Using Distributional Thesaurus To Enhance Transformer-based Contextualized Representations for Low Resource Languages

Gopalakrishnan Venkatesh
International Institute of Information Technology Bangalore
Gopalakrishnan.V@iiitb.ac.in

Abhik Jana
Universität Hamburg
abhik.jana@uni-hamburg.de

Steffen Remus
Universität Hamburg
steffen.remus@uni-hamburg.de

Özge Sevgili
Universität Hamburg
oezge.sevgili.ergueven@studium.uni-hamburg.de

Gopalakrishnan Srinivasaraghavan
International Institute of Information Technology Bangalore
gsr@iiitb.ac.in

Chris Biemann
Universität Hamburg
christian.biemann@uni-hamburg.de

ABSTRACT

Transformer-based language models recently gained large popularity in Natural Language Processing (NLP) because of their diverse applicability in various tasks where they reach state-of-the-art performance. Even though for resource-rich languages like English, performance is very high, there is still headroom for improvement for low resource languages. In this paper, we propose a methodology to incorporate Distributional Thesaurus information using a Graph Neural Network on top of pretrained Transformer models to improve the state-of-the-art performance for tasks like semantic textual similarity, sentiment analysis, paraphrasing, and discourse analysis. In this study, we attempt various NLP tasks using our proposed methodology for five languages – English, German, Hindi, Bengali, and Amharic – and show that by using our approach, the performance improvement over transformer models increases as we move from resource-rich (English) to low-resource languages (Hindi, Bengali, and Amharic).

CCS CONCEPTS

• **Computing methodologies** → **Lexical semantics**;

KEYWORDS

Distributional Thesaurus, Graph Convolution Network, Transformers, Low Resource Language, Semantics

ACM Reference Format:
Gopalakrishnan Venkatesh, Abhik Jana, Steffen Remus, Özge Sevgili, Gopalakrishnan Srinivasaraghavan, and Chris Biemann. 2022. Using Distributional Thesaurus To Enhance Transformer-based Contextualized Representations for Low Resource Languages. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event,* . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3477314.3507077

1 INTRODUCTION

With the evolution of transformer-based language models [10, 27], NLP researchers are applying these in a variety of NLP tasks. For example, the state-of-the-art models for the standard Natural Language Understanding (NLU) benchmark tasks, like GLUE [40], are different variants of BERT [10] based models. Note that, all these models are based on the Transformer [38] architecture and are extensively pretrained on large corpora. Considering their commendable performance on these benchmarks, prior works [7, 16, 36, 44] have extensively investigated the knowledge encoded in the representations of these models. In this direction, some of the recent works [7, 16, 36, 44] have investigated such transformer-based architectures from a linguistic point of view. From such investigations, one conclusive finding conveys that, unlike syntax, semantics and general world knowledge are not brought to the surface by the representations obtained from such models [3].

Researchers attempt to bridge such a gap in transformer-based models in various ways. Efforts have been made to incorporate semantics from resources like pretrained semantic parsers like Semantic Role Labelers [48] and DELPH-IN MRS-derived dependencies (DM) [44]. Semantics-aware BERT (SemBERT) [48] leverages a pretrained Semantic Role Labeler to fetch multiple predicate derived structures of explicit semantics. These semantic structures are further encoded and fused with the BERT embeddings to generate the final semantic embedding. Along a similar direction, Semantics Infused Finetuning (SIFT) explicitly encodes DELPH-IN MRS-derived semantic dependency parses, using Relational Graph Convolutional Networks, and this knowledge is used to enrich the representations from the pretrained transformer-based models.

All of these approaches discussed above leverage some resources that are available for resource-rich languages like English. For medium-resource (German) or low-resource languages (Hindi, Bengali, Amharic), those resources are not available to compensate for the weakness in terms of the lack of training data of transformer-based models. Therefore, we propose the idea of incorporating Distributional Thesaurus (DT) information as an additional knowledge source. DT is a sparse similarity graph that encodes lexical-semantic relationships between words [19]. It can be built from medium-sized corpora in an unsupervised way and serves as a distilled semantic

SAC '22, April 25–29, 2022, Virtual Event,
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507077

lexicon, making it suitable for research in the low-resource language paradigm. To fuse the knowledge of DT on top of pretrained transformer architecture, we use Graph Neural Networks (GNN). We try with two variants of GNN namely, Graph Convolution Network [24] and Local Extrema convolution [33] for this study. We experiment with a wide variety of tasks for five languages namely English, German, Hindi, Bengali, Amharic. As transformer-based language models, we use BERT, RoBERTa, mBERT, and XLM-R as per the availability of models in respective languages. From the extensive set of experiments, we see that the proposed approach produces similar performance for the tasks attempted for the English language. On the other hand, for medium-resource language like German, it shows an average performance gain of 0.5-3% whereas for low-resource languages like Hindi, Bengali, and Amharic the performance gain due to our proposed methodology is consistently higher (for Hindi and Bengali it is 0.08-3.13% and 0.87-2.14% respectively while for Amharic it is 4.54-6.01%).

In a nutshell, our contributions are threefold:

- Propose an idea of utilizing knowledge from DT as an additional resource to improve the performance of transformer-based models especially for low-resource languages while experimenting with various NLP tasks.
- Use GNNs to fuse the knowledge from a DT with a pretrained transformer-based model. The approach architecture is simple and requires less training time as the model parameters are finetuned and not trained from scratch.
- By following our proposed approach we produce promising state-of-the-art performances for the NLP tasks attempted in Hindi, Bengali, and Amharic languages. We make all the data and code publicly available[1].

2 RELATED WORK

2.1 Transformer Representations

In a time when Recurrent Neural Networks (RNN) and Convolution Neural Networks (CNN) were the popular neural network architectures for sequence-to-sequence tasks, Vaswani et al. [38] presented the transformer architecture, which solely relies on attention-based mechanisms. Their proposed method leverages a self-attention scheme, which addresses the issue of long-range dependencies encountered in RNNs. Furthermore, the transformer model recorded the state-of-the-art performance in the machine translation task beating the previous ensemble methods by a large margin [38].

2.1.1 BERT. Based on the successful usage of self-attention in transformers, Devlin et al. [10] presented the Bidirectional Encoder Representations from Transformers (BERT) model. Unlike the conventional auto-regressive Language Models (LM), BERT is pretrained using the masked language modeling objective and the next sentence prediction objective. Here, the masked language modeling objective forces the model to estimate the probability of a masked token by jointly conditioning on both the left and right contexts. For pretraining this model, the authors used the BooksCorpus (800M words) [49] and English Wikipedia (2,500M words). Later, these pretrained representations are to be adapted to the downstream task of interest by finetuning the parameters of the model. Over the years, this *pretrain-finetune* approach has been shown to be successful in a wide variety of NLU tasks [39, 40].

2.1.2 mBERT. Recently, the authors of BERT released the multilingual version of their model, where it has been pretrained on 104 languages. The evaluation performed on a wide variety of tasks in diverse languages presents promising results.

2.1.3 RoBERTa. The RoBERTa model [27] builds on BERT's language masking strategy, wherein the model learns to predict intentionally hidden sections of text within, otherwise, unannotated language examples. RoBERTa modifies certain hyperparameters in the BERT architecture including the removal of the next sentence prediction objective while pretraining the network with much larger mini-batch sizes and learning rates. The authors claim that these changes enable RoBERTa to improve on the masked language modeling objective and thereby leads to improved performance on a wide variety of downstream tasks [39, 40].

2.1.4 XLM-RoBERTa. XLM-RoBERTa [8] is a large multilingual model, based on the RoBERTa architecture, pretrained on 2.5TB of filtered CommonCrawl data. Experiments confirm that it outperforms mBERT on a wide range of cross-lingual tasks.

Alongside the development of the transformer-based encoder representations, recent works [7, 16, 36, 44] have also investigated the representations obtained from these models to better understand the knowledge encoded within them. The authors of these independent works conclude that these models produce powerful representations that capture the syntax from the text but the semantics is not brought to the surface.

SemBERT [48] was proposed to incorporate explicit contextual semantics from a pretrained semantic role labeler and to thereby enrich the underlying language representation model. This approach is of interest as it permits the convenient usability of the pretrained weights of the BERT model while permitting the introduction of semantics in a light finetuning manner without substantial task-specific modifications. Following the *pretrain-finetune* methodology, Semantics-Infused Finetuning (SIFT) [44] incorporates the semantics from DELPH-IN MRS-derived (DM) dependencies [17], by leveraging Graph Neural Networks, in the finetuning stage. Overall, the above methods hold promise as they permit introducing structural semantics while allowing the re-utilization of the knowledge presented in these models. Unfortunately, these methods rely on parsers that are available only for resource-rich languages like English, and hence adapting these methods to other languages would be challenging.

VGCN-BERT [29] incorporates information from a vocabulary co-occurrence graph into a pretrained BERT model. The scheme encodes the graph information by applying Graph Convolutional Neural Network to generate graph embeddings. These representations are then added to the pretrained BERT embeddings, which are then subjected to the transformer block's expensive self-attention scheme. The authors claim that this scheme, in spite of it being computationally expensive, gives consistent improvements in the downstream task of sentiment analysis. Similarly, recent methods [46, 47] encode corpus level co-occurrence information using variants of GNNs to boost the performance on downstream tasks like sentiment analysis. The overall theme from the above methods

[1] https://github.com/uhh-lt/DT2ContextRep

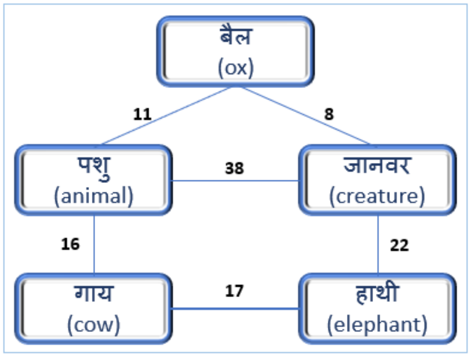

Figure 1: Snapshot of a Distributional Thesaurus (Hindi).

concurs with the notion that word co-occurrence information or word similarity information enriches the LM representations.

2.2 Graph Neural Networks

Leveraging traditional machine learning techniques on graphs is a challenging task due to its arbitrary and highly complex structure. Thereby, well-established neural models like RNNs or CNNs would struggle to generalize on these sparse structures. To tackle this challenge, Graph Neural Networks (GNN) were introduced that aim at learning features on a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. These algorithms take the following as input:

(1) A feature description x_i for every node i; summarized in a $\mathbf{N} \times \mathbf{D}$ feature matrix $\mathbf{X}$, where $\mathbf{N}$ is the number of nodes and $\mathbf{D}$ is the number of input features.
(2) A representative description of the graph structure in a matrix form, which is typically presented in the form of an adjacency matrix $\mathbf{A}$.

The GNN network is expected to produce a node-level output $\mathbf{Z}$: an $\mathbf{N} \times \mathbf{F}$ feature matrix where $\mathbf{F}$ is the number of output features per node. Thereby, every GNN network layer can then be written as a non-linear function:

$$\mathbf{X}' = f(\mathbf{X}, \mathbf{A}) \tag{1}$$

From these node-level representations, a graph-level output can be computed by applying a pooling operation. Please refer to the recent survey by Wu et al. [42] for the recent advancements in GNNs. We further elaborate on the GNNs leveraged by us in our work in Sections 3.2.1 and 3.2.2.

3 METHODOLOGY

The core of our methodology relies on fusing information from a Distributional Thesaurus (DT) with the transformer architecture. Therefore, we first discuss the details of a DT and the data sources used to compute it in different languages.

3.1 Distributional Thesaurus

In this work, following the definition by Biemann and Riedl [5], a Distributional Thesaurus (DT) is a semantic count-based similarity graph constructed by relying on the distributional hypothesis [13, 15]. The nodes of this graph are words and the edge weight between a pair of nodes corresponds to the count of common significant contexts between them. Earlier, the computation of these sparse count-based models used to be inefficient, however, in this era of high-speed processors and storage, attempts are being made to streamline the computation with ease. Riedl and Biemann [34] introduce a highly scalable approach for computing distributional thesauri by incorporating pruning techniques and by leveraging a distributed computation framework. Here, the authors compute Lexicographer's Mutual Information (LMI) [23] for each bigram which thereby gives a measure of the collocational strength of the bigram. Each bigram is thereafter broken into a word and a feature where the feature consists of the bigram relation and the related word. Later, the top 1000 ranked features for each word are considered and for each word pair, an intersection of their corresponding feature set is obtained. The word pairs having their count of overlapping features above a certain threshold are retained in the network. A sample snapshot of a subset of the DT computed on a Hindi corpus [14, 25], comprising of text obtained by crawling newspapers, generic web pages, and Wikipedia, is shown in Figure 1.

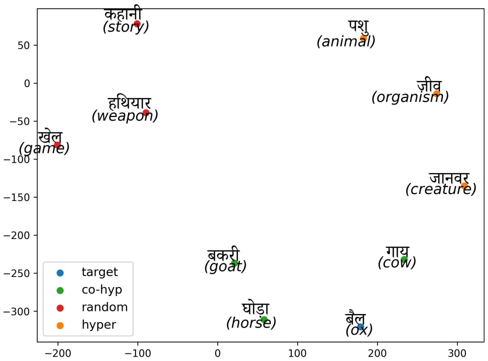

Figure 2: t-SNE plot of DeepWalk embeddings computed from Hindi DT. We can see, for a target word 'Ox' the semantically related words (co-hyponymys, hypernyms) are closer to it; on the other hand random unrelated words are far apart in the embedding space.

Recent work of Jana and Goyal [18] has confirmed that the information present in the DT complements the knowledge encoded in the traditional static word embeddings like GloVe [31]. To substantiate this claim, the authors first convert the sparse DT graph structure into dense vector representations by applying network embedding algorithms [32]. Later, the DT embeddings are fused with the conventional embeddings in their analysis. Experiments on downstream tasks, like word similarity and relatedness, confirm that the fused representations encode more semantics than the conventional embeddings.

Furthermore, the authors of [20] state that embeddings obtained from a DT, by applying standard network embedding algorithms [32], encodes lexical-semantic properties like hypernymy, co-hyponymy, and meronymy. Complementing the encouraging quantitative results, the authors also qualitatively investigate the clustering pattern noticed in the vector space. Figure 2 presents the t-Distributed Stochastic Neighbour (t-SNE) [37] plot of the DeepWalk network embedding [32] computed on the Hindi DT. Qualitative analysis of Figure 2 reveals that the tokens that are co-hyponyms of the target word are closer than the words with the hypernym relation, while the words without any meaningful relationship reside far away from it. Additionally, the analysis performed by Jana and Goyal [19] reveals that quantifiable *cohesion indicating* network properties, including shortest path length, can discriminate lexical relations between word pairs.

The publicly available English and German DTs used in our experiments are taken from the JoBimText Visualizer framework [4]. The DTs of the Indian languages, namely Hindi and Bengali, have been provided by Kumar et al. [25]. In this work, the authors leverage the Leipzig Corpora Collection [14], which comprises of texts collected from crawling newspapers, generic webpages, and Wikipedia, to prepare both the DTs. Finally, the recently released Amharic DT [45], computed on a text source including news articles, tweets, and YouTube comments, is utilized in our framework for the experiments on the Amharic datasets. In our experiments, the large English DT is pruned and only the noun tokens are retained. For the other languages, all tokens are retained in our framework.

Next, our goal is to prepare contextual representation by incorporating Distributional Thesaurus information via Graph Neural Network on top of the Transformer architecture.

3.2 Application of Graph Neural Network

The input sentence is first encoded using a pretrained transformer-based encoder. This generates a sentence level representation, CLS embedding, and subword representations for the given input sequence. As the semantics conveyed by the DT operates at a word-level construct, the subword level information is appropriately mean-pooled into its word-level concept [44]. At this phase, a graph $G = (V, E)$ is induced, where V and E are the node-set and edge-set of the graph respectively. Formally, V is the words seen in the input sentence while $E = \{(u, w, v) | u \in V, v \in V, w = e^{-pathlength(u,v)}\}$. Here, the undirected edges introduce the lexical properties by modeling the unweighted shortest path length information in an exponentially decaying function. As discussed in prior work, the shortest path length in a DT is a *cohesion-indicating* property, which is capable of discriminating lexical-semantic relations between words. Over this induced graph, a Graph Neural Network (GNN) is applied, which further contextualizes the representations while incorporating the lexical semantics knowledge from the DT. In the case of sentence pair tasks, following [44], the sentences are jointly encoded in the transformer-based model and the GNN operates on the two sentences independently. Furthermore, a biaffine [44] graph attention is employed to model the relationship between the sentences. The vanilla architecture, for single sentence tasks, is presented in Figure 3. In our investigation, we attempt the following two GNN approaches.

3.2.1 Graph Convolutional Network (GCN). Graph Convolutional Network (GCN) [24] proposes the idea of aggregating both the self features and the neighbor's features of a node. Using this idea, Equation 1 can be formulated as follows:

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\Theta \quad (2)$$

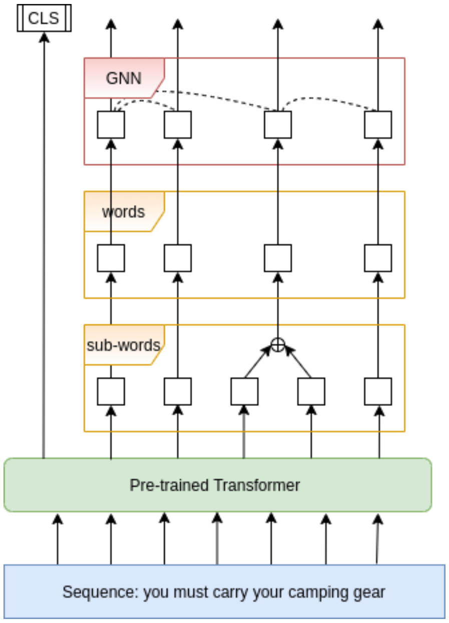

Figure 3: Our proposed architecture: Graph Neural Network on top of a pretrained transformer model.

Here, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix with inserted self-loops and $\hat{\mathbf{D}}_{ii} = \Sigma_{j=0}\hat{\mathbf{A}}_{ij}$ is its diagonal degree matrix. In equations 2 and 3, Θ represents the trainable parameters. Its node-wise formulation can be given by:

$$\mathbf{x}'_i = \Theta \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{e_{j,i}}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j \quad (3)$$

In the above equation, $\hat{d}_i = 1 + \Sigma_{j \in \mathcal{N}(i)} e_{j,i}$ where $e_{j,i}$ denotes the edge weight from the source node j to target node i.

3.2.2 Local Extrema Convolution (LEConv). Local Extrema Convolution (LEConv) [33] presents a graph convolution method that is capable of capturing local extremum information. This graph convolution operator finds the importance of nodes using the difference operator. The node-wise update formulation is given by the following equation:

$$\mathbf{x}'_i = \mathbf{x} \cdot \Theta_1 + \sum_{j \in \mathcal{N}(i)} e_{j,i} \cdot (\Theta_2 \mathbf{x}_i - \Theta_3 \mathbf{x}_j) \quad (4)$$

In the above equation, $e_{j,i}$ denotes the edge weight from the source node j to target node i and Θ_1, Θ_2, and Θ_3 represent the trainable parameters.

Thus, the node/word embeddings are obtained after the GNN fuses the lexical semantics knowledge from the DT with the contextual structural information from the pretrained transformer. For sentence-level NLU tasks, the mean pooled node representation from the GNN is concatenated with the CLS embedding and thereafter used in feedforward networks, as done in prior works [10, 27], for the downstream task of interest. After obtaining the contextualized representation by our proposed architecture we evaluate those

Table 1: Statistics of datasets used in our experiments.

Code	Dataset	train/test
en	SST-2	67349/872
	STS-B	5749/1500
	MRPC	3668/408
de	GermEval T1	5009/3532
	GermEval T2	5009/3532
hi	Product Reviews	4182/523
	MIDAS Discourse	7934/991
	WNLI	635/71
	Amrita S1	2500/900
bn	BEmoC	4994/625
	Sentiment Analysis	6889/1532
am	ASAB All	7511/939
	ASAB Cleaned	6885/861

using several tasks for different languages. The datasets used in the evaluation of our method are discussed in the following section.

4 DATASETS

Considering the flexibility presented in the method presented in Section 3, we evaluate our framework on a wide variety of tasks using datasets from typologically diverse languages. The proposed approach is first validated in English over the GLUE NLU benchmark [40]. Followed by this preliminary analysis, we further evaluate the proposed approach on other languages like German, Hindi, Bengali, and Amharic. The languages that we consider in this study have various amounts of resources and this would help us understand the versatility of the knowledge leveraged from the DT. The details of the used datasets are described below. Following Wu et al. [44], the English models are evaluated over the `dev` splits provided in the GLUE benchmark. For the other languages, the official `train/test` split has been used in our experiments. The statistics of the datasets used are presented in Table 1.

English (*en*) datasets:

- **SST-2**: The Stanford Sentiment Treebank [35] comprises sentences that have been collected from movie reviews and human annotations of their sentiment. This dataset presents the task of predicting the sentiment of a given sentence. Following [40], we use the two-way class (*positive/negative*) split and use only sentence-level labels.
- **STS-B**: The well-studied Semantic Textual Similarity Benchmark [6] presents sentence pairs that have been collected from news headlines, video and image captions, and natural language inference data, each with a human-annotated similarity score from 1 to 5. This dataset presents the regression task of estimating the human-annotated score.
- **MRPC**: The Microsoft Research Paraphrase Corpus [12] presents sentence pairs that have been automatically been extracted from online news sources along with human annotations which indicates whether the sentences in the pair are semantically equivalent.

German (*de*) datasets:

- **GermEval T1**: The GermEval Task 1 dataset [41] presents the coarse-grained binary classification task of deciding if the given tweet includes some form of offensive language.
- **GermEval T2**: The GermEval Task 2 dataset [41] presents the fine-grained multi-class classification task of categorizing the given tweet into ‘profanity’, ‘insult’, ‘abuse’, and ‘other’ based on the conveyed intent.

Hindi (*hi*) datasets:

- **Product Reviews**: The publicly available Product Sentiment Analysis dataset [1] presents the sentiment analysis task as a three-way classification challenge.
- **MIDAS Discourse**: The MIDAS Hindi Discourse Analysis dataset [11] presents the task of classifying sentences into one of the following discourse categories: ‘argumentative’, ‘descriptive’, ‘dialogic’, ‘informative’, and ‘narrative’.
- **WNLI**: The Winograd NLI dataset [26] consists of pairs of sentences wherein the second sentence is constructed from the first sentence by replacing an ambiguous pronoun with a possible referent within the sentence.
- **Amrita S1**: The Amrita Paraphrase Subtask 1 dataset [2] presents sentence pairs and the task is to classify them as Paraphrases (P) or Not Paraphrases (NP).

Bengali (*bn*) datasets:

- **BEmoC**: The Bengali Emotion Corpus [9] presents texts along with human annotations of one of the six basic emotions: ‘anger’, ‘fear’, ‘disgust’, ‘sadness’, and ‘surprise’.
- **Sentiment Analysis**: This dataset [22] presents the sentiment analysis task where the given input sentence has to be classified between two classes.

Amharic (*am*) datasets:

- **ASAB All**: The ASAB dataset [45] presents the sentiment analysis challenge on tweets. The dataset comprises tweets with their corresponding human-annotated sentiment label. In this variant of the dataset, the task is to classify the given tweet into ‘positive’, ‘negative’, ‘neutral’, and ‘mixed’.
- **ASAB Cleaned**: This dataset [45] presents the sentiment analysis task, like in ASAB All dataset, as a 3-way classification challenge into ‘positive’, ‘negative’, and ‘neutral’.

5 EXPERIMENTAL RESULTS AND ANALYSIS

The performed experiments intend to investigate the impact of the prior knowledge conveyed through the induced graphs from the DT. As discussed in Section 3, the proposed flexible framework allows us to encode the information from the DT using various GNNs. In our experiments, we investigate the performance of our framework using two different GNN algorithms namely GCN [24] and LEConv [33]. The operations utilized in the proposed scheme are differentiable and hence the complete model can be optimized in an end-to-end manner while further finetuning the pretrained transformer encoder. Tasks with regression objectives are trained using Mean Squared Error while tasks with classification objectives are trained using Cross-Entropy Loss.

As presented in Section 3, our framework leverages a GNN module over the output of a pretrained transformer model. The

Table 2: Results of proposed models on English datasets. For SST-2 and MRPC, we report accuracy while for STS-B, we report Pearson correlation. For all these measures we report mean and standard deviation over 5 independent runs.

Model	SST-2	STS-B	MRPC
BERT	91.01 ± 1.04	**89.00** ± 0.29	**87.16** ± 0.64
+ GCN	**91.54** ± 0.75	88.45 ± 0.37	86.57 ± 1.23
+ LEConv	91.15 ± 0.55	87.51 ± 0.18	86.57 ± 0.74
RoBERTa	90.48 ± 0.57	90.12 ± 0.28	89.31 ± 0.62
+ GCN	90.71 ± 0.71	**90.20** ± 0.32	**89.75** ± 0.44
+ LEConv	**90.99** ± 0.54	89.69 ± 0.31	88.82 ± 0.99

Table 3: Results of proposed models on German datasets. For both these tasks, we report macro F1 (mean and standard deviation over 5 independent runs).

Model	Germeval T1	Germeval T2
mBERT	72.17 ± 0.63	42.00 ± 0.61
+ GCN	71.95 ± 0.35	41.53 ± 0.75
+ LEConv	**72.59** ± 0.66	**42.92** ± 0.72
XLM-R	72.80 ± 3.82	44.08 ± 1.87
+ GCN	75.40 ± 0.53	45.30 ± 0.29
+ LEConv	**75.95** ± 0.85	**45.71** ± 0.27

GNN module comprises two GNN layers which further contextualizes the semantic information with the DT knowledge. After each graph convolution operation, a non-linear ReLU [30] activation is applied followed by a dropout. The entire model is trained in an end-to-end manner using the AdamW [28] optimizer. On the other hand, as done in prior works [8, 10, 21, 27], the baseline transformer-based representations are adapted to the downstream task of interest by finetuning them based on the sentence representation (`CLS` embedding). In our experiments on tasks in English, we use `bert-base-uncased` as BERT and `roberta-base` as RoBERTa. For the tasks in other languages, based on the availability of the models, we leverage `bert-base-multilingual-cased` and `xlm-roberta-base` as mBERT and XLM-R respectively. In all our experiments, we run the models across 5 seeds and report the mean performance of the independent runs.

We first validate our framework on tasks from the English GLUE benchmark. Following prior work [44], we evaluated the effectiveness of our framework by comparing it against the vanilla transformer representations obtained from BERT and RoBERTa. As discussed in Section 4, we follow [44] and only report development set results due to restricted access to the GLUE test set. From Table 2, it is clear the proposed framework presents comparable results to the baseline transformer models with improvements of up to 0.5% in the respective measure.

Based on the encouraging results noticed in English, we further evaluate our framework on German. The size of the German data used to pre-train mBERT is less than half of the English corpus size used to pre-train the same [43]. From Table 3 it is evident that the knowledge from the DT consistently assists in the performance of the model. On both the GermEval tasks, our framework gives absolute performance gains of up to 3.15% in macro F1 scores. Thereby, this validates that our framework provides a viable approach to efficiently enrich existing transformer embeddings.

Table 4: Results of proposed models on Bengali datasets. For both these tasks, we report accuracy (mean and standard deviation over 5 independent runs).

Model	BEmoC	Sentiment Analysis
mBERT	55.17 ± 2.44	**70.74** ± 1.38
+ GCN	**57.15** ± 1.54	70.37 ± 1.09
+ LEConv	56.86 ± 1.02	69.92 ± 0.78
XLM-R	65.86 ± 1.54	72.94 ± 2.49
+ GCN	**68.00** ± 1.06	**73.81** ± 1.53
+ LEConv	67.90 ± 0.86	71.24 ± 1.01

We further extend our experiments to popular Indian languages namely Hindi and Bengali. The corpus size used to pre-train the representations of these Indian languages accounts for less than 2% of data as compared to English. Recent work in the domain of contextual representations for Indian languages includes IndicBERT [21]. This presents an ALBERT model, which has been pretrained exclusively on Indian languages. Here [21] the authors attempt to boost the performance on the IndicGLUE tasks [21] by collecting additional data and by resorting to the expensive process of pre-training. On the other hand, our framework presents a viable alternative as it incorporates the semantic knowledge from the DT during the cheap and easy fine-tuning stage. Results presented in Table 5 show that our framework achieves state-of-the-art performance on the Hindi benchmark tasks with improvements of up to 2.33%. Similarly, we notice consistent improvements in Bengali tasks with improvements of up to 2.14%. The results for the Bengali language are presented in Table 4.

Next, we apply our framework to Amharic datasets. Unlike the other languages that we have analyzed, Amharic is covered only in XLM-R model as it is extremely low on resources. Despite the challenge in scarcity of resources, Table 6 shows that our proposed framework can give consistent improvements in the downstream task of sentiment analysis. We observe an absolute improvement of up to 4.54% in 'ASAB All' task and over 6% in 'ASAB Cleaned' task.

The exhaustive experiments performed by us confirm that our framework gives consistent improvements across typologically diverse languages. Furthermore, the performance gains noticed from our efficient framework relatively increase as we move from high-resource to low-resource languages. As DTs can be built from a corpus without any human intervention or expertise, we now have an alternative to enrich the transformer models instead of using manually built external sources.

6 CONCLUSION

In this work, we have proposed a method to utilize Distributional Thesaurus information through Graph Neural Networks to enrich the performance of transformer-based language models like BERT and RoBERTa. We show that the proposed approach produces comparable performance on a resource-rich language like English,

Table 5: Results of proposed models on Hindi datasets. For all these tasks, we report accuracy (mean and standard deviation over 5 independent runs).

Model	Product Reviews	WNLI	MIDAS Discourse	Amrita S1
mBERT*	74.57	56.34	71.20	93.81
XLM-R*	78.97	55.87	79.94	93.02
IndicBERT*	71.32	56.24	78.44	93.75
mBERT	**74.76** ± 0.38	56.34 ± 0.00	79.90 ± 0.58	92.51 ± 0.49
+ GCN	73.80 ± 0.52	56.34 ± 0.00	**79.98** ± 0.52	**92.76** ± 0.25
+ LEConv	74.38 ± 0.59	**57.18** ± 1.61	79.37 ± 0.39	92.47 ± 0.56
XLM-R	77.90 ± 5.71	56.34 ± 0.00	80.65 ± 0.57	92.71 ± 0.54
+ GCN	**81.03** ± 0.58	**56.90** ± 0.77	**81.07** ± 0.75	92.64 ± 0.31
+ LEConv	80.61 ± 0.77	55.77 ± 1.61	80.97 ± 0.66	**93.42** ± 0.25

* These results have been collected from the IndicGLUE benchmark [21].

Table 6: Results of proposed models on Amharic datasets. For both these tasks, we report accuracy (mean and standard deviation over 5 independent runs).

Model	ASAB All	ASAB Cleaned
XLM-R	48.71 ± 2.48	51.92 ± 0
+ GCN	50.91 ± 3.04	57.68 ± 3.33
+ LEConv	**53.25** ± 0.87	**57.93** ± 1.09

whereas it consistently improves the performance of mBERT and XLM-R by a significant margin for medium-resource languages like German and low-resource languages like Hindi, Bengali, and Amharic. This study shows a promising alternative direction to improve the pretrained transformer models' performance on low-resource languages without re-training the transformer on large corpora. The experiments performed confirm that consistent performance gains can be obtained by relying on existing publicly available DTs. In the future, we aim to extend this work to more languages and to experiment with other transformer-based models.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for suggestions on the submission. We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) foundation under the "JOIN-T 2" project and a Deutscher Akademischer Austauschdienst (DAAD) doctoral stipend.

REFERENCES

[1] Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A Hybrid Deep Learning Architecture for Sentiment Analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 482–493. https://aclanthology.org/C16-1047

[2] M Anand Kumar, Shivkaran Singh, B Kavirajan, and KP Soman. 2018. DPIL@ FIRE 2016: Overview of shared task on detecting paraphrases in Indian Languages (DPIL). In *CEUR Workshop Proceedings*. 233–238. http://ceur-ws.org/Vol-1737/T6-1.pdf

[3] Meriem Beloucif and Chris Biemanns. 2021. Probing Pre-trained Language Models for Semantic Attributes and their Values. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and in the Dominican Republic.

[4] Chris Biemann, Bonaventura Coppola, Michael R. Glass, Alfio Gliozzo, Matthew Hatem, and Martin Riedl. 2013. JoBimText Visualizer: A Graph-based Approach to Contextualizing Distributional Similarity. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 6–10. https://aclanthology.org/W13-5002

[5] Chris Biemann and Martin Riedl. 2013. Text: now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling* 1, 1 (2013), 55–95. https://doi.org/10.15398/jlm.v1i1.60

[6] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[7] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 276–286. https://doi.org/10.18653/v1/W19-4828

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[9] Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion Classification in a Resource Constrained Language Using Transformer-based Approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Online, 150–158. https://doi.org/10.18653/v1/2021.naacl-srw.19

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, Debanjan Mahata, Rakesh Gosangi, Haimin Zhang, Rajiv Ratn Shah, and Amanda Stent. 2020. An Annotated Dataset of Discourse Modes in Hindi Stories. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1191–1196. https://aclanthology.org/2020.lrec-1.149

[12] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://aclanthology.org/I05-5002

[13] John R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).

[14] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 759–765. http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

[15] Zellig S. Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162. https://doi.org/10.1080/00437956.1954.11659520

[16] John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 4129–4138. https://doi.org/10.18653/v1/N19-1419

[17] Angelina Ivanova, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger. 2012. Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies. In *Proceedings of the Sixth Linguistic Annotation Workshop*. Association for Computational Linguistics, Jeju, Republic of Korea, 2–11. https://www.aclweb.org/anthology/W12-3602

[18] Abhik Jana and Pawan Goyal. 2018. Can Network Embedding of Distributional Thesaurus Be Combined with Word Vectors for Better Representation?. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, USA, 463–473. https://doi.org/10.18653/v1/N18-1043

[19] Abhik Jana and Pawan Goyal. 2018. Network Features Based Co-hyponymy Detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1006

[20] Abhik Jana, Nikhil Reddy Varimalla, and Pawan Goyal. 2020. Using Distributional Thesaurus Embedding for Co-hyponymy Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5766–5771. https://www.aclweb.org/anthology/2020.lrec-1.707

[21] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4948–4961. https://doi.org/10.18653/v1/2020.findings-emnlp.445

[22] Md. Rezaul Karim, Bharathi Raja Chakravarthi, John P. McCrae, and Michael Cochez. 2020. Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. 390–399. https://doi.org/10.1109/DSAA49011.2020.00053

[23] Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology* 105 (2004), 116.

[24] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*. Toulon, France. https://openreview.net/forum?id=SJU4ayYgl

[25] Ayush Kumar, Sarah Kohail, Asif Ekbal, and Chris Biemann. 2015. IIT-TUDA: System for sentiment analysis in indian languages using lexical acquisition. In *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 684–693. https://doi.org/10.1007/978-3-319-26832-3_65

[26] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning (KR'12)*. AAAI Press, 552–561. https://dl.acm.org/doi/10.5555/3031843.3031909

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:cs.CL/1907.11692

[28] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. New Orleans, USA. https://openreview.net/forum?id=Bkg6RiCqY7

[29] Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 369–382. https://doi.org/10.1007/978-3-030-45439-5_25

[30] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. Omnipress, Madison, WI, USA, 807–814.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[32] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, USA, 701–710. https://doi.org/10.1145/2623330.2623732

[33] Ekagra Ranjan, Soumya Sanyal, and Partha Talukdar. 2020. ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 5470–5477. https://doi.org/10.1609/aaai.v34i04.5997

[34] Martin Riedl and Chris Biemann. 2013. Scaling to Large^3 Data: An Efficient and Effective Method to Compute Distributional Thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 884–890. https://www.aclweb.org/anthology/D13-1089

[35] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170

[36] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJzSgnRcKX

[37] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., California, USA. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[39] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Vancouver, Canada. https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[41] Michael Wiegand and Melanie Siegel. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval 2018 Workshop*. Austrian Academy of Sciences, 1–10. https://doi.org/10.1553/0x003a105d

[42] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. 2021. Graph Neural Networks for Natural Language Processing: A Survey. arXiv:cs.CL/2106.06090

[43] Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Online, 120–130. https://doi.org/10.18653/v1/2020.repl4nlp-1.16

[44] Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing Finetuning with Semantic Dependencies. *Transactions of the Association for Computational Linguistics* 9 (2021), 226–242. https://doi.org/10.1162/tacl_a_00363

[45] Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic Sentiment Analysis from Social Media Texts: Building Annotation Tools and Classification Models. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1048–1060. https://doi.org/10.18653/v1/2020.coling-main.91

[46] Mi Zhang and Tieyun Qian. 2020. Convolution over Hierarchical Syntactic and Lexical Graphs for Aspect Level Sentiment Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3540–3549. https://doi.org/10.18653/v1/2020.emnlp-main.286

[47] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 334–339. https://doi.org/10.18653/v1/2020.acl-main.31

[48] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2020)*. https://aaai.org/ojs/index.php/AAAI/article/view/6510

[49] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Santiago, Chile, 19–27. https://doi.org/10.1109/ICCV.2015.11