Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction

Christoph Alt, Marc Hübner, Leonhard Hennig

https://github.com/DFKI-NLP/DISTRE

{firstname.lastname}@dfki.de

Summary

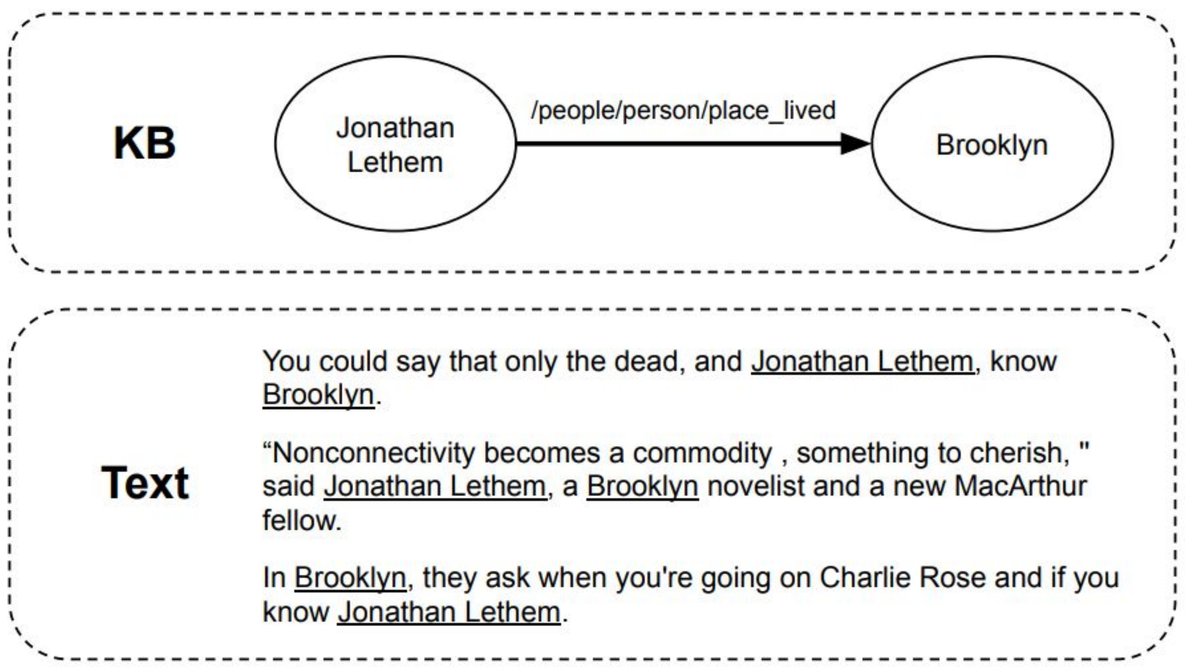
We present a method that combines **language model pre-training** and **fine-tuning** with **distantly supervised relation extraction**, where labels are noisy, to extract a more diverse set of relational facts from text. We utilize a pre-trained language model providing supporting linguistic and contextual information to more efficiently guide the relation classification, which we show to be important for recognizing a more diverse set of relations. By **extending the language model to the distantly supervised setting**, and fine-tuning it on the **NYT10 dataset**, we show that it predicts a larger set of distinct relation types with high confidence. **Manual and automated evaluation** of our model shows that it achieves a state-of-the-art AUC score of 0.422 on the NYT10 dataset, and **performs especially well at higher recall levels**.

Motivation

- Current distantly supervised RE methods are often biased towards recognizing a limited set of relations, while ignoring those in the long tail
- We improve recall for long tail relations by combining:
  - fine-tuning of pre-training language models
  - and bag-level multi-instance learning

The Task (Relation Extraction with Distant Supervision)

The following are typical inputs to a distantly supervised relation extraction system:

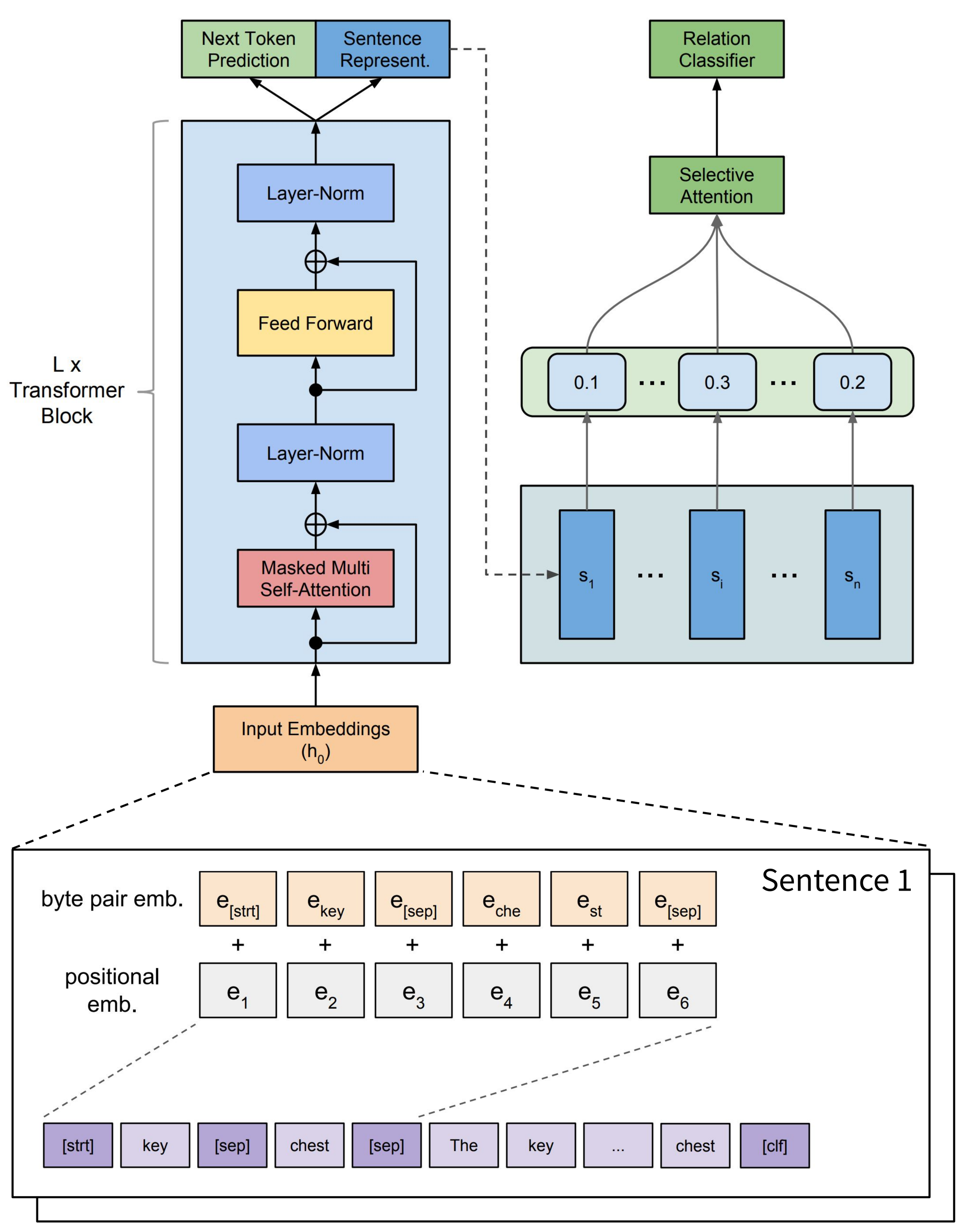


Relation instances for distant supervision, e.g. available in a **knowledge base**

Annotate text corpus with labels assigned by distant supervision heuristic to generate **noisy labels**

Method

Fine-tune Transformer Language Model (OpenAI GPT [Radford et al., 2018]) on a task-specific input format using multi-task learning with auxiliary loss as regularizer.



Classify the relation(s) between a pair of entities, given the combined representation of a bag

Select a combined bag-level representation, given the distinct representations of each sentence in a bag. E.g. via selective attention.

Represent each sentence in a bag by encoding it with a pre-trained transformer language model.

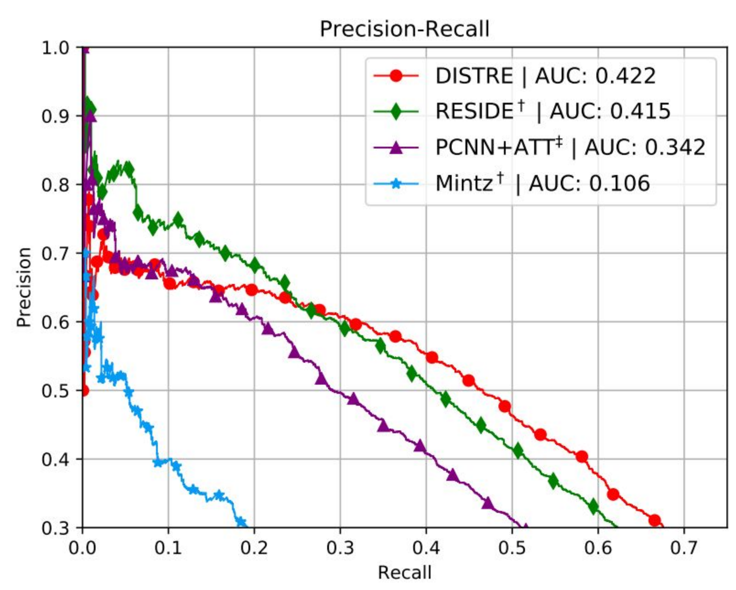
Group all sentences mentioning an entity pair, e.g. (Jonathan Lethem, Brooklyn) into a **bag**.

Results

We evaluate on the **NYT10 dataset** [Riedel et al., 2010]. The dataset contains **53 relation types**, including NA if no relation holds for a given sentence and entity pair. We report **Precision@N** (precision scores for the top 100, top 200, and top 300 extracted relation instances). In addition, three human annotators manually rated the top 300 predicted relation instances for each model.

Held-Out Evaluation

System	AUC	P@100	P@200	P@300	P@500	P@1000	P@2000
Mintz†	0.107	52.3	50.2	45.0	39.7	33.6	23.4
PCNN+ATT‡	0.341	73.0	68.0	67.3	63.6	53.3	40.0
RESIDE†	0.415	**81.8**	**75.4**	**74.3**	**69.7**	59.3	45.0
DISTRE	**0.422**	68.0	67.0	65.3	65.0	**60.2**	**47.9**



Precision-Recall Curve on the NYT dataset

- DISTRE with selective attention achieves a new state-of-the-art AUC value of 0.422
- Our method outperforms RESIDE and PCNN+ATT at higher recall levels, while precision is lower for top predicted relation instances
- Our method yields a more balanced overall performance without using any additional linguistic features, such as paraphrases, relation aliases or entity types

Manual Evaluation

System	P@100	P@200	P@300	Avg Prec
PCNN+ATT	97.3	94.7	90.8	94.3
RESIDE	91.3	91.2	91.0	91.2
DISTRE	88.0	89.8	89.2	89.0

Precision evaluated manually for the top 300 relation instances, averaged across 3 human annotators

relation	DIS	RES	PCNN
location/contains	168	182	214
person/nationality	32	65	59
person/company	31	26	19
person/place_lived	22	–	–
country/capital	17	–	–
admin_div/country	13	12	6
neighborhood/nbhd_of	10	3	2
location/town	3	–	–
company/founders	2	6	–
team/location	2	–	–
person/children	–	6	–

Distribution over the top 300 predicted relations for each method

- DISTRE achieves performance comparable to RESIDE, while predicting a more diverse set of relations with high confidence
- Most errors among the top predictions arise from wrongly labeled *"/location/country/capital"* instances, which the other models do not predict among the top 300 relations
- PCNN's high-confidence predictions are biased towards a set of only four relation types with a strong focus on: *"/location/contains"* and *"/person/nationality"*
- PCNN ranks short and simple patterns higher than more complex patterns where the distance between the arguments is larger

Future Directions

- Distantly supervised relation extraction with document level information
- Integrate external background and commonsense knowledge

German Research Center for Artificial Intelligence

DAY STREAM

BERLIN BIG DATA CENTER

Sponsored by:

Federal Ministry of Transport and Digital Infrastructure

Federal Ministry of Education and Research

Supported through projects DEEPLEE (01IW17001) , BBDC2 (01IS18025E), and DAYSTREAM (19F2031A)