Traducción automática

Mateo Gonzalez Ocampo Juan Alejandro Uribe Ramirez

Junio 2020

1. Introducción

La traducción automática se refiere al uso de computadores para realizar traducciones de un lenguaje SL (*Source Language*) a un lenguaje TL (*Target Language*). Los origenes de esta area del conocimiento se remontan a los años 30, con los trabajos de Georges Artsrouni and Petr Trojanskij, en los que patentaron sistemas que podrían ser usados como traductores [1]. Luego, durante los años 40, con la creación de los primeros computadores, se consideró el uso de estos como traductores, lo cual concluyo en lo que se conoce hoy como el experimento de Georgetown[2], en el que fue posible traducir mas de 60 oraciones del Ruso al Ingles, lo que genero altas expectativas dentro de la comunidad cientifica y en el gobierno americano, por su posible uso en la guerra fria.

Hoy en día, el uso de software que usa traducción automatica esta ampliamente extendido en la población no solo cientifica, sino en general. Un ejemplo claro de este es el traductor de Google.

En el siguiente trabajo se realiza una revisión de algunos de los métodos que se han usado históricamente para resolver el problema de la traducción automática. Los métodos seleccionados fueron la traducción basada en reglas, basada en ejemplos y basada en redes neuronales. Para cada método se presenta una pequeña introducción histórica, un análisis de como funciona y aplicaciones destacadas.

2. Métodos

2.1. Reglas

Los sistemas basados en reglas constituyen los primeros intentos usados para la creación de traductores automáticos y fueron ampliamente usados entre la decada de las 60 y la primera parte de los 80. Estos siste-

mas usan un conjunto de reglas escritas a mano para la transformación de SL a TL, ademas de un diccionario para la traducción de palabras.

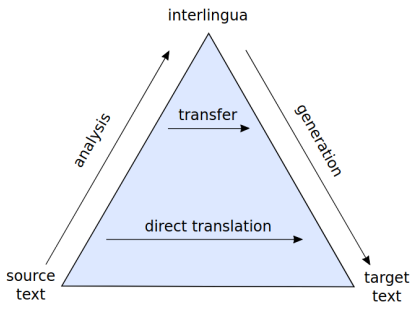

Figura 1: Triangulo de Vauquois.

Se pueden diferenciar los métodos usados en este tipo de sistemas en dos generaciones[1]: La primera, se caracteriza porque la transformación de SL a TL es directa, es decir, se traduce cada palabra por separado, usando un diccionario de SL a TL, y al resultado de esta transformación se le aplican un conjunto de reglas para hacer que la nueva oración sea coherente en TL, lo que puede involucrar una reorganización de las palabras del texto generado. A este tipo de métodos se les conoce como métodos de traducción directa.

La segunda generación de métodos, conocidos como de traducción indirecta, transforman inicialmente el texto en SL a una representación intermedia a partir de la cual se transforma a TL. Existen varios modelos para la generación de dicha representación, siendo los mas conocidos los métodos de interlingua y de transferencia [3][1][4]. En el primero se asume que la representación intermedia es unica para todos los lenguajes, por lo que solo es necesario definir reglas para transformar a este estado intermedio y desde este estado intermedio. Esto, basado en las ideas planteadas por Descartes en el siglo XVII acerca de la existencia de un lenguaje universal. En el segundo método cada lenguaje tiene su propia representación intermedía, lo que implica que ademas de las reglas relacionadas con la representación intermedia, es necesario un conjunto de reglas adicional para transformar de una representación a otra. En la figura 1 se muestra el triangulo de Vauquois, el cual sirve para entender los diferentes niveles de análisis entre los métodos mencionados anteriormente.

SYSTRAN, APERTIUM y GramTrans son algunos ejemplos de aplicaciones que usan sistemas basados en reglas.

2.2. Ejemplos

Los sistemas basados en ejemplos surgieron durante la decada de los 80, debido principalmente al trabajo de Nagao[5], el cual planteaba una nueva forma de traducción automática basada en la idea de que el ser humano no traduce partiendo de análisis lingüísticos complejos, sino des-

componiendo el texto a traducir en fragmentos los cuales son luego traducidos y reorganizados para formar el nuevo texto. De manera similar a los sistemas basados en reglas, los sistemas basados en ejemplos hacen uso de un diccionario de SL a TL junto con una base de datos de ejemplos, la cual se usa para como base para la traducción (En contraste con las reglas usadas en los sistemas basados en reglas). También suele usarse un Tesauro como parte de la base de datos. Un ejemplo esta constituido por dos textos en dos lenguajes, siendo ambos textos traducciones de si mismos[6].

Existen tres etapas en el proceso de traducción mediante ejemplos [5][7]:

1. Como primer paso, se realiza un análisis del texto, para identificar los fragmentos que mejor se relacionen con los textos de la base de datos.

2. Luego, usando los fragmentos obtenidos, se extraen de la base de datos los ejemplos que mas se asemejen a los fragmentos. Los ejemplos se seleccionan con base en alguna métrica que toma ambos textos y asigna un grado de similaridad entre ellos.

3. Finalmente, con base en los ejemplos obtenidos de la base de datos, se genera el texto traducido.

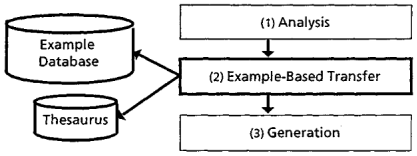

Figura 2: Configuración de un sistema basado en ejemplos Tomado de [7].

Los sistemas basados en ejemplos resuelven algunos de los problemas presentes en los sistemas basados en reglas[7], principalmente la dificultad y complejidad de añadir nuevas reglas a un sistema ya establecido, dado que pueden haber conflictos con las reglas previamente establecidas. Cunei y CMU-EBMT son probablemente de los sistemas basados en ejemplos mas reconocidos.

2.3. Redes neuronales

Los sistemas basados en redes neuronales usan metodos de aprendizaje profundo para intentar obtener la traducción de SL a TL. Si bien las bases matematicas de las redes neuronales se conocen desde finales del siglo XX no fue sino hasta después del año 2013, influenciados principalmente por el trabajo de *Kaichbrenner et al.* [8], que estas empezaron a ser ampliamente estudiadas para su use en el campo de la traducción automática.

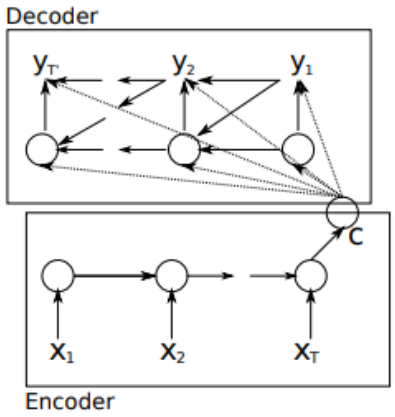

Figura 3: Arquitectura propuesta simulataneamente por *Cho et al.*[9] y *Sutskever et al.*[10]. Tomada de [9]

El metodo propuesto por *Kaichbrenner et al.* consiste en el uso de dos redes neuronales, llamadas codificadora y decodificadora, que a partir de un conjunto de datos de entrenamiento (Similar a la base de datos que se usaba en los sistemas basados en ejemplos) pueden ser usadas para traducir de SL a TL. La red codificadora se encarga de transformar el texto entrada, en SL, a una representación en un espacio vectorial. Este vector resultanto es luego alimentado a la red decodificadora, y esta lo transforma a la traducción en TL. En el trabajo de Kaichbrenner la red neuronal codificadora era una red convolucional y la decodificadora un red recurrente, pero con el paso de los años se han hecho modificaciones al sistema original propuesto por Kaichbrenner [9] [10] (Ver 3). Muchas de las grandes compañias de software internacional como Google, Facebook y Amazon usan este tipo de sistemas dentro de sus productos .

Referencias

[1] W. John Hutchins. Machine translation: A brief history. In E.F.K. KOERNER and R.E. ASHER, editors, *Concise History of the Language Sciences*, pages 431 – 445. Pergamon, Amsterdam, 1995.

[2] John Hutchins. The first public demonstration of machine translation : the georgetown-ibm system , 7 th january 1954. 2006.

[3] Musatafa Albadr, S. Tiun, and Fahad Al-Dhief. Evaluation of machine translation systems and related procedures. *ARPN Journal of Engineering and Applied Sciences*, 13:3961–3972, 06 2018.

[4] Cheragui Mohamed Amine. Theoretical overview of machine translation. *CEUR Workshop Proceedings*, 867:160–169, 01 2012.

[5] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the International NATO Symposium*

on Artificial and Human Intelligence, page 173–180, USA, 1984. Elsevier North-Holland, Inc.

[6] Haihua Pan. Example-based machine translation : A new paradigm. 2002.

[7] Eiichiro Sumita, Hitoshi Iida, and Hideo Kohyama. Translating with examples: A new approach to machine translation. In *Proceedings Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, 1990.

[8] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215, 2014.