ISSN 2395-8618

Interpretabilidad del modelo BERT en el contexto de la similitud semántica

Diana A. Ledesma Roque, Olga Kolesnikova, Ricardo Menchaca Méndez

***Resumen*—Este trabajo aborda la cuestión de la interpretabilidad lingüística del modelo transformador BERT-Base. A diferencia de los modelos tradicionales que dependen de clasificadores independientes y conjuntos de datos especializados, esta investigación propone una alternativa basada en el aprendizaje no supervisado, específicamente en la reducción de dimensionalidad mediante autocodificadores y algoritmos de agrupamiento. Los hallazgos resaltan la relevancia de la longitud de la secuencia en las capas iniciales, con una disminución gradual a lo largo de las capas, mientras que la atención a la similitud semántica se concentra en las capas intermedias y superiores, particularmente en las capas 6, 8, 9 y 10. Además, se observa una tendencia de agrupamiento de secuencias con estructuras gramaticales similares a lo largo de las capas del modelo. Estos resultados se obtuvieron al abordar la tarea de similitud semántica utilizando el conjunto de datos STS-Benchmark y el conjunto SICK-R.**

***Palabras clave*—BERT, transformador, token cls, autoatención, interpretabilidad lingüística, aprendizaje no supervisado.**

Interpretability of the BERT model in the context of semantic similarity

***Abstract*—This work addresses the issue of the linguistic interpretability of the BERT-Base transformer model. In contrast to traditional models that rely on probes and specialized datasets, this research proposes an alternative based on unsupervised learning, specifically on dimensionality reduction through autoencoders and clustering algorithms. The findings highlight the relevance of sequence length in the early layers, with a gradual decrease across layers, while attention to semantic similarity concentrates in the intermediate and upper layers, particularly in layers 6, 8, 9, and 10. Additionally, a clustering trend of sequences with similar grammatical structures is observed across the model layers. These results were obtained by addressing the task of semantic similarity using the STS-Benchmark dataset and the SICK-R dataset.**

***Index Terms*—BERT, transformer, cls token, self-attention, linguistic interpretability, unsupervised learning.**

I. Introducción

Los modelos transformadores, como BERT, han demostrado ser altamente efectivos en la resolución de tareas de procesamiento de lenguaje natural. Sin embargo, debido a su complejidad, a menudo se consideran "cajas negras", lo que dificulta su interpretación. Explorar la interpretabilidad de estos modelos podría proporcionarnos una comprensión más profunda de por qué funcionan. La interpretación de estos modelos ofrece varios beneficios, como la identificación de vulnerabilidades, áreas de mejora y oportunidades de optimización.

Se han llevado a cabo varios trabajos en busca de la interpretabilidad del modelo BERT, algunos de los cuales se han centrado en el análisis de cabezales de atención del transformador como en [1, 2]. Otros enfoques han implicado tomar las representaciones del modelo transformador BERT y someterlas a análisis mediante la creación de sondas o clasificadores independientes para evaluar tareas específicas de procesamiento de lenguaje natural como lo hicieron en [3-6]. Además, se han realizado esfuerzos para aportar interpretabilidad a través de herramientas de visualización específicas para el modelo BERT. Un ejemplo destacado es BERTVIZ, desarrollado por [7], que proporciona una vista detallada de la atención a nivel de cabezal, una vista del modelo completo y una vista a nivel de neurona.

En este trabajo se explora como el modelo BERT aborda la similitud semántica mediante sus mecanismos de autoatención. En concreto, se busca comprender que aspectos lingüísticos se abstraen de las representaciones de autoatención en cada una de las capas del modelo, cuando se enfrenta al desafío de resolver la similitud semántica. Se plantea la pregunta de si estas representaciones de autoatención pueden revelar patrones relacionados con el grado de similitud semántica entre las oraciones y si estas además contienen información respecto al tamaño de secuencia y similitud de estructura gramatical.

Buscar la interpretabilidad lingüística es un desafío, y la complejidad del problema condujo a desarrollar una herramienta de visualización que automatizara ciertos procesos y permitiera comprender qué estaba ocurriendo con las representaciones de la atención de BERT. Esta herramienta se centra en el enfoque del aprendizaje no supervisado y tiene como objetivo visualizar y analizar patrones y agrupaciones en las autoatenciones del modelo, relacionándolas con aspectos lingüísticos.

II. Antecedentes

El modelo transformador ha revolucionado diversas tareas de procesamiento de lenguaje natural. En comparación con las redes recurrentes, sus predecesoras, el transformador desarrollado por [8] ofrece ventajas clave, como el procesamiento altamente paralelo y la capacidad para manejar secuencias largas sin sufrir problemas de desvanecimiento del gradiente.

A partir de la arquitectura del bloque codificador del transformador, presentado en el trabajo de [8], la idea de generar representaciones contextualizadas de palabras, como se propuso en el enfoque de [9] con la creación de ELMo

Manuscript received on 14/05/2024, accepted for publication on 20/97/2024.

Los autores trabajan en el Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), México ({dledesmar2022, kolesnikova, ric}@cic.ipn.mx).

ISSN 2395-8618

(Embeddings from Language Models), y la visión de un entrenamiento en dos etapas que incluye una fase inicial de preentrenamiento generativo no supervisado y una fase posterior de ajuste fino y supervisado, como se materializó en el modelo GPT (Generative Pre-Training) desarrollado por [10], surgieron las bases conceptuales para el desarrollo del modelo BERT (Bidirectional Encoder Representations from Transformers).

BERT, desarrollado por [11], introduce el concepto de un Modelo de Lenguaje Enmascarado (MLM) durante la fase de pre-entrenamiento. En la tarea MLM, algunos tokens de entrada se enmascaran de forma aleatoria, y el modelo se entrena para predecir el identificador de estos tokens enmascarados basándose en su contexto. Además de la tarea de MLM, BERT es capaz de generar representaciones contextualizadas para pares de oraciones, a menudo denominada "predicción de la siguiente oración". Sin embargo, BERT es versátil y puede adaptarse para generar representaciones de una o varias sentencias.

La arquitectura de BERT se basa en el modelo transformador de [8], pero se centra en la utilización de bloques codificadores. Para el preentrenamiento de BERT, se utilizan tokens especiales, como el token de clasificación CLS y el token separador SEP. El token SEP se utiliza para separar tareas que involucran múltiples sentencias, mientras que el token CLS se coloca al principio de la secuencia y se utiliza para obtener una representación contextualizada de toda la secuencia, especialmente útil para tareas de clasificación o regresión. El modelo BERT superó todas las tareas de comprensión de lenguaje del conjunto de evaluación GLUE con un aumento promedio entre 4.5 y 7.0 puntos porcentuales con respecto a su estado del arte.

III. Estado del arte

Se ha investigado la interpretabilidad del modelo BERT mediante el empleo de sondas o clasificadores independientes para evaluar tareas lingüísticas específicas. No obstante, en la comunidad científica existe controversia acerca de cómo BERT abstrae aspectos lingüísticos, especialmente en relación con la sintaxis y la semántica.

Algunos estudios, como el llevado a cabo por [3], sugieren que las capas inferiores del modelo codifican información superficial, las capas intermedias contienen información sintáctica, y las capas superiores albergan información semántica. En contraste, [4] argumenta que BERT realiza tareas lingüísticas convencionales, como el etiquetado de partes del discurso, la identificación de constituyentes, las dependencias sintácticas, los roles semánticos y la resolución de la correferencia de manera interpretativa y adaptable. Según esta perspectiva, la información sintáctica emerge en las capas iniciales, mientras que los aspectos semánticos se manifiestan en las capas superiores.

Por otro lado, la propuesta de [5] sugiere que las tareas de superficie como la longitud de secuencia, se desarrollan en las capas bajas del modelo, mientras que las tareas sintácticas y semánticas están intrínsecamente relacionadas en las capas intermedias y superiores. Estos autores argumentan que las tareas semánticas alcanzan su mejor desempeño en las capas intermedias (capas 6-9) y señalan que las capas individuales de BERT no encapsulan completamente la supuesta tubería de procesamiento sugerida por [4].

El trabajo de [6] se enfoca en analizar la estructura geométrica de los vectores y embeddings de palabras en el modelo BERT. Este estudio revela la presencia de subespacios sintácticos y semánticos. Los resultados muestran que los sentidos de las palabras forman grupos claramente distinguibles en el espacio vectorial. En base a estos hallazgos, se plantea la hipótesis de que la geometría interna de BERT puede descomponerse en subespacios lineales, cada uno destinado a la información sintáctica y semántica de manera separada.

En el estudio de [2], se analizaron en detalle los cabezales de atención de BERT. Se observó que algunos cabezales se centraban en aspectos lingüísticos específicos, como la atención al token anterior, al token siguiente, la resolución de la correferencia, entre otros. Sin embargo, más del 50% de los cabezales no mostraban un enfoque claro hacia una característica o relación lingüística particular. Se notó que, en las primeras capas de BERT, las distribuciones de atención presentaban una alta entropía, que disminuía en capas posteriores. Observaron un cambio en la atención del modelo, inicialmente hacia el token CLS y luego hacia el token de separación SEP. Estos patrones de atención se mantuvieron consistentes entre las cabezas de la misma capa del modelo.

Un estudio que adoptó el enfoque del aprendizaje no supervisado es el de [12]. En este trabajo, proyectaron las incrustaciones contextualizadas de BERT utilizando PCA y examinaron en qué medida se forman regiones semánticas. Sus experimentos revelaron que el subespacio de BERT entrelaza a la semántica con aspectos como la sintaxis y el sentimiento.

Trabajos más recientes han optado por implementar clasificadores binarios para mejorar la interpretabilidad de los modelos BERT. En el estudio [13], se evaluaron y analizaron los modelos BERT en función de ciertos aspectos lingüísticos, concluyendo que el modelo se centra en el uso de pronombres. Asimismo, en la investigación [14], se exploró el espacio vectorial de las representaciones de palabras del modelo BERT mediante grafos de conocimiento, demostrando que las incrustaciones contienen información sobre la estructura de un grafo de conocimiento sin necesidad de realizar ajuste fino.

IV. Metodología

Con el propósito de llevar a cabo un análisis lingüístico e interpretativo del modelo BERT, se utilizó la versión base de BERT, que comprende 12 bloques codificadores, 12 cabezales de atención en cada capa y una dimensión oculta de 768.

Para llevar a cabo el análisis de los cabezales de atención, se emplearon los siguientes conjuntos de datos focalizados en la tarea de similitud semántica para pares de oraciones en el idioma inglés:

- STS-Benchmark [15]: Este conjunto, denominado *Semantic Textual Similarity Benchmark*, presenta una escala continua de 0 a 5, donde 0 indica ausencia de similitud y 5 representa similitud perfecta.
- SICK-R [16]: Conocido como *Sentences Involving Compositional Knowledge - Relatedness*, este conjunto en inglés se caracteriza por una escala continua de 1 a 5, donde 1 denota ausencia de similitud y 5 refleja similitud perfecta.
- El análisis de autoatención, se divide en tres fases:
- Fase 1: resolución de la tarea de similitud semántica mediante ajuste fino.
- Fase 2: obtención y procesamiento de atenciones para crear representaciones fijas mediante una reducción dimensional.
- Fase 3: análisis de aspectos lingüísticos mediante agrupamiento.

ISSN 2395-8618

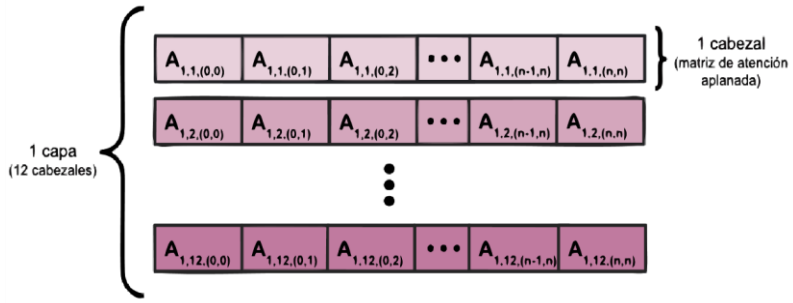

Fig. 1. Atenciones de cabezales apilados para cada capa del modelo BERT. Cada capa contiene los 12 cabezales con matrices de atención aplanadas

El análisis de agrupamiento se llevó a cabo con el respaldo de diagramas de caja y pruebas estadísticas. Se tomaron en cuenta características lingüísticas de interés, tales como la longitud de la secuencia (LS), la similitud semántica (SS) y la estructura gramatical (EG). A continuación, se describen en detalle cada una de estas fases.

A. Resolución de la similitud semántica

Para abordar la tarea de similitud semántica, se implementó un ajuste fino mediante un regresor lineal con el propósito de obtener el grado de similitud semántica. Se tomó el token de clasificación CLS como entrada para el regresor.

La implementación de los algoritmos de aprendizaje automático se realizó utilizando las bibliotecas PyTorch y *Transformers* de *Hugging Face*. La optimización de hiperparámetros se llevó a cabo mediante el *framework* Optuna con 30 ejecuciones. Se definió un tamaño de lote de 32, 5 épocas de entrenamiento, una longitud de sentencia de 128. Se empleó la función de pérdida del error cuadrático medio (MSELoss) y propagación de la raíz cuadrada media (RMSprop) como optimizador. Para el conjunto STS-Benchmark, se estableció una tasa de aprendizaje de 3.1e-5 y de 2.85e-5 para el conjunto SICK-R.

Para la medición de desempeño del modelo se utilizó el coeficiente de correlación de Pearson (ρ_p) y el coeficiente de correlación de Spearman (ρ_s).

B. Reducción dimensional de autoatenciones

Dado que el modelo BERT base consta de 12 capas y 12 cabezales de atención en cada capa, y considerando que las matrices de atención tienen tamaños diversos, se llevó a cabo un preprocesamiento de las atenciones. Se implementó un autocodificador mediante una red recurrente de memoria a largo y corto plazo LSTM, con el propósito de reducir la dimensión de la matriz a un vector fijo de 2 dimensiones, facilitando así su visualización.

Se extrajeron las atenciones de todos los cabezales de cada capa del modelo BERT-Base, para cada secuencia de los conjuntos de prueba STS-Benchmark y SICK-R, una vez que se abordó la tarea de similitud semántica. Para obtener la representación de una capa en particular, se aplanaron las atenciones de sus 12 cabezales. En la Fig. 1, el primer subíndice de la matriz de atención *A* denota la capa *l*, el segundo subíndice representa el número de cabezal *h*, y el tercer subíndice se refiere a la posición de un peso de atención específico (*i*,*j*), es decir, $A_{l,h,(i,j)}$. Además, *n* indica el número de tokens en una secuencia.

Una vez aplanadas las autoatenciones para cada capa, se configuró el tamaño del lote en 12, que corresponde al número de capas en el modelo BERT base. La dimensión de la secuencia se definió conforme a la longitud de la atención aplanada de cada cabezal, mientras que el número de características que se esperan en cada paso de tiempo de la red LSTM es igual al número de cabezales presentes en una capa del modelo BERT base, que es 12. Finalmente, la estructura de los datos de entrada para el modelo autocodificador se ilustra en la Fig. 2.

El modelo autocodificador se entrenó utilizando la función de pérdida del error cuadrático medio, el optimizador Adam y una tasa de aprendizaje de 1e-3. Esta misma configuración y preprocesamiento se utilizó para los dos conjuntos de datos STS-Benchmark y SICK-R.

C. Análisis de autoatenciones mediante agrupamientos

Tras llevar a cabo la reducción dimensional, se procedió a realizar un análisis de agrupamiento mediante diagramas de dispersión. Este análisis se aplicó a los vectores de 2 dimensiones de todas las secuencias en los conjuntos de datos (conjunto de prueba) para cada capa. Dado el considerable número de vectores, la complejidad de las estrategias de agrupamiento, la diversidad de algoritmos y métricas de similitud o disimilitud, así como los hiperparámetros asociados a cada algoritmo, se desarrolló una herramienta que simplificó la ejecución de las pruebas. Esta herramienta permitió encontrar el valor k óptimo de agrupamientos, así como la evaluación de los mismos.

Dicha herramienta facilitó la realización de pruebas con diversos algoritmos de agrupamiento, como k-means, DBSCAN, agrupamiento jerárquico, algoritmo espectral y mezcla gaussiana. También se creó una herramienta visual que posibilitó la observación del comportamiento de las características lingüísticas (longitud de secuencia, similitud semántica y similitud de estructura gramatical) mediante la definición de rangos de valores. Esto mejoró la comprensión del comportamiento de los datos.

Para evaluar la calidad del agrupamiento, se emplearon métricas intrínsecas, tales como el índice de silueta (S), el índice de Davies-Bouldin (DB) y el índice de Calinski-Harabasz (CH). Estas métricas resultaron ser útiles para determinar el número óptimo de agrupamientos. La identificación del número óptimo implicó la realización de ejecuciones con diferentes valores de K.

ISSN 2395-8618

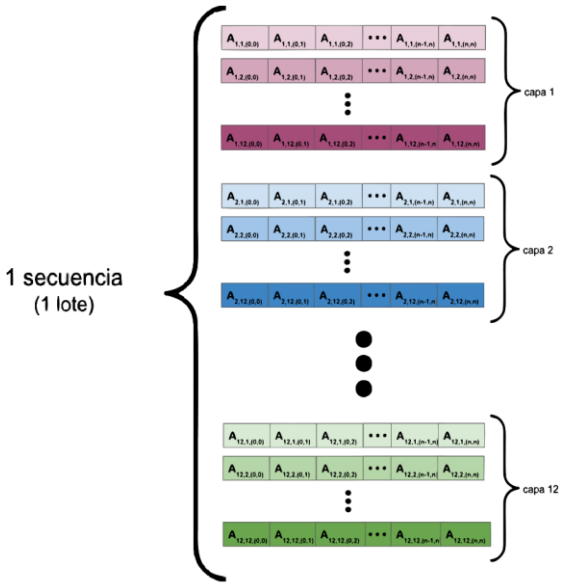

Fig. 2. Estructura de las representaciones de atención como entrada para el modelo autocodificador

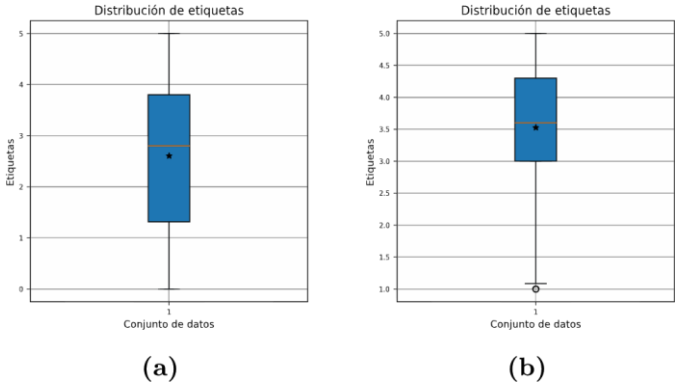

Fig. 3. Distribución de los datos en el conjunto de prueba STS-Benchmark (a) y el conjunto de prueba SICK-R (b) con respecto a sus etiquetas de similitud semántica

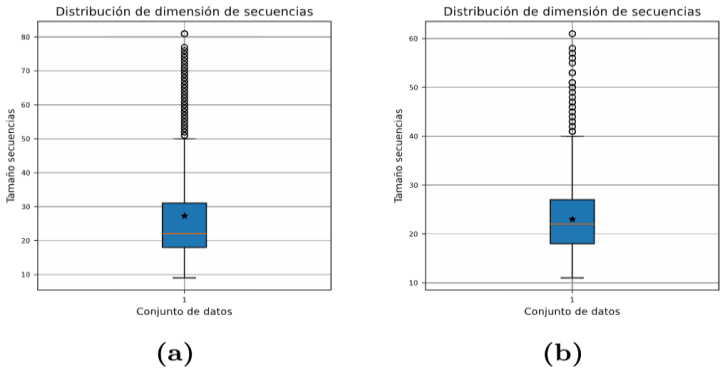

Fig. 4. Distribución de los datos (a) en el conjunto de prueba STS-Benchmark y (b) del conjunto de prueba SICK-R con respecto al tamaño de las secuencias

La decisión sobre el número de agrupamientos óptimos se basó en el consenso de las métricas y en el contexto del análisis de datos.

Para obtener más detalles sobre los algoritmos y los hiperparámetros utilizados, se proporciona información detallada en la Tabla 5 del anexo A.

Se observa que la distribución de las etiquetas del conjunto STS-Benchmark no presenta valores atípicos, pero exhibe un ligero sesgo hacia la izquierda, lo que indica que hay etiquetas con valores de similitud semántica más altos en la mitad superior del rango intercuartil.

En cuanto al conjunto de datos SICK-R se observa mayor dispersión de los datos para valores bajos en similitud semántica. Por otro lado, la dispersión de los datos en el rango intercuantil es menor en el conjunto SICK-R con respecto al conjunto STS-Benchmark.

ISSN 2395-8618

TABLA I
RESULTADOS DE LA EVALUACIÓN DE SIMILITUD SEMÁNTICA DEL CONJUNTO DE PRUEBA DE STS-BENCHMARK Y EL CONJUNTO SICK-R

STS-Benchmark		**SICK-R**	
Spearman	Pearson	Spearman	Pearson
84.8	86.1	83.2	88.6

TABLA II
RESULTADOS DE LA EVALUACIÓN DE LA CALIDAD DE AGRUPAMIENTO MEDIANTE ÍNDICE DE SILUETA (S), ÍNDICE DE DAVIES-BOULDIN (DB) Y CALINSKI-HARABASZ (CH) PARA EL CONJUNTO STS-BENCHMARK Y EL CONJUNTO SICK-R

	STS-Benchmark			**SICK-R**		
Capa	**S**	**DB**	**CH**	**S**	**DB**	**CH**
1	0.387	0.7616	1093	0.377	0.787	3788
2	0.357	0.784	932	0.501	0.705	7162
3	0.435	0.695	1394	0.418	0.556	2596
4	0.4312	0.6736	1909	0.514	0.662	7690
5	0.3471	0.7919	1072	0.4541	0.706	9009
6	0.3493	0.8452	1076	0.385	0.802	4322
7	0.5484	0.5935	2057	0.523	0.648	7983
8	0.5019	0.6184	2917	0.514	0.589	10895
9	0.5345	0.6285	2428	0.522	0.652	8346
10	0.3765	0.8431	1704	0.537	0.625	9157
11	0.678	0.3308	3760	0.896	0.465	10957
12	0.7838	0.201	5959	0.91	0.407	17662

Para evaluar si existían diferencias o similitudes en la distribución de características entre los grupos de datos y comprender la distribución en general, se utilizaron diagramas de caja y pruebas estadísticas como Kruskal-Wallis y Mann-Whitney U, considerando que no existen diferencias significativas entre los grupos como hipótesis nula, en contraste con la hipótesis alternativa que sugiere la presencia de diferencias significativas entre al menos dos grupos.

Similitud semántica. El primer análisis que se realizó tuvo como objetivo identificar patrones generados por el modelo en función de la similitud semántica de las muestras.

La distribución de los datos etiquetados según su similitud semántica, se muestra en el diagrama de caja de la Fig. 3 (a) para el conjunto STS-Benchmark y la Fig. 3(b) para el conjunto SICK-R.

Tamaño de secuencia. Se extrajeron las dimensiones de cada secuencia de los conjuntos de prueba de STS-Benchmark y SICK-R en función del tamaño de la tokenización *Word Piece* utilizada por BERT. La distribución de los datos según el tamaño de la secuencia se presenta en el diagrama de caja de la Fig. 4(a) para el conjunto STS-Benckmark y en la Fig. 4(b) para el conjunto SICK-R.

Tanto el conjunto de datos STS-Benchmark como el conjunto de datos SICKR muestran una asimetría positiva, que se atribuye a la presencia de secuencias con dimensiones notablemente altas.

Estructura gramatical. En este análisis, se procedió a identificar las estructuras gramaticales presentes en todas las secuencias de sus respectivos conjuntos de prueba. Se calculó la similitud coseno de la estructura gramatical más frecuente con respecto al resto de las estructuras, y estos valores se utilizaron como variable de análisis.

Cada estructura está conformada por dos oraciones que se evalúan en términos de su similitud semántica según los conjuntos de datos STS-Benchmark y SICK-R. La estructura gramatical más frecuente del conjunto STS-Benchmark se comparte en 32 secuencias del conjunto de prueba, mientras que la estructura gramatical más frecuente del conjunto SICK-R se comparte 82 veces. La estructura más frecuente del conjunto de datos STS-Benchmark es la misma que la del conjunto SICK-R. Ambas oraciones de la secuencia comparten las mismas etiquetas de dependencia. Las sentencias tienen la siguiente forma:

[’dep’,’det’,’nsubj’,’aux’,’ROOT’,’det’,’dobj’,’punct’].

V. ANÁLISIS Y RESULTADOS

Para la primera fase que respecta a resolver la similitud semántica mediante un regresor lineal como ajuste fino, se pueden observar los resultados de los conjuntos de prueba del STS-Benchmark y SICK-R en la Tabla 1.

En cuanto a la reducción dimensional del autocodificador LSTM, se encontró una pérdida de entrenamiento de 0.045 para

ISSN 2395-8618

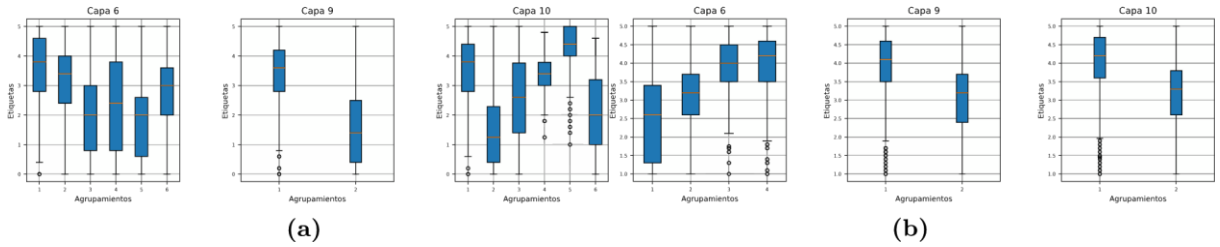

Fig. 5. Diagrama de caja para análisis de la similitud semántica para (a) el conjunto de datos STS-Benchmark y (b) el conjunto SICK-R

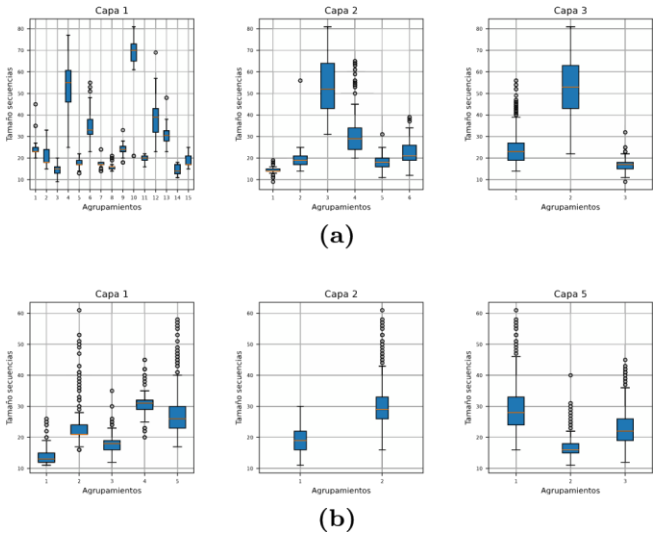

Fig. 6. Diagrama de caja para el análisis de tamaño de secuencia para (a) el conjunto STS-Benchmark y (b) conjunto SICK-R

el conjunto de datos STSBenchmark y de 0.05 para el conjunto SICK-R.

Una vez que se obtuvieron los vectores bidimensionales, se procedió a visualizar las representaciones mediante diagramas de dispersión. Según estos diagramas, se aprecia que los datos tienden a mostrar una mayor dispersión en las primeras capas, mientras que las dos últimas capas presentan agrupamientos más compactos y definidos (consultar Fig. 8 y Fig. 9 del anexo B).

Estas observaciones se respaldan con la evaluación de la calidad del agrupamiento mediante métricas intrínsecas, como se detalla en la Tabla 2 para el conjunto STS-Benchmark y el conjunto SICK-R.

A. Similitud semántica

Según el análisis de la similitud semántica, se observa una relevancia significativa a partir de las capas intermedias. Este patrón puede visualizarse mediante diagramas de caja. En la Fig. 5(a), se presentan los diagramas de caja que muestran una mayor relevancia de la similitud semántica para el conjunto STS-Benchmark. En la Fig. 5(b), se exhiben los diagramas de caja correspondientes a las capas con mayor relevancia en similitud semántica para el conjunto SICK-R.

En el caso del conjunto STS-Benchmark, la capa 9 presenta rangos intercuartiles a diferentes niveles, indicando agrupamientos distintos en cuanto a sus tendencias centrales. En la capa 10, se observa que el agrupamiento 5 exhibe valores de similitud semántica notablemente altos, mientras que el agrupamiento 2 concentra más del 50% de sus datos en un rango de valores entre 0.4 y 2.2. Estos hallazgos sugieren que en las capas 9 y 10, la similitud semántica adquiere relevancia.

En el conjunto SICK-R, se destacan resultados significativos en las capas 9 y 10, donde las cajas exhiben alturas distintas. En la capa 6, se observa que el agrupamiento 1 abarca un rango de valores desde muy bajos hasta niveles intermedios de similitud, mientras que los agrupamientos 3 y 4 comprenden valores altos de similitud.

El análisis de similitud semántica no se basó únicamente en la observación de diagramas de caja, sino que también se respaldó con pruebas estadísticas, como el análisis de Kruskal-Wallis y Mann-Whitney U. Se calculó la estadística y el valor de probabilidad (p) para cada capa. No se definió un nivel de significancia; simplemente, se observó el valor p. En las capas 6, 8, 9 y 10, se observa un valor muy bajo de p, sugiriendo diferencias significativas entre al menos dos de los grupos. En otras palabras, los agrupamientos presentan valores de similitud diferentes, indicando que el modelo atiende la similitud semántica en estas capas. Para obtener más detalles sobre los resultados de los valores p, se puede consultar la Tabla 3, columna SS.

En la Fig. 10 y la Fig. 11 del anexo C se puede consultar los diagramas de caja de todas las capas para el conjunto STS-Benchmark y del conjunto SICK-R, para el aspecto de similitud semántica. También se pueden apreciar los diagramas de dispersión (Fig. 16 y Fig. 17 del anexo D) usando la herramienta visual que muestra el comportamiento de las muestras

ISSN 2395-8618

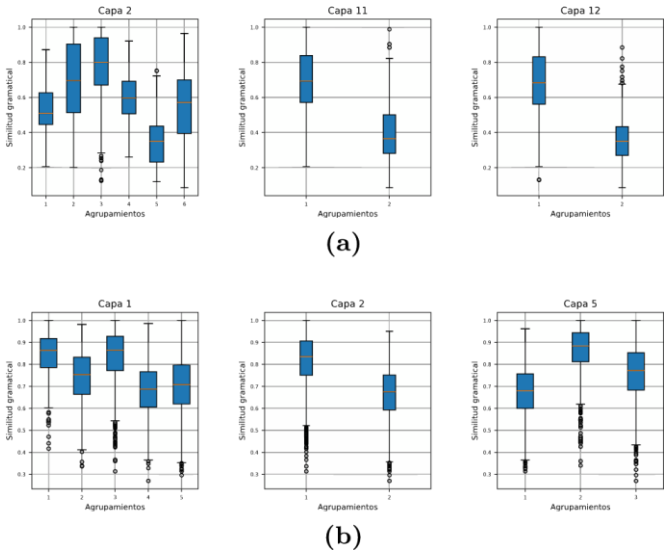

Fig. 7. Diagrama de caja para analizar la similitud estructural de la secuencia más frecuente con respecto a otras estructuras para (a) el conjunto STS-Benchmark y (b) el conjunto SICK-R

TABLA III
RESULTADOS DEL VALOR P DE KRUSKAL-WALLIS Y MANN-WHITNEY U PARA SIMILITUD SEMÁNTICA (SS), TAMAÑO DE SECUENCIA (LS) Y ESTRUCTURA GRAMATICAL (EG) PARA EL CONJUNTO STS-BENCHMARK Y EL CONJUNTO SICK-R

	STS-Benchmark			SICK-R		
Capa	**SS(p)**	**LS(p)**	**EG(p)**	**SS(p)**	**LS(p)**	**EG(p)**
1	0.00188	3.36e-243	3.38e-46	0.0048	0.0	4.1e-268
2	0.00239	1.06e-206	4.62e-85	0.45	0.0	0.0
3	0.00058	9.769e-168	1.079e-08	0.26	4.7e-84	2.5e-11
4	0.00086	7.57e-121	1.63e-34	0.511	4.3e-292	2e-210
5	0.00101	6.207e-120	9.22e-64	3.9e-5	0.0	0.0
6	2.31e-47	2.96e-105	3.75e-24	0.0	0.0	1.1e-244
7	0.1377	1.34e-140	1.58e-12	1.8e-76	0.0	2.8e-167
8	4.545e-34	1.32e-162	1.99e-09	5.8e-92	0.0	4.6e-221
9	1.21e-107	3.69e-20	1.23e-22	1.6e-259	4.2e-10	3.1e-05
10	1.94e-113	3.94e-113	2.48e-47	6.6e-279	3.3e-36	0.198
11	0.00075	8.87e-29	8.82e-95	0.237	4.3e-44	1.7e-13
12	0.0353	5.06e-26	8.82e-93	0.0798	1.3e-58	1.8e-21

conforme a su valor de similitud semántica para el conjunto STS-Benchmark y SICK-R.

B. Tamaño de secuencia

El modelo muestra una atención significativa hacia el tamaño de la secuencia, especialmente en las capas iniciales. Sin embargo, según la herramienta visual diseñada para representar los rangos de longitud de las sentencias (consultar Fig. 18 y Fig. 19 del anexo D), se observa que este enfoque disminuye en capas más avanzadas.

Este comportamiento se refleja en los diagramas de caja presentados en la Fig. 6, donde se destacan las dimensiones de las secuencias en las capas donde la longitud de la secuencia fue más relevante.

Para la capa 1 del conjunto STS-Benchmark, se observa que algunos agrupamientos muestran intersecciones en el mismo rango intercuartil, mientras que otros presentan claras diferencias en los valores. Los rangos intercuartiles que se superponen en los diagramas de caja es porque están bastante cercanos entre sí. En la capa 3 es el comportamiento es más notable ya que las cajas se encuentran a diferente nivel. Para el conjunto de datos SICK-R, se observa de forma más clara como los rangos intercuantiles se encuentran a diferente altura.

Los resultados del valor p de Kruskal-Wallis y Mann-Whitney U se describen en la Tabla 3, en la columna LS. Estos resultados reflejan concordancia con los diagramas de caja. Las capas inferiores se fijan particularmente en el tamaño de secuencia. Sin embargo, en la Tabla 3, es posible apreciar gradualidad en la pérdida de atención del modelo por la

ISSN 2395-8618

dimensión de la secuencia para el conjunto STS-Benchmark. Para el conjunto SICK-R, no se aprecia tanta gradualidad, sin embargo, el enfoque que el modelo pone en el tamaño de la secuencia es menor en las últimas 4 capas. En base a estas observaciones, los conjuntos STS-Benchmark y SICK-R tienen un bajo enfoque en el tamaño de secuencia en las capas 9, 11 y 12. Para mayor detalle, en las Fig. 12 y 13 del anexo C se pruede encontrar los diagramas de caja para el análisis de tamaño de secuencia de todas las capas para los conjuntos de datos utilizados.

C. Estructura gramatical

Al igual que en las secciones anteriores se realizó el análisis de la estructura gramatical mediante el apoyo de diagramas de caja. De todos los análisis este fue el que mostró mayor variabilidad en la mayoría de las capas. En la Fig. 7, se aprecia los diagramas de caja de las capas que mostraron mayor enfoque en la estructura gramatical, tanto para el conjunto STS-Benchmark, como para el conjunto SICK-R. Ambos conjuntos de datos tienen un mayor enfoque en la estructura gramatical en la capa 2. Si se desea visualizar los diagramas de caja de similitud de estructura gramatical para todas las capas del modelo se puede ver las Fig. 14 y 15 del anexo C.

Los resultados del valor p de Kruskal-Wallis y Mann-Whitney U mostrados en la Tabla 3 (columna EG), revelan que las capas del modelo muestran una atención existente a lo largo de sus capas, pero menos definida y poco clara con respecto al resto de los análisis. Para el conjunto STS-Benchmark las capas 11 y 12 parecen tener una mayor tendencia a enforcarse en la estructura gramatical, mientras que para el conjunto SICK-R la capa 11 y 12 no muestran esta tendencia. Por otro lado, las capas 3 y 9 tienden a tener agrupaciones más similares entre sí en ambos conjuntos de datos, indicando que el enfoque en la agrupación de estructuras gramaticales más similares es débil.

La herramienta visual revela que las muestras con estructuras similares tienden a mantenerse cercanas entre sí, cuyo detalle puede verse en las Fig. 20 y 21 del anexo D. Aunque se observa un patrón o tendencia en las muestras con estructuras gramaticales similares, este patrón no está claramente definido y en ocasiones presenta dificultades al intentar agruparlas de manera coherente.

VI. Discusión

Los resultados del análisis de la atención concuerdan con observaciones previas mencionadas por [2], donde se señala una mayor dispersión en las autoatenciones en las capas iniciales del modelo, en contraste con una menor dispersión en las capas finales. Esta variabilidad se refleja en la calidad de los agrupamientos que pueden ser obtenidos.

Coincidimos con la observación de [3] respecto a que elementos como el tamaño de la secuencia son destacados en las capas iniciales del modelo, aunque su impacto tiende a reducirse a medida que avanzamos a capas más profundas. No obstante, es crucial señalar que el modelo aún conserva cierto grado de atención hacia este aspecto, aunque de manera significativamente menor.

En relación con la similitud semántica, se observa actividad en las capas 9 y 10 del modelo para ambos conjuntos de datos, así como en capas intermedias, como las capas 6 y 8. Este hallazgo respalda afirmaciones previas realizadas por [5].

En relación a la sintaxis, los experimentos realizados no proporcionan claridad sobre qué capas del modelo centran su atención en aspectos sintácticos. No obstante, nuestras observaciones revelan una tendencia consistente hacia la agrupación de estructuras gramaticales similares en al menos el 75% de las capas. Para obtener conclusiones más precisas sobre la estructura gramatical, se requiere un análisis más detallado y la exploración de enfoques adicionales.

Además, sería beneficioso realizar un análisis semántico más profundo para determinar si los agrupamientos están abstrayendo conceptos y si el modelo puede discernir homonimias y sinonimias en función del contexto.

Es crucial resaltar que todos los experimentos presentados en este trabajo incorporan una fase de ajuste fino. Sería sumamente relevante replicar estos mismos experimentos prescindiendo del ajuste fino para evaluar cómo afecta a los resultados. Dado que las representaciones utilizadas en este estudio son bidimensionales, existe la posibilidad de una pérdida significativa de información. Por tanto, sería beneficioso realizar pruebas y evaluaciones de agrupamientos con una dimensionalidad más alta para verificar estos mismos hallazgos, utilizando técnicas como PCA o T-SNE para su visualización.

VII. Conclusiones

El enfoque de interpretación lingüística de las atenciones del modelo BERT a través del aprendizaje no supervisado proporciona una herramienta valiosa para la investigación. Los hallazgos obtenidos mediante este método concuerdan de manera significativa con las conclusiones establecidas en el estado del arte, con matices diferentes. Se observa que la longitud de las sentencias recibe una atención destacada en las capas iniciales del modelo. Por otro lado, el modelo muestra enfoque en la similitud semántica en las capas 6, 8, 9 y 10. En cuanto a la sintaxis, se puede ver que las estructuras gramaticales similares tienden a agruparse.

Tanto el método utilizado para generar abstracciones a partir de las atenciones del modelo BERT como el diseño del modelo autocodificador lograron una efectiva abstracción, a pesar de la considerable reducción en la dimensionalidad.

Referencias

[1]. E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned,” *arXiv preprint* arXiv:1905.09418, 2019. DOI: 10.18653/v1/P19-1580.

[2]. K. Clark, U. Khandelwal, O. Levy, and C.D. Manning, “What does Bert look at? An analysis of Bert’s attention,” *arXiv preprint* arXiv:1906.04341 (2019. DOI: 10.18653/v1/W19-4828.

[3]. G. Jawahar, B. Sagot, and D. Seddah, “What does Bert learn about the structure of language?” in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019. DOI: 10.18653/v1/P19-1356.

[4]. I. Tenney, D. Das, and E. Pavlick, “Bert rediscovers the classical NLP pipeline,” *arXiv preprint*, arXiv:1905.05950, 2019. DOI: 10.48550/arXiv.1905.05950.

[5]. J. Niu, W. Lu, and G. Penn, “Does Bert rediscover a classical NLP pipeline?” in *Proceedings of the 29th International*

ISSN 2395-8618

Conference on Computational Linguistics, pp. 3143–3153, 2022. DOI: 10.48550/arXiv.1905.05950.

[6]. E. Reif, A. Yuan, M. Wattenberg, F.B. Viegas, A. Coenen, A. Pearce, and B. Kim, "Visualizing and measuring the geometry of Bert," *Advances in Neural Information Processing Systems*, Vol. 32, 2019. DOI: 10.48550/arXiv.1906.02715.

[7]. J. Vig, "Bertviz: A tool for visualizing multihead self-attention in the Bert model," in *ICLR workshop: Debugging machine learning models*, Vol. 23, 2019.

[8]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol. 30, 2017. DOI: 10.48550/arXiv.1706.03762.

[9]. M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 2018. DOI: 10.48550/arXiv.1802.05365.

[10]. P. Matthew, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations." in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 2227–2237, Association for Computational Linguistics, 2018. DOI: 10.18653/v1/N18-1202.

[11]. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., *Improving language understanding by generative pre-training*, 2018.

[12]. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Computation and Language*, 2018. DOI: 10.48550/arXiv.1810.04805.

[13]. D. Yenicelik, F. Schmidt, and Y. Kilcher, "How does Bert capture semantics? A closer look at polysemous words," in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 156–162, 2020.

[14]. J. Novikova and K. Shkaruta, "Deck: Behavioral tests to improve interpretability and generalizability of Bert models detecting depression from text," *Computation and Language*, 2022. DOI: 10.48550/arXiv.2209.05286.

[15]. V.W. Anelli, G.M. Biancofiore, A. De Bellis, T. Di Noia, and E. Di Sciascio, "Interpretability of Bert latent space through knowledge graphs," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3806–3810, 2022. DOI: 10.1145/3511808.3557617.

[16]. D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," in *International Workshop on Semantic Evaluation, Computer Science, Linguistics*, 2017. DOI: 10.18653/v1/S17-2001.

[17]. M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th international workshop on semantic evaluation*, pp. 1–8, 2014.