Traducción automática neuronal

Francisco Casacuberta Nolla
Álvaro Peris Abril

Francisco Casacuberta Nolla
fcn@prhlt.upv.es
ORCID:
0000-0002-8497-5598

Álvaro Peris Abril
lvapeab@fiv.upv.es
ORCID:
0000-0002-1818-8377

Resumen

La traducción automática estuvo dominada desde el principio por los sistemas basados en el conocimiento lingüístico, posteriormente se abrieron paso otras aproximaciones, tales como las memorias de traducción y los sistemas estadísticos de traducción, que extraían el conocimiento de corpus paralelos. Recientemente, los modelos neuronales constituyen el estado de vanguardia de la traducción automática. Numerosas empresas de traducción y conocidas páginas web están utilizando estas tecnologías con éxito. Un modelo neuronal es un tipo de modelo estadístico formado por un conjunto de unidades de proceso simple densamente conectadas entre si. Los parámetros de estos modelos se estiman a partir de corpus paralelos gracias a eficientes algoritmos de aprendizaje automático y a potentes procesadores gráficos. La aplicación de los modelos neuronales a la traducción automática obliga a que las palabras se representen en forma de vectores y que para procesar frases se utilicen redes neuronales recurrentes.

Palabras clave: traducción automática; traducción asistida; traducción automática neuronal

Resum

La traducció automàtica va estar dominada des del començament pels sistemes basats en el coneixement lingüístic, posteriorment es van obrir pas altres aproximacions, tals com les memòries de traducció i els sistemes estadístics de traducció, que extreien el coneixement de corpus paral·lels. Recentment, els models neuronals constitueixen l'estat de vanguarda de la traducció automàtica. Nombroses empreses de traducció i conegudes pàgines web estan fent servir aquestes tecnologies amb èxit. Un model neuronal és un tipus de model estadístic format per un conjunt d'unitats de procés simple densament connectades entre elles. Els paràmetres d'aquests models s'estimen a partir de corpus paral·lels gràcies a algoritmes d'aprenentatge automàtic eficients i a processadors gràfics potents. L'aplicació dels models neuronals a la traducció automàtica obliga a què les paraules es representen en forma de vectors i que per tal de processar frases s'utilitzen xarxes neuronals recurrents.

Paraules clau: traducció automática; traducció assistida; traducció neuronal

Rebuda: 28 de juliol de 2017 / Acceptació: 16 de novembre de 2017 / Publicació: 22 de desembre de 2017

Abstract

From the outset, automatic translation was dominated by systems based on linguistic information, but then later other approaches opened up the way, such as translation memories and statistical machine translation which draw on parallel language corpora. Recently the neuronal machine translation (NMT) models have become the cutting edge in automatic translation and many translation agencies and well-known web pages are successfully using these technologies. One NMT model is a kind of statistical model comprising a group of simple deeply interconnected process units. The parameters of these models are estimated from parallel corpora using efficient automatic learning algorithms and powerful graphic processors. Applying these neural models to automatic translation requires words to be represented in the form of vectors and use recurrent neural networks in order to process phrases.

Keywords: machine translation; computer assisted translation; neural translation.

1. Introducción

En la actualidad se producen cantidades ingentes de texto que deben ser traducidos a otras lenguas con cierta rapidez. Estas tareas de traducción pueden ser llevadas a cabo por profesionales, los cuales garantizan la calidad de las traducciones, pero el proceso es generalmente lento y de elevado coste. Esto ha obligado a la (semi)automatización de estos procesos mediante el uso de sistemas de {*traducción automática* (TA o MT del inglés "machine translation"). Un ejemplo de la importancia de la TA reside en el volumen de palabras que genera la industria de la TA y que es de más de 100 mil millones de palabras diarias.[1]

La TA es, según la Wikipedia, "Un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro". El diccionario Oxford en español la define como "traducción realizada por computadores o máquinas adecuadas para este fin". En éstas y en cualquier otra definición, aparece de forma implícita o explícita el uso de los computadores y por tanto la necesidad de disponer de unos modelos precisos y de unos algoritmos eficientes.

En los años 50, se propusieron modelos estadísticos sencillos para abordar el problema de la TA, pero la (muy) limitada potencia de los computadores de la época no permitió el desarrollo de estos modelos. Más tarde, apareció la *traducción basada en el conocimiento* (TBC) lingüístico, en la que los modelos son reglas de un "sistema experto" y el proceso de traducción consiste en aplicar las reglas necesarias sobre una frase dada para obtener la traducción. El principal problema de las tecnologías basadas en reglas reside en la gran dificultad de formalizar el conocimiento lingüístico humano mediante reglas precisas. Otro problema es el elevado coste que implica la construcción y mantenimiento de tales sistemas así como la correspondiente adaptación a nuevos dominios o pares de lenguas.

Una alternativa a la TBC la constituyó la traducción basada en ejemplos o en corpus, siendo las *memorias de traducción* (Lagoudaki, 06) la aproximación más

[1] "Ten years of Google Translate", <https://www.blog.google/products/translate/ten-years-of-google-translate/ >

simple. Por otra parte, la *traducción automática estadística* (TAE o SMT del inglés "statistical machine translation") resurgió en los 90 con modelos estadísticos complejos que podían ser implementados en los computadores de esos años (Koehn, 2009). Estos modelos probabilísticos requieren grandes corpus bilingües y han constituido el estado del arte de la TA hasta hace poco tiempo. La hipótesis de trabajo en este marco es que cualquier frase de la lengua destino es una posible traducción de una frase origen dada. El reto consiste en utilizar los datos del corpus de entrenamiento para conseguir que los modelos probabilísticos asignen una probabilidad alta a aquellas frases destino que son traducciones reales de la frase dada y una probabilidad baja o nula a aquellas frases que no lo son.

Matemáticamente, dada una frase origen **e**, la TAE consiste en encontrar una frase destino **ŝ** que verifique la ecuación (1):

$$\hat{s} = \underset{s}{\operatorname{argmax}} \Pr_\theta(s \mid e) \qquad \textit{Ecuación (1).}$$

esto es, la frase destino **s** que maximice la probabilidad de que **s** sea una traducción de la frase origen dada **e**. Esta probabilidad se calcula utilizando unos parámetros Θ que se estiman automáticamente a partir de un corpus de entrenamiento (Koehn, 2009).

Una de las herramientas más populares para la construcción y uso de sistemas de TAE ha sido Moses (Koehn, 2009), siendo muy utilizada por empresas de servicios de traducción.

Otra alternativa en la TA está basada en *redes neuronales artificiales* (RNA) y *aprendizaje profundo* para calcular la probabilidad de la ecuación (1). Aunque esta aproximación fue inicialmente muy prometedora (Castaño y Casacuberta, 1997) no ha sido hasta la actualidad que la denominada *traducción automática neuronal* (TAN o NMT del inglés "neural machine translation") ha demostrado su eficacia gracias a las nuevas arquitecturas neuronales, nuevos componentes y nuevos procesadores (Kalchbrenner y Blunson, 2013; Cho *et al.*, 2014; Sutskever, Vinyals y Le, 2014; Koehn, 2017). A continuación se describe brevemente la tecnología en la que se basa la TAN.

2. Traducción automática neuronal

La TAN es una aproximación a la TA basada en corpus que, en muchos casos, está proporcionando mejores resultados que la TAE clásica. Una característica de la TAN es que las palabras y las frases son representadas de forma numérica mediante vectores (Bengio *et al.*, 2003) mientras que en las otras aproximaciones la representación era discreta. Este hecho ha permitido el uso de potentes técnicas de aprendizaje automático (ML del inglés "machine learning") como las redes neuronales.

Una red neuronal se compone de un conjunto de unidades de procesamiento simple (o neuronas artificiales) densamente conectadas entre sí y cuya función es realizar un producto escalar entre las entradas a la neurona y un vector de pesos (asociado a cada neurona) seguido de una función no lineal de activación (Bishop, 2006). La red neuronal más popular se conoce como *perceptrón multicapa* y se compone de capas de neuronas de forma que las salidas de una capa son las entradas de la capa siguiente. Por otra parte, existen las denominadas *redes neuronales recurrentes* (Elman, 1990; Jordan, 1990), en las que las neuronas se realimentan con sus propias salidas de forma directa o indirecta. Esto convierte a las redes neuronales recurrentes en potentes modelos de secuencias temporales. Los pesos de las conexiones de estas redes se estiman a partir de corpus de entrenamiento mediante una extensión de un algoritmo muy popular denominado *algoritmo de retro-propagación del error* (Rumelhart, Hilton y Williams, 1986).

Formalmente, la idea básica de la TAN parte de la siguiente descomposición de la ecuación de la ecuación (2):

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} \prod_{i=1}^{|\mathbf{s}|} \Pr_\theta(s_i \mid s_1^{i-1}, c(\mathbf{e}))$$ *Ecuación (2).*

donde cada palabra traducida $\mathbf{s_i}$ depende de las palabras $\mathbf{s}_1^{i-1}$ previamente traducidas de una cierta representación de la frase origen **e** (denotada por $c(\mathbf{e})$) y de los parámetros del modelo de traducción Θ.

La arquitectura más utilizada en la actualidad para implementar la ecuación (2) está basada en un codificador seguido de un decodificador (ver fig. 1) (Bahdanau, Cho y Bengio, 2014). El codificador es una red neural (recurrente) que analiza de izquierda a derecha y de derecha a izquierda (Schuster y Paliwal, 1997) la frase origen para producir una representación vectorial de la misma. El decodificador genera la frase destino condicionado por la frase origen. Esta última parte es otra red neuronal recurrente que, en un instante dado, genera una palabra destino tomando como entrada la palabra previamente generada, el estado de la red neuronal en el instante anterior y una representación de la frase origen ($c(\mathbf{e})$ en la ecuación (2)) obtenida por un alineador (denominado *modelo de atención*) a partir de la información suministrada por el codificador. El modelo de atención es a su vez otra red neuronal, en este caso un perceptrón multicapa, que permite alinear las palabras destino con las palabras de la frase origen (Bahdanau, Cho y Bengio, 2014}. Tanto el codificador como el decodificador son redes recurrentes sofisticadas, del tipo LSTM ("long short term memory") (Hochreiter y Schmidhuber, 1997) o GRU ("gated recurrent unit") (Bahdanau, Cho y Bengio, 2014).

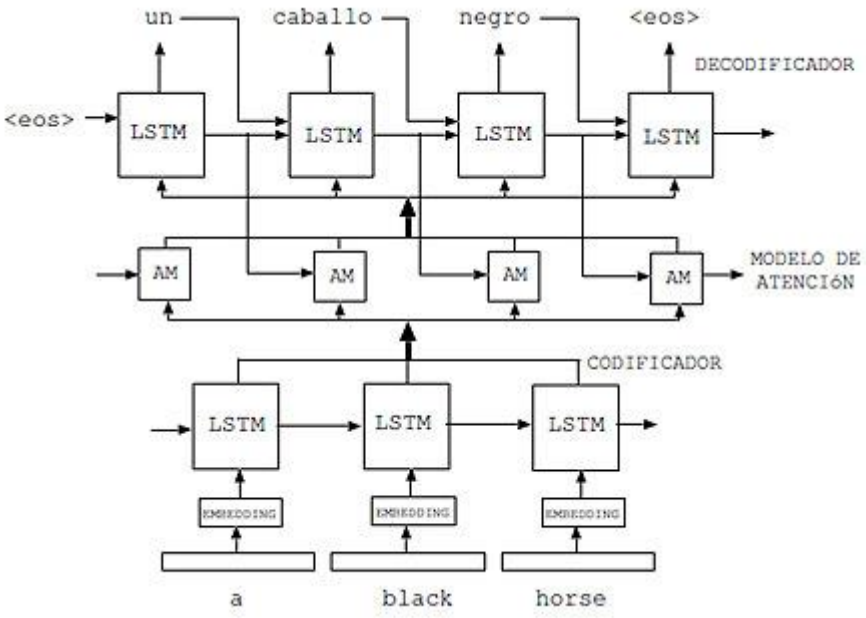

Fig. 1: Ejemplo de arquitectura codificadora-decodificadora para traducir "a black horse" a "un caballo negro".

El codificador en la figura representa tres instancias de la misma red recurrente para analizar la frase de izquierda a derecha (en este caso solo una dirección para simplificar la figura). El decodificador es otra red recurrente que genera la traducción. Cada palabra traducida se obtiene a partir de las palabras traducidas anteriores y de una cierta representación de la frase de entrada dada por el modelo de atención a partir de la representación generada por el codificador.

Una de las características fundamentales de esta arquitectura es que los parámetros de la misma, Θ (los pesos asociados a las conexiones de las unidades), se estiman (entrenan) conjuntamente (los de la conversión de las palabras a vectores, los del codificador, los del decodificador y los del modelo de atención) mediante la extensión del algoritmo de retro-propagación del error.

Una vez estimados los parámetros del TAN, el problema consiste en generar la traducción de una frase origen dada. La solución comúnmente adoptada consiste en ir generando, en cada instante y según la ecuación (2), varias palabras alternativas lo que da lugar a un espació de búsqueda de la traducción en forma de árbol. Este espacio de búsqueda puede crecer de forma prohibitiva, pero se puede controlar mediante técnicas denominadas de búsqueda en haz (esto es eliminando aquellas ramas de árbol que no sean prometedoras) (Sutskever, Vinyals y Le, 2014).

El proceso de entrenamiento y el de traducción son computacionalmente muy costosos y es necesario el uso de procesadores gráficos (GPUs) que permiten realizar operaciones matriciales a gran velocidad. También se dispone de herramientas software para el entrenamiento y la traducción como Tensorflow, GroundHog, Keras, etc. (Abadi *et al.*, 2016).

En una importante conferencia sobre traducción como es la del AMTA de 2016 se presentó una comparativa entre la TAE y la TAN en la que se observó que los resultados de esta última superan a la primera en muchos pares de lenguas (Sennrich, Birch y Junczys-Dowmunt, 2016). También es de destacar que importantes empresas como Systran, Google o Wipo han anunciado la construcción de sus propios traductores neuronales (Crego *et al.*, 2016; Wu *et al.*, 2016).

3. Aplicaciones de los traductores neuronales

La arquitectura neuronal descrita anteriormente tiene multitud de aplicaciones además de la propia en TA. Por un lado está la *traducción asistida por ordenador* (TAO) donde la TA es utilizada como una herramienta más. La idea más elemental consiste en utilizar técnicas de TA para producir traducciones que posteriormente son revisadas y corregidas por un traductor humano en un proceso de *posedición* (PE). Una idea más sofisticada es la *traducción interactiva* (TI) donde el traductor humano influye en el proceso de traducción mediante la corrección de los errores que va detectando y que el sistema aprovecha para mejorar la traducción en curso y las posteriores (Barrachina *et al.*, 2008; Casacuberta *et al.*, 2008). Originalmente se utilizó la TAE en la TI, pero el uso de la TAN ha permitido diseñar sistemas que superan claramente a los primeros (Peris, Domingo y Casacuberta, 2017). Por otra parte, cada vez que el humano valida una traducción se dispone de una nueva muestra de aprendizaje que puede servir para mejorar los modelos de traducción y evitar la repetida aparición de un mismo error. Esto se realiza mediante un proceso de adaptación por *aprendizaje en* línea (Cettolo *et al.*, 2013; Peris, Cebrián y Casacuberta, 2017).

Otra de las aplicaciones de la arquitectura neuronal codificadora-decodificadora es la traducción multimodal, donde una frase origen se traduce con la ayuda de una imagen relacionado con el texto a traducir (Caglayan, Barrault y Bougares, 2016}. Recientemente también se ha aplicado a la traducción del habla (Bansal *et al.*, 2017).

Fuera del ámbito de la TA, también se puede encontrar aplicaciones de la mencionada arquitectura en otros ámbitos como la descripción de imágenes y vídeos (Xu *et al.*, 2015; Yao *et al.*, 2015) (en este caso se sustituye la red recurrente del codificador por una red convolucional) o la generación de respuestas a preguntas sobre los contenidos en imágenes o videos (Antol *et al.*, 2015), etc.

4. Conclusiones

Aunque no se ha cumplido el sueño de que la TA de alta calidad sea una realidad hoy en día, lo cierto es que las tecnologías desarrolladas, y sobre todo las basadas en TAN, han avanzado lo suficiente como para poder disponer de sistemas de TA de utilidad práctica en muchas aplicaciones. No obstante, en los casos en los que se exige una calidad alta, también disponemos de tecnologías TAC basadas en TAN en las que el humano y la máquina colaboran para conseguir la calidad exigida con un

esfuerzo humano significativamente inferior al que sería necesario si el humano tuviera que realizar su trabajo sin ningún tipo de asistencia.

Bibliografía

Abadi, M. [et al.] (2016). *Tensorflow: Large-scale machine learning on heterogeneous distribut ed systems*. Cornell University Library. <arXiv:1603.04467>. [Consulta: 24 de julio de 2017].

Antol, S. [et al.] (2015). "VQA: Visual question answering". En: Balasubramanian, R. [et al.]. (eds.). *Proceedings of the International Conference on computer computer vision and image processing*. New York: Springer, p. 2425–2433.

Barrachina, S. [et al.] (2009). "Statistical approaches to computerassisted translation," *Computational Linguistics,* v. 35, n. 1 (March), p. 3–28. <http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2008.07-055-R2-06-29>. [Consulta: 13 de octubre de 2017].

Bansal, S.; Kamper, H.; Lopez, A.; Goldwater, S. (2017). "Towards speech-to-text translation without speech recognition". *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Cornell University Library.* <arXiv:1702.03856v1>. [Consulta: 19 de octubre de 2017].

Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. (2003). "A neural probabilistic language model". *Journal of Machine Learning Research,* v. 3, p. 1137–1155. <http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>. [Consulta: 13 de octubre de 2017].

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer. (Information science and statistics).

Casacuberta, F. [et al.] (2008). "Human interaction for high quality machine translation". *Communications of the ACM,* v. 52, n. 10, p. 135–138. <https://dl.acm.org/citation.cfm?id=1562798>. [Consulta: 13 de octubre de 2017].

Castaño, M. A.; Casacuberta, F. (1997). "A connectionist approach to machine translation". En: *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation.* <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=96F8792ED52CC84E61BBFEC9513C6DAD?doi=10.1.1.466.8488&rep=rep1&type=pdf>. [Consulta: 23 de julio de 2017].

Caglayan, O.; Barrault, L.; Bougares, F. (2016). *Multimodal attention for neural machine translation*. Cornell University Library. <arXiv:1609.03976v1>. [Consulta: 18 de octubre de 2017].

Cettolo, M. [et al.] (2013). "Issues in incremental adaptation of statistical mt from human post-edits". En: *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, p. 111–118. https://hal.archives-ouvertes.fr/hal-01158054/document>. [Consulta: 18 de octubre de 2017].

Cho, K. [et al.] (2014). "Learning phrase representations using RNN encoder–decoder for statistical machine translation". En: *Proceedings of the 2014 Conference on Empirical*

Methods in Natural Language Processing, p. 1724–1734. Cornell University Library. <https://arxiv.org/abs/1406.1078>. [Consulta: 18 de octubre de 2017].

Crego, J. [et al.] (2016). *SYSTRAN's: pure neural machine translation systems*. Cornell University Library. <arXiv:1610.05540>. [Consulta: 17 de mayo de 2017]

Elman, J. L. (1990). "Finding structure in time". *Cognitive Science*, v. 14, n. 2, p. 179–211. <http://psych.colorado.edu/~kimlab/Elman1990.pdf>. [Consulta: 18 de octubre de 2017].

Hochreiter, S.; Schmidhuber, J. (1997). "Long short-term memory". *Neural Computation*, v. 9, n. 8, p. 1735–1780. <http://www.bioinf.jku.at/publications/older/2604.pdf>. [Consulta: 18 de octubre de 2017].

Jordan, M. I. (1990). "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine". En: Diederich, J. (ed.). *Artificial neural networks.* IEEE Press, p. 112-127.

Kalchbrenner, N.; Blunsom, P. (2013). "Recurrent continuous translation models". En: *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* p. 1700–1709. <https://dl.acm.org/citation.cfm?id=1613715&picked=prox>. [Consulta: 18 de octubre de 2017].

Koehn, P. (2009). *Statistical Machine Translation.* Cambridge [etc.]: Cambridge University Press.

Koehn, P. (2017). "Statistical Machine Translation". [Draft of chapter 13] *Neural Machine Translation.* <https://arxiv.org/pdf/1709.07809v1.pdf>. [Consulta: 18 de octubre de 2017].

Lagoudaki, E. (2006). *Translation memory systems: Enlightening users' perspective.* Imperial College London.

Peris, A.; Domingo, M.; Casacuberta, F. (2017). "Interactive neural machine translation". *Computer Speech and Language,* v. 45 (September), p. 201–220. <http://www.sciencedirect.com/science/article/pii/S0885230816301000>, <https://doi.org/10.1016/j.csl.2016.12.003>. [Consulta: 18 de octubre de 2017].

Peris, A.; Cebriań, L.; Casacuberta, F. (2017). *Online learning for neural machine translation post-editing*. Cornell University Library. <arXiv:1706.03196v1>. [Consulta: 18 de octubre de 2017].

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. (1986). "Learning Internal Representations by Error Propagation". En: Rumelhart, D. E.; McClelland, James. J.; PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge [etc.]: MIT Press. (Computational models of cognition and perception), v. 1, p. 318-362.

Sennrich, R.; Birch, A.; Junczys-Dowmunt, M. (2016). *Advances in neural machine translation*. [Slides to] The Twelfth Conference of The Association for Machine Translation in the Americas (AMTA). <http://homepages.inf.ed.ac.uk/rsennric/amta2016-tutorial.pdf>. [Consulta: 18 de octubre de 2017].

Sutskever, I.; Vinyals, O.; Le, V. (2014). "Sequence to sequence learning with neural networks". En: *[Electronic Proceedings of] Advances in Neural Information Processing*

Systems, n. 27, p. 3104–3112. <https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>. [Consulta: 18 de octubre de 2017].

Schuster, M.; Paliwal, K. K. (1997). "Bidirectional recurrent neural networks". *IEEE Transactions on Signal Processing*, v. 45, n. 11 (November), p. 2673–2681. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.9441&rep=rep1&type=pdf>. [Consulta: 22 de octubre de 2017].

Theano Development Team. *Theano: A Python framework for fast computation of mathematical expressions*. Cornell University Library. <arXiv:1605.02688v1>. [Consulta: 22 de octubre de 2017].

Wu, Y. [et al.] (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.* Cornell University Library <arXiv:1609.08144>. [Consulta: 22 de octubre de 2017].

Xu, K. [et al.]. (2015). *Show, attend and tell: Neural image caption generation with visual attention*. Cornell University Library. <arXiv:1502.03044>. [Consulta: 22 de octubre de 2017].

Yao, L. [et al.]. (2015). *Describing videos by exploiting temporal structure*. Cornell University Library. <arXiv:1502.08029>. [Consulta: 22 de octubre de 2017].