El corpus ROBOT-TALK para el reconocimiento del origen robótico de textos en español

Ana M.ª Fernández-Pampillón Cesteros (Universidad Complutense de Madrid, Spain)

Lara Alonso Simón (Universidad Complutense de Madrid, Spain)

ROBOT-TALK es un corpus monitor comparable de textos humanos en español y su contrapartida escrita por grandes modelos generativos del lenguaje (en adelante «modelos del lenguaje»). Su objetivo es permitir el estudio de posibles rasgos lingüísticos diferenciadores entre textos generados automáticamente y los generados por las personas en el marco del PID2022-140897OB-I00 (https://www.ucm.es/robottalk/).

Los modelos del lenguaje producen textos gramaticalmente correctos y con alto nivel de coherencia, por lo que es difícil distinguirlos de los humanos (Uchendu, Le y Lee, 2023). Esto plantea desafíos de alto impacto político, económico y social cuando se usan con fines maliciosos (Pizarro, 2019; Pavlyshenko, 2022; Cardenuto, Yang, Padilha, Wan, Moreira, Li, Wang, Andaló, Marcel y Rocha, 2023; Crothers, Japkowicz y Viktor, 2023). Entonces, resulta clave conocer si el autor es una persona o una máquina (Maloyan, Nutfullin y Ilyushin, 2022). Así, ROBOT-TALK proporciona el primer recurso lingüístico en español para el reconocimiento de autoría humana vs. «robótica» de textos.

Contiene textos humanos y contrapartidas «robóticas». Cubre tres géneros de diferentes niveles de formalidad en la lengua escrita: artículos científicos, noticias y reseñas. Cada par de textos, de longitud similar, trata el mismo tema. Los modelos del lenguaje se seleccionaron aplicando dos criterios: que generen textos de muy alta calidad y sean accesibles mediante API o una interfaz. Se recogen muestras de Claude de Anthropic, Falcon de Technology Innovation Institute, ChatGPT-3.5-turbo y ChatGPT-4 de OpenAI, Gemini de Google y Mixtral-8x7B-Instruct-v0.1 de Mixtral AI. El periodo de recogida abarca desde julio de 2023 hasta febrero de 2025. La tabla 1 muestra la distribución por tipo de autor y género textual.

corpus	huma	bard	clau	g35t	gpt4	mxit	total por
artículos	144	90	0	90	90	90	504
noticias	171	171	60	111	151	111	775

reseñas	160	160	65	95	160	95	735
total por autor	475	421	125	296	401	296	**2014**

Tabla 1. Composición del corpus

Actualmente, el corpus consta de 2014 textos: 475 humanos (144 artículos, 171 noticias y 160 reseñas) y 1539 generados por modelos del lenguaje.

El método de construcción consta de cuatro pasos:

(1) búsqueda del texto humano, asegurando que la autoría es humana;

(2) almacenamiento del texto humano en formato xml para describir los metadatos y la estructura del contenido;

(3) generación del texto robótico comparable mediante *prompts* que determinan que el contenido tenga el mismo registro, longitud y temática que el texto humano comparable; y

(4) almacenamiento del texto robótico en formato xml con los metadatos y la estructura del contenido.

El etiquetado en xml de los textos del corpus permite su consulta con cualquier herramienta de análisis textual (ej. SketchEngine) que soporte este estándar de marcado. En este sentido, ROBOT-TALK se ha utilizado con la herramienta SketchEngine para realizar (1) un análisis lingüístico profundo para encontrar los rasgos más salientes que caracterizan los textos generados por los modelos de lenguaje; (2) un análisis estadístico de rasgos lingüísticos propios de los modelos del lenguaje frente a un posible estilo general humano en español (Alonso Simón, Fernández-Pampillón Cesteros, Fernández Trinidad y Márquez Cruz, 2024); y (3) la construcción de clasificadores automáticos binarios y multiclase basados en aprendizaje automático para distinguir textos róticos y humanos. El corpus no está todavía publicado, pero se puede consultar una muestra en https://www.ucm.es/robottalk/corpus-robot-talk.

Palabras clave: Corpus monitor, corpus comparable, corpus de texto en español, grandes modelos del lenguaje, identificación de textos automáticos.

Referencias

Alonso Simón, L., Fernández-Pampillón Cesteros, A.M., Fernández Trinidad, M. y Márquez Cruz, M. (2024). ¿Tienen GPT-3.5 y GPT-4 un estilo de escritura

diferente del estilo humano? Un estudio exploratorio para el español. *RAEL: Revista Electrónica de Lingüística Aplicada, 23*, 34-54. https://doi.org/10.58859/rael.v23i1.666

Cardenuto, J. P., Yang, J., Padilha, R., Wan, R., Moreira, D., Li, H., Wang, S., Andaló, F., Marcel, S. y Rocha, A. (2023). The Age of Synthetic Realities: Challenges and Opportunities. *APSIPA Transactions on Signal and Information Processing, 12*(1), 1–62. https://doi.org/10.1561/116.00000138

Crothers, E. N., Japkowicz, N. y Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *arXiv:2210.07321*, Oct. 2023. https://doi.org/10.1109/ACCESS.2023.3294090

Maloyan, N., Nutfullin, B. y Ilyushin, E. (2022). DIALOG-22 RuATD Generated Text Detection. En *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* (2022) (pp. 396–401). Moscú: RSUH. *arXiv.2206.08029*. https://doi.org/10.48550/arXiv.2206.08029

Pavlyshenko, B. M. (2022). Methods of Informational Trends Analytics and Fake News Detection on Twitter. *arXiv:2204.04891v1*. https://doi.org/10.48550/arXiv.2204.04891

Pizarro, J. (2019). Using n-grams to detect bots on Twitter, notebook for PAN at CLEF 2019. En L. Cappellato, N. Ferro, D. E. Losada, and H. Müller (Eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*. https://ceur-ws.org/Vol-2380/paper_183.pdf

Uchendu, A., Le, T. y Lee, D. (2023). Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter, 25*(1), 1–18. https://doi.org/10.1145/3606274.3606276