LREC-COLING '24

Leveraging Linguistically Enhanced Embeddings for Open Information Extraction

Authors

Fauzan Farooqui, Thanmay Jayakumar, Pulkit Mathur, Mansi Radke

Department of Computer Science and Engineering, **Visvesvaraya National Institute of Technology**, India

Open Information Extraction

OIE extracts information present in a sentence in the form of tuples, mainly triples.

Example:

Input: “The cat sat on the mat.”

Output: (cat;sat;mat)

OIE finds use in open-domain, large-scale IE needs. OIE triples naturally form a knowledge graph, which after entity disambiguation and similar tools, can become a valuable resource for reasoning over data in downstream tasks like link prediction.

Neural-based Approaches to OIE

- Discriminative - Tagging-based (Encoder) Models
- Generative - Seq2Seq-based (Encoder-Decoder) Models

We use the generative approach as it easily allows extracting multiple predicates with multiple triples from a single sentence.

TANL

Until recently, each structured prediction (SP) task was studied separately. TANL (Paolini et. al., 2021 - AWS) was the first paper to train a model on all such models.

- This can be interpreted as a “multilingual” model that has a unique output format for each task.
- Instead of training task-specific discriminative classifiers, it’s treated as a translation task between augmented natural languages, from which the task-relevant information is easily extracted.
- Uses the same architecture (T5) and hyperparameters for all tasks and even when training a single model to solve all tasks at the same time (multi-task learning).
- **Primary result:** They observe similar or better scores for the tasks, concluding that a general Seq2Seq SP model is advantageous.
- However, they **didn’t work on OIE.** Following this work, we use T5 and extend TANL to OIE, contributing the first known study on how SP pre-training affects OIE.

Linguistic Features in OIE

- Past work has looked towards including linguistic features such as PoS and SynDP for the OIE task.
- RnnOIE (Stanovsky et al., 2018), SenseOIE (Roy et al., 2019) and SpanOIE (Zhan and Zhao, 2020) explore concatenation of such feature embeddings along with source word embeddings. The findings from these papers suggest that including information that is directly relevant to the source word could be enough, and a better way to include so should be explored.
- In the direction closer to our work, (Mtumbuka and Lukasiewicz, 2022) echo the results of past work on the benefits of including linguistic information. However, this work doesn't investigate how the choice of tag embedding size affects the model. More importantly, the embeddings have been taken from PLMs but they have not used the full Seq2Seq architecture. We believe to be the first to incorporate linguistic features while using a PLM for the task.
- We also explore including a linguistic tag that has not been studied before for OIE: Semantic Dependency Parsing (SemDP), which significantly reduces computing overheads compared to its counterparts.

Token	Part-of-Speech (PoS)	Syntactic Dependency Parse (SynDP)	Semantic Dependency Parse (SemDP)
info_extract	<pad>	<pad>	<pad>
:	<pad>	<pad>	<pad>
The	DT (determiner)	study: det (determiner)	_
study	NN (common noun)	published: nsubj:pass (passive nominal subject)	The: BV | published: ARG2 | Change: ARG1
was	VBD (past tense verb)	published: aux:pass (passive auxiliary)	_
published	VBN (past participle verb)	0: root	0: root | in: ARG1 | yesterday: loc
in	IN (preposition)	journal: case (case marking)	_
journal	JJ (adjective)	published: obl (oblique nominal)	_
Nature	NNP (singular proper noun)	Change: compound	_
Climate	NNP	Change: compound	_
Change	NNP	journal: appos (appositional modifier)	0: root | in: ARG2 | Climate: compound | yesterday: loc
yesterday	NN	published: obl:tmod (temporal modifier)	_
.	. (sentence terminator)	published: punct (punctuation)	_
</s> (EOS token)	</s>	</s>	</s>

Table 10: Example of linguistic tags for the sentence *"The study was published in journal Nature Climate Change yesterday."* (Token indices were replaced by the actual token in the SynDP and SemDP tags)

Literature Review - DeepStruct

DeepStruct performs task-agnostic structural pre-training by formatting various structured prediction tasks, including OIE, as triples, treating the tuple format as the structure itself.

During their structural pre-training, they train on a subset of the large OPIEC dataset. MinIE shifts information from ClausIE's extractions to tuple annotations. Though meta-data rich, its special SpaTe format cannot be directly used by existing OIE approaches.

Hence, whether DeepStruct still achieves comparable performance on recent OIE benchmarks like CaRB, remains unclear as they have not given instructions for OIE-usage of their currently available model.

However, they rely on structure from the form of the input.

Our key idea: Featured Embedding Enhancements with PLMs

- PLMs can greatly boost a task performance, but how do we enhance them while taking advantage of linguistic structure?
- We contribute **two novel word embedding enhancement techniques**:
  - **Weighted Addition (WA)**
  - **Linearized Concatenation (LC)**
- To the best of our knowledge, we are the first to apply such embedding enhancement using PLMs (specifically **T5**) for the task of OIE.
- Our other novel contributions will be explained as we move along, and will be summarized by the end.

Embedding Enhancement using WA and LC

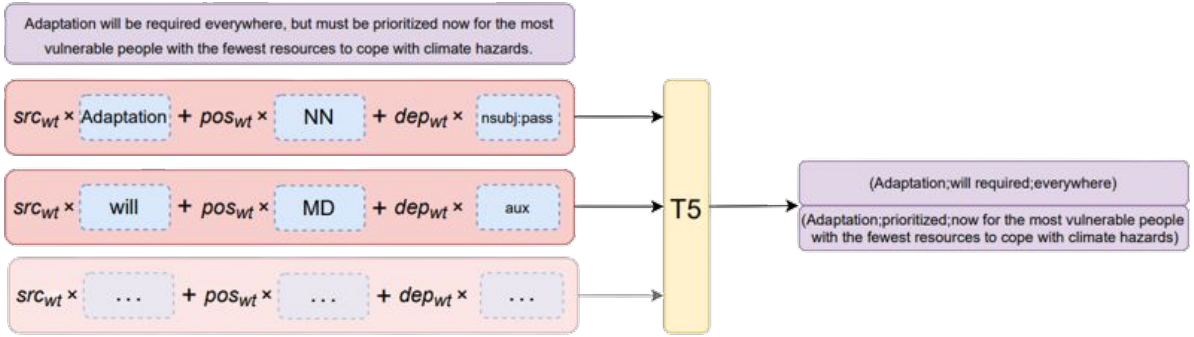

Figure 1: Structure Embedding Addition (Source sentence from the United Nation's website)

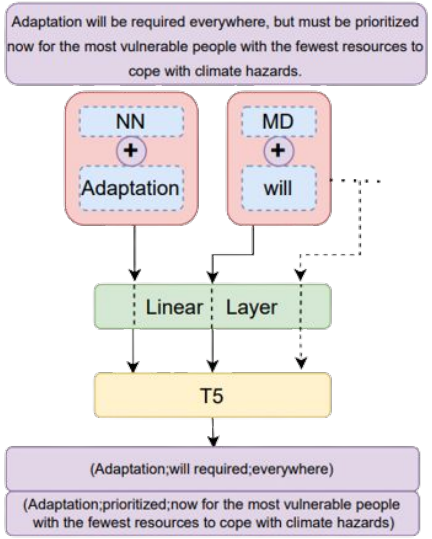

Figure 2: Structure Embedding Concatenation (Source sentence from the United Nation's website)

Dataset and Preprocessing

- We convert a processed version of LSOIE (a large OIE dataset) from tag-based extractions to a Seq2Seq format.
- However, LSOIE is also quite noisy. To remedy this, we create a synthetic dataset by using ClausIE's extractions for the LSOIE inputs, rather than the original LSOIE extractions. This gives an almost instant remedy to the low scores seen from the original dataset.
- Besides, as noted earlier, we extend the TANL format to OIE.
- A subjective analysis shows that our model's output is often better than LSOIE's test extractions themselves.

Dataset and Preprocessing

LSoIE Sentence	LSoIE Labels	Generated Labels
Akerson will also relinquish his chairman role, to be replaced by current director Theodore Solso.	(Akerson;will relinquish;his chairman role)	(Akerson;will relinquish;his chairman role) (current director Theodore Solso;will replaced;Akerson)
Road accidents killed 8,600 on the nation's roads last year.	(on the nation's roads last year;killed;8,600)	(Road accidents;killed;8,600 on the nation's roads last year)
He said the world and the Paralympic movement is aware of the situation in the Ukraine, but the IPC needs to stay true to its mission.	(the IPC;needs;to stay true to its mission) (the IPC;stay;true to its mission)	(the IPC;should stay;true to its mission)

Table 1: Examples where LSoIE data is not clean

Experiments

- We created a baseline for each model that trains without any linguistic feature.
- On our main models, the feature tag embeddings are learned during training.
- For Weighted Addition, we assign a fractional weight to the input word embeddings and linguistic features. We experiment with various weights.
- For concatenation, we fix a embedding size for the linguistic features. We experiment with various such sizes.
- Training is performed on the Wiki split of the LSOIE dataset for domain independence.
- We evaluate on the CaRB benchmark (Bhardwaj et al. 2019), a crowdsourced benchmark for OIE.
- **Main result:** Linguistic features significantly enhance performance!

Results

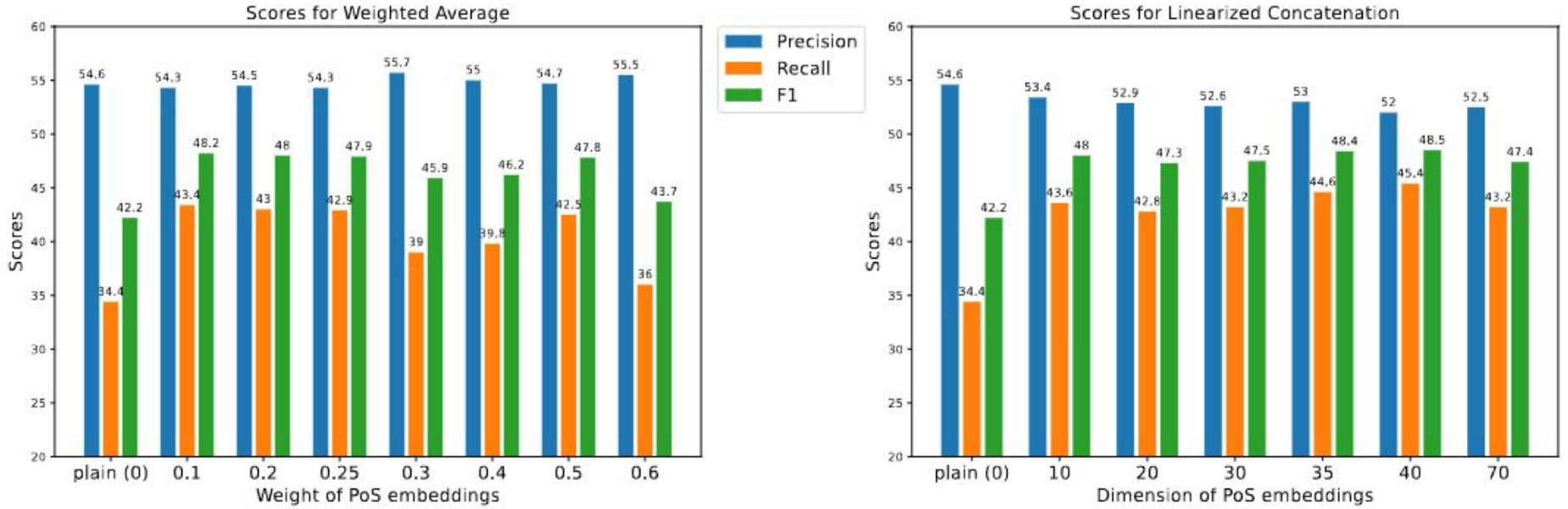

Figure 3: Scores for different wt_{pos} for Weighted Addition (left) and different PoS embedding dimensions for LC (right), trained on the ClausIE-extracted dataset.

Results

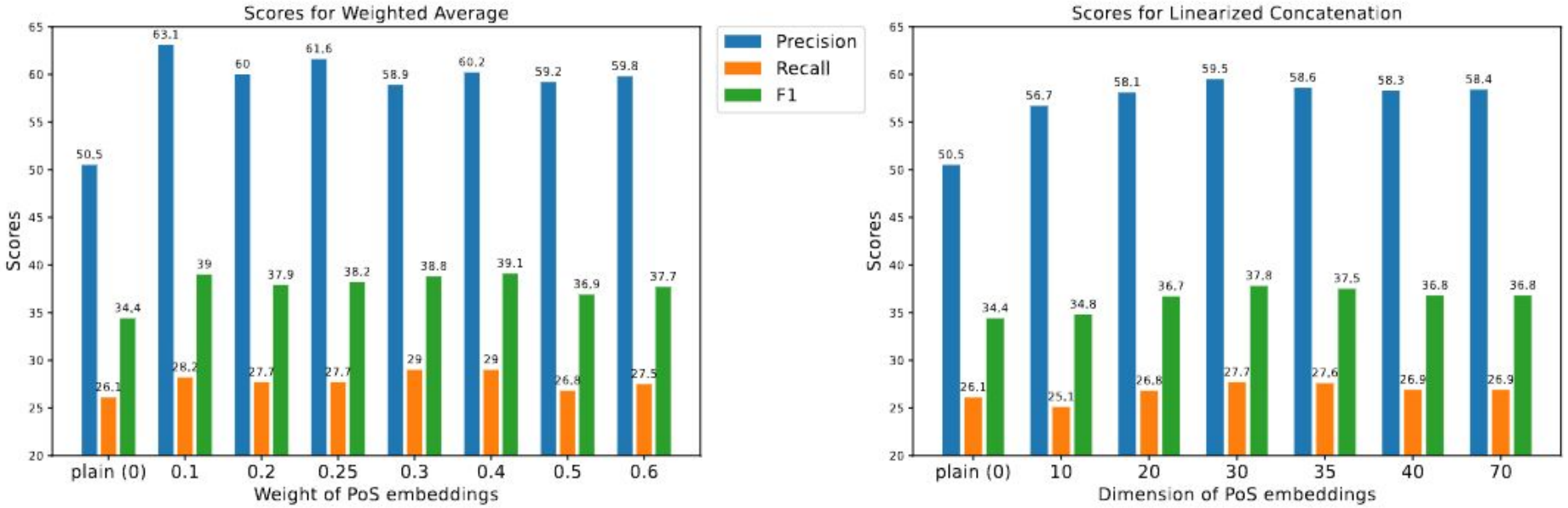

Figure 4: Scores for different wt_{pos} for Weighted Addition (left) and different PoS embedding dimensions for LC (right), trained on the LSOIE-extracted dataset.

Ablations

We conducted various ablation studies on different parameters:

- Fine-tuning TANL on TANL-formatted LSOIE marginally improves the F1 score, indicating that OIE is indeed a harder task than other SP tasks.
- Freezing the source embeddings slightly decreases F1 scores, indicating that a holistic learning of source with linguistic embeddings helps the model.
- We trained a model on the LSOIE Science dataset along with the Wiki dataset, and got similar results to the Wiki only model.
- We combined both SynDP and PoS tags and observed that scores slightly dip, indicating that the linguistic tags may interfere in each others' learning process.

Analysis

How do WA and LC perform with respect to each other? We note that both have advantages in differing settings. For example, WA performs beter for the LSOIE-extracted dataset, but LC performs better for its TANL-format. For the ClausIE-extracted dataset, there is little difference in performance boost.

Choice of dataset: We observe a general trend that if a jump in one parameter is high, the other parameter's increase isn't as significant. Recall in the original dataset is very limited. Both our datasets (ClausIE-extracted and TANL-format) give an immediate solution with the large jump in recall scores. Thus, we attribute the low recall scores to the unclean LSOIE extractions. However, that at least one parameter does indeed improve hugely, given the contrasting natures of the datasets, establishes the usefulness of both our methods on any type of dataset.

Analysis

Which linguistic features better help the model? There seems to be no clear winner among the three features. Each perform better than the other in different settings, showing the utility of all tags.

How does the new SemDP feature fare? SemDP shows as much improvement as other features, cementing its usefulness. Because SemDP uses the smallest tagset - with just about three tags being most frequently used - to gain equivalent performance, it could mean lesser training time and lower energy cost. Thus, we consider SemDP to be the best amongst the features.

Limitations & Future Directions

- We have covered an exhaustive set of experiments for one type of OIE model. As our work is model-agnostic, future work can look towards replicating our experiments in other OIE models. This indeed becomes a large task as each model comes with their own unique codebases to work with.
- A further study of available linguistic annotating tools and how they affect performance needs to be carried out in the context of OIE.
- Our TANL format depends on our tagging strategy of predicates. Future work can focus on improving this strategy to increase precision.
- The choice of a PLM can be studied on how it affects OIE.
- Our SemDP and SynDP information only considers the linguistic tag. To further test previous works' conclusions, it would be useful to experiment with combining the embeddings of the head word or tag too.

Summary – Contribution 1

To demonstrate the usefulness of linguistic structure in boosting performance for the OIE task, we propose two distinct novel word embedding enhancement techniques - Weighted Addition and Linearized Concatenation - that highly increase performance. We are thus the first to successfully integrate features with a PLM (T5) in OIE, while also being the first to exploit Seq2Seq PLMs for the generative approach to OIE.

We believe this to be an important direction in the field, as this can give any neural OIE architecture the power of both PLMs and linguistic tags in one go.

Contribution 2

We empirically study the effects of using three important word-level linguistic information from the sentence alone: PoS, Syntactic DP (SynDP) and Semantic DP (SemDP) tags. We are the first to exploit SemDP tags, which is also the strongest contender among single linguistic features. They reduce computing overheads by using 72% less tags compared to its SynDP counterpart, while maintaining the same performance boost. We thus believe SemDP to be a crucial novel step for incorporating useful, scalable linguistic features.

Contribution 3

We contribute a synthetic dataset (built from ClausIE) that boosted performance by 73.7% and 37.9% on Recall and F1 scores over the Seq2Seq version of the best-existing dataset (LSOIE), the latter which we show to be largely unclean and flawed. We believe researchers in the field will find this to be an integral resource, which includes extracted linguistic tags and processed LSOIE outputs too.

Contribution 4

We are the first to study how a model trained on all other SP tasks, TANL, affects OIE performance, contributing novel insights along the wider SP research direction.