Preprints.org

Article

Not peer-reviewed version

Advanced Machine Translation with Linguistic-Enhanced Transformer

Richard Patricia *, Judith Sadok , Rodolfo Patel

Posted Date: 4 December 2023

doi: 10.20944/preprints202312.0186.v1

Keywords: enhanced machine translation; linguistic syntax; transformer model

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Advanced Machine Translation with Linguistic-Enhanced Transformer

Richard Patricia *, Judith Sadok and Rodolfo Patel

Briar Cliff University

Abstract: Recent advancements in neural language models, particularly attention-based architectures like the Transformer, have substantially surpassed traditional methods in various natural language processing tasks. These models adeptly generate nuanced token representations by considering contextual relations within a sequence. However, augmenting these models with explicit syntactic knowledge, such as part of speech tags, has been found to remarkably bolster their effectiveness, especially under constrained data scenarios. This study introduces the Linguistic Enhanced Transformer (LET), which integrates multiple syntactic features, showing a notable increase in translation accuracy, evidenced by an improvement of up to 1.99 BLEU points on subsets of the WMT '14 English-German dataset. Furthermore, this paper demonstrates that enriching BERT models with syntax-aware embeddings enhances their performance on several GLUE benchmark tasks.

Keywords: enhanced machine translation; linguistic syntax; LET model

1. Introduction

In the realm of natural language processing (NLP), attention-driven deep learning models have demonstrated remarkable efficacy in a wide array of tasks, including machine translation, paraphrasing, and summarization. These models excel in sequence-to-sequence challenges by enabling a comprehensive representation of each token (e.g., words or subwords) within a sequence through a collective consideration of all tokens in the sequence [1]. The Transformer model, a pinnacle of this approach, employs a sophisticated multi-headed self- and cross-attention mechanism. This mechanism facilitates the parallel learning of varied token representations, enriching the contextuality of language understanding [2]. Its architecture comprises multiple identical layers in both encoder and decoder sections, each layer further refining the embeddings for tokens in source and target languages [3–7].

Despite its success in diverse language pairs, the Transformer's effectiveness is heavily contingent on vast datasets and extensive training to grasp the subtle syntactic and semantic intricacies of languages [8]. Prior research highlights that integrating the grammatical structure of sentences significantly bolsters an NLP model's capacity to comprehend and generate language [9–12]. To address this challenge and enhance the Transformer's efficiency, particularly in data-scarce scenarios, we propose a novel approach that leverages the innate syntactic elements of natural language.

The Linguistic Enhanced Transformer (LET) model is a groundbreaking adaptation of the Transformer architecture. It revolutionizes the embedding process by enabling tokens within the encoder to be informed not only by other tokens but also by rich syntactic features [13] such as POS, grammatical case, and subword position. These features are either deduced via a dedicated model (for POS) or directly defined (for case and subword position) and are integrated into each token's one-hot vector encoding. This hybrid embedding, which merges token and syntactic feature representations, is cohesively refined throughout the Transformer's layers, resulting in a more nuanced and context-aware token representation.

We rigorously evaluate the LET model in the context of English to German (EN-DE) translation using the comprehensive WMT'14 dataset. This evaluation process is distinctive, as it encompasses an array of syntactic features, namely POS, case, and subword tags. These tags are particularly instrumental in maintaining the structural integrity of complex words during tokenization processes. For instance, in a word like "amalgamation," subword tags ensure that its segmentation into smaller

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

tokens does not dilute its semantic and syntactic significance. Our extensive experiments substantiate that incorporating these syntactic features into the Transformer markedly enhances its translation performance across various training data sizes. The LET model's superiority is further evidenced by its consistent outperformance of the baseline Transformer, a testament to the value of integrating syntactic features.

Additionally, we extend this syntactic enhancement approach to the $\text{BERT}_{\text{BASE}}$ model. Our modified version, dubbed $\text{BERT}_{\text{BASE + POS}}$, is infused with syntax information, predominantly POS embeddings. This enriched model demonstrates notable improvements in several tasks within the General Language Understanding Evaluation (GLUE) benchmarks, surpassing the performance of the standard $\text{BERT}_{\text{BASE}}$ model on multiple fronts.

In summary, the primary contributions of our research are:

1. The development and validation of the LET model, which innovatively incorporates explicit syntax information into the Transformer's embeddings, showing significant performance enhancements in the EN-DE machine translation task.
2. The enhancement of the $\text{BERT}_{\text{BASE}}$ model through the integration of syntax information, culminating in the $\text{BERT}_{\text{BASE + POS}}$ model, which demonstrates superior performance across a range of GLUE benchmark tasks.

2. Related Works

The evolution of self-attention mechanisms in NLP models has been a focus of recent research. A notable development by [14] introduced a Gaussian bias to the self-attention module, aiming to refine the model's capacity for local context comprehension while preserving its ability to handle long-range dependencies. An innovative approach by [19] involved the use of relative positional embeddings, which consider the distance between sequence elements. This method significantly outperformed traditional absolute positional embeddings, leading to substantial performance enhancements.

Incorporating linguistic structures into models like the Transformer is an effective strategy for refining the attention mechanism. The use of POS and subword tags, for instance, serves as a kind of relative positional embedding, reinforcing sentence structure and thereby improving attention accuracy. The work of [25] introduced disagreement regularization to encourage different attention heads to focus on varied types of information, such as positional context and token representation. Furthermore, [26] proposed an approach to model word-to-word local dependencies more effectively by integrating the concept of distance into the self-attention mechanism.

The integration of syntax into deep learning models [27] for NLP has also been extensively explored. For example, [28,29] utilized syntax dependency tree information in conjunction with bidirectional RNNs for translation systems, applying Graph Convolutional Networks (GCNs) [30] to model these trees. This syntax-centric approach has shown significant benefits in tasks such as Chinese-English translation [31], where linearizing parse trees led to notable performance improvements. Additionally, imposing a syntax-based distance constraint on the attention mechanism to generate a more semantically rich context vector has proven effective in translation tasks, particularly in Chinese-English and English-German language pairs [32]. Collectively, these studies confirm the substantial benefits of incorporating syntactic information into NLP models. This integration not only enhances the models' translation capabilities but also contributes to improved performance metrics across various language tasks.

3. Preliminary

3.1. Transformer

The Transformer, a cornerstone in modern NLP frameworks, is structured into encoder and decoder modules. Each module comprises multiple layers that work in tandem to produce intricate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

representations of words in both source and target sequences [2]. As an initial step, words are segmented into subwords, maintaining a length at or below the original word's length [33], fostering vocabulary sharing between the source and target languages.

Let us consider any $m \in \{1, 2, \ldots, M\}$, with M representing the source sequence's length. The encoder's embedding layer transforms subwords $\mathbf{x}_m$ into embeddings $\mathbf{e}_m$:

$$\mathbf{e}_m = \mathbf{E}\mathbf{x}_m \tag{1}$$

In this equation, $\mathbf{E} \in \mathbb{R}^{D\times N}$ denotes a learnable matrix, where N stands for the total count of subwords in the shared vocabulary, and $\mathbf{x}_m \in \{0,1\}^N : \sum_i x_{mi} = 1$ is a one-hot vector representing subword m. These embeddings sequentially traverse through six encoder layers. Each layer is equipped with a self-attention mechanism, enabling a comprehensive representation of each subword as an amalgamation of the entire input sequence. The self-attention mechanism operates through three weight matrices: key ($\mathbf{K}$), query ($\mathbf{Q}$), and value ($\mathbf{V}$). These matrices interact to score and weigh each subword, thus determining the output embeddings. The process can be mathematically depicted as:

$$\begin{aligned} \mathbf{K} &= \mathbf{H}\mathbf{W}_K \\ \mathbf{Q} &= \mathbf{H}\mathbf{W}_Q \\ \mathbf{V} &= \mathbf{H}\mathbf{W}_V \\ \mathbf{A} &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\rho}}\right)\mathbf{V} \end{aligned} \tag{2}$$

Here, $\mathbf{H} = [\mathbf{h}_1\, \mathbf{h}_2\, \cdots\, \mathbf{h}_M]^\top \in \mathbb{R}^{M\times D}$ represents the embeddings for the sequence of M subwords. The projection matrices $\mathbf{W}_K$, $\mathbf{W}_Q$, and $\mathbf{W}_V$ all lie in $\mathbb{R}^{D\times P}$, and ρ, a scaling factor (chosen as P), modulates the attention mechanism. Each encoder layer generates a unique attention-weighted representation $\mathbf{A} \in \mathbb{R}^{M\times P}$, subsequently passed to the next layer, with $\mathbf{h}_m = \mathbf{e}_m$ for the initial layer.

The encoder's final layer relays its output to the decoder, comprising six analogous subunits. The decoder, while mirroring the encoder's structure, introduces an additional cross-attention layer. This enables the decoder to consider both the entire input sequence and the progressively decoded words.

3.2. BERT: Bidirectional Encoder Representations from Transformers

While the Transformer excels in crafting elaborate word sequences through its attention mechanism, the decoder's specificity limits its general applicability. However, the encoder's learned word embeddings can be adeptly adapted for various downstream tasks. The BERT (Bidirectional Encoder Representations from Transformers) model, an evolution of the Transformer concept, facilitates such adaptability. BERT essentially functions as a Transformer encoder, pre-trained through two methods: masked language modeling (MLM) and next-sentence prediction (NSP), followed by the addition of a softmax layer for task-specific applications, such as classification, sequence labeling, and question answering. [34] reported that BERT achieves remarkable success, setting new benchmarks across eleven different NLP tasks in the GLUE benchmark [35].

4. The Proposed Model

4.1. Linguistic Enhanced Transformer (LET)

Syntax plays a pivotal role in shaping the grammatical structure of sentences, influencing how different parts of speech (POS) interact to convey meaning. Studies like [36] have demonstrated that NLP models, when sufficiently trained, can implicitly grasp syntactic and semantic patterns, although imperfectly. Research by [9,11] suggests that even a rudimentary understanding of syntax

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

can significantly enhance model performance over baseline. This leads us to posit that integrating syntactic information explicitly could further augment NLP models in various tasks.

To enable the Transformer to efficiently leverage syntactic information for improved translation, we introduce the LET model, which integrates syntactic elements at different levels. We use a pretrained model[1] to tag words in the source sequence with POS, identify the case of each word, and pinpoint the relative position of each subword within a word (subword tagging). These syntax embeddings are then appended to the token embeddings, creating a fusion of syntactic and semantic information.

Initially, each source sequence word is tagged with its POS label. Upon segmenting words into subwords, these subwords inherit the POS label of their parent word. For instance, the word `sunshine` split into subwords `sun`, `sh`, and `ine` would all receive the POS tag `NOUN`. Subsequently, we extract POS embeddings from a trainable matrix, similar to subword embeddings. The resulting POS embeddings $\mathbf{f}_m^p$ for each subword m are then amalgamated with the subword embeddings $\mathbf{e}_m \in \mathbb{R}^{D-d}$, where d is the feature embedding dimension, forming a composite embedding.

Case and subword position features are also incorporated. A binary indicator $z_m^c \in \{0,1\}$ determines the feature embedding $\mathbf{f}_m^c$ for each subword, depending on the capitalization of the original word. For subword positioning, a categorical variable $z_m^s \in \{B, M, E, O\}$ assigns a feature embedding $\mathbf{f}_m^s$ based on the subword's position (beginning, middle, end, or sole occupant of the word). The syntactic features are synthesized as:

$$[\mathbf{e}_m^\top \, \mathbf{f}_m^\top]^\top = \mathbf{h}'_m \in \mathbb{R}^D$$

Here, $\mathbf{f}_m = \mathbf{f}_m^p \oplus \mathbf{f}_m^c \oplus \mathbf{f}_m^s \in \mathbb{R}^d$ represents the learned syntactic feature embeddings for subword m, with $\oplus$ denoting concatenation or summation.

The LET model is hypothesized to enhance translation performance by integrating these syntactic relationships directly into its architecture.

4.2. Syntax-infused BERT (LET Variant)

Extending the syntax-infusion concept to BERT is a logical progression. BERT, with its versatile embeddings, can benefit from syntactic feature integration for a range of downstream tasks, particularly those that rely heavily on semantic structure.

Considering the varying resource availability across datasets, we focus on POS tagging as the primary syntactic feature for BERT, creating the BERT_{LET} variant. We explore two methods for merging POS features with pre-trained BERT embeddings: (1) adding the POS embedding vector to the token embedding, and (2) concatenating the POS embedding with the token embedding, followed by an affine transformation to match the input dimension of the $\text{BERT}_{\text{BASE}}$ model ($D = 768$). Given the complexity of learning a large matrix and the limited training data in some tasks, we opt for the first approach for consistency and ease of training. Thus, the input representation of a token in BERT_{LET} is constructed by summing the corresponding BERT token embeddings with the POS embeddings (illustrated in Figure **??**).

Mathematically, for token m, the input tokens $\mathbf{h}'_m \in \mathbb{R}^D$ are represented as $\mathbf{h}'_m = \mathbf{e}_m + \mathbf{f}_m^p$, where $\mathbf{e}_m$ is the BERT token embedding and $\mathbf{f}_m^p$ is the POS embedding. For classification tasks, the final representation $\hat{\mathbf{y}}_{CLS}$ of the first token (denoted as [CLS]) is inputted into a softmax classifier to predict the output label.

---

[1] https://spacy.io/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

5. Experiments

5.1. Experimental Setup

Datasets

In our machine translation experiments, we focus on the WMT'14 EN-DE dataset, comprising 4.5 million training sentences. Validation is conducted on the newstest2013 dataset (3,000 sentences), and testing is carried out on the newstest2014 dataset (2,737 sentences) [37]. Considering the robustness of syntax parsers trained on English, we chose to translate from English to German. While integrating linguistic features generally benefits NLP models, we need to consider the trade-off with the allocation of dimensions for word embeddings, which still play a crucial role.

Table 1. BLEU scores comparing baseline Transformer and LET on various data fractions for EN-DE translation on newstest2014.

Data Fraction	Number of Sentences	Baseline Transformer	LET Transformer
1%	45k	1.10	**1.67**
5%	225k	8.51	**10.50**
10%	450k	16.28	**17.28**
25%	1.1M	22.72	**23.24**
50%	2.25M	25.41	**25.74**
100%	4.5M	28.94	**29.64**

Machine Translation with LET

In our LET approach, we experimented with both concatenation and summation of syntactic features onto word embeddings. Especially in scenarios with limited training data, adding too many features might reduce the allocation of dimensions for word embeddings. For low-resource machine translation, we opted for concatenation of only POS features, which implicitly encompasses elements of case and subword position, offering a balanced feature set.

Both the baseline and LET models were trained for 100,000 steps. The LET model uses 512-dimensional embedding vectors, with 492 dimensions allocated for word embeddings and 20 for feature embeddings in the encoder, as determined through hyperparameter tuning. In contrast, the decoder utilizes all 512 dimensions for word embeddings, focusing solely on decoding words. The architecture comprises six encoder and decoder layers, each with eight heads for multi-headed attention. Parameters are initialized following Glorot [38], with a dropout rate of 0.1 and a batch size of 4096. The Adam optimizer is employed with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and a label-smoothing factor of 0.1.

The choice of including POS, case, and subword tag features in LET is driven by the context and size differences of the EN-DE dataset compared to typical POS training datasets. The feature dimension d for $\mathbf{f}_m$ is set to 20 after a grid search over 8 to 64.

Natural Language Understanding with LET

We utilize the GLUE benchmark [35] to evaluate LET in natural language understanding tasks. This benchmark comprises eight datasets including MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, and RTE. For details on these datasets, see [34]. In these tasks, we use POS as the primary syntactic feature for LET. Standard hyperparameters are used for fine-tuning both $\text{BERT}_{\text{BASE}}$ and LET, including a batch size of 32 and three training epochs for all tasks. The learning rate for each task is selected from $\{5, 4, 3, 2\} \times 10^{-5}$.

In LET, a separate POS embedding $\mathbf{H}_{\text{POS}}$ is learned for each task, combined with the final hidden vector $\mathbf{C}$ corresponding to the first input token ([CLS]). The classification layer weights $\mathbf{W}$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

$\in \mathbb{R}^{K \times H}$, where **K** is the number of labels, use a standard classification loss with **C** and **W**, i.e., $log(softmax(\mathbf{CW}^T))$.

Table 2. GLUE test results from the GLUE evaluation server. Training example numbers are listed below each task. Scores in bold indicate tasks where LET outperforms $\text{BERT}_{\text{BASE}}$.

System	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
LET	84.4/83.3	**71.4**	90.4	**93.9**	**52.9**	85.5	88.8	**66.9**	**79.7**

Table 3. Translation examples comparing baseline Transformer and LET on the EN-DE dataset. Text in blue indicates accurate predictions by LET not captured by the baseline model.

Reference	Baseline Transformer	LET
Parken in Frankfurt könnte bald empfindlich teurer werden .	Das Personal war sehr freundlich und hilfsbereit .	Parken in Frankfurt könnte bald spürbar teurer sein .
Die zurückgerufenen Modelle wurden zwischen dem 1. August und 10. September hergestellt .	Zwischen August 1 und September 10.	Die zurückgerufenen Modelle wurden zwischen dem 1. August und 10. September gebaut
Stattdessen verbrachte Bwelle Jahre damit , seinen Vater in überfüllte Kliniken und Hospitäler zu begleiten , um dort die Behandlung zu bekommen , die sie zu bieten hatten .	Stattdessen verbrachte Bwelle Jahre damit , seinen Vater mit über füllten Kliniken und Krankenhqusern zu beherbergen .	Stattdessen verbrachte Bwelle Jahre damit , seinen Vater zu überfüllten Kliniken und Krankenhäusern zu begleiten , um jede Behandlung zu bekommen , die sie bekommen konnten .
Patek kann gegen sein Urteil noch Berufung ein legen .	Patek kann noch seinen Satz an rufen .	Patek mag sein Urteil noch Berufung ein legen .

5.2. Machine Translation Results with LET

Our evaluation focuses on the EN-DE translation task, exploring the efficacy of integrating syntax into the baseline Transformer. We incorporate three syntactic features—POS, subword tags, and case—aiming to aid the Transformer in deciphering sentence patterns more effectively.

When embedding multiple features, we face the challenge of balancing the allocation of dimensions between word and feature embeddings. A total of 512 dimensions are used for both, with the key tradeoff being that an increase in feature dimensionality leads to a decrease in the dimensions available for word embeddings. Given the limited range of values for POS, case, and subword tags, allocating excessive dimensions to each feature is counterproductive. Through grid search, we find that an optimal total feature dimension of 20 maximizes BLEU score gains. This can be achieved either by summing 20-dimensional embeddings for each feature or concatenating feature embeddings to reach a total dimensionality of 20. We discover that the summation of 20-dimensional embeddings for each feature and their subsequent concatenation with 492-dimensional word embeddings results in peak performance.

Table 1 demonstrates the performance variations between the baseline and LET models across different training data proportions. The LET model consistently surpasses the baseline, with improvements ranging from 0.57 to 1.99 BLEU points, depending on the data fraction. Notably, the greatest relative improvement occurs when training with only 10% of the dataset, highlighting the particular effectiveness of explicit syntax integration in limited data scenarios. We illustrate that LET captures semantic connections between distant yet related tokens more effectively, enhancing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

translation quality. Additionally, Table 3 provides examples of German translations produced by both the baseline and LET models.

5.3. Natural Language Understanding Results with LET

The performance of LET in natural language understanding is evaluated using the GLUE benchmark test set, as shown in Table 5. LET outperforms $\text{BERT}_{\text{BASE}}$ in 4 out of the 8 tasks, with improvements ranging from marginal to significant, peaking at a 0.8 point increase over $\text{BERT}_{\text{BASE}}$ in CoLA, a task assessing sentence linguistic structure. Notably, LET excels in tasks evaluating semantic relatedness, benefiting from the integration of POS embeddings. Table 4 presents examples from the RTE dataset, showcasing the improved classification accuracy of LET over $\text{BERT}_{\text{BASE}}$.

Table 4. Selected sentences from the RTE dataset demonstrating the improved entailment classification by LET compared to $\text{BERT}_{\text{BASE}}$.

Sentence 1	Sentence 2	True label
The Qin (from which the name China is derived) established the approximate boundaries and basic administrative system that all subsequent dynasties were to follow .	Qin Shi Huang was the first Chinese Emperor .	Not entailment
In Nigeria, by far the most populous country in sub-Saharan Africa, over 2.7 million people are infected with HIV .	2.7 percent of the people infected with HIV live in Africa .	Not entailment

Table 5. Comparative results on the GLUE benchmark test set for $\text{BERT}_{\text{BASE}}$ and LET. Training example numbers are shown below each task. Scores in bold indicate where LET outperforms $\text{BERT}_{\text{BASE}}$.

System	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
LET	84.4/83.3	**71.4**	90.4	**93.9**	**52.9**	85.5	88.8	**66.9**	**79.7**

6. Conclusion

In this study, we have successfully integrated syntactic information into the Transformer network, thereby creating the Linguistic Enhanced Transformer (LET) for machine translation. Our results demonstrate that LET notably excels in translation accuracy, particularly when trained on subsets of the training data. This indicates that explicit syntax integration is highly beneficial in scenarios where training resources are limited. Additionally, we provided insights into the differences between LET and the baseline Transformer by analyzing attention visualizations. These analyses reveal that LET achieves a more nuanced understanding of sentence structure, leading to improved translation quality. Furthermore, we extended the concept of syntax infusion to the BERT model, resulting in the LET variant of BERT, which shows enhanced performance across several downstream tasks in the GLUE benchmark. An intriguing avenue for future research is the exploration of architectures that inherently model linguistic structures more effectively than current end-to-end models. The potential for further efficiency improvements in sophisticated NLP models lies in creating architectures that deeply integrate linguistic knowledge into their fundamental design, moving beyond the current paradigm of end-to-end learning. Future work might focus on developing such inherently language-aware architectures, potentially leading to further advancements in the field of natural language processing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

References

1. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* **2015**.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 2017, pp. 5998–6008.
3. Xu, W.; Sun, H.; Deng, C.; Tan, Y. Variational autoencoder for semi-supervised text classification. Thirty-First AAAI Conference on Artificial Intelligence, 2017.
4. Fei, H.; Ren, Y.; Zhang, Y.; Ji, D.; Liang, X. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics* **2021**, *22*.
5. Yang, Z.; Hu, Z.; Salakhutdinov, R.; Berg-Kirkpatrick, T. Improved variational autoencoders for text modeling using dilated convolutions. Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 3881–3890.
6. Fei, H.; Ren, Y.; Ji, D. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management* **2020**, *57*, 102311.
7. Shen, D.; Zhang, Y.; Henao, R.; Su, Q.; Carin, L. Deconvolutional latent-variable model for text sequence matching. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
8. Fei, H.; Ren, Y.; Ji, D. Retrofitting Structure-aware Transformer Language Model for End Tasks. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 2151–2161.
9. Kuncoro, A.; Dyer, C.; Hale, J.; Yogatama, D.; Clark, S.; Blunsom, P. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1426–1436.
10. Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; Li, F. Unified Named Entity Recognition as Word-Word Relation Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 10965–10973.
11. Linzen, T.; Dupoux, E.; Goldberg, Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* **2016**, *4*, 521–535.
12. Li, J.; Xu, K.; Li, F.; Fei, H.; Ren, Y.; Ji, D. MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 1359–1370.
13. Fei, H.; Wu, S.; Ren, Y.; Li, F.; Ji, D. Better Combine Them Together! Integrating Syntactic Constituency and Dependency Representations for Semantic Role Labeling. Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, 2021, pp. 549–559.
14. Yang, B.; Tu, Z.; Wong, D.F.; Meng, F.; Chao, L.S.; Zhang, T. Modeling localness for self-attention networks. *arXiv preprint arXiv:1810.10182* **2018**.
15. Wu, S.; Fei, H.; Li, F.; Zhang, M.; Liu, Y.; Teng, C.; Ji, D. Mastering the Explicit Opinion-Role Interaction: Syntax-Aided Neural Transition System for Unified Opinion Role Labeling. Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, 2022, pp. 11513–11521.
16. Shi, W.; Li, F.; Li, J.; Fei, H.; Ji, D. Effective Token Graph Modeling using a Novel Labeling Strategy for Structured Sentiment Analysis. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4232–4241.
17. Fei, H.; Zhang, Y.; Ren, Y.; Ji, D. Latent Emotion Memory for Multi-Label Emotion Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 7692–7699.
18. Wang, F.; Li, F.; Fei, H.; Li, J.; Wu, S.; Su, F.; Shi, W.; Ji, D.; Cai, B. Entity-centered Cross-document Relation Extraction. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9871–9881.
19. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* **2018**.
20. Shang, M.; Li, P.; Fu, Z.; Bing, L.; Zhao, D.; Shi, S.; Yan, R. Semi-supervised Text Style Transfer: Cross Projection in Latent Space. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4939–4948.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 December 2023 doi:10.20944/preprints202312.0186.v1

21. Wu, S.; Fei, H.; Ren, Y.; Ji, D.; Li, J. Learn from Syntax: Improving Pair-wise Aspect and Opinion Terms Extraction with Rich Syntactic Knowledge. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021, pp. 3957–3963.
22. Fei, H.; Li, F.; Li, B.; Ji, D. Encoder-Decoder Based Unified Semantic Role Labeling with Label-Aware Syntax. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 12794–12802.
23. Fei, H.; Wu, S.; Li, J.; Li, B.; Li, F.; Qin, L.; Zhang, M.; Zhang, M.; Chua, T.S. LasUIE: Unifying Information Extraction with Latent Adaptive Structure-aware Generative Language Model. Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022, 2022, pp. 15460–15475.
24. Wu, S.; Fei, H.; Ji, W.; Chua, T.S. Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 2593–2608.
25. Li, J.; Tu, Z.; Yang, B.; Lyu, M.R.; Zhang, T. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183* **2018**.
26. Im, J.; Cho, S. Distance-based self-attention network for natural language inference. *arXiv preprint arXiv:1712.02047* **2017**.
27. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2023.
28. Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; Sima'an, K. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675* **2017**.
29. Fei, H.; Wu, S.; Ren, Y.; Zhang, M. Matching Structure for Dual Learning. Proceedings of the International Conference on Machine Learning, ICML, 2022, pp. 6373–6391.
30. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**.
31. Li, J.; Xiong, D.; Tu, Z.; Zhu, M.; Zhang, M.; Zhou, G. Modeling source syntax for neural machine translation. *arXiv preprint arXiv:1705.01020* **2017**.
32. Fei, H.; Liu, Q.; Zhang, M.; Zhang, M.; Chua, T.S. Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 5980–5994.
33. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* **2015**.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
35. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* **2018**.
36. Jawahar, G.; Sagot, B.; Seddah, D.; Unicomb, S.; Iñiguez, G.; Karsai, M.; Léo, Y.; Karsai, M.; Sarraute, C.; Fleury, É.; others. What does BERT learn about the structure of language? 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019.
37. Zhang, B.; Titov, I.; Sennrich, R. Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention. *arXiv preprint arXiv:1908.11365* **2019**.
38. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.