existe una aplicación para **Google Chrome** que intenta simular la experiencia de sumergirte en un lenguaje extranjero utilizando **Google Translate** (ver [7]). ¡Las posibilidades son ilimitadas!

Sin embargo, la herramienta de traducción **no es perfecta** y, por lo tanto, uno de los **riesgos** es que las traducciones podrían ser **imprecisas** o **erróneas** y producir **malentendidos** y/o **confusiones**. Las personas que la usen deberían ser **conscientes** de ello para evitar tomarse al pie de la letra las traducciones obtenidas y mantener un estado **crítico**.

4.2. Sexismo

En varias ocasiones, **Google** ha sido acusado de sexismo debido a ciertas **sugerencias** y/o **traducciones** realizadas bajo **Google Translate**. Veamos algunos ejemplos de esto:

- Hace un tiempo si introducías "*Men are men and men should clean the house*", entonces el traductor te sugería que quizás querías decir "*Men are men and **women** should clean the house*".
- Algo similar ocurre, todavía hoy en día, cuando introduces "*Men and women are **now** equal*", sugiriéndote "*Men and women are **not** equal*".
- Al traducir frases sin género como *"a nurse"* o *"a cooking teacher"* de idiomas como el **inglés** a idiomas como el **español** donde el **determinante indefinido** expresa género, se producen resultados como *"una enfermera"* o *"una profesora de cocina"*.

Este hecho llegó a generar mucha controversia en su día. Sin embargo, como ya se ha explicado en el apartado 2, **Google Translate** no sabe **realmente** diferenciar entre géneros. Como bien explicó Xi Cheng, administrador de aplicaciones de **Google** (ver [9]):

> Google Translate is **fully** automated by machine; **no one is explicitly imposing any rules**; the translation is generated according to the **statistical nature** of the corpus we have.

Por lo tanto, si el material y los datos usados para la traducción contienen cierta **tendencia** o **parcialidad**, ésta se verá reflejada en las **sugerencias** y/o **traducciones**.

5. Referencias

[1] Peter F. Brown y col. «The Mathematics of Statistical Machine Translation: Parameter Estimation». En: (1993). [Internet]. URL: http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf.

[2] Ed Burnette. «Google pulls the rug out from under web service API developers, nixes Google Translate and 17 others». En: *ZDNet* (27 de mayo de 2011). [Internet]. URL: http://www.zdnet.com/blog/burnette/google-pulls-the-rug-out-from-under-web-service-api-developers-nixes-google-translate-and-17-others/2284 (visitado 10-12-2013).

[3] Jeff Chin. «An experiment in cross-language communication with the BBC». En: *The Official Google Translate Blog* (25 de mar. de 2010). [Internet]. URL: http://googletranslate.blogspot.com.es/2010/03/experiment-in-cross-language.html (visitado 10-12-2013).

[4] *Inside Google Translate*. [Internet; vídeo]. Google. URL: http://www.youtube.com/watch?v=Rq1dow1vTHY (visitado 10-12-2013).

[5] *Investor Relations*. [Internet]. Google. URL: http://investor.google.com/ (visitado 10-12-2013).

[6] Kevin Knight. *A Statistical MT Tutorial Workbook*. [Internet]. 1997. URL: http://www.isi.edu/natural-language/mt/wkbk.rtf.

[7] *Language Immersion for Chrome*. [Google Chrome; Plugin]. Use All Five. 1 de mayo de 2012. URL: https://chrome.google.com/webstore/detail/language-immersion-for-ch/bedbecnakfcpmkpddjfnfihogkaggkhl (visitado 10-12-2013).

[8] Franz Och. «Breaking down the language barrier—six years in». En: *Google Official Blog* (26 de abr. de 2012). [Internet]. URL: http://googleblog.blogspot.com/2012/04/breaking-down-language-barriersix-years.html (visitado 10-12-2013).

[9] Neal Ungerleider. «Google Translate's Gender Problem (And Bing Translate's And Systran's...)» En: *Fast Co.Labs* (28 de mayo de 2013). [Internet]. URL: http://www.fastcolabs.com/3010223/google-translates-gender-problem-and-bing-translates-and-systrans (visitado 10-12-2013).

[10] Wikipedia. *Google Translate*. [Internet]. Wikipedia, The Free Encyclopedia. URL: http://en.wikipedia.org/w/index.php?title=Google_Translate&oldid=576358172 (visitado 10-12-2013).

[11] Wikipedia. *reCAPTCHA*. [Internet]. Wikipedia, la enciclopedia libre. URL: http://es.wikipedia.org/w/index.php?title=ReCAPTCHA&oldid=70439248 (visitado 10-12-2013).

[12] Wikipedia. *Statistical machine translation*. [Internet]. Wikipedia, The Free Encyclopedia. URL: http://en.wikipedia.org/w/index.php?title=Statistical_machine_translation&oldid=583124831 (visitado 10-12-2013).