Organizando Historias a partir de un Conjunto Masivo de Noticias Utilizando Inteligencia Artificial

Emanuel Orler[1], Ana G. Maguitman[1,2], Mariano Maisonnave[1,3], Axel J. Soto[1,2], Juan M. Suárez[1], Carlos A. Díaz[4], Santiago García Eleisequi[4], Franco Jofre[4], Sebastián Venturino[4]

[1] Departamento de Cs. e Ing. de la Computación, Universidad Nacional del Sur, Bahía Blanca, Argentina
[2] Instituto de Cs. e Ing. de la Computación - ICIC (CONICET - UNS), Bahía Blanca, Argentina
[3] Department of Management Science, Shannon School of Business, Cape Breton University, Sydney, NS, Canada
[4] GlobalNews Group, Buenos Aires, Argentina

Resumen En este artículo se presenta un proyecto en ejecución en el marco de un Servicio Tecnológico de Alto Nivel (STAN) desarrollado por el Instituto de Ciencias e Ingeniería de la Computación (ICIC CONICET-UNS) para GlobalNews Group. Esta empresa, especializada en monitoreo y evaluación de medios, busca investigar soluciones innovadoras para analizar gran volumen de noticias recolectadas. El proyecto tiene como objetivo abordar desafíos en la organización de historias a partir de un conjunto masivo de noticias, utilizando técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático. Se exploran distintas estrategias de agrupación a nivel de historia y capítulo mediante el modelado semántico de noticias y técnicas de clustering. Además, se trabaja en la creación de un dataset de alta calidad, con el propósito de entrenar y mejorar los modelos desarrollados y aumentar la precisión y eficacia del sistema de organización de noticias.

Palabras claves: Procesamiento de lenguaje natural· Agrupación de noticias· Corpus de noticias.

1. Introducción

GlobalNews Group (`https://www.globalnewsgroup.com`) es una empresa de monitoreo y análisis de medios que provee servicios de seguimiento y análisis de noticias y redes sociales a través de la auditoría y la evaluación de contenido. El Instituto de Ciencias e Ingeniería de la Computación (ICIC - `https://icic.conicet.gov.ar`) es un instituto de doble dependencia CONICET-UNS dedicado a la investigación y transferencia tecnológica en diversas áreas de las ciencias de la computación. Desde el año 2022, se desarrolla un Servicio Tecnológico de Alto Nivel (STAN) entre ICIC y GlobalNews Group que tiene como objetivo organizar grandes volúmenes de noticias en historias y capítulos. Para ello, se aplican técnicas de procesamiento de lenguaje natural y aprendizaje automático que permiten realizar un modelado semántico de la información, detectar entidades relevantes y agrupar noticias a diferentes niveles de granularidad temática. Por otro lado, junto con un equipo de anotadores de GlobalNews Group, se está trabajando en la creación de un dataset de noticias con etiquetas. Este dataset será fundamental para entrenar y validar los métodos desarrollados.

En este artículo, ofrecemos un breve resumen del trabajo llevado a cabo como parte de este STAN. En particular, proporcionamos una serie de definiciones para los conceptos implicados, describimos las tareas realizadas hasta el momento y delineamos las conclusiones y el trabajo a futuro.

2. Marco Conceptual y Tareas Realizadas

Para abordar los desafíos planteados, se ha desarrollado un conjunto de definiciones clave que proporcionan el marco conceptual necesario para este proyecto. Estas definiciones son las siguientes:

- **Evento:** Se refiere a algo que ocurre en un momento y lugar específico. Los eventos son reportados en las noticias y una sola noticia puede hacer referencia a uno o más eventos. Por ejemplo, "Detienen en Argentina un avión venezolano vinculado a Irán".
- **Historia:** Es una secuencia temáticamente cohesiva de eventos (es decir, eventos que pertenecen al mismo tema o tópico), ordenados en el tiempo y relacionados de manera significativa.
- **Capítulo:** Representa un conjunto de noticias directamente relacionadas con un evento principal. Los capítulos dividen una historia en subconjuntos de noticias cercanas en el tiempo que hacen referencia al mismo evento principal. Por ejemplo, "La toma de Mariúpol" es un *capítulo* dentro de la *historia* invasión rusa a Ucrania iniciada en 2022.
- **Noticia de tipo opinión/perspectiva:** Son noticias que ofrecen un resumen de eventos pasados. También incluimos en esta categoría a aquellas noticias que presentan un punto de vista sobre un tema específico.
- **Noticia de tipo evento:** Son noticias que se centran en anunciar un evento puntual.

Las tareas realizadas en el marco de este proyecto están orientadas a desarrollar un sistema capaz de organizar las noticias relacionadas con un mismo evento en capítulos, para luego agrupar estos capítulos y formar historias. Además, se busca distinguir entre noticias de tipo opinión/perspectiva y noticias de tipo evento. El objetivo es automatizar y mejorar el proceso de clasificación y organización de noticias, lo que facilitará la gestión y comprensión de grandes volúmenes de información para GlobalNews Group. A continuación se describen los métodos en proceso de desarrollo y la tarea de etiquetado que se está llevando a cabo sobre el dataset de noticias.

2.1. Modelos de Similitud entre Noticias

Para llevar a cabo las tareas propuestas, es necesario calcular la similitud entre noticias. En este sentido, se han investigado y evaluado modelos que ofrecen representaciones semánticas del contenido, abarcando aspectos como las entidades nombradas y la similitud de *embeddings* (i.e. representaciones vectoriales de texto). Inicialmente, se investigaron definiciones y tareas propuestas en el trabajo seminal "Topic Detection and Tracking – Event-based Information Organization" [1], así como una serie de publicaciones sobre el proyecto "Topic Detection and Tracking" (TDT - [2]). Posteriormente, se investigaron formulaciones y tareas más recientes. En particular, la tarea de "Similitud de artículos en noticias multilingues" propuesta como *SemEval-2022 (Tarea 8)* [3] resultó especialmente relevante para abordar el problema de computar la similitud entre

noticias. Además, el dataset ofrecido como parte de esta tarea guarda similitudes con la estructura de organización sobre noticias deseada por el equipo de GlobalNews Group.

El principal desafío identificado en el desarrollo de modelos de similitud de noticias radica en que la similitud esperada no se limita únicamente al contenido textual, sino que también implica considerar otras dimensiones de análisis, como la geográfica y la temporal, así como menciones a personas, organizaciones y productos, entre otras. Los modelos analizados se basan en el uso de *Sentence-BERT* [4], una técnica que permite codificar oraciones completas en vectores semánticamente ricos y contextualizados, para generar *embeddings* utilizando tanto el título como el cuerpo de las noticias. Este proceso permite representar cada noticia en un espacio vectorial denso. Además, los modelos calculan puntajes de similitud entre pares de noticias, basados en anotaciones proporcionadas por *SemEval-2022* que indican si las noticias pertenecen o no a la misma historia. Estos puntajes se utilizan para realizar un ajuste fino (*fine-tuning*) de los modelos de similitud, con el objetivo de mejorar la capacidad de determinar la similitud entre las noticias y así contribuir a la tarea de agrupamiento de noticias relacionadas.

2.2. Clustering para Detección de Tópicos

Se examinaron estrategias orientadas a la aplicación del tradicional algoritmo de clustering basado en densidad *DBSCAN* [5] y su adaptación conocida como *STDBSCAN* [6], que agrupa objetos (en este caso, noticias) en clusters (tópicos en este caso) incorporando aspectos espaciales y temporales. A la vez, se examinaron propuestas que permitiesen extender los métodos de clustering basados en densidad a una versión jerárquica donde fuese posible definir un dendrograma para el análisis de tópicos definidos a diferentes niveles de granularidad. En este sentido, se analizó el modelado de tópicos a través de *BERTopic* [7]. *BERTopic* ofrece mecanismos para organizar un gran volumen de datos en tópicos y subtópicos dividiendo el conjunto de datos en lotes menores. Sin embargo, esta solución no resulta escalable para la tarea objeto de este proyecto ya que las representaciones de embeddings y tópicos utilizadas no son adecuadas para la aplicación de métodos de búsqueda que permitan una comparación eficiente basada en similitud semántica. Para sobreponernos a esta limitación, se comenzó a explorar el uso de *FAISS* (Facebook AI Similarity Search) [8], una biblioteca desarrollada por Meta's Fundamental AI Research para la búsqueda y clustering eficiente de vectores densos.

2.3. Etiquetado de Noticias

Para llevar adelante la tarea de etiquetado por parte de los anotadores de GlobalNews Group, se configuró la herramienta *Doccano* [9] en un servidor de ICIC. Las noticias fueron elegidas estratégicamente para optimizar el aprovechamiento de las anotaciones. Por cada tarea de etiquetado se analizaron los resultados, brindando directrices y comentarios a los anotadores. Actualmente se está trabajando en una nueva iteración de la tarea de etiquetado en busca de conseguir anotaciones de mayor calidad para el entrenamiento de futuros modelos.

3. Conclusiones y Trabajo a Futuro

Como resultado del análisis realizado, es fundamental resaltar la importancia de utilizar métodos escalables, aprovechando las capacidades proporcionadas por *FAISS* para el procesamiento semántico eficiente de grandes conjuntos de noticias. Asimismo, se

destaca la relevancia del ajuste fino de los modelos de puntuación de similitud semántica mediante la utilización de conjuntos de datos etiquetados. Por último, tras identificar dificultades por parte de los anotadores en la tarea de etiquetado, resulta crucial seguir colaborando con ellos para asegurar la obtención de un conjunto de datos con etiquetas confiables. Esto permitirá avanzar de manera efectiva con el entrenamiento y la validación de los métodos, garantizando la calidad y la precisión de los resultados del análisis.

Como actividades a futuro se propone llevar adelante una fase de entrenamiento con los anotadores seguida de una tarea de anotación a mayor escala para lograr un dataset con mayor volumen de noticias y de mejor calidad. También se planea utilizar facilidades ofrecidas por *FAISS* para implementar índices especializados que permitan identificar de manera eficiente y efectiva tópicos de noticias, considerando tanto la similitud semántica de las noticias como las entidades nombradas en ellas. Se pondrá especial énfasis en analizar la escalabilidad de *FAISS* en el contexto de un gran número de noticias. Otra línea de trabajo a futuro es la generación de pares de noticias sintéticas pertenecientes al mismo tópico utilizando grandes modelos de lenguaje (LLMs). Estas noticias sintéticas serán usadas para aumentar el tamaño del conjunto de datos etiquetados disponible para entrenar modelos de aprendizaje automático para la detección de historias, lo que puede ayudar a mejorar la generalización y rendimiento de los modelos. Para ello se analizarán diversas opciones de LLMs de código abierto, tales como *LLaMA 2* [10], *Mistral* [11], *BLOOM* [12] y *Gemma* [13].

Referencias

1. James Allan, editor. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, USA, 2002.
2. James Allan et al. Topic Detection and Tracking Pilot Study Final Report, 6 2018.
3. Xi et al. Chen. SemEval-2022 Task 8: Multilingual news article similarity. In Emerson Guy, et al, editor, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States, July 2022. Association for Computational Linguistics.
4. Nils Reimers et al. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019.
5. Martin et al. Ester. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.
6. Derya Birant et al. ST-DBSCAN: An Algorithm for Clustering Spatial–Temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007. Intelligent Data Mining.
7. Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022.
8. Matthijs Douze et al. The Faiss library, 2024.
9. Hiroki Nakayama et al. Doccano: Text Annotation Tool for Human, 2018. Software available at https://github.com/doccano/doccano.
10. Hugo Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.
11. Albert Q. Jiang et al. Mistral 7B, 2023.
12. BigScience Workshop et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2023.
13. Gemma Team and Thomas Mesnard et al. Gemma: Open Models Based on Gemini Research and Technology, 2024.