CSSL: Contrastive Self-Supervised Learning for Dependency Parsing on Relatively Free Word Ordered and Morphologically Rich Low Resource Languages

Anonymous ACL submission

Abstract

Significant advancements have been made in the domain of dependency parsing, with researchers introducing novel architectures to enhance parsing performance. However, the majority of these architectures have been evaluated predominantly in languages with a fixed word order, such as English. Consequently, little attention has been devoted to exploring the robustness of these architectures in the context of relatively free word-ordered languages. In this work, we examine the robustness of graph-based parsing architectures on 7 relatively free word order languages. We focus on investigating essential modifications such as data augmentation and the removal of position encoding required to adapt these architectures accordingly. To this end, we propose a contrastive self-supervised learning method to make the model robust to word order variations. Furthermore, our proposed modification demonstrates a substantial average gain of 3.03/2.95 points in 7 relatively free word order languages, as measured by the Unlabelled/Labelled Attachment Score metric when compared to the best performing baseline.

1 Introduction

Morphologically rich languages (MRLs) tend to have sentences which follow a relatively free word order. Instead of relying on the word ordering, such languages prefer encoding the structural information of a sentence using inflectional morphology. Majority of the pretrained models tend to include (relative or absolute) position encoding in their pretraining stage, which may not be ideal for several of the MRLs. Moreover, simply dropping the position encoding of the encoder for such models, during fine-tuning, often would lead to sub-optimal performances for parsing tasks (Krishna et al., 2019; Ghosh et al., 2024). In this work, we propose a self-supervised contrastive learning based module that makes a model agnostic to word order variations within a sentence.

We propose a novel Contrastive Self-Supervised Learning (CSSL) module, inspired by He et al. (2020), to accommodate variations in word order within the model architecture. Moreover, the modular nature of our approach allows for seamless integration with any encoder architecture, without necessitating alterations to pretraining decisions. In self supervised contrastive learning, for a given input, one needs to find positive samples, whose embedding level similarity with the input needs to be increased, and negative samples, whose embedding similarity with the original input, needs to be decreased. As shown in Figure 1, the original sentence serves as an anchor point, while its permutations represent positive examples, juxtaposed with randomly generated sentences serving as negative examples.

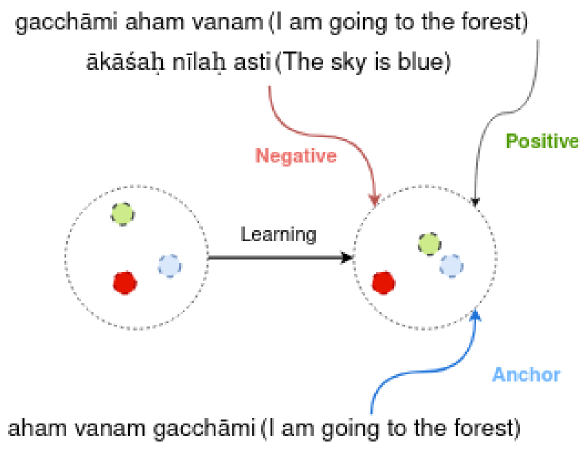

Figure 1: The Contrastive Loss minimizes the distance between an anchor (blue) and a positive (green), both of which have a similar meaning, and maximizes the distance between the anchor and a negative (red) of a different meaning.

The self-supervised contrastive learning objective aims to minimize the distance between positive examples and the anchor point, while simultaneously maximizing the distance from negative examples. In essence, this objective fosters the robustness of the encoder to accommodate word order variations. Our approach, to the best of our knowledge, is the first to use a contrastive learning technique for dependency parsing to overcome challenges caused by a lack of set word order and limited data resources.

MRLs rely less on word order and instead use morphological markers to encode structural information of a sentence. Given the comprehensive morphological marking system inherent in MRLs, the core semantic essence of the sentence remains unaltered, rendering the permuted counterpart as a suitable positive pairing for contrastive learning. Several MRLs have demonstrated that permutations of word order following weak projectivity generally retain semantic equivalence of the original (Sapir, 1921; Kulkarni et al., 2015; Kuboň et al., 2013; Ghosh et al., 2024).

Moreover, preference for certain word order typology in these languages is often not due to the limitations of the morphology, but are attributed to cognitive, psycho-linguistic, and information theoretic aspects of communication (Krishna et al., 2019; Clark et al., 2023; Dyer et al., 2023; Xu and Futrell, 2024). For instance, Sanskrit, a classical language, predominantly consists of sentences written as verses in its pre-classic and classic literature. Here, such sentences prefer to adherence to metrical constraints in prosody over any word ordering constructions, resulting in arbitrary word orderings (Krishna et al., 2020, §2). In our experiments, hence we treat the permutations of a given sentences as their semantic equivalents.

Substantial progress has been made in dependency parsing, including for low-resource languages and MRLs (Ji et al., 2021a,b; Dozat and Manning, 2017; Kulmizev et al., 2019), aimed at augmenting parsing efficacy. Our proposed approach is agnostic of the encoder architecture and does not necessitate the need for changes in pretraining. Moreover, our objective is to leverage recent advancements in parsing literature and further augment them by adding our CSSL module that would make these models more robust to word order variations. In this work, we start by examining the robustness of graph-based parsing architectures (Ji et al., 2019; Mohammadshahi and Henderson, 2020, 2021) on 7 relatively free word order languages. We believe, graph-based parsing architectures could be a natural choice to model flexible word order. We then focus on investigating essential modifications such as data augmentation (Şahin and Steedman, 2018) and the removal of position encoding required to adapt these architectures accordingly. We finally show the efficacy of our approach on the best baseline (Mohammadshahi and Henderson, 2021, RNGTr) model by integrating CSSL with it and report an average performance gain of 3.03/2.95 points (UAS/LAS) improvement over 7 MRLs.

Our main contributions are as follows:

- We propose a novel contrastive self-supervised learning (CSSL) module to make dependency parsing robust for free word order languages.
- Empirical evaluations of CSSL module affirm its efficacy for 7 free word-ordered languages
- We demonstrate statistically significant improvements with an average gain of 3.03/2.95 points over the best baseline on 7 MRLs.

2 Contrastive Self-Supervised Learning

CSSL enables joint learning of representation, via contrastive learning, with the standard classification loss for dependency parsing. Here, via CSSL, we identify sentences which are word-level permutations of each other as similar sentences, and others as dissimilar sentences. The similar sentences are brought closer while pushing dissimilar examples apart (van den Oord et al., 2019; Tian et al., 2020). For a given input, when selecting a dissimilar sample, we choose a random sentence that clearly differs significantly from any permutation of the given sentence.

Formally, for a sentence X_i (anchor example), its representation should be similar to the permuted instance X_i^+ as permutation does not alter the meaning of a sentence belonging to MRL. However, the representation will differ from a random sentence X_i^- (negative example). Therefore, the distance between the appropriate representations of X_i and X_i^+ is expected to be small. Thus, we can develop a contrastive objective by considering (X_i, X_i^+) a positive pair and $N-1$ negative pairs (X_i, X_i^-) :

$$\mathcal{L}_{\text{CSSL}} \quad = \quad -\log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{i+}/\tau\right)}{\sum_{a \in N} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)}$$

where N represents a batch, z_i represents the representation vector of the anchor sample, z_i^+ denotes

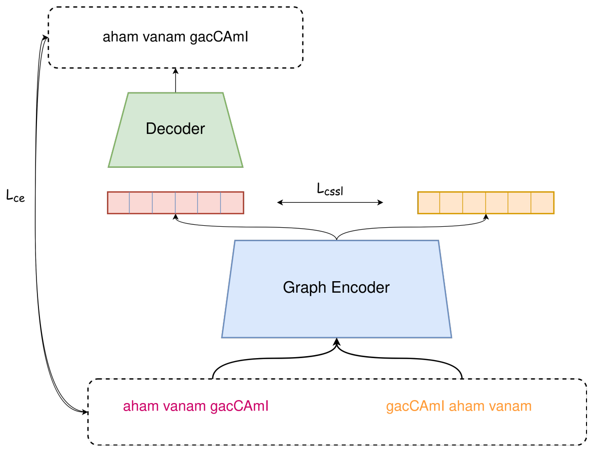

Figure 2: Schematic illustration of the proposed approach. Starting from an input sentence (bottom) (Translation: "I am going to the forest."), two embeddings are produced: (1) original and (2) permuted sentence. Self-contrastive loss is imposed on the embeddings (**center**). A decoder uses a cross-entropy objective for predicting the dependency tree.

the representation vector for the positive sample (permuted sample), z_a represents the representation vector for a sample in the batch (N different samples), and τ is a temperature parameter that controls the concentration of the distribution. We employ pooled sentence embedding of the original and permuted sentences for CSSL loss. Therefore, our final loss is:

$$\mathcal{L} = \mathcal{L}_{\text{cssl}} + \mathcal{L}_{\text{ce}} \tag{1}$$

The classification loss L_{ce} is applied only to token-level labels of the original training input.

3 Experiment

3.1 Dataset and metric

As our primary benchmark dataset, we utilize the Sanskrit Treebank Corpus (Kulkarni, 2013, STBC). From STBC, we use a train and dev split of 2,800 and 1,000, respectively. Further, we employ a test set comprising 300 sentences, drawn from the classical Sanskrit work, *Śiśupāla-vadha* (Ryali, 2016).

Moreover, from Universal Dependencies (de Marneffe et al., 2021, UD-2.13), we choose 6 additional languages, namely, Turkish, Telugu, Gothic, Hungarian, Ancient Hebrew, and Lithuanian.[1] Please note that all the seven languages are chosen from diverse language families and are typologically diverse. Our experiments are primarily focused on a low-resource setting. Here, the largest training set size we use is of 3,435 sentences for Turkish. For Turkish, we simulate a low-resource scenario by considering from the Turkish-IMST treebank. We also experiment with English which is a fixed-ordered high-resource language. Here, we use a training set of 12,544 sentences. We use standard UAS/LAS metrics (McDonald and Nivre, 2011) for evaluation.

[1]the statistics of each of the treebanks used for our experiments is mentioned in Table 4 in the Appendix.

Model	UAS	LAS
G2GTr (Transition-based)	85.75	82.21
GNN (Graph-based)	88.01	82.8
RNGTr (Graph-based)	89.62	87.43
RNGTr (NoPos)	80.78	78.37
RNGTr (DA)	90.38	88.46
Prop. System (CSSL)	**91.86**	**89.38**
CSSL + DA	92.43	90.18

Table 1: Comparison of graph-based parsers on Sanskrit STBC dataset. We modify the best baseline RNGTr by integrating the proposed method (CSSL) to compare against variants, removing position encoding (NoPos) and data augmentation (DA). The best performances are bold-faced. The results (CSSL vs DA) and (CSSL vs DA+CSSL) are statistically significant as per the t-test with a p-value < 0.01 for the LAS metric.

Baselines: We utilize Mohammadshahi and Henderson (2020, **G2GTr**), a transition-based dependency parser. Furthermore, we explore Ji et al. (2019, **GNN**) a graph neural network-based model that captures higher-order relations in dependency trees. Finally, we examine Graph-to-Graph Non-Autoregressive Transformer proposed by Mohammadshahi and Henderson (2021, **RNGTr**) which iteratively refines arbitrary graphs through recursive operations.

Hyper-parameters: We implement our CSSL module in RNGTr architecture which uses a pre-trained mBERT model (110M parameters) from Huggingface transformers (Wolf et al., 2020). For RNGTr model, we use the same architecture as Mohammadshahi and Henderson (2020) with pre-trained mBERT as the encoder and an MLP and biaffine followed by softmax for the decoder. We adopt the RNGTr codebase with hyperparameter settings as follows: the batch size is 16, the learning rate as 2e-5, the number of transformer blocks as 12 and for the decoder 2 Feed Forward Layers with dropout as 0.33 having bi-affine attention, and the remaining hyperparameters are the same.

Language	RNGTr		RNGTr + DA		RNGTr + CSSL	
	UAS	LAS	UAS	LAS	UAS	LAS
Turkish-IMST	72.86	71.99	74.18	72.96	**78.21**	**74.69**
Telugu-MTG	90.02	80.34	91.86	81.51	**93.79**	**85.67**
Gothic-POIEL	86.59	81.28	88.61	82.93	**89.15**	**84.19**
Hungarian-SZEGED	88.13	84.93	90.02	86.65	**91.65**	**87.28**
Ancient Hebrew-PTNK	90.76	86.42	91.43	87.12	**92.35**	**88.68**
Lithuanian-ALKSNIS	87.63	83.27	88.41	84.79	**89.82**	**86.45**
English-EWT	*92.08*	*90.23*	***93.76***	***92.16***	*93.19*	*90.71*

Table 2: Performance comparison on the RNGTr model, RNGTr + DA (Data Augmentation) and RNGTr + CSSL module. The best performances are bold-faced. Our results (CSSL) are statistically significant compared to both RNGTr and RNGTr + DA for each language as per the t-test with a p-value < 0.01 for the LAS metric

3.2 Results

In Table 1, we benchmark graph-based parsers on the Sanskrit STBC dataset. Our proposed contrastive loss module is standalone and could be integrated with any parser.[2] Thus, we modify the best baseline RNGTr by integrating the proposed method (CSSL) and comparing it against variants, removing position encoding (NoPos), and augmenting data augmentation (DA). Table 1 illustrates that the proposed framework adds a complementary signal making robust word order representations to RNGTr by improving 2.24/1.95 points in UAS/LAS scores. The performance significantly drops (8.8/9.0 UAS/LAS) when position embeddings are removed (vs. Pos kept) from RNGTr due to train-test mismatch in pretraining and fine-tuning steps. Moreover, our method outperforms data augmentation technique (Şahin and Steedman, 2018) by 1.48/0.92 points (UAS/LAS) when integrated with the RNGTr baseline. We integrate CSSL on top of an RNGTr+DA system and observe statistically significant improvements of 0.57/0.80 points (UAS/LAS), suggesting the proposed method complements the data-augmentation technique.

Results on multilingual experiments: In this section, we investigate the efficacy of CSSL module in multi-lingual settings. Table 2 reports results on 6 other morphologically rich languages in low-resource settings. Our approach averages 3.16/3.12 higher UAS/LAS scores than the usual cross-entropy-based RNGTr baseline. Our system outperforms the rotation-based DA technique with an average increase of 1.74/1.83 in UAS/LAS scores. Here, as expected, our proposed CSSL approach outperforms the standard RNGTr and DA approaches for all the languages, except English. English is not an MRL and it relies heavily on configurational information of the words to understand sentence structure. The DA approach performs better by 0.57/1.45 UAS/LAS scores than our framework. However, it is interesting to note that CSSL still outperforms the RNGTr baseline by 1.11/0.48 UAS/LAS, possibly due to robustness of permutation invariant representation learning we employ in CSSL. As illustrated in Table 1, it is evident that combining CSSL with DA surpasses CSSL alone by approximately 0.5 points, exhibiting a 2-point enhancement over DA.

4 Conclusion

In this work, we investigated the robustness of graph-based parsing architectures across 7 languages characterized by relatively flexible word order. We introduced a self-supervised contrastive learning module aimed at making encoders insensitive to variations in word order within sentences. Additionally, the modular nature of our approach enables seamless integration with any encoder architecture without necessitating modifications to pretraining decisions. To the best of our knowledge, our approach represents the first utilization of contrastive learning techniques for dependency parsing to address challenges arising from variable word order in low-resource settings. Finally, we demonstrate the effectiveness of our approach by integrating it with the RNGTr model (Mohammadshahi and Henderson, 2021), reporting an average performance improvement of 3.03/2.95 points (UAS/LAS) across the 7 MRLs.

Limitations We could not evaluate on complete UD due to limited available compute resources (single GPU); hence, we selected 7 representative

[2] Refer to Appendix A.1 for empirical evidence.

languages for our experiments.

Ethics Statement We do not foresee any ethical concerns with the work presented in this manuscript.

References

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

William Dyer, Charles Torres, Gregory Scontras, and Richard Futrell. 2023. Evaluating a Century of Progress on the Cognitive Science of Adjective Ordering. *Transactions of the Association for Computational Linguistics*, 11:1185–1200.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Poulami Ghosh, Shikhar Vashishth, Raj Dabre, and Pushpak Bhattacharyya. 2024. A morphology-based investigation of positional encodings.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

Tao Ji, Yong Jiang, Tao Wang, Zhongqiang Huang, Fei Huang, Yuanbin Wu, and Xiaoling Wang. 2021a. A unified encoding of structures in transition systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4121–4133, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Ji, Yong Jiang, Tao Wang, Zhongqiang Huang, Fei Huang, Yuanbin Wu, and Xiaoling Wang. 2021b. Word reordering for zero-shot cross-lingual structured prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4109–4120, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Ji, Yuanbin Wu, and Man Lan. 2019. Graph-based dependency parsing with graph neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2475–2485, Florence, Italy. Association for Computational Linguistics.

Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal. 2020. A graph-based framework for structured prediction tasks in Sanskrit. *Computational Linguistics*, 46(4):785–845.

Amrith Krishna, Vishnu Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. 2019. Poetry to prose conversion in Sanskrit as a linearisation task: A case for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166, Florence, Italy. Association for Computational Linguistics.

Vladislav Kuboň, Markéta Lopatková, and Jiří Mírovský. 2013. A case study of a free word order. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 222–231, Taipei, Taiwan. Department of English, National Chengchi University.

Amba Kulkarni. 2013. A deterministic dependency parser with dynamic programming for Sanskrit. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 157–166, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Amba Kulkarni, Preethi Shukla, Pavankumar Satuluri, and Devanand Shukl. 2015. How free is free word order in sanskrit. *The Sanskrit Library, USA*, pages 269–304.

Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano, and Joakim Nivre. 2019. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited. 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 2755–2768, Hong Kong, China. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Alireza Mohammadshahi and James Henderson. 2020. Graph-to-graph transformer for transition-based dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3278–3289, Online. Association for Computational Linguistics.

Alireza Mohammadshahi and James Henderson. 2021. Recursive non-autoregressive graph-to-graph transformer for dependency parsing with iterative refinement. *Transactions of the Association for Computational Linguistics*, 9:120–138.

Anupama Ryali. 2016. Challenges in developing sanskrit e-readers:semi-automatically using online analyser saṂsĀdhanĪ:with special reference to ŚiŚupĀlavadha of mĀgha. In *Workshop on Bridging 4797the Gap Between Sanskrit CL Tools Management of Sanskrit DL, ICON2016.*

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Edward Sapir. 1921. An introduction to the study of speech. *Language*, 1:15.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation.

Weijie Xu and Richard Futrell. 2024. Syntactic dependency length shaped by strategic memory allocation. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–9, St. Julian's, Malta. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2022. Pairwise supervised contrastive learning of sentence representations.

A Appendix

A.1 Integration of CSSL with another encoder

The modular nature of CSSL framework allows for seamless integration with any encoder architecture, without necessitating alterations to pretraining decisions. We have shown its effectiveness for the best-performing baseline. We are also showing results with one more baseline (for Sanskrit). Our supplementary results indicate that activating contrastive loss for the G2GTr baseline on the STBC treebank for Sanskrit leads to an approximate 2-point enhancement in performance measured by UAS/LAS.

	CE		CSSL	
	UAS	LAS	UAS	LAS
G2GTr	87.16	85.68	89.05	87.05

Table 3: Contrastive Loss with G2GTr on STBC dataset.

A.2 Treebank Statistics

Table 4 provides the detailed statistics for the languages used in the experiments.

A.3 Related Work

Contrastive learning has been the pinnacle of recent successes in sentence representation learning. In order to optimize the appropriately designed contrastive loss functions, (Gao et al., 2021; Zhang et al., 2022) uses the entailment sentences in NLI as positive pairs, significantly improving upon the prior state-of-the-art results. To this end, a number of methods have been put forth recently in which the augmentations are obtained through back-translation (Fang et al., 2020), dropout (Yan et al., 2021; Gao et al., 2021), surrounding context sampling (Logeswaran and Lee, 2018; Giorgi et al., 2021), or perturbations carried out at different semantic-level (Wu et al., 2020; Yan et al., 2021).

Treebank	Language Family	train	dev	test
Sanskrit-STBC	Indo-Aryan	2,800	1,000	300
UD-Turkish_IMST	Turkic	3,435	1,100	1,100
UD-Gothic_Proeil	Germanic	3,387	985	1,029
UD-Telugu_MTG	Dravidian	1,051	131	146
UD-Hungarian_Szeged	Uralic	910	441	449
UD-Ancient_Hebrew_PTNK	Semitic	730	439	410
UD-Lithuanian_ALKSNIS	Baltic	2,341	617	684
UD-English_EWT	Roman	12,544	2,001	2,077

Table 4: Treebank Statistics. The number of sentences in train, dev and test for each language.