Lab06 - Artigo de Dataset Público

Aluno

- 218733: Joao Pedro de Moraes Bonucci

Análise do Artigo EventNarrative: A large-scale Event-centric Dataset for Knowledge Graph-to-Text Generation

campo	valor
referência	COLAS, Anthony; SADEGHIAN, Ali; WANG, Yue; WANG, Daisy Zhe. EventNarrative: A large-scale Event-centric Dataset for Knowledge Graph-to-Text Generation. NeurIPS 2021 Datasets and Benchmarks Track (Round 1), University of Florida, p. 1-13, 7 jun. 2021. Disponível em: https://openreview.net/forum?id=3ZQqjt_Q6b. Acesso em: 25 set. 2021.
link	https://openreview.net/forum?id=3ZQqjt_Q6b
dataset	www.kaggle.com/dataset/551460c9e6dc73dfdf5bafa1b8a3ac8217c13b3845a602a68f049d1d08237d47
formato	JSON

Resumo

> O artigo apresenta um banco de dados de conhecimento que seja paralelo, publico e na estrutura graph-to-text. Ou seja, um banco deados onde input e output sao firmemente acoplados e possui estrutura de grafos. O artigo tem como objetivo suprir a demanda dentro do campo de Processamento de Linguagem Natural que procura uma quantidade massiva de dados de qualidade, nao esparsos para melhorarem seus modelos e tecnicas. Para isso, os autores focaram em reduzir o numero de desconexoes entre os dados e linguagem natural de texto. Outro objetivo do banco porposto eh trazer uma forma de coleta diferente, baseada em eventos, para que assim consigam unir diferentes componentes como inviduos, tempo, relacoes, etc; tornando a informacao contida nos grafos mais rica e util. As fontes utilizadas para a criacao do banco sao: Wikidata, Wikipedia, e EventKG

Perguntas de pesquisa/análises

> Este data set permite conseguir de forma rapida e estruturada os dados presentes em plataformas de conhecimento geral como WikiPedia. Apesar de por ser utilizado de varias formas, ele foi pensado para dar melhores insumos a fim de treinar e melhorar tecnicas e modelos na area de Processamento de Linguagem natural

Trabalhos relacionados

iniciativas parecidas

> Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. Reading the manual: Event extraction as definition comprehension. In Proceedings of the Fourth Workshop on Structured Prediction for NLP, pages 74–83, 2020.

> Xinya Du and Claire Cardie. Event extraction by answering (almost) natural questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 671–683, 2020.

> Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 894–908,2021.

bancos de dados relacioandos

> CNN/DM dataset
>
> - Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1, pages 1693–1701, 2015.
> - Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstrac-tive text summarization using sequence-to-sequence rnns and beyond. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290, 2016.

> Gigaword Dataset
>
> - Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractivesentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, 2015.

> Persona-Chat Dataset
>
> - Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston.Personalizing dialogue agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),pages 2204–2213, 2018.

> DSTC7 Dataset
>
> - Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. Grounded response generation task at dstc7. In AAAI Dialog System Technology Challenges Workshop, 2019.

> CoQA Dataset
>
> - Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266,2019.

> AGENDA Dataset
>
> - Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text generation from knowledge graphs with graph transformers. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2284–2293, 2019.

> WebNLG Challange dataset
>
> - Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creat- ing training corpora for nlg micro-planning. In 55th annual meeting of the Association for Computational Linguistics (ACL), 2017.

> GenWiki dataset
>
> - Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. Genwiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2398–2409, 2020.