Applied Intelligence, pp.711-1-711-18, .

Enhancing keyphrase extraction from long scientific documents using graph embeddings

A.J. López López; D. Mahata; J. Portela González; R. Martínez Cruz

Abstract-

This study explores the integration of graph neural network (GNN) representations with pre-trained language models (PLMs) to enhance keyphrase extraction (KPE) from lengthy documents. We demonstrate that incorporating graph embeddings into PLMs yields richer semantic representations, especially for long texts. Our approach constructs a co-occurrence graph of the document, which we then embed using a graph convolutional network (GCN) trained for edge prediction. This process captures non-sequential relationships and long-distance dependencies, both of which are often crucial in lengthy documents. We introduce a novel graph-enhanced sequence tagging architecture that combines PLM-based contextual embeddings with GNN-derived representations. Through evaluations on benchmark datasets, our method outperforms state-of-the-art models, showing notable improvements in F1 scores. Beyond performance on standard benchmarks, this approach also holds promise in domains such as legal, medical, and scientific document processing, where efficient handling of long texts is vital. Our findings underscore the potential for GNNs to complement PLMs, helping address both technical and real-world challenges in KPE for long documents.in KPE for long documents

Index Terms- Keyphrase extraction · Graph embeddings · Long documents · Pre-trained language models · Natural language processing · Deep learning

Citation:
López López, A.J.; Mahata, D.; Marti-nez-Cruz, R.; Portela, J. "Enhancing keyphrase extraction from long scientific documents using graph embeddings", Applied Intelligence, pp.711-1-711-18, .