Centro de Investigación en Matemáticas

Unidad Monterrey

Proyecto final.

Temas selectos de ciencia de datos

Machine Translation para lenguajes con pocos recursos.

Ricardo Cruz Sánchez
Rolando Corona Jiménez

Índice

1. Introducción.

Una de las tareas más relevantes en el procesamiento del lenguaje natural (NLP) es la traducción automática o machine translation (MT), la cual consiste en poder traducir textos de una lengua (origen) a otro idioma (objetivo) a través de software especializado.

El enfoque que se tenía antes del desarrollo del deep learning, era basado en reglas (RBMT) que posteriormente fue remplazado por el enfoque estadístico (SMT) el cual resulto bastante exitoso en los años 90's.

El insumo principal de un modelo de MT es un corpus paralelo, el cual representa textos en dos (o más) idiomas. Estos textos deben presentarse en el idioma origen y su correspondiente traducción al idioma objetivo. Esta traducción puede ser por frase, capítulos o documentos completos.

Los modelos más sofisticados en la actualidad y que pueden considerarse como estado del arte utilizan el enfoque SMT. Usualmente utilizan la estructura conocida como *encoder-decoder* y algunas de sus variantes. Encoder-decoder se sustenta en subestructuras de redes recurrentes neuronales (RNN) y long-short term memory (LSTM)

Sin embargo, el desempeño de estos modelos recae mucho sobre la cantidad de datos disponibles.

Hoy en día, a pesar de que existen miles de lenguas en el mundo, se estima que hasta el 90 % de ellas no están contenidas en la web, la cual representa la principal fuente para la creación de un corpus paralelo, además de que la existencia de lenguas predominantes, por ejemplo, el ingles genera que los trabajos solo se desarrollen para esos idiomas.

Por lo anterior, crear un traductor de ciertas lenguas puede tornarse una tarea sumamente complicada.

El primer obstáculo es encontrar un corpus paralelo, el cual en caso de no existir, se podría llegar a crear utilizando más fuentes de información, pero, la complejidad de está tarea puede resultar tan grande como la traducción misma.

La segunda problemática y en la que se realizará un énfasis en el presente trabajo es, tener un corpus paralelo, pero no lo suficientemente grande para poder tener resultados exitosos si se implementará el estado del arte.

Existen diversas alternativas para la solución de esta problemática, que van desde el uso de más recursos hasta recurrir a especialistas para el análisis más detallado de la lengua.

Este documento centrará la atención en el uso de *transfer learning* como posible solución a la traducción automática con pocos recursos y se ejemplificará con un caso de estudio para facilitar el entendimiento del tema.

2. Machine Translation.

Neural machine translation (NMT) es el estado del arte en tareas de traducción, ha sido ampliamente estudiado y muestra un gran desempeño cuando se posee un número considerable de recursos.

La estructura principal en esta rama de NLP se conoce como encoder-decoder. La figura 2 muestra las componentes de la red. Originalmente el encoder-decoder, no poseía el mecanismo de atención, pero se ha mostrado que incluirlo puede significar mejoras en el desempeño.

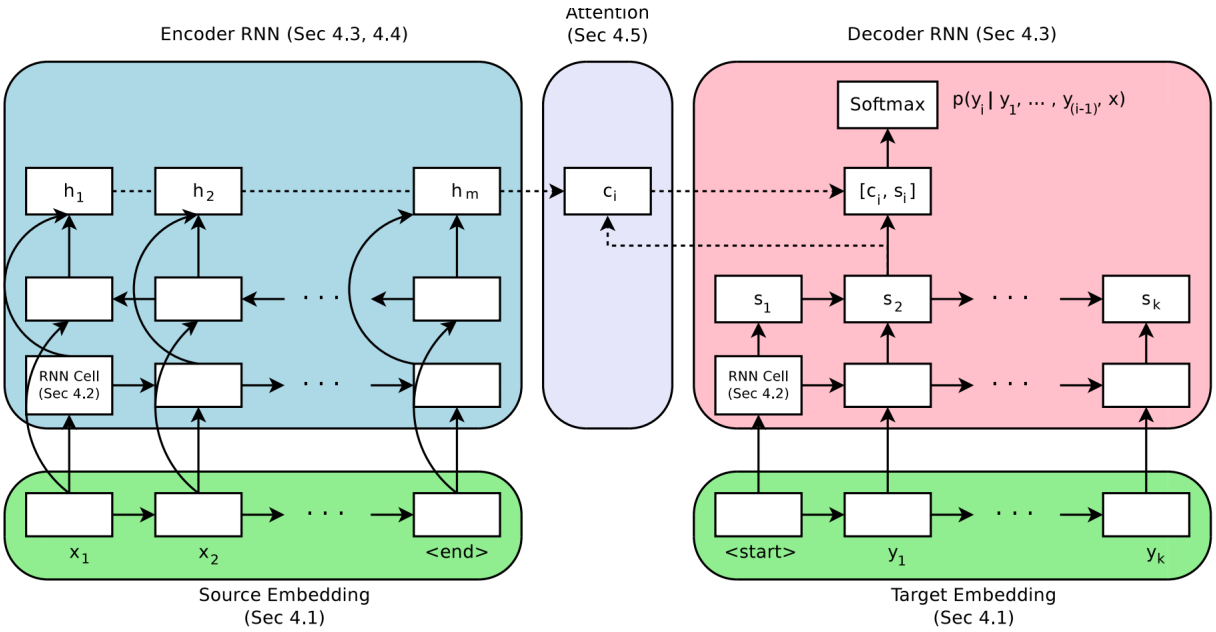

Figura 1: Modelo encoder-decoder attention.

- **Embeddings:** Las componentes mostradas en color verde corresponden a los embeddings del lenguaje origen y el lenguaje objetivo. A grandes rasgos, los embeddings representan a cada palabra de un idioma como un vector de dimensión fija.
- **Encoder:** La caja de color azul es el encoder. Es una red neuronal recurrente, la cual se alimenta de la representación vectorial de las palabras en el idioma origen. Usualmente se utiliza una capa LSTM para el encoder. Su tarea es representar de manera vectorial a toda la frase que represente las características de cada frase. A estos vectores, de longitud fija, se les conoce como vector resumen
- **Decoder:** La caja de color rosa es el decoder. Al igual que el encoder, es una red neuronal recurrente, la cual se alimenta de la representación vectorial de las palabras en el idioma objetivo y del output del encoder. Usualmente se utiliza una capa LSTM para el encoder. El objetivo del decoder es generar una frase a partir del vector resumen y las palabras del idioma objetivo.

- **Attention:** La caja de color morado es la capa conocida como atenttion. Este mecanismo permite cambiar el vector de contexto por múltiples vectores de anotaciones, también de longitud fija. Los *annotation vectors* representan a cada palabra de la frase original, en lugar de ocupar el vector que resume a toda la frase. Los vectores de anotación generan un vector nuevo que se actualiza cada que se conoce una palabra, este vector se llama vector de contexto y alimenta al decoder.

El desempeño de este modelo se ha centrado en corpus en los cuales se poseen millones de frases y vocabularios bastante extensos.

Cuando el corpus es escaso o incluso ni siquiera existe, pueden surgir varias alternativas, en las que no puede determinarse cual es la mejor, pues cada una dependerá de los datos disponibles y la naturaleza de las lenguas que se manejen.

Las alternativas más conocidas son:

- **Transfer learning:** A grandes rasgos, consiste en encontrar un corpus de un tercer idioma, del que si se posean recursos suficientes, además de que comparta el idioma objetivo, si se satisface las condiciones se procede a realizar MT con el idioma con recursos, a este idioma se le conocerá como idioma padre. Posteriormente, los parámetros del encoder se heredan al idioma con pocos recursos (idioma hijo) y se realiza MT.

- **Pivoteo:** La idea general de este procedimiento, nace al querer realizar MT sobre un idioma origen y un objetivo del cual no se poseen corpus, pero, existe un tercer idioma (pivote) del que se poseen corpus extensos con el idioma origen y el objetivo, por lo que, primero se realiza MT del idioma origen al pivote y después MT del idioma pivote al objetivo

- **Zero shot** Es un enfoque similar al pivoteo, pero en lugar de pasar por la pseudo-traducción con el idioma pivote intermedio, solo ocupa el idioma para generar un corpus extenso y poder realizar MT de manera *directa*

- **Estudio más detallado de la lengua con pocos recursos:** En ocasiones se sugiere descomponer las frases en las formas más simples posibles, para así poder realizar una correspondencia efectiva entre el idioma origen y el objetivo. Esto puede mejorar la traducción pero requiere de mayor conocimiento del los lenguajes y de su propia naturaleza, tareas que corresponden directamente a lingüistas y traductores.

- **Modificación de parámetros y capas:** En la actualidad, existen enfoque que pretenden mostrar que a diferencia de los anteriores, no es necesario recurrir a un idioma auxiliar, sino que, basta con modificar los parámetros de la estructura encoder-decoder para generar un traductor eficiente. Las estrategias incluyen considerar tamaños de batches pequeños, dropouts agresivos, regularización en las capas y tomar embeddings diferentes, como el node2vec.

En el presente trabajo se abordará solo transfer learning como posible solución a los problemas de bajos recursos. Las razones de esta elección se basan en la cantidad de información

disponible para esta metodología, además de no requerir de algún especialista.

Información de cualquier otra metodología se puede encontrar en la red y en la bibiliografía recomendada.

3. Transfer learning

4. Descripción del conjunto de datos.

5. Conclusiones.

Referencias

[1] Lorrie Faith Cranor and Brian A. LaMacchia. 1998. Spam!. Commun. ACM 41, 8 (August 1998), 74-83.

[2] Emilio Ferrara. 2019. The history of digital spam. Commun. ACM 62, 8 (July 2019), 82-91.

[3] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..

[4] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (July 2009), 427-437.