EMBEDDIA at SemEval-2022 Task 8: Investigating Sentence, Image, and Knowledge Graph Representations for Multilingual News Article Similarity

Elaine Zosa*
University of Helsinki
Helsinki, Finland
elaine.zosa@helsinki.fi

Emanuela Boros*
University of La Rochelle
La Rochelle, France
emanuela.boros@univ-lr.fr

Boshko Koloski*
Jožef Stefan Institute
Jožef Stefan Institute IPS
Ljubljana, Slovenia
boshko.koloski@ijs.si

Lidia Pivovarova
University of Helsinki
Helsinki, Finland
lidia.pivovarova@helsinki.fi

Abstract

In this paper, we present the participation of the EMBEDDIA team to the SemEval 2022 Task 8 (*Multilingual News Article Similarity*). We cover several techniques and propose different methods for finding the multilingual news article similarity by exploring the dataset in its entirety. We take advantage of the textual content of the articles, the provided metadata (e.g., titles, keywords, topics), the translated articles, the images (those that were available), and knowledge graph-based representations for entities and relations present in the articles. We, then, compute the semantic similarity between the different features and predict through regression the similarity scores. Our findings show that, while our researched methods obtained promising results, exploiting the semantic textual similarity with sentence representations is unbeatable. Finally, in the official SemEval 2022 Task 8, we ranked fifth in the overall team ranking cross-lingual results, and second in the English-only results.

1 Introduction

Detecting news stories related to a single theme and combining them into news clusters has been an increasing interest in the creation of news aggregators that consolidate thousands of articles from different publishers and websites (Pranjić et al., 2020). Tracking similarity of news coverage between different outlets or regions has also been urgent and challenging. For example, whether previously with Ebola or recently with the COVID-19 pandemic, monitoring and containment of infectious disease outbreaks has remained a key component of public health strategy to contain the diseases. The ability to track disease outbreaks in an accurate manner is critical in the deployment of efficient intervention measures. As such reports may not only be in English, there is also a need for effective multilingual systems. Hence, recent research has been focused on the area of identifying similarities between documents, phrases, stories, etc.

Semantic textual similarity (STS) deals with determining how similar two groups of sentences are by measuring their semantic similarity. Over the years, several solutions were proposed to assess STS. The most general approach is pre-training on massive datasets before fine-tuning on subsequent downstream tasks (Jiang et al., 2020; Raffel et al., 2019; Lan et al., 2019; Yang et al., 2019; Liu et al., 2019; Sanh et al., 2019). Other works considered finding the similarity by classifying texts using BERT-based models (Devlin et al., 2019) with a pair of sentences packed together as input (Yang et al., 2019; Liu et al., 2019; Sanh et al., 2019; Wang et al., 2019).

The SemEval 2022 Task 8 (*Multilingual News Article Similarity*) aimed at developing systems that identify multilingual news articles that provide similar information by rating them on a real-valued $[1-4]$ scale, from most to least similar.

In this paper, we cover several techniques and propose different methods for finding the multilingual news article similarity by exploring the dataset in its entirety. We consider that the textual content, the provided metadata (e.g., title, keywords, topics), representative images corresponding to the news articles, and knowledge graph-based representations for entities and relations present in the articles, would draw on a multiplicity of modes, all of which contribute to the meaning and the main story of the news articles. Moreover, we also translate

* Equal contribution from all the authors.

the articles in a high-resource language (English) in order to assess the ability of our models in an English-only context. Therefore, we investigate the multimodality of the data by experimenting with sentence, image, and knowledge graph embeddings in two scenarios: (1) by directly computing the semantic similarity between the different features and (2) by learning through regression and predicting the similarity scores.

2 Data

The training data has 4,964 article pairs from seven languages (English, German, Spanish, Arabic, Polish, Turkish, and French) and gold standard similarity scores for six dimensions (*Geography, Entities, Time, Narrative, Style, Tone*), plus the *Overall* score. The final evaluation data has 4,902 pairs and three "surprise" languages that were not present in the training data (Chinese, Italian, and Russian).

	Train	**Eval**
Monolingual pairs	4,387	3,462
Cross-lingual pairs	577	1,440
Unseen language pairs	NA	2,000
Total	4,964	4,902
Top image	6,755	7,569

Table 1: Training and evaluation data statistics.

Moreover, the metadata includes the article titles, several specific topics and keywords, and links to representative images. The statistics of the training and final evaluation data are in Table 1. Since some of our methods use images, we also report in the table a total number of images we were able to download for the datasets. We use only images from the URL specified as *top_image* in the JSON files of the articles.

3 Experiments

Next, we detail all our approaches and perform a detailed error analysis[1]. The evaluation is performed in terms of Pearson correlation. Our results are presented in Table 2. Each type of approach is detailed with the corresponding pre-trained models[2]. Also, each type of model has an id corresponding to the subsection number is detailed (1a, 2b, etc.).

[1]Our code is available at https://github.com/bkolosk1/semeval-2022-MNS

[2]All models are available at https://huggingface.co/.

3.1 Semantic Textual Similarity

A straightforward solution for finding the similarity between two texts is approaching it with sentence embeddings. Thus, we start our experimental setup by encoding the articles with Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a modified pre-trained BERT (Devlin et al., 2019) that uses a siamese and triplet network structure to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. We explore this approach by encoding the articles with SBERT and and using the cosine similarity of articles pairs as the predicted *Overall* score. For these experiments, we used the default hyperparameters provided by Reimers and Gurevych (2019).

Similarity based We first concatenate the title and the textual content of each article, and due to the multilingual characteristic of the data, we encode the textual sequence with a pre-trained multilingual SBERT model and compute the Pearson correlation between the cosine similarity of these sentence embeddings and the gold labels, results presented in Table 2 (1a). Then, we experiment with machine translating all the non-English articles to English using Google Translate and use an English SBERT model. The results are presented in Table 2 (1b).

Regression based We fine-tune the SBERT model on the multilingual pairs, results presented in Table 2 (1c) and on the machine-translated articles, results presented in Table 2 (1d). For fine-tuning, we use only the *Overall* score as the target similarity score. Since the similarity scores provided in the training data are in the range [1-4] from *most to least* similar, we normalize the *Overall* scores (the scores provided by cosine similarity are in the $[0, 1]$ range from *least to most* similar).

3.2 Image Similarity & Regression

We download the images from the *top_image*, and as observed in Table 1, out of 9,928 articles (4,964 pairs), only 6,755 articles has a viable image in the train set. Out of 9,804 articles in the test set, only 7,567 were downloaded. For both, only around 60% of the articles has an image that could be used. Moreover, only around half of the pairs in both sets have representative images for both articles. Nonetheless, we attempt at using them in our approaches. We experiment with two recent pre-trained models, CLIP (Radford et al., 2021)

Model		Pearson-r
Semantic Textual Similarity & Regression		
(1a) SBERT (PARAPHRASE-MULTILINGUAL-MPNET)	Similarity	0.6713
(1b) SBERT (ALL-MPNET) - Google Translate	Similarity	0.7139
(1c) SBERT (PARAPHRASE-MULTILINGUAL-MPNET)	Regression	0.7396
(1d) SBERT (ALL-MPNET) - Google Translate	Regression	**0.7835**
Image Similarity & Regression		
(2a) Images (CLIP-VIT-PATCH32)	Similarity	0.2991
(2b) Cross-images (CLIP-VIT-PATCH32)	Similarity	0.2607
(2c) Images (CLIP-VIT-PATCH32)	Regression	0.1043
(2d) Images (VIT-LARGE-PATCH32)	Regression	0.1124
Knowledge Graph Similarity & Regression		
(3a) KGm+LSA+SBERT (DISTILBERT+XLM-ROBERTA+ROBERTA)	Similarity	0.7128
(3b) KGm+LSA+SBERT (DISTILBERT)	Regression	0.5134
Text & Image Regression		
(4a) Text+metadata (XLM-ROBERTA-LARGE)	Regression	0.7773
(4b) Text+metadata+images (XLM-ROBERTA-BASE+CLIP-VIT-PATCH32)	Regression	0.7020
(4c) Text+metadata+images (XLM-ROBERTA-LARGE+VIT-LARGE-PATCH32)	Regression	0.7335

Table 2: Correlation between similarity scores from different proposed models and the *Overall* score.

and ViT (Dosovitskiy et al., 2020).

Similarity based As for texts, we generate the image embeddings using CLIP, compute the cosine similarity between the paired images, and report the Pearson correlation between the obtained similarities and the gold labels. The results are presented in Table 2 (2a). For the missing images, we assign the default cosine similarity of 0.5. We also experiment with an alternative strategy, which takes advantage of the fact that CLIP is a multimodal model and produces images and text embeddings in the same space, *Cross-images*. In this strategy, we compute all possible similarities between data points: image-to-image, text-to-text, and image-to-text. In the best case, when both images are available, this results in a total of four similarity values. In the worse case, only the similarity between texts is used. If only one image is available, the strategy results in two similarities: text-to-image and text-to-text. The final score is obtained by averaging the similarities available. Surprisingly, this strategy works slightly worse than an approach based solely on images, as it can be seen in Table 2 (2b).

Regression based This method is detailed in Section 3.4. The results are presented in Table 2 (2c and 2d).

3.3 Knowledge Graph Similarity & Regression

We use the Wikidata5m (Wang et al., 2021) knowledge graph (KG) in order to retrieve knowledge-based features as used by Koloski et al. (2022). Similarly, we exploit six different knowledge graph embeddings: transE (Bordes et al., 2013), rotatE (Sun et al., 2019), complEx (Trouillon et al., 2016), distmult (Yang et al., 2015), simplE (Kazemi and Poole, 2018), and quate (Zhang et al., 2019). We use GraphVite (Zhu et al., 2019), a system for training node embeddings, pre-trained on aforementioned embeddings of the Wikidata KG. For these experiments, we use the translated articles. We concatenate the title and the body of the articles to search n-grams of sizes 1, 2, and 3, as potential concepts appearing in the KG. After extracting potential candidates, we extract the embeddings of the candidates from the KG. In addition, we generate latent semantic analysis (LSA), SBERT and stats representations as done by Koloski et al. (2021). The results are in Table 2 (3a and 3b).

Similarity based First, we generate all ten feature spaces. Next, we generate combinations of feature spaces (1024 combinations in total), we concatenate and normalize them (KGm). Finally, we find thresholds to estimate the similarity scores, with respect to the *Overall* label. Our best results are presented in Table 2 (3a).

Regression based We utilize all six of the aforementioned KG representations, LSA and DistilBERT (Sanh et al., 2019) SBERT representations. Next, we use a singular value decomposition (SVD) to generate a new latent space of the devised features and we proceed to learn a deep neural network on the whole target space. Our best results are presented in Table 2 (3b).

3.4 Text & Image Regression Models

We also propose a classical approach that considers the task of finding the similarity between two articles by considering it as a regression task, and by predicting the similarity for the *Overall* score. This approach consists of a pre-trained and fine-tuned language model (BERT (Devlin et al., 2019) pre-trained on multilingual data). Because these models expect input data in a specific format, we need a special token, [SEP], to mark the end of a sentence or the separation between two sentences, and [CLS], at the beginning of a text generally used for classification or regression tasks.

Regression based After the pair of articles are tokenized and together encoded with [CLS] at the start and then separated by [SEP], they are passed through the encoder. Similarly, images are passed through a ViT encoder. For the missing images, we generate a *fake* white image. The BERT output token representations are afterward concatenated with the [CLS] representation and Vit output image representation followed by a linear layer for regression. The learning of the model is conducted end-to-end by optimizing an objective corresponding to *Overall* prediction. For these experiments, we utilized AdamW (Kingma and Ba, 2014) with a learning rate of 1×10^{-5} for 2 epochs with mean squared error (MSE) loss. We also consider a maximum sentence length of 512 (the maximum possible accepted by BERT or RoBERTa). These results are presented in Table 2 (from 4a to 4d).

4 Error Analysis

Semantic Textual Similarity We can substantially improve the English-only model (1d) for STS by fine-tuning not just with monolingual English pairs from the training data but by using all the machine-translated pairs. However, we observe some cases where our best performing fine-tuned model is misled by similar turns of phrase even if the article pair covers different events. We show extracts from an article pair in Table 3 that covers a fire and a traffic accident, respectively. The gold *Overall* score for this pair is 4.0 (very dissimilar) but our best-performing model scores it at 3.1 (somewhat dissimilar) due to the similar phrasing that opens the articles and that they both mention the same-named entities.

Article1	Article2
1492472369 (EN): **At least one person has been confirmed dead**, following Saturday's fire that gutted the Mgbuka Obosi Spare Parts Market in Idemili North **Local Government Area of Anambra** . . . Mr **Edwin Okadigbo**, the Public Relations Officer of the **Nigeria Security and Civil Defence Corps (NSCDC)**, Anambra command . . .	1530831511 (EN): **At least, one person has been confirmed dead** . . . in a road mishap that involved a commercial bus and a motorcycle in Mbosi junction, Ihiala **Local Government Area of Anambra** State on Tuesday . . . Spokesperson of the **Nigeria Security and Civil Defence Corps, NSCDC** in Anambra State, **Edwin Okadigbo** said preliminary . . .

Table 3: Extracts from an article pair and their similarity scores predicted by SBERT translated (1b) with an *Overall* of 3.159, while the gold score is 4.0. Similar terms are in bold.

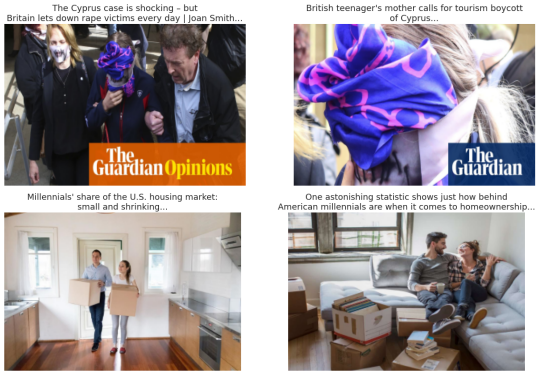

Figure 1: Two pairs of similar English articles (gold score of 1.0 for both) correctly predicted by the image-based model (1.28 & 1.0), and incorrectly predicted by SBERT (1.83 & 1.63).

Figure 2: A pair of marginally similar Russian articles (gold score of 2.0), which is an unseen language during training, correctly predicted by the image-based model (1.64), and incorrectly predicted by SBERT (2.94).

Image Similarity & Regression We analyze the scores predicted by two textual-based methods, (1d) SBERT with the best scores when using only images (2a). Out of $4,902$ pairs in the evaluation

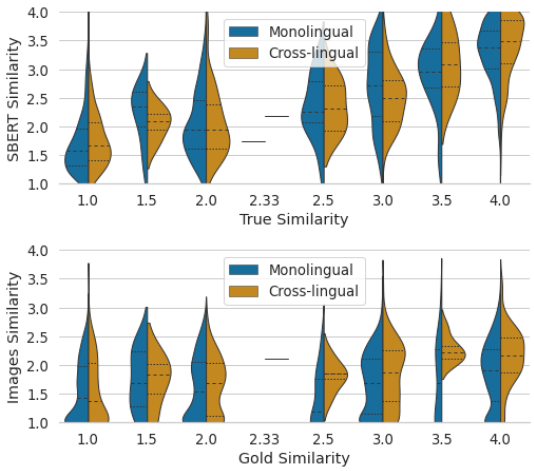

Figure 3: Similarity scores for the article pairs with available images for the image-based model, Images (2a) and Text+metadata (4c).

set (Table 1), only 2,009 have representative images for both news articles. Thus, we look closer at the predictions for these pairs and notice that 13% of them (262 pairs) are correctly predicted by (2a), and not by (1d), all of these being images with either faces or visible and clearly distinguished texts or text boxes, as shown in Figure 1 for two pairs of English articles. We also give an example where this model is able to better distinguish the similarity between two articles in an unseen language (Russian) in Figure 2, where the articles speak of the same topic but describe different events.

Knowledge Graph Similarity & Regression We analyze the representations of articles based on the number of concepts retrieved from the WikiData5m. The top-most appearing concepts include entities such as *government, coronavirus, epidemic, report, information, death, economy, etc.*, showcasing us that most of the articles report about the pandemic, the statistics, and results. The distribution of concepts per document is shown in Figure 4. Originally, the Wikidata5m KG is based only on English concepts. We notice a performance drop for the non-English articles, due to the translation to English, some original concepts are lost and replaced with another. For the training set, we retrieved an average of 55 concepts per article, while for the evaluation set we obtained 54 concepts per article. The lowest amount of retrieved concepts was 1 and the highest was 757.

Text & Image Regression Figure 3 presents the Images (2a) similarity scores in comparison with

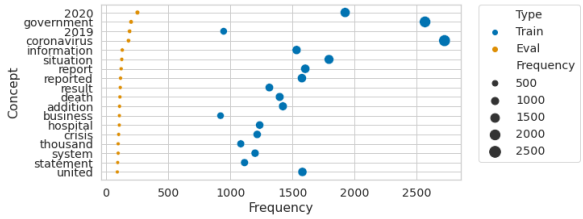

Figure 4: Distribution of KG concepts in the train and eval sets.

Text+metadata (4b) and Text+metadata+images (4d) similarity scores. First, the results for Text+metadata (4a) seem to be rather similarly distributed to those provided by SBERT, with a slight difference in the monolingual pairs with a gold score of 1.5, where SBERT generally predicts a similarity of 2.5. When using image representations, not surprisingly, we notice that the results for Images (2a) are generally staying around an average of 2.0, proving that having only around half of the train and eval sets with images is not enough in helping distinguish news articles.

SemEval-2022 Task 8 In the official SemEval-2022 Task 8, we ranked fifth in the overall team ranking multilingual and cross-lingual results, and second in the English-only results, both with our *Semantic Textual Similarity* with pre-trained multilingual and monolingual SBERT models.

5 Conclusions

In this paper, we covered several techniques for finding the similarity of multilingual and monolingual news articles in the context of SemEval-2022 Task 8 *Multilingual News Article Similarity*. We notice that, even is using images and knowledge graph representations give promising results, approaching STS with sentence embeddings is still unbeatable. However, images, being a language-agnostic medium, could be helpful if they represent people or text boxes. Future work could include an adaptable inclusion of images (for handling missing imags) and the usage of multilingual knowledge graph representations.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (EMBEDDIA), and by the ANNA and Termitrad projects funded by the Nouvelle-Aquitaine Region.

References

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4289–4300.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Boshko Koloski, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. 2021. Identification of covid-19 related fake news via neural stacking. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 177–188, Cham. Springer International Publishing.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marko Pranjić, Vid Podpečan, Marko Robnik-Šikonja, and Senja Pollak. 2020. Evaluation of related news recommendations using document similarity methods. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JDTH2020)*, pages 81–86, Ljubljana, Slovenia. Inštitut za novejšo zgodovino.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-

training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. 2019. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2731–2741.

Zhaocheng Zhu, Shizhen Xu, Meng Qu, and Jian Tang. 2019. Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference*, pages 2494–2504. ACM.