Pipeline Stage	Model	Statistic	Result	Used
OCR	EffOCR-Word (Small) (3)	CER	0.015	*
	Tesseract OCR	CER	0.106	*
	EasyOCR	CER	0.170	
	PaddleOCR	CER	0.304	
Deduplication	Locally Sensitive Hashing	ARI	73.7	
	N-gram Overlap	ARI	75.0	
	Neural biencoder (7)	ARI	91.5	*
	Biencoder + crossencoder	ARI	93.7	
Georeferencing	N-gram matching	Accuracy	94.9	*
	GPT-4o-mini	Accuracy	85.3	
NER	Custom NER (6)	F1	90.4	*
	Roberta-Large tuned on CoNLL03 (4)	F1	77.8	
Entity disambiguation	LinkNewsWikipedia (1)	Accuracy	78.3	*
	BLINK (8)	Accuracy	59.9	
	GENRE (5)	Accuracy	63.4	
	ReFinED (2)	Accuracy	65.4	

Table 1: Models considered for each stage of the pipeline and comparisons of their performance on the test set for each stage. Starred models are those used. For CER, smaller values are better, while for the other statistics, bigger numbers are better. CER = Character Error Rate, ARI = Adjusted Rand Index.

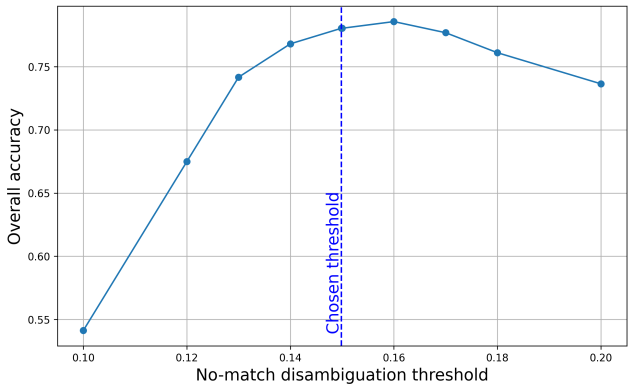

Figure 1: Sensitivity of disambiguation results to choice of no-match threshold.

References

[1] Arora, A., Yang, X., Jheng, S. Y., and Dell, M. Linking representations with multimodal contrastive learning. *arXiv preprint arXiv:2304.03464* (2023).

[2] Ayoola, T., Tyagi, S., Fisher, J., Christodoulopoulos, C., and Pierleoni, A. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *NAACL* (2022).

[3] Bryan, T., Carlson, J., Arora, A., and Dell, M. Efficientocr: An extensible, open-source package for efficiently digitizing world knowledge". *Empirical Methods on Natural Language Processing (Systems Demonstrations Track)* (2023).

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[5] De Cao, N., Izacard, G., Riedel, S., and Petroni, F. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904* (2020).

[6] Franklin, B., Silcock, E., Arora, A., Bryan, T., and Dell, M. News deja vu: Connecting past and present with semantic search, 2024.

[7] Silcock, E., D'Amico-Wong, L., Yang, J., and Dell, M. Noise-robust de-duplication at scale. In *The Eleventh International Conference on Learning Representations* (2022).

[8] Wu, L., Petroni, F., Josifoski, M., Riedel, S., and Zettlemoyer, L. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814* (2019).