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

Christoph Alt **Marc Hübner** **Leonhard Hennig**

German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab
{christoph.alt, marc.huebner, leonhard.hennig}@dfki.de

Task

- Distantly supervised relation extraction is widely used to extract relational facts from text, but suffers from noisy labels.
- Pretrain + distance supervised
- By extending the GPT to the distantly supervised setting, and fine-tuning it on the NYT10 dataset,

GPT transformer-Decoder

$$h_0 = TW_e + W_p$$
$$h_l = tf\_block(h_{l-1}) \; \forall \; l \; \in \; [1, L] \quad (1)$$

Where T is a matrix of one-hot row vectors of the token indices in the sentence, W_e is the token embedding matrix, W_p is the positional embedding matrix, L is the number of Transformer blocks, and h_l is the state at layer l. Since the Trans-

2.2 Unsupervised Pre-training of Language Representations

Given a corpus $\mathcal{C} = \{c_1, \ldots, c_n\}$ of tokens c_i, the language modeling objective maximizes the likelihood

$$L_1(\mathcal{C}) = \sum_i \log P(c_i | c_{i-1}, \ldots, c_{i-k}; \theta), \quad (2)$$

where k is the context window considered for predicting the next token c_i via the conditional probability P. The distribution over the target tokens is modeled using the previously defined Transformer model as follows:

$$P(c) = softmax(h_L W_e^T), \quad (3)$$

where h_L is the sequence of states after the final layer L, W_e is the embedding matrix, and θ are the model parameters that are optimized by stochastic gradient descent. This results in a probability distribution for each token in the input sequence.

Model

1. Enabling bag-level multi-instance learning on distantly supervised datasets
2. A description of our task-specific input representation for relation extraction

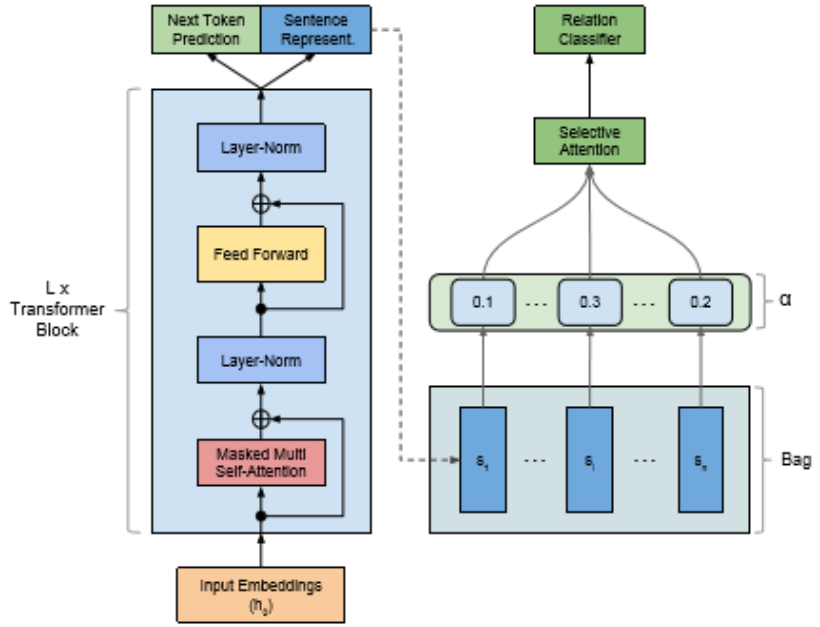

Figure 2: Transformer-Block architecture and training objectives. A Transformer-Block is applied at each of the L layers to produce states h_1 to h_L. After encoding each sentence in a bag into its representation s_i, selective attention informs the relation classifier with a representation aggregated over all sentences $[s_1, \ldots, s_n]$.

Distantly Supervised Fine-tuning on Relation Extraction

已知dataset:

$\mathcal{D} = \{(x_i, head_i, tail_i, r_i)\}_{i=1}^{N}$, where each example consists of an input sequence of tokens $x_i = [x^1, \ldots, x^m]$, the positions $head_i$ and $tail_i$ of the relation's head and tail entity in the sequence of tokens, and the corresponding relation label r_i, assigned by distant supervision. Due to

---

Label is unreliable -> bag level

is applied on a bag level, representing each entity pair $(head, tail)$ as a set $S = \{x_1, \ldots, x_n\}$ consisting of all sentences that contain the entity pair. A set representation s is then derived as a weighted sum over the individual sentence representations:

$$s = \sum_i \alpha_i s_i, \qquad (4)$$

Distantly Supervised Fine-tuning on Relation Extraction

noise. The weight α_i is obtained for each sentence by comparing its representation against a learned relation representation r:

$$\alpha_i = \frac{exp(s_i r)}{\sum_{j=1}^{n} exp(s_j r)} \quad (5)$$

To compute the output distribution $P(l)$ over relation labels, a linear layer followed by a softmax is applied to s:

$$P(l|S, \theta) = softmax(W_r s + b), \quad (6)$$

where W_r is the representation matrix of relations r and $b \in R^{d_r}$ is a bias vector. During fine-tuning we want to optimize the following objective:

$$L_2(\mathcal{D}) = \sum_{i=1}^{|S|} \log P(l_i|S_i, \theta) \quad (7)$$

According to Radford et al. (2018), introducing language modeling as an auxiliary objective during fine-tuning improves generalization and leads to faster convergence. Therefore, our final objective combines Eq. 2 and Eq. 7:

$$L(\mathcal{D}) = \lambda * L_1(\mathcal{D}) + L_2(\mathcal{D}), \quad (8)$$

where the scalar value λ is the weight of the language model objective during fine-tuning.

L1是语言模型loss

Input Representation

1. **BPE Encoding**
2. **Entity + sentence**

The classification token signals the model to generate a sentence representation for relation classification.

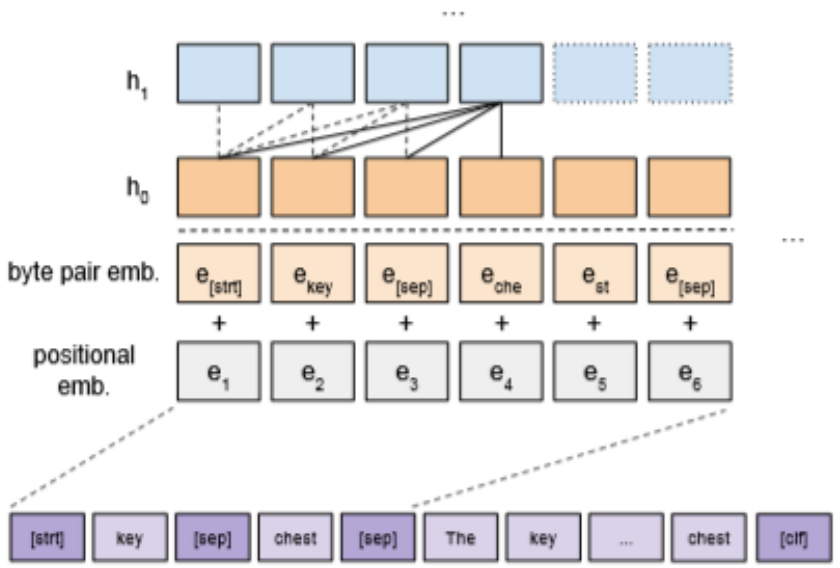

Figure 3: Relation extraction requires a structured input for fine-tuning, with special delimiters to assign different meanings to parts of the input. The input embedding h_0 is created by summing over the positional embedding and the byte pair embedding for each token. States h_l are obtained by self-attending over the states of the previous layer h_{l-1}.

Experiment

- Dataset：NYT10Dataset
- a standard benchmark for distantly supervised relation extraction. It was generated by aligning Freebase relations with the New York Times corpus, with the years 2005–2006 reserved for training and 2007 for testing.
- The training data contains 522,611 sentences, 281,270 entity pairs and 18,252 relational facts. The test data contains 172,448 sentences, 96,678 entity pairs and 1,950 relational facts. There are 53 relation types, including NA if no relation holds for a given sentence and entity pair.

Baseline

- The piecewise convolutional neural network (PCNN) segments each input sentence into parts to the left, middle, and right of the entity pair, followed by convolutional encoding and selective attention to inform the relation classifier with a bag level representation.
- RESIDE, on the other hand, uses a bidirectional gated recurrent unit (GRU) to encode the input sentence, followed by a graph convolutional neural network(GCN) to encode the explicitly provided dependency parse tree information. This is then combined with named entity type information to obtain a sentence representation that can be aggregated via selective attention and forwarded to the relation classifier

Experiment

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**

Table 1: Precision evaluated automatically for the top rated relation instances. † marks results reported in the original paper. ‡ marks our results using the OpenNRE implementation.

Experiment

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/team	3	–	–
company/founders	2	6	–
team/location	2	–	–
person/children	–	6	–

Table 3: 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. PCNN+ATT shows a strong focus on two relations: */location/location/contains* and */people/person/nationality*.

shows that for these relations, the PCNN+ATT model picks up on entity type signals and basic syntactic patterns, such as "LOC, LOC" (e.g., "Berlin, Germany") and "LOC in LOC" ("Green Mountain College in Vermont") for */location/location/contains*, and "PER of LOC" ("Stephen Harper of Canada") for */people/person/nationality*. This suggests that the

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

Table 2: Precision evaluated manually for the top 300 relation instances, averaged across 3 human annotators.

Experiment

Sentence	Relation
Mr. Snow asked, referring to Ayatollah **Ali Khamenei**, **Iran**'s supreme leader, and Mahmoud Ahmadinejad, **Iran**'s president.	/people/person/nationality
In **Oklahoma**, the Democratic governor, **Brad Henry**, vetoed legislation Wednesday that would ban state facilities and workers from performing abortions except to save the life of the pregnant woman.	/people/person/place_lived
Jakarta also boasts of having one of the oldest golf courses in **Asia**, Rawamangun , also known as the Jakarta Golf Club.	/location/location/contains
Cities like New York grow in their unbuilding: demolition tends to precede development, most urgently and particularly in **Lower Manhattan**, where **New York City** began.	/location/location/contains

Table 4: Examples of challenging relation mentions. These examples benefit from the ability to capture more complex features. Relation arguments are marked in bold.