Marco Teórico (Cap 2)

Los modelos de tópicos neuronales (*Neural Topic Models, NTM*) representan un avance significativo frente a los modelos de tópicos tradicionales al optimizar directamente los parámetros sin depender de derivaciones específicas del modelo. Esta característica los hace altamente escalables y flexibles.

En particular, los NTMs permiten inferir eficientemente los parámetros mediante *back-propagation*, aprovechando redes neuronales profundas para modelar tópicos latentes. Un ejemplo destacado es el *Variational Autoencoder (VAE)*, ampliamente utilizado en este ámbito. Tal flexibilidad posibilita que los investigadores ajusten las estructuras de los modelos a las necesidades específicas de diversas aplicaciones. Asimismo, los NTMs gestionan con eficacia grandes volúmenes de datos, aprovechando las capacidades de computación paralela de las GPU. Estas ventajas han fomentado el desarrollo de nuevos métodos y aplicaciones basados en NTMs (Wu, Nguyen, and Luu 2020).

Para formalizar el problema de modelado de tópicos, consideremos un corpus de documentos y un vocabulario predefinido. Sea una colección de N documentos con un vocabulario de tamaño V, donde cada documento se representado como x, el objetivo de los modelos tópicos es identificar K temas latentes en esta colección. El número de temas K se define como un hiperparámetro que los investigadores eligen manualmente en función de las características del conjunto de datos y las tareas a resolver.

Cada tema se representa mediante una distribución sobre el vocabulario, conocida como *distribución tema-palabra*, denotada por $\beta_k \in \mathbb{R}^V$. La matriz de distribuciones tema-palabra para todos los temas se define como:

$$\beta = (\beta_1, \beta_2, \ldots, \beta_K) \in \mathbb{R}^{V \times K}. \tag{1}$$

Adicionalmente, los modelos infieren la *distribución documento-tema*, denotada como $\theta \in \Delta^K$, que describe la proporción en la que cada tema está presente en un documento. Aquí, θ_k representa la proporción del Tema k en el documento, y el simplex de probabilidad Δ^K se define como:

$$\Delta^K = \left\{ \theta \in \mathbb{R}^K \mid \sum_{k=1}^{K} \theta_k = 1 \right\}. \tag{2}$$

Para un modelo tópico tratar las palabras de manera estática es una simplificación que es consistente con el objetivo de identificar los temas semánticos dentro de cada

documento. Sin embargo, para muchas colecciones de interés, el supuesto implícito de tópicos estáticos no es apropiado. Colecciones de documentos como revistas académicas, correos electrónicos, artículos de noticias y registros de consultas de búsqueda reflejan contenido en evolución. Por ejemplo, el artículo de Science "The Brain of Professor Laborde" puede estar en el mismo camino científico que el artículo "Reshaping the Cortical Motor Map by Unmasking Latent Intracortical Connections", pero el estudio de la neurociencia se veía muy diferente en 1903 que en 1991. Los temas en una colección de documentos evolucionan con el tiempo, y es de interés modelar explícitamente la dinámica de los temas subyacentes (Blei and Lafferty 2006).

Los NTMs dinámicos se inspiran en los modelos de tópicos dinámicos Blei and Lafferty 2006 y Wang, Blei, and Heckerman 2008. Estos modelos permiten que los tópicos evolucionen con el tiempo para capturar los cambios en los documentos secuenciales. Específicamente, estos modelos asumen que los documentos se dividen en intervalos de tiempo donde cada intervalo tiene K temas latentes. Los temas asociados con el intervalo t evolucionan a partir de los temas asociados con el intervalo $t-1$ (Wu, Nguyen, and Luu 2020).

La representación de variables (palabras y tópicos) es fundamental en el proceso de identificación de tópicos latentes. Los modelos de tópicos neuronales, que utilizan *embeddings* para representar tanto palabras como tópicos en un espacio vectorial común, han revolucionado este campo al permitir capturar relaciones semánticas más profundas y flexibles entre las palabras y los temas. Un modelo de tópicos con representación en *embeddings* es un enfoque en el que cada palabra del vocabulario está representada por un vector numérico (*embedding*) en un espacio de dimensión D, cada documento se considera como un conjunto de *embeddings* que representan las palabras y cada tópico se modela como un vector numérico (*embedding*) en el mismo espacio de diemnsión D, es decir tópicos y palabras están representados o proyectados por vectores en un mismo espacio vectorial. Al proyectar las palabras y los tópicos en el mismo espacio vectorial se define un método para medir la diferencia semántica entre los *embeddings* de las palabras de un documento y los de los tópicos. En lugar de las técnicas tradicionales que se centran en co-ocurrencias de palabras, estos modelos miden la diferencia semántica entre los vectores de embeddings de las palabras de un documento y los vectores de los tópicos.

Dieng, Ruiz, and Blei 2020 al construir Embedding Topic Model (ETM) descomponen los tópicos en dos parámetros de *embeddings*:

$$\beta = W^{\top} T. \tag{16}$$

Aquí, $W \in \mathbb{R}^{D \times V}$ representa los V *embeddings* de palabras (*word embeddings*), y $T \in \mathbb{R}^{D \times K}$ denota los K *embeddings* de tópicos (*topic embeddings*), donde D es la

dimensión del espacio de *embeddings*. EMT mejoró la eficiencia al inicializar los embeddings de palabras con modelos preentrenados como Word2Vec y GloVe, y permitió una representación más flexible y escalable de grandes corpus de texto.

A medida que la investigación avanzaba, surgieron variantes de modelos tópicos con *embeddings* para superar limitaciones como el manejo de la distancia y la diversidad de los temas. Variantes que utilizaban distancias de transporte óptimo para mejorar la coherencia entre temas y documentos (Zhao et al. 2021), distancias de transporte condicional (wang et al. 2022) , *embeddings* globales de temas adaptados a tareas específicas (Zhibin Duan and Zhou 2022) y *embeddings* hiperbólicos para capturar jerarquías temáticas (Xu et al. 2022) presedieron al modelo ECRTM (Wu et al. 2023) que usa una fórmula de distancia euclidiana y un método de regularización de *clustering* para evitar la repetición de temas y mejorar su diversidad. Este modelo representa la matriz de distribución palabra-tópico de la siguiente forma:

$$\beta_{jk} = \frac{\exp\left(-\frac{\|w_j - t_k\|^2}{\tau}\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{\|w_j - t_{k'}\|^2}{\tau}\right)}.$$

Aquí, β_{jk} indica la correlación entre la palabra j y el tópico k, τ es un hiperparámetro de temperatura, w_j corresponde a la embedding de la palabra j en W, y t_k a la embedding del tópico k en T. Este modelo calcula la distancia euclidiana entre las embeddings de palabras y tópicos, normalizando los valores mediante una función *softmax*. Además, incorpora una regularización basada en agrupamiento: considera las embeddings de los tópicos como centros de los clústeres y las embeddings de palabras como muestras. A través de transporte óptimo, esta regularización fuerza a los tópicos a agrupar palabras de manera efectiva, evitando el problema de colapso de tópicos, donde diferentes tópicos resultan redundantes.

La técnica *Constrastive Learnig*, como una técnica de aprendizaje auto-supervisado, ha sido utilizado para mejorar los Modelos de Tópicos Neuronales (NTMs). Esta técnica se centra en medir las relaciones de similitud y diferencia entre pares de muestras en un espacio de representación, buscando que las muestras similares estén más cerca entre sí y las diferentes más alejadas. En el contexto de NTMs, el *contastative learnig* ha permitido mejorar la coherencia y la separación entre tópicos al trabajar sobre distribuciones documento-tema y embeddings semánticos, en algunos casos creando pares positivos y negativos mediante el muestreo de palabras relevantes o documentos similares, utilizando estas relaciones para ajustar las representaciones del modelo.

La evolución del *contrastative learnig* en NTMs ha ido desde enfoques centrados en distribuciones documento-tema, como el de Research 2021, hasta métodos que incorporan relaciones semánticas y datos aumentados, como el propuesto por Wu, Luu, and Dong 2022. Posteriormente, Zhou et al. 2023 lo aplicaron en embeddings para mejorar la separación entre tópicos, y Han et al. 2023 introdujeron técnicas de agrupamiento

y refinamiento con modelos preentrenados. El avance más reciente, de Nguyen et al. 2024, expande el aprendizaje contrastivo al nivel de los temas, integrando un método multiobjetivo que aborda simultáneamente relaciones a nivel de documento y tema, marcando un enfoque más completo y sofisticado.

En el mundo real, la mayoría de los documentos tienen atributos no textuales, como el autor, la marca de tiempo o la calificación, información a la que se le llama metadatos. Mientras que los NTMs comunes aprenden tópicos de manera no supervisada (solo usando documentos), existen otros que pueden aprovechar los metadatos de los documentos para guiar el modelado de tópicos, similar al LDA supervisado (Mcauliffe and Blei 2007).

En detalle Card, Tan, and Smith 2018 introducen SCHOLAR, un NTM que puede incorporar diversos metadatos de documentos. Codifica un documento junto con sus etiquetas (por ejemplo, sentimiento) y covariables (por ejemplo, año de publicación), y genera el documento condicionado a las covariables. Iryna Korshunova and Theis 2019 modelan la generación de documentos y etiquetas juntos de manera discriminativa; luego entrenan su modelo con inferencia variacional de campo medio. También pueden incorporar una variedad de modalidades de datos como imágenes. Wang and YANG 2020 modelan conjuntamente los tópicos y entrenan un clasificador RNN para predecir las etiquetas de los documentos. Están conectados por un mecanismo de atención. Yiming Wang and Jihong 2021 incorporan redes de documentos en un NTM y reconstruyen conjuntamente documentos y redes.

La evolución de los grandes modelos de lenguaje ha transformado significativamente el campo del procesamiento del lenguaje natural, empleando métodos matemáticos para generalizar las leyes y el conocimiento del lenguaje con el fin de predecir y generar texto. A través de extensas investigaciones que abarcan décadas, el modelado del lenguaje ha evolucionado desde los modelos de lenguaje estadísticos iniciales hasta el panorama contemporáneo de los grandes modelos de lenguaje (LLM). Notablemente, la rápida evolución de los LLM ha alcanzado la capacidad de procesar, comprender y generar texto a nivel humano.

Los investigadores suelen combinar los NTMs con modelos de lenguaje preentrenados. Los modelos de lenguaje preentrenados basados en Transformers (Ashish Vaswani 2017) han sido muy utilizados en el campo del procesamiento del lenguaje natural, y se preentrenan en grandes corpus para capturar características lingüísticas contextuales. Múltiples estudios aprovechan las características contextuales de estos modelos preentrenados para proporcionar información más rica que el tradicional modelo de bolsa de palabras (BoW). Por ejemplo, Federico Bianchi and Hovy 2021 introducen la concatenación del BoW y los embeddings contextuales de documentos de Sentence-BERT (Nils Reimers 2019), y luego reconstruyen el BoW como en trabajos anteriores.

Referencias

Ashish Vaswani Noam Shazeer, Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez Łukasz Kaiser Illia Polosukhin (2017). "Attention is all you need." In: *Advances in neural information processing systems*.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 113–120. DOI: 10.1145/1143844.1143859. URL: https://doi.org/10.1145/1143844.1143859.

Card, Dallas, Chenhao Tan, and Noah A. Smith (2018). "Neural Models for Documents with Metadata". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. DOI: 10.18653/v1/p18-1189. URL: http://dx.doi.org/10.18653/v1/P18-1189.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei (July 2020). "Topic Modeling in Embedding Spaces". In: *Transactions of the Association for Computational Linguistics* 8, pp. 439–453. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00325. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00325/1923074/tacl_a_00325.pdf. URL: https://doi.org/10.1162/tacl_a_00325.

Federico Bianchi, Silvia Terragni and Dirk Hovy (2021). "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence". In: *Proceedings of*

the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 2, 759– 766.

Han, Sungwon et al. (May 2023). "Unified Neural Topic Model via Contrastive Learning and Term Weighting". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1802–1817. DOI: 10.18653/v1/2023.eacl-main.132. URL: https://aclanthology.org/2023.eacl-main.132.

Iryna Korshunova Hanchen Xiong, Mateusz Fedoryszak and Lucas Theis (2019). "Discriminative topic modeling with logistic lda." In: *In Advances in Neural Information Processing Systems* 32.

Mcauliffe, Jon and David Blei (2007). "Supervised Topic Models". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt et al. Vol. 20. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf.

Nguyen, Thong Thanh et al. (2024). "Topic Modeling as Multi-Objective Contrastive Optimization". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HdAoLSBYXj.

Nils Reimers, Iryna Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks." In: *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992.

Research, Thong Nguyen VinAI (2021). "Contrastive Learning for Neural Topic Model". In.

Wang, Chong, David Blei, and David Heckerman (2008). "Continuous Time Dynamic Topic Models". In: *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press. Princeton, NJ, USA, pp. 579–586.

wang, dongsheng et al. (2022). "Representing Mixtures of Word Embeddings with Mixtures of Topic Embeddings". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=IYMuTbGzjFU.

Wang, Xinyi and YI YANG (2020). "Neural Topic Model with Attention for Supervised Learning". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1147–1156. URL: https://proceedings.mlr.press/v108/wang20c.html.

Wu, Xiaobao, Anh Tuan Luu, and Xinshuai Dong (Dec. 2022). "Mitigating Data Sparsity for Short Text Topic Modeling by Topic-Semantic Contrastive Learning". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2748–2760.

DOI: 10.18653/v1/2022.emnlp-main.176. URL: https://aclanthology.org/2022.emnlp-main.176.

Wu, Xiaobao, Thong Nguyen, and Anh Tuan Luu (2020). "A Survey on Neural Topic Models: Methods, Applications, and Challenges". In: *Nanyang Technological University and National University of Singapore*.

Wu, Xiaobao et al. (2023). "Effective Neural Topic Modeling with Embedding Clustering Regularization". In.

Xu, Yishi et al. (2022). *HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding*. arXiv: 2210.10625 [cs.IR]. URL: https://arxiv.org/abs/2210.10625.

Yiming Wang, Ximing Li and Jihong (2021). "Layer-assisted neural topic modeling over document networks". In: *IJCAI*.

Zhao, He et al. (Aug. 2021). "Topic Modelling Meets Deep Neural Networks: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, pp. 4713–4720. DOI: 10.24963/ijcai.2021/638. URL: https://doi.org/10.24963/ijcai.2021/638.

Zhibin Duan Yishi Xu, Jianqiao Sun Bo Chen Wenchao Chen Chaojie Wang and Mingyuan Zhou (2022). "Bayesian deep embedding topic meta-learner". In: *In International Conference on Machine Learning*, 5659–5670.

Bibliografía

Wu, Xiaobao, Thong Nguyen, and Anh Tuan Luu (2020). "A Survey on Neural Topic Models: Methods, Applications, and Challenges". In: *Nanyang Technological University and National University of Singapore*.