University of Amsterdam

Passing Parser Uncertainty to the Transformer

Dongqi Liu X Khalil Sima'an

ABSTRACT

Does passing both **parser uncertainty** and **labeled syntactic knowledge** to the Transformer improve its translation performance?

We contribute a novel method for infusing the whole **labeled dependency distributions (LDD)** of the source sentence's dependency forest into the Transformer. Experimental results demonstrate that our approach outperforms both the vanilla Transformer as well as the single best–parse Transformer model across several evaluation metrics.

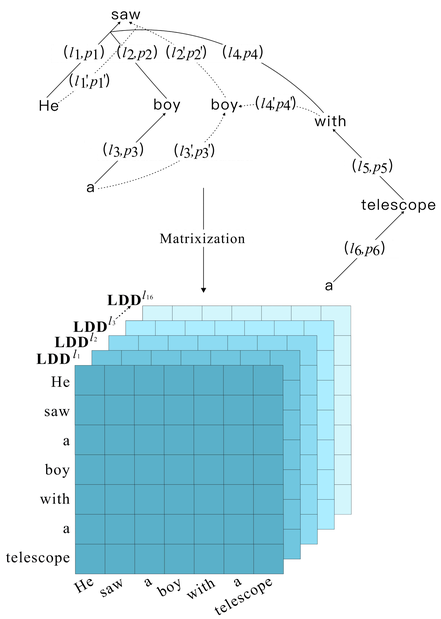

Figure 1: Labeled dependency distributions

CONTACT

Dongqi Liu
Email: dongqi.me@gmail.com
Homepage: https://dongqi.me/

INTRODUCTION

There are three note–worthy gaps in the literature addressing source syntax:

- **None** of the existing works conditions on the probability distributions over source syntactic relations.
- **None** of the existing approaches conditions on the dependency labels, thereby conditioning only on the binary choice whether there is an unlabeled dependency relation.
- **Other methods** (data manipulation, linearization, or embeddings) can be explained as a mere regularization effect of the model, which does not help the Transformer to exploit the actual syntactic knowledge.

METHOD

Dependency Distributions

Our primary idea is to exert a soft influence by dependency distribution on the self–attention in the encoder of the Transformer to allow it to fit its parameters with both syntax and translation awareness together.

- A dependency distribution in the form of conditional probabilities, which could be taken to represent the degree of confidence of the parser in the individual dependency relations.
- Each dependency relation type (label), provides a more granular local probability distribution that could assist the Transformer model in making a more accurate estimation of the context vector.

Parser–Infused Self–attention

We propose a novel Transformer NMT model that incorporates the LDD into the first layer of the encoder side:

METHOD

1. We infuse the resulting self–attention weight matrix $\mathbf{S}^{h_i}$ for head h_i with the specific $\mathbf{LDD}$ matrix $\mathbf{LDD}^{l_i}$ for label l_i using element–wise multiplication. Assuming that $d^{l_i}_{p,q} \in \mathbf{LDD}^{l_i}$, this is to say:

$$n^{hi}_{p,q} = s^{hi}_{p,q} \times d^{li}_{p,q}, \text{ for } p,q = 1,\ldots,\mathrm{T}$$

2. Next, the resulting weights are softmaxed to obtain the final syntax–infused distribution matrix for head h_i and the label attached to this head l_i :

$$\boldsymbol{N}^{hi} = \mathrm{softmax}(\boldsymbol{S}^{hi} \odot \mathbf{LDD}^{li})$$

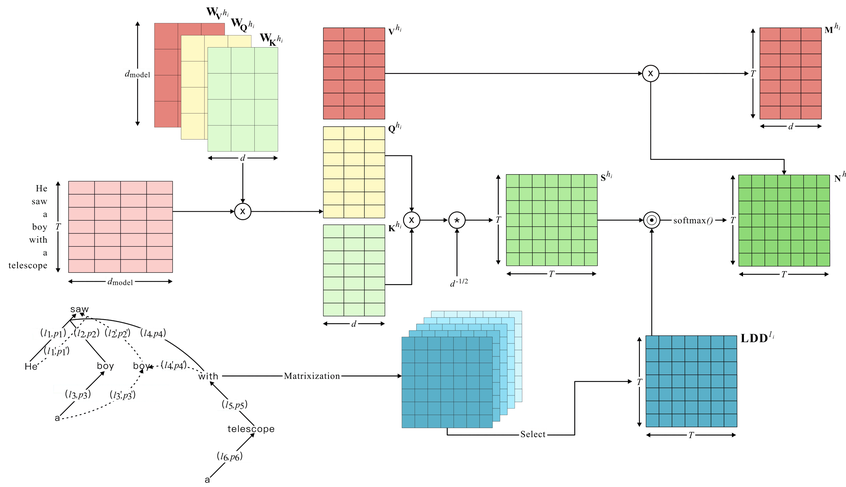

Figure2: Labeled dependency distribution sub–layer ($\mathbf{LDD}^{l_i}$ for head h_i)

RESULTS

- As compared to the baseline model, either form of modeling the syntactic knowledge of the source language could be beneficial to the NMT models. Whether it was in the choice of lexical (BLEU–1) or in the order of word (RIBES).
- The proposed model achieved the best score in at least three of the five different evaluation metrics, regardless of the language translation tasks, and consistently reached the highest results on BLEU–4.
- In most translation experiments, incorporating labeled dependency distributions provided better outcomes than the 1–best unlabeled dependency parse system.
- Simply incorporating LDD (replacing the K and Q matrices in the attention matrices) as dependency attention outperformed the baseline model on average, which can drastically decrease the number of parameters and computing requirements.

ANALYSIS

- The model we propose has higher scores than the baseline model and the 1–best parse model in the BLEU–4 score distribution.
- When the sentence length exceeds 50, the BLEU–4 scores of our method remained significantly different from both the baseline model and the 1–best parse model.
- The proposed model is better than the baseline model and 1–best parse model in terms of attention alignment, which demonstrates that the syntactic knowledge contained in LDD can guide the weight computation of the attention mechanism to pay more attention to words with syntactic relations.

CONCLUSION

We presented a novel supervised conditional labeled dependency distributions Transformer network (LDD–Seq):

- Our method primarily improves the self–attention mechanism in the Transformer model by converting the dependency forest to conditional probability distributions.
- Each self–attention head in the Transformer learns a dependency relation distribution, allowing the model to learn the source language's dependency constraints, and generates attention weights that are more in line with the syntactic structures.
- The method could improve the Transformer's translation performance without increasing the complexity of the network or interfering with the highly parallelized characteristic of the model.