BLCU-ICALL at SemEval-2022 Task 1: Cross-Attention Multitasking Framework for Definition Modeling

Cunliang Kong[1], **Yujie Wang**[2], **Ruining Chong**[1], **Liner Yang**[1]*,
Hengyuan Zhang[1], **Erhong Yang**[1], **Yaping Huang**[2]
[1]School of Information Science, Beijing Language and Culture University
[2]School of Computer and Information Technology, Beijing Jiaotong University
cunliang.kong@outlook.com

Abstract

This paper describes the BLCU-ICALL system used in the SemEval-2022 Task 1 Comparing Dictionaries and Word Embeddings, the Definition Modeling subtrack, achieving 1st on Italian, 2nd on Spanish and Russian, and 3rd on English and French. We propose a transformer-based multitasking framework to explore the task. The framework integrates multiple embedding architectures through the cross-attention mechanism, and captures the structure of glosses through a masking language model objective. Additionally, we also investigate a simple but effective model ensembling strategy to further improve the robustness. The evaluation results show the effectiveness of our solution. We release our code at: https://github.com/ blcuicall/SemEval2022-Task1-DM.

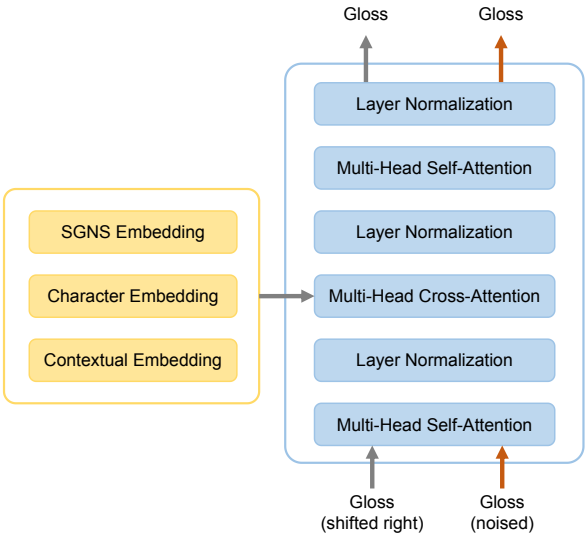

Figure 1: Architecture of the Cross-Attention Multitasking Framework.

1 Introduction

Word embeddings (Mikolov et al., 2013a; Pennington et al., 2014; Yogatama et al., 2015) are dense and low dimensional vectors used in many NLP tasks because they are found to be useful representations of words and often lead to better performance in various tasks. In recent years, large pretrained language models (PLMs), such as BERT (Devlin et al., 2019) and GPT (Petroni et al., 2019) families of models, have taken the NLP field by storm, achieving state-of-the-art performance on many tasks (Min et al., 2021). The contextual embeddings generated by PLMs are proven to capture syntax and semantic features of words (Jawahar et al., 2019; Turton et al., 2020). But for human beings, word embeddings containing these information is still a *black box* and unexplainable.

There have been many efforts devoted to evaluating the word embeddings' lexical information, such as the word similarity (Landauer and Dumais, 1997; Downey et al., 2007) and analogical relation (Mikolov et al., 2013c) tasks. However, these tasks can only serve as indirect evaluation methods. In light of this, Noraset et al. (2017) proposed the task of definition modeling to evaluate whether a word embedding can be employed to generate a dictionary gloss. Since the gloss is a direct and explicit statement of word meaning, this task provides a more transparent view.

The SemEval-2022 Task 1 Comparing Dictionaries and Word Embeddings (Mickus et al., 2022) aims at comparing the two types of semantic descriptions: dictionary glosses and word embeddings. The subtrack 1 is a definition modeling task, which requires models to generate glosses from word embeddings. The task provides data from 5 languages (English, Spanish, French, Italian, Russian) as well as static, character, and contextual embeddings.

Our team propose a transformer-based (Vaswani et al., 2017) Cross-Attention Multitasking Framework to explore the task and apply the framework to all 5 languages. We integrate the multiple embed-

*Corresponding author: Liner Yang.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 23 - 28
July 14-15, 2022 ©2022 Association for Computational Linguistics

	Train	Dev.	Test	SGNS Emb.	Character Emb.	Electra Emb.	Gloss Len.
English	43,608	6,375	6,221	✓	✓	✓	11.73
Spanish	43,608	6,375	6,221	✓	✓	✗	14.84
French	43,608	6,375	6,221	✓	✓	✓	14.31
Italian	43,608	6,375	6,221	✓	✓	✗	13.58
Russian	43,608	6,375	6,221	✓	✓	✓	11.32

Table 1: Detailed statistics of the dataset. The last column lists the average length of glosses in the training set.

ding architectures through a cross-attention mechanism, which allows the model to query all the embeddings at each time step during generation. To better capture the structure of glosses, we employ an additional masking language model (MLM) (Devlin et al., 2019) into the framework. We also investigate the ensemble strategies to further enhance the robustness.

Therefore, the contributions of our system lie in:

- We propose the Cross-Attention Multitasking Framework as a novel solution to the definition modeling task.
- The evaluation results show the effectiveness of our solution. Our system achieves 1st on Italian, 2nd on Spanish and Russian, and 3rd on English and French.

2 Background

The definition modeling subtrack provides participants with a multilingual dataset in the form of $\{E, \boldsymbol{g}\}$, where E is a set including SGNS (Mikolov et al., 2013b), character (Kim et al., 2016), and Electra (Clark et al., 2020) embeddings, and $\boldsymbol{g}$ is a dictionary gloss. This task takes E as the input, and requires models to generate $\boldsymbol{g}$. Note that all the embeddings have 256 dimensions, and the Electra embeddings are only available for 3 of the 5 languages. More detailed statistics of the dataset are listed in Table 1.

Many previous work used additional data to improve the performance of generation, such as example sentences (Gadetsky et al., 2018; Chang et al., 2018; Ishiwatari et al., 2019; Kong et al., 2020) and semantic features (Yang et al., 2020). Some studies also investigated how to employ PLMs for this task (Reid et al., 2020; Bevilacqua et al., 2020; Huang et al., 2021; Kong et al., 2022).

Differently, to keep the results linguistically significant and easily comparable, the SemEval-2022 Task 1 prohibits the usage of external data and PLMs. Therefore, our system focuses on effectively integrating all given embeddings and modeling the glosses.

3 System Overview

Figure 1 illustrates the entire architecture of our system, which is a Cross-Attention Multitasking Framework based on transformer. The framework consists of two objectives, namely the generation and reconstruction objectives. This section introduces the system in detail.

3.1 The Generation Objective

The generation objective serves as a standard transformer decoder, which generates the gloss as the following language model:

$$P(\boldsymbol{g}|E;\boldsymbol{\theta}) = \prod_t P\left(\boldsymbol{g}_t|\boldsymbol{g}_{<t}, E;\boldsymbol{\theta}\right), \quad (1)$$

where $\boldsymbol{g}_t$ is the t-th token in the gloss, and $\boldsymbol{\theta}$ is the set of parameters. The model is then optimized using the following loss function:

$$\mathcal{L}_{gen}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{g}\in D} \log P(\boldsymbol{g}|E;\boldsymbol{\theta}), \quad (2)$$

where D is the training dataset.

In the above operations, a crucial challenge is to integrate multiple embeddings corresponding to one word. We assume that the SGNS, character, and Electra embeddings contain different lexical features, and better results can be obtained by comprehensively considering all the information. To achieve that, we feed the set E, including all these embeddings, into the cross-attention mechanism:

$$\text{Cross-Attn}(H, E, E) = \text{softmax}(\frac{HE^T}{\sqrt{d_h}})E \quad (3)$$

where H is the hidden-states obtained from by self-attention, and d_h is the dimension of the hidden-states. This operation ensures the given embeddings are adaptively integrated at each time-step.

3.2 The Reconstruction Objective

Our system is a language model specially designed for dictionary glosses. We further enhance this model by incorporating a reconstruction objective.

We corrupt each gloss $\boldsymbol{g}$ by randomly substituting or blanking some words. And then we obtain a corrupted version $\tilde{\boldsymbol{g}}$. We input $\tilde{\boldsymbol{g}}$ into our system and obtain $\boldsymbol{g}$ by solving a self-supervised task of:

$$P(\boldsymbol{g}|\tilde{\boldsymbol{g}};\boldsymbol{\theta}) = \prod_t P(\boldsymbol{g}_t|\boldsymbol{g}_{<t},\tilde{\boldsymbol{g}};\boldsymbol{\theta}). \quad (4)$$

Note that we share exactly the same parameters $\boldsymbol{\theta}$ as in the generation objective. The model is optimized by the following loss function:

$$\mathcal{L}_{rec}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{g}\in D} \log P(\boldsymbol{g}|\tilde{\boldsymbol{g}};\boldsymbol{\theta}), \quad (5)$$

The goal of the reconstruction objective is to better model the glosses. Therefore, we don't use the given embeddings in this operation. In practice, we feed a zero vector into the cross-attention mechanism to mask it out as $\text{Cross-Attn}(H, \mathbf{0}, \mathbf{0})$.

3.3 Training and Ensembling

We train the entire multitasking framework by jointly minimizing the weighted sum of both loss functions:

$$\mathcal{L} = \mathcal{L}_{gen} + \lambda\mathcal{L}_{rec}, \quad (6)$$

where λ is a hyper-parameter.

Model ensembling is proven to be effective to improve the robustness (Allen-Zhu and Li, 2020). In our work, we adopt a simple but effective model ensembling strategy. We train a series of models initialized by different random seeds, and then vote with the trained models during inference.

4 Experimental Setup

4.1 Implementation Details

Many neural network-based generation systems struggle with the OOV (out-of-vocabulary) problem. To alleviate the problem, we apply the SentencePiece algorithm (Kudo and Richardson, 2018) to glosses to reduce the vocabulary size. We use the tokenizers[1] toolkit for implementation and set the size to 10k for all 5 languages.

Our system is a 3-layer, 8-head transformer-based model implemented by the Pytorch library (Paszke et al., 2019). We use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. We adopt the Noam Optimizer proposed by (Vaswani et al., 2017) with an initial learning rate of $1e-7$, a maximum learning rate of $1e-3$, and a minimum learning rate of $1e-9$. We set the warmup steps to 4000 and batch size to 128. The maximum epochs is set to 500. And we set an early stop strategy in the patience of 5 epochs. To avoid gradient exploding, we clipped the gradient norm within 0.1. We also employ label smoothing technique (Pereyra et al., 2017) with a smoothing value of 0.1 during training. For the gloss corruption in the reconstruction objective, we follow Devlin et al. (2019) to randomly delete and blank words with a uniform probability of 0.2. And the λ (in Equation 6) is set to 1. For model ensembling, we train 5 models with different seeds. Due to the time constraints, our official submission has a result of ensembling three models on English, and results of single models on the reset of 4 languages. We submitted the results of ensembling 5 models in the post-evaluation phase.

For each language, we use the development set released by organizers for model selection. We select the best epoch using the summary of BLEU (Papineni et al., 2002) and MoverScore (Zhao et al., 2019) on the development set.

4.2 Evaluation Metrics

The definition modeling subtrack uses three metrics, which are MoverScore (Zhao et al., 2019), BLEU (Papineni et al., 2002) , and lemma-level BLEU respectively. Readers can refer to the task paper (Mickus et al., 2022) for more details.

5 Results and Analysis

In this section, we present the evaluation results and discuss our analysis of the generated definitions.

5.1 Main Results

Table 2 presents the evaluation scores on all 5 languages. Results show that our system significantly outperforms the baseline models in terms of the sentence BLEU and lemma-level BLEU. This indicates the effectiveness of our proposed cross-attention multitasking framework. However, the SGNS and Char are strong baselines in terms of the MoverScore, and our system only outperforms the baselines on English. We speculate that our results have more coincide words with references, but are not fluent enough, which leads to a low score from

[1]tokenizers: https://github.com/huggingface/tokenizers.

	Models	S-BLEU	L-BLEU	MvSc.
EN	SGNS	0.00125	0.00250	0.10339
	Char	0.00011	0.00022	0.08852
	Electra	0.00165	0.00215	0.08798
	CAMF	**0.03127**	**0.03957**	**0.13475**
	Ensemble	0.03106	0.03906	0.13273
ES	SGNS	0.01536	0.02667	**0.20130**
	Char	0.01505	0.02471	0.19933
	CAMF	0.03914	0.05606	0.12778
	Ensemble	**0.03925**	**0.05624**	0.13121
FR	SGNS	0.00351	0.00604	0.18478
	Char	0.00280	0.00706	**0.18579**
	Electra	0.00219	0.00301	0.17391
	CAMF	0.02679	0.03691	0.04193
	Ensemble	**0.02700**	**0.03738**	0.04455
IT	SGNS	0.02591	0.04081	**0.20527**
	Char	0.00640	0.00919	0.15920
	CAMF	0.06646	0.09926	0.11717
	Ensemble	**0.06812**	**0.10147**	0.12233
RU	SGNS	0.01520	0.02112	**0.34716**
	Char	0.01313	0.01847	0.32307
	Electra	0.01189	0.01457	0.33577
	CAMF	0.04843	0.06548	0.14820
	Ensemble	**0.05192**	**0.07074**	0.15702

Table 2: Evaluation results of different models in 5 languages. The SGNS, Char, Electra are baseline models provided by the organizers. The CAMF (Cross-Attention Multilingual Framework) is the model of official submission. And the Ensemble is an ensemble of 5 models submitted in the post-evaluation. Bold and underline mark the best and second scores, respectively.

(1) Redundancy and overusing common phrases: 42.00%	
word	explosion
reference	A sudden outburst.
hypothesis	A sudden, sudden, or destruction.
(2) Self-reference: 2.00%	
word	discover
reference	To reveal (information); to divulge, make known.
hypothesis	To make a conclusion of; to discover.
(3) Wrong Part-Of-Speech: 5.50%	
word	genius
reference	ingenious, brilliant, very clever, or original.
hypothesis	A person or thing that is extraordinary.
(4) Under-specified: 23.50%	
word	mayor
reference	The leader of a city.
hypothesis	A person who is a member of authority.
(5) Opposite: 2.00%	
word	solid
reference	Excellent , of high quality , or reliable.
hypothesis	Having no size or value.
(6) Close Semantics: 17.00%	
word	bed
reference	The time for going to sleep or resting in bed.
hypothesis	The state or quality of being a room.
(7) Incorrect: 52.00%	
word	smooth
reference	Lacking projections or indentations; not serrated.
hypothesis	Having the shape of a tree.

Table 3: Error types and examples.

the pretrained model used by MoverScore.

We also observe that model ensembling has brought the improvement of performance. It can be seen from the table that the Ensemble model outperforms the CAMF on 4 of the 5 languages, except for a slight decline on English. This may be due to the randomness of the parameter initialization. We also argue that better performance can be obtained by applying hyper-parameter searching algorithms and ensembling more models.

5.2 Error Analysis

In order to qualitatively analyze the definitions generated by our system, we randomly select several items from the English test set and manually annotate the error types following Noraset et al. (2017). In total, we extract 200 items, of which 197 contain some degree of error. We illustrate the error types and examples in Table 3. Note that each item may contain multiple errors, so the sum of the percentages in the table is greater than 100%.

From the table, we observe that the quality of English definitions generated by our system still need to be improved. Error types (1) to (3) are problems from the system, and types (4) to (6) are shortcomings in the embeddings. As we can see, the former accounts for a much larger proportion than the latter. The 52% incorrectness indicated by type (7) shows that many glosses generated by our system are irrelevant to the word. And the dataset released in this task will support significant future work on the definition modeling task.

6 Conclusion

In this paper, we present the implementation of the BLCU-ICALL system submitted to the SemEval-2022 Task 1, Definition Modeling subtrack. We propose a Cross-Attention Multitasking Framework that leverages multiple embedding architectures and jointly trains two objectives. We also investigate a simple but effective ensembling strategy to enhance the robustness. In future efforts, we plan to further improve our system to better handle the problems of redundancy and incorrect glosses.

Acknowledgements

This work was supported by the funds of Beijing Advanced Innovation Center for Language Resources (No. TYZ19005), Research Project of the National Language Commission (No. ZDI135-105, No. ZDI135-131), and BLCU supported project for young researchers program (supported by the Fundamental Research Funds for the Central Universities)(No. 20YCX142). We would like to thank all anonymous reviewers for their valuable comments and suggestions on this work.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Ting-Yun Chang, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xsense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *CoRR*, abs/1809.03348.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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.

Doug Downey, Stefan Schoenmackers, and Oren Etzioni. 2007. Sparse information extraction: Unsupervised language models to the rescue. In *ACL*.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. 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 3467–3476. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI conference on artificial intelligence*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. Toward cross-lingual definition generation for language learners. *arXiv preprint arXiv:2010.05533*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2022. SemEval-2022 Task 1: Codwoe – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6331–6344.

Jacob Turton, David Vinson, and Robert Elliott Smith. 2020. Deriving contextualised semantic features from bert (and other transformer model) embeddings. *arXiv preprint arXiv:2012.15353*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah Smith. 2015. Learning word representations with hierarchical sparse coding. In *International Conference on Machine Learning*, pages 87–96. PMLR.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *CoRR*, abs/1909.02622.