Generación Automática de Resúmenes Abstractivos de Noticias en Español

Renso Bernoldi and Gabriel Tolosa

Departamento de Ciencias Básicas, Universidad Nacional de Luján
rensobernoldi@gmail.com, tolosoft@unlu.edu.ar

Resumen Una tarea desafiante es la extracción automática de información relevante en documentos escritos. Los últimos esfuerzos, principalmente para el idioma inglés, toman ventaja de las arquitecturas basadas en *Transformers*, actualmente utilizadas para diferentes tareas relacionadas con el lenguaje. En este trabajo se propone y evalúa un pipeline de procesamiento basado en BERT para la generación de resúmenes abstractivos sobre noticias en español. Los resultados preliminares muestran una performance competitiva con los modelos para el idioma inglés, estableciendo una referencia para potenciales mejoras.

Keywords: Resumenes abstractivos · español · transformers.

1. Introducción

Una de las tareas cognitivas mas desafiantes para el ser humano es la comprensión y extracción de información relevante en documentos escritos. Además, la creciente proliferación de documentos digitales vislumbra la necesidad de contar con técnicas automáticas que permitan condensar la información de los documentos para agilizar la lectura, manteniendo el significado del mismo. Producto del esfuerzo que conlleva elaborar resúmenes, tiempo atrás, se comenzó a diseñar modelos que permiten resumir un texto de una manera automática [1].

Existen dos enfoques principales para realizar resúmenes. Por un lado, de manera *extractiva*, que consiste en la selección y ordenamiento de aquellas oraciones relevantes y, por otro lado, *abstractiva* que tiene como finalidad identificar la idea principal y generar el resumen evitando reproducir, en lo posible, las mismas palabras/oraciones del documento de partida.

En los últimos años, los avances en representaciones densas de palabras basadas en el concepto de *embeddings* [10] permiten establecer relaciones entre éstas a partir de operar en espacios vectoriales. Además, con el avance y accesibilidad a recursos de computación en la nube se posibilita la ejecución de nuevos modelos de redes neuronales (Ejemplo: Long-Short-Term-Memory [7]) sobre datos masivos. En particular, los modelos para procesar lenguaje se benefician de la arquitectura denominada *Transformers* (Vaswani et al. [14]) y su mecanismo de *atención*, que en la actualidad, son considerados estado del arte.

La utilización de Transformers muestra avances en varias tareas relacionadas con el procesamiento del lenguaje natural, en especial, con la propuesta de BERT

(Bidirectional Encoder Representations from Transformers) [4] el cual avanza sobre los modelos de representación del lenguaje entrenados en una dirección a una bidireccionalidad a través de dos estrategias denominadas *Masked LM* y *Next Sentence Prediction*. Una característica de BERT es que se puede utilizar un modelo pre-entrenado general y luego hacer ajustes particulares (*fine tuning*) propios de la tarea a resolver. Entre las tareas que se benefician del uso de BERT se encuentra el área de resúmenes automáticos. Diversos trabajos han abordado el problema de diseñar una arquitectura para esta tarea, principalmente sobre corpus de noticias para textos en inglés [6, 9, 17].

En este trabajo, se examina el uso de modelos basados en BERT pre-entrenados aplicados a la generación de resúmenes abstractivos automáticos de noticias en español. Este tipo de documentos son habitualmente utilizados para evaluar esta tarea ya que presentan una dificultad extra debido a su longitud reducida. La arquitectura general del modelo se basa en un *encoder* de oraciones que captura las relaciones entre palabras/frases dentro del texto y genera resúmenes que pueden contener palabras que no aparecen en el texto original. La generación de resúmenes abstractivos amplía el uso de estas técnicas a múltiples dominios que se beneficien de la posibilidad de generación de texto, por ejemplo, agentes conversacionales o recortes de prensa. De acuerdo a nuestro conocimiento este es el primer trabajo que aborda el problema de los resúmenes abstractivos automáticos sobre noticias en español utilizando modelos de lenguaje pre-entrenados. Las contribuciones de este trabajo son las siguientes:

- Se define un pipeline de procesamiento basado en la utilización de una arquitectura de Transformers (pre-entrenados) para la generación de resúmenes abstrativos sobre un corpus de noticias en español.
- Se construye un corpus de noticias en español con características similares a aquellos en inglés utilizados para la evaluación de la tarea en cuestión.
- Se evalúa la performance siguiendo la metodología estándar, obteniendo resultados comparables con el estado del arte para noticias en inglés.

2. Modelo Propuesto

Una de las dificultades al trabajar sobre resúmenes abstractivos en español es la ausencia de corpus de entrenamiento, por lo que se lleva adelante la construcción de un corpus de noticias en español. Los resúmenes sobre noticias son comúnmente utilizados en este tipo de tareas, generalmente en inglés [11]. Una noticia es un texto relativamente corto, compuesto por un título, opcionalmente un subtítulo y un cuerpo. El título junto al subtítulo son considerados una forma de resumen por lo que habitualmente se los utiliza para la evaluación (*target*) del resumen sobre el cuerpo.

BERTSUM [9] es un framework basado en BERT para la tarea de resúmenes extractivos y abstractivos. Introduce un novedoso mecanismo de representación de oraciones para tal fin, que fue evaluado utilizando noticias en inglés. En este trabajo se realizan las modificaciones necesarias para utilizar BERTSUM sobre

BETO [2], un modelo pre-entrenado exclusivamente en Español, en lugar de BERT original. Sobre esta base se ajustan los parámetros del modelo con un corpus de noticias en español. El *pipeline* de procesamiento propuesto parte de un corpus de noticias en español, y consta de los siguientes pasos:

- Separar las noticias en tres subconjuntos (entrenamiento, validación y test), siguiendo particiones similares a las usadas en el dataset de noticias CNN (98 %, 1 % y 1 %, respectivamente).
- Tokenizar los documentos.
- Realizar las modificaciones en el código de BERTSUM para utilizar BETO y su vocabulario.
- Ajustar los parámetros de BERTSUM para la tarea especifica de resumen abstractivo.
- Testear la performance del modelo.

3. Experimentos y Resultados

Datos: El corpus de noticias se construye a partir del sitio web del periódico español *20 Minutos*, el cual cuenta con un historial de noticias de libre acceso y descarga. Se recolectaron noticias transcurridas en el año 2019, previo al inicio de la pandemia, suponiendo que el volumen de noticias generadas sobre esta temática podía sesgar los temas. Dado que el periódico cubre distintos países de Europa con diferentes idiomas, se limitó a aquellas en Español. Considerando que el subtítulo de una noticia es parte del *target* y que en varias de ellas solamente se mencionaba el lugar donde transcurría el hecho, se descartaron. Los datos básicos del corpus se presentan en la Table 1.

Corpus *20 Minutos*	**Train**	**Valid**	**Test**
# documentos	93,913	964	960
Promedio # tokens en documentos	2,023	2,092	1,892
Desvío # tokens en documentos	1,339	1,396	1,379
Promedio # tokens en target	355	350	309
Desvío # tokens en target	123	121	123

Tabla 1: Composición del corpus de noticias generado para los experimentos.

Para la descomposición de frases en *tokens* se usa CoreNLP, desarrollado por la Universidad de Standford[1]. De la misma forma que en el trabajo de Liu [9], las entradas al modelo fueron truncadas a 512 tokens.

Métricas: La evaluación de la performance del modelo se basa en las métricas standard ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [16],

[1] https://stanfordnlp.github.io/CoreNLP/

	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**Media**
PointerGenerator+Coverage [13]	39.53	17.28	36.38	31.06
ML+RL+intra-attn [12]	39.87	15.82	36.90	30.87
inconsistency loss [8]	40.68	17.97	37.13	31.93
Bottom-Up Summarization [5]	41.22	18.68	38.34	32.75
DCA [3]	41.69	**19.47**	37.92	**33.11**
Modelo BERT-Transformer [17]	39.50	17.87	36.65	31.34
BERTSUM [9]	**41.72**	19.39	**38.76**	32.58
BERTSUM_ESP	40.00	18.37	34.98	31.12

Tabla 2: Resultados de las métricas de los modelos más competitivos para el inglés sobre el conjunto de datos CNN/Daily Mail (mejores valores en negrita). La última fila presenta los resultados de este trabajo para el español con el corpus *20 Minutos*. Si bien los resultados no son directamente comparables, se muestran como una primera referencia.

que tratan de correlacionar un resumen de referencia y el generado automáticamente. La métrica ROUGE-N se basa en el solapamiento a nivel de n-gramas de ambos resúmenes, es decir, cuántos coinciden respecto del total. Es una medida orientada a la exhaustividad. En este caso, se usan ROUGE-1 y ROUGE-2 que calculan el solapamiento a nivel de unigramas y bigramas, respectivamente. Por otro lado, ROUGE-L utiliza la subsecuencia de palabras común más larga para calcular el *score*. La intuición detrás de esta métrica es que mientras más larga sea la secuencia coincidente entre dos resúmenes, más similares son. Una ventaja que presenta es que no requiere de coincidencias consecutivas sino *en secuencia*, es decir, coincidencias que reflejen el orden de las palabras.

Resultados: Los resultados se presentan en la Tabla 2. Como punto de comparación se incluye la performance de los modelos más recientes para esta tarea sobre el conjunto de datos standard CNN/Daily Mail (en idioma inglés). Como se puede apreciar los resultados obtenidos para la tarea en español son comparables con los modelos para el inglés. Para la métrica ROUGE-1, la performance es alrededor de $-4.1\,\%$. Aquí se puede considerar que el modelo aprende palabras relevantes del texto aunque el orden para generar oraciones debe mejorar aún ($-5.6\,\%$) de la misma manera que los modelos para el inglés. La diferencia mayor de performance se obtiene en ROUGE-L ($-9.7\,\%$), es decir, generando secuencias largas de texto. En este caso, una posible explicación sobre esta diferencia puede estar en las diferencias gramaticales entre el inglés y el español. Si bien estos resultados no son directamente comparables por tratarse de conjuntos de datos e idiomas diferentes, sirven para obtener una primera referencia del desempeño esperables de estos modelos para la tarea de resumen abstractivo sobre noticias. De forma complementaria, considerando que los resúmenes abstractivos tienen su fundamento en la generación de palabras "nuevas"no vistas en el documento original, se calcula la proporción de n-gramas nuevos generados por el modelo (Figura 1).

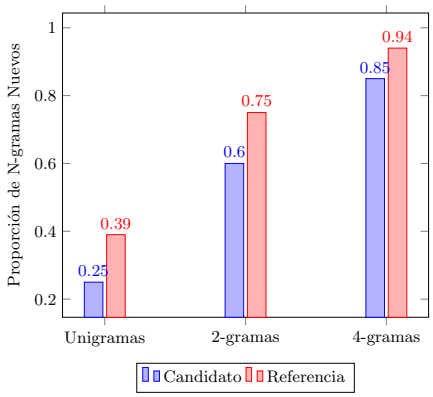

Figura 1: Número de n-gramas nuevos generados por el modelo (Candidato) y el resumen de referencia.

Como se puede apreciar, considerando que la métrica se calcula sin excluir preposiciones, se obtienen valores aceptables en cuanto a unigramas. Por otra parte, a medida que aumenta el número de n-gramas, validando la idea de evitar extracciones directas sobre la fuente, es esperable que disminuya la intersección con el texto de referencia.

4. Conclusiones y Trabajos Futuros

En este trabajo se propone y evalúa un pipeline de procesamiento basado en BERT que genera resúmenes abstractivos sobre noticias en español. Los resultados preliminares muestran valores próximos al estado del arte para el idioma inglés, considerando las diferencias gramaticales entre ambas lenguas. Entre los siguientes pasos, se propone estudiar la relación entre los diferentes parámetros de entrenamiento del modelo y las métricas, considerando mejorar la generación de frases consistentes en la lengua (ROUGE-L). De igual manera, se propone avanzar con métodos que superen la limitación en el tamaño de la entrada.

Referencias

1. Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E., Gutierrez J., Kochut K. “Text Summarization Techniques: A Brief Survey”. International Journal of Advanced Computer Science and Applications, (2017).
2. Cañete J., Chaperon G., Fuentes R. and Perez J. “Spanish Pre-Trained Bert Model and Evaluation Data”. International Conference on Learning Representations, (2020).
3. Celikyilmaz A., Bosselut A., He X., and Choi Y. “Deep Communicating Agents for Abstractive Summarization”. Conference of the North American Chapter of the Association for Computational Linguistics, ACL, (2018).
4. Devlin J., Chang M., Lee K., Toutanova K. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. North American Chapter of the Association for Computational Linguistics, (2019).

5. Gehrmann S., Deng Y., and Rush A. “Bottom-Up Abstractive Summarization”. Conference on Empirical Methods in Natural Language Processing, ACL, (2018).
6. Gupta H. and Patel M. “Method Of Text Summarization Using LSA and Sentence Based Topic Modelling With BERT”. International Conference on Artificial Intelligence and Smart Systems, (2021).
7. Hochreiter, S. and Schmidhuber J. “LSTM can Solve Hard Long Time Lag Problems”. MIT Press, (1996).
8. Hsu W., Lin C., Lee M., Min K., Tang J., and Sun M. “A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss”. 56th Annual Meeting of the Association for Computational Linguistics, ACL, (2018).
9. Liu Y. and Lapata M. “Text Summarization with Pretrained Encoders”. Conference on Empirical Methods in Natural Language Processing, (2019).
10. Mikolov T., Chen K., Corrado G., Dean J. “Efficient Estimation of Word Representations in Vector Space”. International Conference on Learning Representations, (2013).
11. Nallapati R., Zhou B., dos Santos C., Gulcehre C. and Xiang B. “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond”. Conference on Computational Natural Language Learning, (2016).
12. Paulus R., Xiong C. and Socher R. “A Deep Reinforced Model for Abstractive Summarization”. International Conference on Learning Representations, (2018).
13. See A., Liu P., and Manning C. “Get To The Point: Summarization with Pointer-Generator Networks”. 55th Annual Meeting of the Association for Computational Linguistics, ACL, (2017).
14. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L. and Polosukhin I. “Attention Is All You Need”. Advances in Neural Information Processing Systems, (2017).
15. Wang A., Singh A., Michael J., Hill F., Levy O., Bowman S. “Glue: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. Conference on Empirical Methods in Natural Language Processing, (2018).
16. Lin, C.-Y. “Rouge: A package for automatic evaluation of summaries”. Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain, (2004).
17. Zhang H., Cai J., Xu J., and Wang J.“Pretraining-Based Natural Language Generation for Text Summarization”. Conference on Computational Natural Language Learning, (2019).