Uso de BERT para sistemas Preguntas Respuestas.

Amalia Ibarra Rodríguez
Gabriela B. Martínez Giraldo
Sandra Martos Llanes

Universidad de La Habana,
La Habana, Cuba
Grupo C-512

1 Introducción.

Dado el reciente crecimiento acelerado del internet y, con él, la existencia de una cantidad exorbitante de información disponible en la web, se hace cada vez más y más necesaria la implementación de herramientas, como sistemas de búsqueda y de preguntas respuestas automáticos, que permitan obtener resultados válidos para preguntas dadas en lenguaje natural.

Dicho esto, responder preguntas puede ser una tarea trivial para los humanos, pero no lo es tanto para una máquina. Para responder a cualquier pregunta, las máquinas deben superar muchos desafíos diferentes, como la brecha léxica, la resolución de correferencias, la ambigüedad del idioma, etc.

Este tema ha sido ampliamente estudiado en los últimos años. Numerosos sistemas han sido desarrollos y perfeccionados, a tal punto que ya se hace difícil incursionar en este ámbito y realizar aportes considerables.

Fundamentalmente se consideran dos tipos de sistemas: los de dominio cerrado que, por lo general, se centran en un tema en particular, dígase medicina, comida, deportes, etc; y los de dominio abierto, que permiten hacer preguntas más generales. Los segundos son probablemente los que más complejos resulten de abordar, dado el volumen tan elevado de datos que requieren procesar para lograr un buen desempeño.

Comúnmente, para tratar este tipo de problemas, se combinan técnicas de recuperación de información (information retrieval aka. IR), extracción de información (information extraction aka. IE) y procesamiento de lenguaje natural o PLN (natural language procesing aka. NLP)

En el presente trabajo se pretende hacer un acercamiento a los sistemas de preguntas respuestas automáticos. Se pretende con él hacer una caracterización general de los mismos y analizar las principales técnicas que se utilizan en la actualidad para atacar este tipo de problemas.

2 Sistemas Preguntas Respuestas

Para los problemas de preguntas respuestas automáticos se utilizan generalmente modelos de machine learning que pueden responder preguntas en cierto contexto

y, a veces, sin ningún contexto necesario (como por ejemplo hacen los sistemas de dominio abierto). Pueden extraer respuestas de párrafos, parafrasear la respuesta generativamente, elegir una opción de una lista de opciones dadas, y así sucesivamente. Todo depende del conjunto de datos en el que se entrenó, o el problema para el que se entrenó, o hasta cierto punto, la arquitectura de la red neuronal.

Dichos modelos deben entender la estructura del idioma, tener una comprensión semántica del contexto y las preguntas, tener la capacidad de ubicar la posición de una frase de respuesta y mucho más. Entonces, sin duda, es difícil entrenar modelos que realicen estas tareas. Afortunadamente, el concepto de atención en las redes neuronales ha sido un salvavidas para tareas tan difíciles. Desde su introducción para tareas de modelado de secuencias, muchas redes RNN con mecanismos de atención sofisticados han mostrado una gran mejora en las tareas de control de calidad. Sin embargo, una arquitectura de red neuronal completamente nueva basada en la atención, específicamente la autoatención, llamada Transformer, ha sido el verdadero cambio de juego en el PNL.

Las diferentes variantes de Transformers, con su capacidad para procesar tokens en paralelo y su impresionante rendimiento debido al mecanismo de autoatención y diferentes objetivos de preentrenamiento, han hecho que el entrenamiento de modelos grandes (y a veces modelos realmente grandes), que entiendan el lenguaje natural realmente posible. Diferentes modelos de lenguaje basados en Transformer, con pequeños cambios en su arquitectura y objetivo de entrenamiento previo, se desempeñan de manera diferente en diferentes tipos de tareas. BERT (Representaciones de codificador bidireccional de transformadores) es uno de esos modelos. BERT ha sido entrenado utilizando la arquitectura Transformer Encoder, con Masked Language Modelling (MLM) y el objetivo de entrenamiento previo Next Sentence Prediction (NSP).

3 Datasets

En aras de lograr que los sistemas que se desarrollen alcancen una preparación adecuada a la tarea requerida, históricamente se han utilizado grandes conjuntos de datos poblados con valores reales para el entrenamiento o bien datos sintéticos, que permitan entrenar los modelos en el área de la comprensión de textos. Entre los datasets más utilizados para la construcción de sistemas preguntas respuestas robustos es posible nombrar MCTest, SQuAD y CoQA.

3.1 SQuAD

El Stanford Question Answering Dataset (SQuAD) ([2] [5]) es, como su nombre lo indica, un conjunto de datos para la comprensión de textos que ha sido muy bien acogido por la comunidad en los últimos años. Este está conformado por un grupo de preguntas confeccionadas manualmente a partir de artículos de Wikipedia y de respuestas que consisten en fragmentos del propio texto base.

Dado que la palabra grupo no da una noción clara de la extensión de dicho dataset, es válida la aclaración de que, tan sólo en su primera versión, SQuAD logró agrupar cerca de 100000 preguntas.

SQuAD fue novedoso en su momento por la proposición de encontrar la respuesta en el conjunto de datos en lugar de elegir entre posibles soluciones, lo cual complejiza el trabajo de los modelos a elaborar.

3.2 CoQA

CoQA ([3]) es un conjunto de datos a gran escala para construir sistemas de respuesta a preguntas conversacionales. El objetivo del desafío CoQA es medir la capacidad de las máquinas para comprender un pasaje de texto y responder una serie de preguntas interconectadas que aparecen en una conversación.

CoQA contiene más de 127000 preguntas con respuestas recopiladas de más de 8000 conversaciones. Cada conversación se obtiene emparejando a dos trabajadores para conversar sobre un pasaje en forma de preguntas y respuestas. Las características únicas de CoQA incluyen:

1. las preguntas son conversacionales;
2. las respuestas pueden ser texto de forma libre;
3. cada respuesta también viene con una subsecuencia de evidencia resaltada en el pasaje;
4. los pasajes se recopilan de siete dominios diversos. CoQA tiene muchos fenómenos desafiantes que no están presentes en los conjuntos de datos de comprensión de lectura existentes, por ejemplo, correferencia y razonamiento pragmático.

4 BERT

BERT es el acrónimo para Bidirectional Encoder Representations from Transformers (Representaciones de Codificador Bidireccional de Transformadores). Técnicamente, BERT es un modelo de Redes Neuronales Artificiales (RNA) aplicado al campo del Natural Language Processing (NLP), específicamente al subcampo del Natural Language Understanding (NLU).

Históricamente, los modelos de lenguaje solo podían leer la entrada de texto secuencialmente, ya sea de izquierda a derecha o de derecha a izquierda, pero no podían hacer ambas cosas al mismo tiempo. BERT es diferente porque está diseñado para leer en ambas direcciones a la vez. Esta capacidad, habilitada por la introducción de Transformers, se conoce como bidireccionalidad.

Usando esta capacidad bidireccional, BERT está pre-entrenado en dos tareas de NLP diferentes, pero relacionadas: Masked Language Model (MLM) y predicción de la siguiente oración.

El objetivo del entrenamiento del Modelo de lenguaje enmascarado (MLM) es ocultar una palabra en una oración y luego hacer que el programa prediga qué palabra se ha ocultado (enmascarado) según el contexto de la palabra oculta. El

objetivo del entrenamiento de predicción de la siguiente oración es hacer que el programa prediga si dos oraciones dadas tienen una conexión secuencial lógica o si su relación es simplemente aleatoria.

Debido a su bidireccionalidad, tiene un sentido más profundo del contexto y el flujo del lenguaje y, por lo tanto, hoy en día es uno de los modelos de PLN más populares y ampliamente utilizados.

5 Implementación

La propuesta de este trabajo fue desarrollar un pequeño sistema de preguntas respuestas. Como se ya ha expuesto utilizando un modelo de Bert. En este caso el modelo escogido fue el *deepset/bert-base-cased-squad2*, un modelo *fined tuned* con *SQuAd*, lo cual lo hace adecuado para este tipo de tareas.

El sistema dimplementado recibe una pregunta y su contexto y los pasa al tokenizer de BERT que los procesa como un par.

El tokenizer de BERT se auxilia de tokens especiales:

1. CLS Significa clasificación y busca representar la clasificación a nivel de oración.
2. SEP Se utiliza para separar las dos piezas de texto.

BERT también utiliza internamente "Segment Embeddings", que no es más que una forma de diferenciar la pregunta del texto. Para representar esto en la práctica, se utiliza un vector de ceros si las palabras pertenecen a la pregunta o un vector de unos si las palabras pertenecen al texto. Lo anterior se envía como entrada al modelo, de cuyo resultado se escogen las palabras iniciales y finales más probables y se devuelve la respuesta solo si el token final está después del token inicial (para evitar posibles errores que el modelo de Bert pudiera cometer).

Para la implementación se utilizaron las bibliotecas de *torch* y *transformers* de *python*. De *transfromers* fue importante el aprovechamiento de *BertForQuestionAnswering* y *BertTokenizer* para obtener un trabajo relevante.

6 Métricas y Resultados

El sitema obtenido parece mostrar resultados bastante buenos para preguntas no muy complejas en un contexto bastante entendible. Pero cómo obtener un sentido más claro de su desempeño?

Se implementó un módulo de validación de resultados que permite obtener las medidas *F1 score* y *exact match* a partir de lo que se indica en la página oficial de *SQuAD* para evaluar estos modelos.

Se considera en los sistemas de preguntas respuestas que el *exact match* está definido por la capacidad que tenga el modelo de devolver respuestas idénticas a las que se esperan. Para el caso de *F1* se definen la *precisión* y el *recobrado* a partir de los tokens de la predicción que coincidan con los esperados (luego de una limpieza de la frase recuperada, extrayendo artículos y elementos irrelevantes).

-	Exact Match	F1 score
SQuAD	60.00	73.33
CoQA	14.80	23.30

Table 1. Análisis del desempeño del sistema.

Teniendo en cuenta estas medidas se desarrollaron numerosas pruebas sobre los *datasets* mencionados, *SQuAD* y *CoQA*, obteniendo los resultados siguientes:

Como ser puede apreciar para *SQuAD* se obtuvieron resultados bastante buenos, lo cual no es de extrañar conociendo que el modelo utilizado está preentrenado sobre este *dataset* y Bert está muy bien capacitado para estas tareas. Sin embargo qué sucede con *CoQA*? Como bien se habí explicado este segundo cojunto de datos tien una forma más coloquial, más conversacional, por lo que sus preguntas normalmente son menos claras y tienen una mayor dependencia entre ellas, de modo que si nos encontramos con una frase cómo "*and what's his name*", probablemente se está haciendo referencia a una persona de la que se hanía discutido antes. Esto complejiza mucho más el problema y se aleja bastante de lo que realmente se trabaja en *SQuAD*, donde se manejan preguntas más concisas y explicativas. Sin embargo cabe destacar que por tratarse de una másquina siguen siendo bastante buenos ambos resutados, por la dificultad del problema.

7 Conclusiones y recomendaciones

En este trabajo se propuso la implementación de un sistema de preguntas respuestas sencillo a partir de un modelo pre entrenado de BERT. Se comprobó su desempeño en *datasets* bien conocidos como son *SQuAD* y *CoQA*. Los resultados obtenidos permitieron reconocer la enorme capacidad que tiene Bert para modelar el rezonamiento y preocesar el lenguaje natural, sobre todo en tareas de este tipo. Las dificultades para tener en cuenta el conocimiento previo o un lenguaje más conversacional que el que se utilizó para su entrenamiento también saltaron a la luz, pero en el futuro pudieran considerarse ciertas modificaciones al modelo que puedan tener en cuenta estos problemas, a pesar de que el primero es un problema abierto aún.

En este sentido se puede agregar que en el estado del arte son numerosas las modificaciones de BERT que se pueden mencionar en problemas de este tipo, como *ALBERT*, *ROBERTA* y *SemBERT*. En los sitios oficiales de *SQuAD*[4] y *CoQA*[1] se puede ver como estos y combiacines que los involucran dominan el *ranking*.

Es relevante destacar además como numerosos de estos sistemas han logrado facilitar al ser humano efectuar preguntas en su lenguaje natural, sin necesidad ya de usar lenguajes específicos para comunicarse con los programas necesarios, y obtener respuestas relevantes, claras, no un simple conjunto de documentos. Aunque el modelo utilizado acá se limitaba al hecho de seleccionar una sentencia

existente, no a generar una, en el futuro pudiera explorarse también la posibilidad de realizar estas generaciones de respuestas con un modelo de BERT.

References

1. CoQA Sitio oficial. https://stanfordnlp.github.io/coqa/
2. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.
3. Reddy, S., Chen, D., & Manning, C. D. (2019). Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7, 249-266.
4. SQuAD Sitio oficial. https://rajpurkar.github.io/SQuAD-explorer/
5. Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.