Contrastive 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 4 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 loss objective to make the model robust to word order variations. Furthermore, our proposed modification demonstrates a substantial average gain of 3.48/3.10 points in 4 relatively free word order languages, as measured by the Unlabelled/Labelled Attachment Score metric when compared to the best performing modifications.

1 Introduction

Substantial progress has been achieved within the realm of dependency parsing, wherein researchers have introduced novel architectures (Chen and Manning, 2014; Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017; Kulmizev et al., 2019) aimed at augmenting parsing efficacy. Nevertheless, the predominant evaluation of these architectures has been limited to languages characterized by a fixed word order, like English. Consequently, inadequate focus has been allocated towards scrutinizing the adaptability of these architectures within the domain of languages featuring a comparatively flexible word order.

In the domain of linguistic typology, it is observed that languages falling within the analytical classification, exemplified by English, typically lack inflectional morphemes. This absence entails a lack of explicit marking for objects or subjects, necessitating a rigid adherence to word order. Conversely, languages categorized as morphologically rich languages (MRLs) diverge from such strictures, as they employ complex systems of marking that afford greater flexibility in word arrangement.

The configurational information in sentences of a free word order language is of limited use. Therefore, the graph-based parsing architectures could be a natural choice to model flexible word order. In this work, we examine the robustness of graph-based parsing architectures (Ji et al., 2019; Mohammadshahi and Henderson, 2020, 2021) on 4 relatively free word order languages. We 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. Existing multi-lingual pretraining opts for default position encoding, however, this decision may not be optimal for low-resource relatively free word order MRLs. Further, building word order agnostic encoder from scratch is not feasible due to data sparsity. Our results suggest that if we simply drop position encoding in the encoder then a mismatch in pretraining and task setting leads to suboptimal performance.

This work introduces a novel Contrastive Self-Supervised Learning (CSSL) module, as inspired by He et al. (2020), to accommodate variations in word order within the model architecture. In languages characterized by relatively flexible word order, the presence of intricate morphology facilitates the relaxation of constraints pertaining to word order, where permutations of word order following weak projectivity retain semantic equivalence. Given the comprehensive morphological marking system inherent in Morphologically Rich Languages, the core semantic essence of the sentence remains unaltered, rendering the

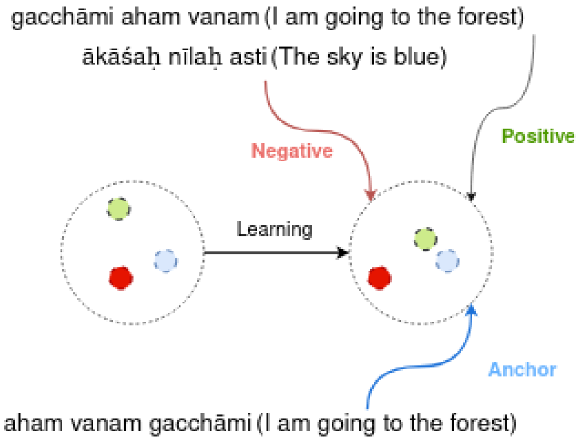
gacchāmi aham vanam (I am going to the forest)
ākāśaḥ nīlaḥ asti (The sky is blue)
Negative
Positive
Learning
Anchor
aham vanam gacchāmi (I am going to the forest)

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.

permuted counterpart a suitable positive pairing for contrastive learning. As depicted 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. 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. Moreover, the modular nature of this approach allows for seamless integration with any encoder architecture, without necessitating alterations to pretraining decisions. 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. The main contributions of our work are as follows:

- We propose a contrastive learning-based novel module to make Dependency Parsing robust for free word order languages.
- Empirical evaluations of our proposed module affirm the efficacy of for 4 free word-ordered languages.
- We demonstrate significant improvements with an average gain of 3.48/3.10 points over the strong baseline.

2 Proposed Method

This section outlines the framework that we have proposed, which incorporates contrastive learning as a novel loss function within the domain of Dependency Parsing. To summarise, we first utilize contrastive learning to get the contrastive objective, which acts as a measure for evaluating the complexity of sentences that are word permutations of each other. Ultimately, the model is trained to utilize both the classification and contrastive objectives.

2.1 Graph Transformers

Following the work by (Mohammadshahi and Henderson, 2020), Graph-to-Graph Transformer architecture was developed by integrating it with an attention-like function for graph connection prediction. This autoregressive parsing model predicts one edge (based on preceding edges) at a time, which uses (Devlin et al., 2019) and it's multilingual variant as the encoder. According to (Mohammadshahi and Henderson, 2021), a new graph prediction framework leverages G2GTr's graph-to-graph capabilities to improve the resulting graph iteratively.

2.2 Contrastive Learning

Contrastive learning seeks to gain meaningful representations by merging thematically comparable examples while separating semantically dissimilar occurrences. In this study, we will investigate sentences written as word order variations of one another, resulting in a higher semantic affinity. As a result, it is essential that the representation of a certain sentence is consistent across both original and permuted samples. In (van den Oord et al., 2019; Tian et al., 2020), connections are made of the contrastive loss to maximization of mutual information between different views of the data.

More specifically, 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^-) :

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

where N represents a batch, z_i represents the representation vector of the anchor sample, z_i^+ denotes the representation vector for the positive sample (permuted sample), z_a represents the representation vector for the negative samples (different samples), and τ is a temperature parameter that controls the concentration of the distribution.
Therefore, our final loss is:

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

3 Experiment

In this section, we evaluate our framework against Mohammadshahi and Henderson (2021, RNGTr). We also show that our framework consistently outperforms the rotation-based DA technique Şahin and Steedman (2018, crop-rotate) on 4 MRLs for dependency parsing.

3.1 Dataset and metric

We utilize the Sanskrit Treebank Corpus (Kulkarni, 2013) as our primary benchmark dataset. For our experiments, we employ a training set consisting of 2800 sentences from the prose domain alongside a development set containing 1000 sentences from the same domain. We employ test set comprising 300 sentences, drawn from the classical Sanskrit work, *Śiśupāla-vadha* (Ryali, 2016). We utilize our model on Universal Dependencies (UD-2.13) (de Marneffe et al., 2021) in order to assess its performance on each of these 4 low-resource languages with extensive morphological structures. With the exception of Sanskrit, we examine three more low-resource MRLs: Turkish, Telugu, and Gothic. These languages belong to various language families, script families, character sets, training sets, morphological complexity, and domains. We use standard UAS/LAS metrics (McDonald and Nivre, 2011) for evaluation.

Language Selection Criteria: We choose low-resource languages from 4 typological families, guaranteeing that each language belongs to a unique family. We choose languages with explicit morphological information, which means they have a thorough marking system and inflectional morphemes, suggesting that they are Morphologically rich languages.

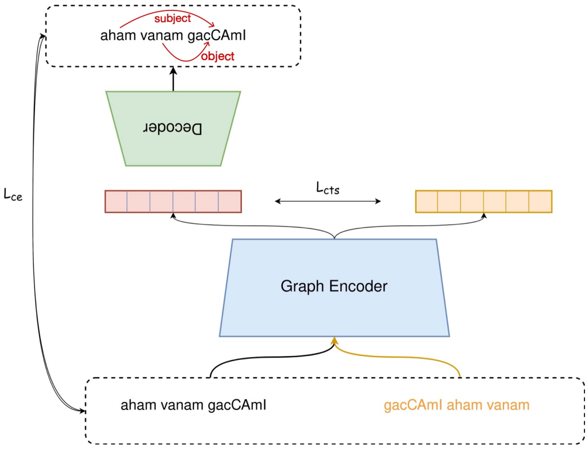

Figure 2: Schematic illustration of the proposed approach. Starting from an input sentence (bottom), two embeddings are produced:(1) original and (2) permuted sentence. Self-contrastive loss is imposed on the embeddings (**center**). A decoder finds the dependency tree for the input from the embedding.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

Baselines: We utilize Mohammadshahi and Henderson (2020, **G2GTr**), a member of the transition-based dependency parsing family. 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.

3.2 Experimental Settings

We implement our framework using RNGTr architecture which uses a pre-trained mBERT Base model from Huggingface which has 110M parametes. During training, for each training example, we permute the word order of the sentence to make its positive example and all other sentences in the batch are considered as negative examples for contrastive learning. The classification loss is calculated based only on the original training example's label.

3.3 Results

On STBC, our suggested method outperforms all other approaches, as Table 2 illustrates. Our method outperforms the RNGTr model Mohammadshahi and Henderson (2021, RNGTr) by 2.24/1.95 points in UAS/LAS scores. Our model outperforms the RNGTr model without position

	CE		DA		CTS	
Language	UAS	LAS	UAS	LAS	UAS	LAS
Sanskrit	89.62	87.43	90.38	88.46	**91.86**	**88.89**
Turkish	72.86	71.99	74.18	72.96	**78.21**	**74.69**
Telugu	90.02	80.34	91.86	81.51	**93.79**	**85.67**
Gothic	86.59	81.28	88.61	82.93	**89.15**	**84.19**
English	92.08	90.23	**93.76**	**92.16**	93.19	90.71

Table 1: Performance comparison on the RNGTr model among standard cross entropy (CE), DA(CE+Data Augmentation) and CTS(CE+Contrastive) techniques. The best performances are bold-faced.

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	**91.86**	**89.38**

Table 2: Performance comparison among different methodologies on Sanskrit STBC dataset. The best performances are bold-faced.

embeddings by 11.08/11.01 UAS/LAS scores. Moreover, our method outshined even the RNGTr model augmented with rotation-based data augmentation techniques pioneered by Şahin and Steedman (2018, crop-rotate). Though the margin was narrower, our method still managed to secure a substantial 1.48-point improvement in UAS and a 0.96-point boost in LAS, reaffirming its superiority in dependency parsing for low-resource MRLs.

4 Analysis

Three other morphologically rich languages besides Sanskrit are compared using our suggested method in Table 1. It is noteworthy that Sanskrit, Telugu, Turkish, and Gothic belong to Indo-Aryan, Dravidian, Turku, and Germanic language families. On average, our approach achieves 3.48/3.10 higher UAS/LAS scores than the usual cross-entropy based technique. Our system outperforms the rotation-based DA technique with an average increase of 1.99/1.89 in UAS/LAS scores. It is also evident that English, not being a morphologically rich language, lacks a comprehensive marking system. Because sentence permutation in English does not always imply a positive meaning, our framework is unable to work. By 0.57/1.95 UAS/LAS scores, the DA approach performs better than our framework. Due to some degree of data augmentation, it is also evident that our methodology outperforms the industry baseline by 1.11/0.48 in the corresponding metrics.

5 Conclusion and Future Work

In conclusion, our study delved into the robustness of graph-based parsing architectures across four languages characterized by relatively free word order. While significant strides have been made in dependency parsing for languages with fixed word order like English, our research illuminates a critical gap in evaluating these architectures across more diverse linguistic structures. By focusing on essential modifications such as data augmentation and the removal of position encoding, we aimed to adapt these architectures to accommodate varied word order patterns effectively. Our findings demonstrate the efficacy of the proposed modifications, particularly the incorporation of a contrastive loss objective, in enhancing parsing performance across the languages under scrutiny.
Future work could consider extending this method for dependency parsing in poetry data, where more intricate word orderliness is found. Future research in this domain could further refine and generalize these modifications to encompass a broader spectrum of languages, ultimately advancing the field of dependency parsing in linguistically diverse contexts.

Limitations We could not evaluate on complete UD due to limited available compute resources (single GPU); hence, we selected 5 representative languages for our experiments.

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

References

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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.

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, 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.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

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.

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.

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.

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

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

B 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).