Linguistic-aided Sentence Representation Learning

Gomez Nalin , Vatsalan Dante , Woods Ali *

Posted Date: 30 November 2023

doi: 10.20944/preprints202311.2002.v1

Keywords: sentence semantics; syntactic enhancement network; language model integration

Article

Linguistic-Aided Sentence Representation Learning

Gomez Nalin, Vatsalan Dante and Woods Ali *

University of Charleston

* Correspondence: aliwoods12345@gmail.com)

Abstract: The realm of natural language processing has always been fascinated by the intricacies of sentence semantics. The emergence of context-aware word representations, especially from pre-trained models like ELMO and BERT, has revolutionized several semantic tasks including but not limited to, question answering, text categorization, and sentiment analysis. Despite these advancements, the integration of supplementary knowledge, particularly syntactic, to augment semantic comprehension in models remains an area ripe for exploration. This paper introduces the Syntactic Enhancement Network (SEN), a pioneering method for synergizing syntactic elements with established pre-trained language models. Our approach encompasses a dual-phase evaluation: initially, we delve into the syntactic augmentation of both RNN-based and Transformer-based language models; subsequently, we test our SEN's proficiency in two specific domains: the task of sentence completion and the extraction of biological relationships. The results are striking, with SEN achieving a stellar 91.2% accuracy in sentence completion—surpassing baseline models by a substantial 37.8%. In the context of biological relation extraction, SEN demonstrates competitive prowess with a 75.1% F_1 score.

Keywords: sentence semantics; syntactic enhancement network; language model integration

1. Introduction

In recent years, the field of natural language processing has experienced an unprecedented interest in the area of pre-trained language models. The focus of contemporary research in this domain has been categorized into three distinct yet interconnected streams: (1) The application of context-aware word representations in a variety of downstream tasks, as evidenced in works like [1–3]; (2) A deep dive into understanding the extensive knowledge that these models garner from vast text corpora [6,7]; (3) The enhancement of word representations by incorporating external, particularly syntactic, information [8,9]. In line with these developments, our research presents the innovative Syntactic Enhancement Network (SEN), a groundbreaking framework that integrates syntactic structures into pre-trained language models with finesse. Unlike conventional models, SEN adopts a unique tripartite approach to dependency syntax, ensuring compatibility with both RNN and Transformer-based models. This novel method stands out for its efficiency and compatibility with mini-batch training methodologies, setting it apart from traditional dependency parser tree models [10–12].

Our exploration of SEN's capabilities is conducted through two distinct sentence-level tasks. The first task, sentence completion, involves choosing the most apt word or phrase from a selection to complete a sentence effectively. For this, we utilize the well-established Microsoft Research Sentence Completion Challenge (MSCC) [13], which is composed of sentences from the classic Sherlock Holmes novels, each with a single missing word and five potential choices. Prior research [15,16] has validated the importance of pre-trained models and syntactic information in this particular context. The second task is centered around relation extraction, with the objective of classifying pairs of entities in sentences. For this, we use the renowned SemEval2013 DDI Extraction challenge dataset, a benchmark dataset for biological relation extraction, again highlighting the crucial role of syntactic information in this area [4,11,12,17–20].

Furthermore, we are excited to introduce the English Sentence Gap-Filling (ESG) dataset, an innovative resource specifically designed for sentence completion tasks, derived from a range of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2023 doi:10.20944/preprints202311.2002.v1

standardized English examinations. This dataset broadens the scope of question types and grammatical phenomena beyond what is available in the MSCC, as demonstrated in Table 1. A comprehensive exploration of the ESG dataset is provided in Section 4.1.

Table 1. Examples of three question formats: single words, contiguous phrases, and brief sentences

1. To ___ his thirst, John drinks two liters of water daily. A. quench B. avoid C. satisfy D. alleviate	2. Not only is the book ___ but also ___ enlightening. A. engaging, insightful B. long, tedious C. useful, practical D. interesting, controversial	3. The author implies that the character is ___ A. often misunderstood B. not as simple as he seems C. more complex than he appears D. misunderstood by most

The contributions of this paper are significant and varied:

- We introduce the Syntactic Enhancement Network (SEN), a pioneering approach that skillfully weaves complex syntactic structures into the fabric of pre-trained language models.
- We rigorously test the SEN across two pivotal sentence-level tasks, employing diverse experimental designs to thoroughly evaluate its performance.
- We develop the extensive English Sentence Gap-Filling dataset, tailored for sentence completion in English examinations, poised to become a benchmark in the field.

2. Related Work

Recent advancements in natural language processing have been significantly influenced by the development of pre-trained language models, notably ELMO and BERT, as introduced by Peters et al. [21] and Devin et al. [22] respectively. ELMO's innovation lies in its utilization of a BiLSTM trained with a dual language model objective, while BERT, diverging from BiLSTM, leverages a bidirectional Transformer encoder as conceptualized by Vaswani et al. [23] for predicting words in a masked context. Building upon these foundational works, subsequent research has explored various applications and enhancements of these models. For instance, Sun et al. [1] and Nogueira et al. [3] applied the BERT model to tasks like sentiment analysis and passage ranking, achieving groundbreaking results. Additionally, studies by Liu et al. [27] and Tenney et al. [6] demonstrated the capability of ELMO and BERT in capturing both syntactic and semantic nuances in language. Furthermore, Zhang et al. [8] and Sun et al. [28] proposed the integration of knowledge graphs and entity information with contextualized embeddings to further refine language representations. In the domain-specific context, Lee et al. [29] and Beltagy et al. [30] showed how fine-tuning BERT with specialized data can significantly enhance performance in specific fields.

Focusing on the sentence completion task, Zweig et al. [31] introduced baseline models based on n-gram language models and latent semantic analysis, achieving accuracies of 39% and 49% respectively. Mikolov et al. [15] enhanced these results by integrating skip-gram with an RNN-based language model, leading to a 58.9% accuracy in the MSCC. Park et al. [35] further improved upon these results with a bidirectional word-level RNN. Beyond architectural changes, there has been a significant interest in incorporating syntactic information into models. For instance, Piotr Mirowski et al. [36] achieved a 10% absolute improvement over previous RNN-based baselines by integrating syntactic information into the RNN language model. Zhang et al. [10] employed a tree-structured LSTM model to infuse dependency path information into the model.

In the context of the DDI2013 Extraction task, Zhang et al. [12] combined shortest dependent path information and sentence sequences using a hierarchical RNN network, surpassing other state-of-the-art methods. Miwa et al. [11] adopted a bidirectional Tree-LSTM to capture substructural information within dependency trees, leading to significant error reductions in F1-score. From these preceding studies on sentence completion and relation extraction tasks, it is evident that models leveraging syntactic information predominantly rely on tree structures. In our proposed Syntactic Enhancement Network (SEN), we linearize this tree structure into a triple form and utilize expanded

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2023 doi:10.20944/preprints202311.2002.v1

syntactic information for improved sentence representations. This approach demonstrates versatility and effectiveness in both sentence completion tasks and relation extraction tasks.

3. The Proposed Framework

In this section, we elaborate on our Syntactic Enhancement Network (SEN) model. The architecture of SEN, encompasses three primary components: (1) Encoding Layer: employing a pre-training language model to generate word embeddings; (2) Dependency Syntax Integration Layer: merging syntactic information with pre-trained word embeddings, followed by a BiLSTM to yield refined sentence representations; (3) Output Layer: determining the most probable answer or label based on the final semantic representations of sentences. We represent a sentence $S = [w_1, w_2, \cdots, w_n]$ as a sequence of n words.

3.1. Encoding Layer

The encoding layer's primary function is to transform each word in the sentence into a fixed-length vector using pre-trained word embeddings. We initially utilize a bidirectional LSTM network (Hochreiter and Schmidhuber [40]), akin to ELMO, and refer to it as biLM. Subsequently, we apply the BERT pre-trained model based on Transformer architecture to acquire a distinct set of contextualized word embeddings. Specifically, for biLM, the forward phase calculates the probability of sentence by modeling the likelihood of word w_k given $(w_1, \ldots, w_{k-1})$, while the backward phase processes the sentence in reverse. The joint objective maximizes the log likelihood of both passes as follows:

$$\begin{aligned} p\left(w_1, w_2, \ldots, w_n\right) &= \prod_{k=1}^{n} p\left(w_k | w_1, w_2, \ldots, w_{k-1}\right) \\ p\left(w_1, w_2, \ldots, w_n\right) &= \prod_{k=1}^{n} p\left(w_k | w_{k+1}, w_{k+2}, \ldots, w_n\right) \end{aligned} \tag{1}$$

$$\begin{aligned} \sum_{k=1}^{n} \Big(& \log p\left(w_k | w_1, \ldots, w_{k-1}; \overleftarrow{\theta}_{LSTM}\right) \\ & + \log p\left(w_k | w_{k+1}, \ldots, w_N; \overrightarrow{\theta}_{LSTM}\right) \Big) \end{aligned} \tag{2}$$

For BERT, given its unique architecture, a modified language modeling objective is adopted, masking some tokens and predicting their likelihood through a softmax layer. During training, BERT's parameters are updated while biLM's remain fixed.

3.2. Dependency Syntax Integration Layer

This subsection describes the fusion of explicit syntax information with pre-trained word embeddings. Dependency syntax, illustrating grammatical relations between words, is traditionally represented by binary trees. In this structure, each word relies on a unique head. These head-dependent relations, essential for applications like coreference resolution and information extraction, are commonly depicted in a triple form: (dependent, relation, head). For example, (fruit, nsubj, is) suggests that *fruit* modifies *is* with a *nsubj* relation. We transform these dependency trees into an expanded form.

Using this triple form, syntax information is effortlessly integrated with contextualized word embeddings to create novel sentence representations. We employ two fusion functions to map a triple into a continuous vector, defined as:

$$x_i = g\left(w_i, r_i, w_i^H\right) \tag{3}$$

Here, w_i and $w_i^H \in \mathbb{R}^d$, with w_i^H being the head of w_i and d the embedding dimension. r_i is a low-dimensional vector representing different dependency relations. The fusion function $g(\cdot)$ has two implementations: vector concatenation and a gating mechanism controlling information flow:

$$g\left(w_i, r_i, w_i^H\right) = w_i \oplus r_i \oplus w_i^H \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2023 doi:10.20944/preprints202311.2002.v1

$$g\left(w_i, r_i, w_i^H\right) = w_i + \sigma\left(r_i\right) \otimes w_i^H \tag{5}$$

The processed sentential sequence is denoted as $S^p = [x_1, x_2, \ldots, x_n]$.

S^p is then inputted into a BiLSTM network. The following equations detail the calculation of each gate and memory cell unit in the BiLSTM:

$$\begin{aligned}
h_i &= \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \\
h_t &= o_t \tanh\left(c_t\right) \\
f_t &= \sigma\left(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f\right) \\
i_t &= \sigma\left(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i\right) \\
g_t &= \tanh\left(W_{xc} \cdot [C_{t-1}, h_{t-1}, x_t] + b_c\right) \\
c_t &= i_t g_t + f_t c_{t-1} \\
o_t &= \sigma\left(W_o \cdot [C_t, h_{t-1}, x_t] + b_o\right)
\end{aligned} \tag{6}$$

The final sentence representation $S^f = \left[x_1^f, x_2^f, \ldots, x_n^f\right]$ combines the forward and backward output at the last position.

3.3. Output Layer

The output layer utilizes a fully connected feed-forward network for final predictions. However, the objective functions differ for sentence completion and relation extraction tasks. In sentence completion, we employ a pairwise approach to rank answer options. The loss function is defined as:

$$\text{loss} = \sum_{i=1,2,3}^{N} \max(0, -f(S_{true}^f) + f(S_{false\_i}^f) + \text{margin}) \tag{7}$$

$f(\cdot)$ computes scores for each candidate sentence, selecting the highest-scoring completion. Here, $u \in \mathbb{R}^k$ is a learnable parameter, and *margin* is a threshold to increase score separation between S_{true}^f and $S_{false\_i}^f$.

For DDI extraction, we use negative log-likelihood as our loss function:

$$\text{loss} = -\log p\left(y_i | S_{p_i}^f\right) = -\log\left(\frac{\exp(y_i)}{\sum_j^5 \exp(y_j)}\right) \tag{8}$$

where $S_{p_i}^f$ is an entity pair in sentence S^f, and y_i is the relation class of p_i. The softmax function computes the probability that an entity pair $S_{p_i}^f$ belongs to class y_i.

4. Experiments

4.1. Settings

ESG dataset We sourced our English Sentence Gap-Filling (ESG) dataset primarily from high school English multiple-choice questions found on online educational platforms. Prior to training, the dataset underwent preprocessing as follows: Initially, to avoid redundancy, questions with an editing distance below 8 were filtered out. Then, utilizing the Stanford Parser, we implemented part-of-speech tagging and dependency parsing. The final dataset, comprising 62,834 questions, was divided into training and test sets in a 9:1 ratio.

DDI dataset The DDIExtraction 2013 dataset, derived from the DrugBank and DrugMedline databases, is accessible at https://www.cs.york.ac.uk/semeval-2013/task9/. This task revolves around correctly classifying drug-drug interaction relations into five categories: advice, effect, mechanism, int, and negative. Table 2 offers a detailed breakdown of this data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2023 doi:10.20944/preprints202311.2002.v1

Table 2. Breakdown of the DDIExtraction 2013 dataset. The term *negative* denotes the absence of any relationship between two drug entities.

	Train			Test		
	DrugBank	MedLine	Overall	DrugBank	MedLine	Overall
Positive	3767	231	3998	884	92	976
Negative	14445	1179	15624	2819	243	3062
Advice	815	7	822	214	7	221
Effect	1517	152	1669	298	62	360
Mechanism	1257	62	1319	278	21	299
Int	178	10	188	94	2	96

4.2. Configurations

In the sentence completion task, we employed the PyTorch implementation of BERT with models provided by Google.[1] Given BERT's use of BPE for tokenization, which splits some words into sub-words, we used the first word piece embedding to maintain consistent sequence lengths with the original sentences. We chose the *bert-base-cased* model ($\text{BERT}_{\text{base}_{\text{u}}\text{ncased}}$), with a maximum sequence length of 128, a mini-batch size of 16, and BertAdam as the optimizer. The learning rate was set to $\alpha = 2e-5$ with $warmup = 0.05$. The relation vector size was fixed at 200 and the hidden size of the BiLSTM at 768. The model was trained for 6 epochs, with evaluations after each epoch. Additionally, we gathered a 26G unannotated text corpus from Wikipedia, Gigaword, and English learning websites to train the biLM model.

For the DDI task, we opted for the BioBERT [29] model, a domain-adaptive version of BERT pre-trained on biomedical corpora. We set the sequence length at 350 and the minibatch size at 32. The learning rate and *warmup* were adjusted to $3e-5$ and 0.1, respectively. The BiLSTM's hidden size was 512, with other hyperparameters mirroring those of the sentence completion task. We replaced entities in the sentence with "drug1" and "drug2" for those requiring prediction, and "drug0" for others. This approach, as shown in prior studies, helps models better grasp semantic entity relationships.

4.3. Results and Analysis

Main results The results on the ESG dataset are summarized in Table 3. Our baseline model, using randomly initialized word vectors and a BiLSTM layer, served as the comparison point. In the table, models marked as 'biLM' and '$\text{BERT}_{\text{base}_{\text{u}}\text{ncased}}$' were trained without the syntax integration layer. Accuracy was the chosen metric for evaluation. As evidenced, the best performance (91.2%) was achieved by the SEN model utilizing BERT. Pre-trained language models substantially enhanced performance, and the SEN model with BERT outperformed the one with biLM. Syntax integration on top of pre-trained LMs improved accuracy by 2% and 0.6% for biLM-based and BERT-based models, respectively. Furthermore, fusion methods (equations 4 and 5) had varying impacts depending on the model. For instance, the gate mechanism proved more effective for the biLM+SEN model, while the reverse held true for the '$\text{BERT}_{\text{base}_{\text{u}}\text{ncased}}$+SEN' model.

Inspired by domain-specific fine-tuning approaches like SciBERT[30] and BioBERT[29], we introduced a BERT variant, '$\text{BERT}_{\text{finetune}}$', pre-trained on an additional 1.6G corpus containing nuanced words and phrases from error samples. This fine-tuned model exhibited stable performance improvements on the ESG dataset.

[1] https://github.com/huggingface/pytorch-pretrained-BERT

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 November 2023 doi:10.20944/preprints202311.2002.v1

Table 3. Performance comparison of various models on the ESG dataset.

Model	Acc(%)
baseline	53.4
biLM	73.0
biLM+SEN(concat)	73.9
biLM+SEN(gate)	**75.9**
$\text{BERT}_{\text{base\_uncased}}$	90.3
$\text{BERT}_{\text{base\_uncased}}$+SEN(gate)	90.7
$\text{BERT}_{\text{base\_uncased}}$+SEN(concat)	90.9
$\text{BERT}_{\text{finetune}}$+SEN(concat)	**91.2**

Table 4. The comparison of model training time between Tree-LSTM and SEN.

Model	Time (seconds per train epoch)
Tree-LSTM	10431
SEN	2367
SEN +Tree-LSTM	2587

Error Analysis and Visualization An error analysis revealed a notable decrease in error rates for verb differentiation, tense consistency, and phrase collocation, attributing this improvement to the integration of syntax information and the fine-tuning step. Additionally, we utilized Bertviz to visualize attention patterns in the '$\text{BERT}_{\text{base}_{\text{u}}\text{ncased}}$' and '$\text{BERT}_{\text{base}_{\text{u}}\text{ncased}}$+SEN' models, highlighting the impact of syntax integration on model learning.

DDI Task Evaluation We benchmarked our model against recent DDI task methods, focusing on precision, recall, and micro F1-score metrics. The results, detailed in Table 5, demonstrate that our model not only achieved competitive performance but also benefited from the integration of syntactic expansion information.

Table 5. Comparative results on the DDI2013 Extraction task. P, R, and F_1 denote precision, recall, and micro F1-scores, respectively.

Method	P	R	F_1
Multichannel CNN (Quan et al. [41])	75.9	62.2	70.2
Hierarchical RNNs (Zhang et al. [12])	74.1	71.8	72.9
One-Stage Model Ensemble (Lim et al. [42])	77.8	69.6	73.5
$\text{BERT}_{\text{base\_cased}}$	72.6	66.3	69.3
$\text{BERT}_{\text{base\_cased}}$+SEN(concat)	79.6	65.3	71.8
$\text{BioBERT}_{\text{base\_cased}}$	75.5	73,2	74.4
$\text{BioBERT}_{\text{base\_cased}}$+SEN(concat)	**77.7**	**72.4**	**75.1**

Ablation Study We performed an ablation study on the inputs of the dependency fusion layer by varying the composition of the triple in equation 4. The results in Table 6 indicate that the dependent relation vector significantly contributes to model performance improvements.

Table 6. Results from the ablation study evaluating different triple forms in the SEN model.

Triple form	ESG dataset(Acc)	DDI dataset (F_1)
$x_i = w_i \oplus r_i \oplus w_i^H$	**91.2**	**75.1**
$x_i = w_i \oplus w_i^H$	90.5	74.6
$x_i = w_i$	90.2	74.1

5. Conclusion and Future Exploration

This study was dedicated to exploring the synergistic integration of syntactic information with sophisticated pre-trained language models. To achieve this, we innovatively developed a triplet-based structure for expanding dependency trees, leading to the creation of the Syntactic Enhancement Network (SEN). This novel approach was rigorously tested across both sentence completion and relation extraction tasks. The results from these experiments were quite illuminating, leading to two primary insights. Firstly, the utilization of an advanced pre-trained language model can significantly

boost the overall performance of the system. Secondly, the fusion of dependency syntax information with contextualized word embeddings is instrumental in achieving more nuanced and accurate sentence representations. Moreover, we constructed a new dataset for sentence completion tasks, tailored to reflect real-world application scenarios more closely. This dataset not only served as a practical testbed for our SEN model but also contributes to the broader research community by providing a resource that closely mirrors actual usage patterns. Looking ahead, our future endeavors will concentrate on further refining the combination of external knowledge with pre-trained language models. We aim to extend the applicability of our SEN model to a broader array of downstream tasks, thereby enhancing its versatility and effectiveness in various natural language processing applications. This ongoing research will contribute to the evolving landscape of language processing technologies, offering more sophisticated tools for understanding and utilizing natural language.

References

1. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. *arXiv preprint arXiv:1903.09588* **2019**.
2. 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.
3. Nogueira, R.; Cho, K. Passage Re-ranking with BERT. *CoRR* **2019**, *abs/1901.04085*, [1901.04085].
4. 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.
5. 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.
6. Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R.T.; Kim, N.; Van Durme, B.; Bowman, S.R.; Das, D.; others. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316* **2019**.
7. Fei, H.; Li, F.; Li, C.; Wu, S.; Li, J.; Ji, D. Inheriting the Wisdom of Predecessors: A Multiplex Cascade Framework for Unified Aspect-based Sentiment Analysis. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, 2022, pp. 4096–4103.
8. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced Language Representation with Informative Entities. *arXiv preprint arXiv:1905.07129* **2019**.
9. 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*.
10. Zhang, X.; Lu, L.; Lapata, M. Top-down Tree Long Short-Term Memory Networks. Proceedings of NAACL-HLT, 2016, pp. 310–320.
11. Miwa, M.; Bansal, M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, Vol. 1, pp. 1105–1116.
12. Zhang, Y.; Zheng, W.; Lin, H.; Wang, J.; Yang, Z.; Dumontier, M. Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics* **2017**, *34*, 828–835.
13. Zweig, G.; Burges, C.J. The microsoft research sentence completion challenge. *Microsoft Research, Redmond, WA, USA, Tech. Rep. MSR-TR-2011-129* **2011**.
14. 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.
15. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**.
16. Monz, K.T.A.B.C. Recurrent Memory Networks for Language Modeling. Proceedings of NAACL-HLT, 2016, pp. 321–331.
17. 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.

18. 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.
19. Wu, S.; Fei, H.; Cao, Y.; Bing, L.; Chua, T.S. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 14734–14751.
20. 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.
21. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2227–2237.
22. 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**.
23. 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.
24. 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.
25. 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.
26. 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.
27. Liu, N.F.; Gardner, M.; Belinkov, Y.; Peters, M.; Smith, N.A. Linguistic Knowledge and Transferability of Contextual Representations. *arXiv preprint arXiv:1903.08855* **2019**.
28. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. ERNIE: Enhanced Representation through Knowledge Integration. *CoRR* **2019**, *abs/1904.09223*, [1904.09223].
29. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746* **2019**.
30. Beltagy, I.; Cohan, A.; Lo, K. SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *arXiv preprint arXiv:1903.10676* **2019**.
31. Zweig, G.; Platt, J.C.; Meek, C.; Burges, C.J.; Yessenalina, A.; Liu, Q. Computational approaches to sentence completion. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012, pp. 601–610.
32. 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.
33. 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.
34. 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.
35. Park, H.; Cho, S.; Park, J. Word RNN as a Baseline for Sentence Completion. 2018 IEEE 5th International Congress on Information Science and Technology (CiSt). IEEE, 2018, pp. 183–187.
36. Mirowski, P.; Vlachos, A. Dependency recurrent neural language models for sentence completion. *arXiv preprint arXiv:1507.01193* **2015**.
37. Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; Chua, T.S. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2023, pp. 1171–1182.

38. 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.
39. Li, B.; Fei, H.; Li, F.; Wu, Y.; Zhang, J.; Wu, S.; Li, J.; Liu, Y.; Liao, L.; Chua, T.S.; Ji, D. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 13449–13467.
40. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
41. Quan, C.; Hua, L.; Sun, X.; Bai, W. Multichannel convolutional neural network for biological relation extraction. *BioMed research international* **2016**, *2016*.
42. Lim, S.; Lee, K.; Kang, J. Drug drug interaction extraction from the literature using a recursive neural network. *PloS one* **2018**, *13*, e0190926.