Combining Distant and Direct Supervision for Neural Relation Extraction
远程监督与直接监督相结合的神经关系提取

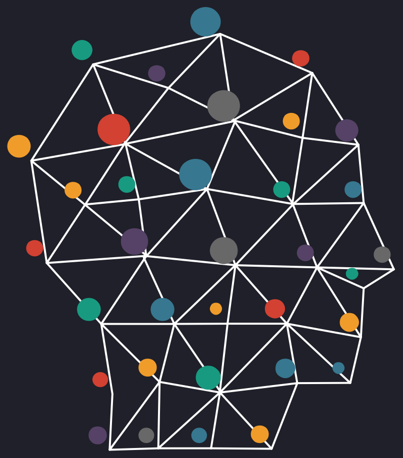
NAACL 2019

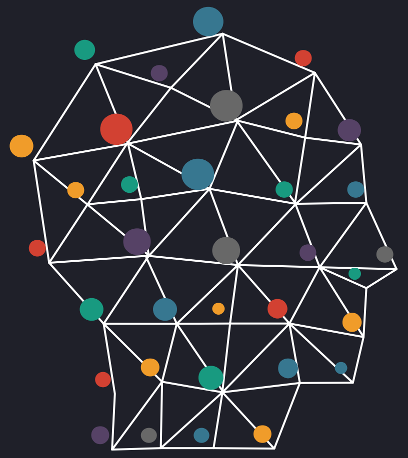
王 伟 51194501180

East China Normal University

School of Computer Science and Technology

CONTENTS





1. Introduction

1.1 Relation Extraction

Information Extraction:

Its main purpose is to transform unstructured or semi-structured natural language text into structured content, including extracting specified types **of entities, relations and events**.

Relation Extraction :

It is mainly responsible **for identifying entities from unstructured text** and **extracting semantic relations** between entities, which is widely used in information retrieval, question answering system and knowledge graph.

Traditional Method:

- **As a supervised multi classification problem.**
- **Shortcoming:** need a large number of high-quality labeled training data. So the **Distant Supervision** method is proposed.

1.2 Distant Supervision

What is Distant Supervision?

Assumption: if there is a certain relation between two entities in the knowledge base, then all unstructured sentences containing these two entities in a specific corpus can represent this relation.

Advantages: Uses the knowledge in the existing knowledge base to label the corpus, which effectively solves the problem of data annotation in relation extraction.

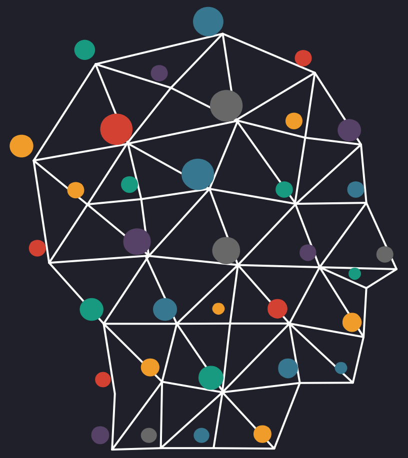
Disadvantages: Facing **noise problems.**

(**Alice**, Spouse, **Bob**)

- ☑ Soon, **[Alice]** got married with **[Bob]**.
- ☒ **[Bob]** and **[Alice]** are my primary school classmates.
- ☒ **[Bob]** is three years older than **[Alice]**.

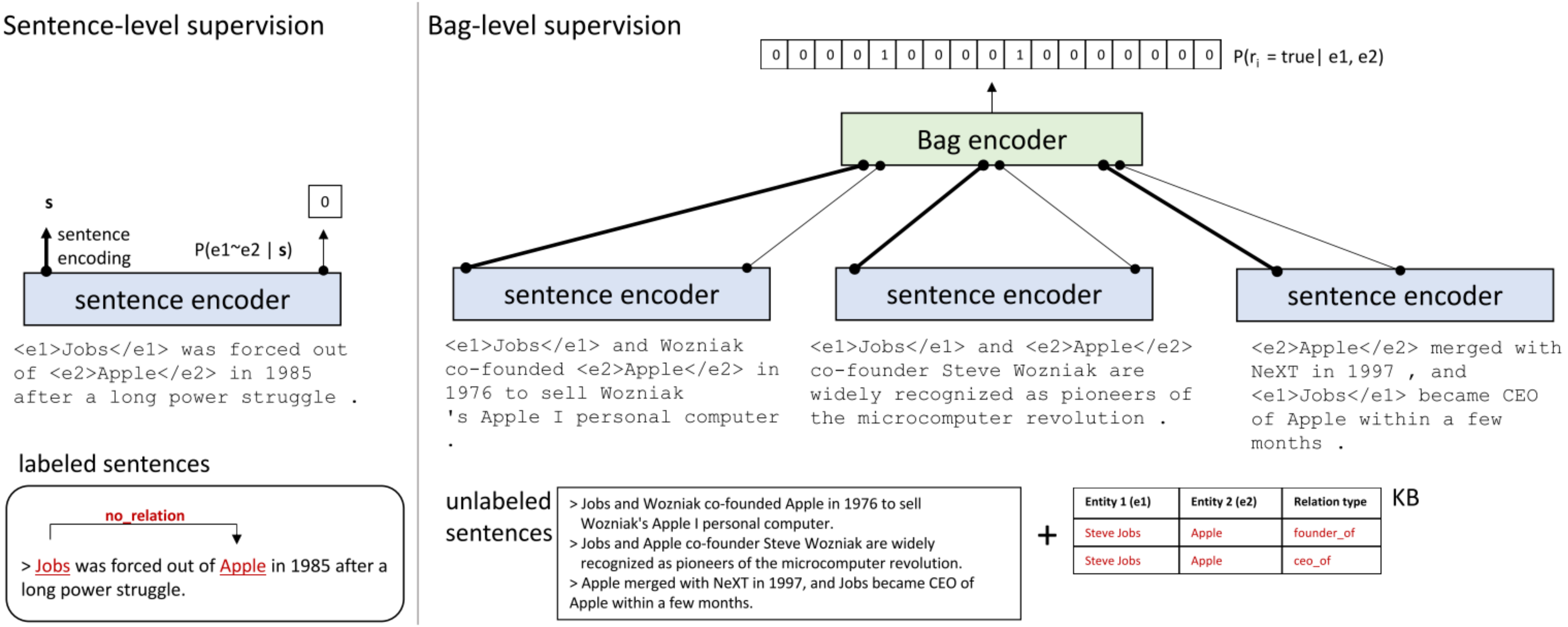
(**Barack Obama**, BornIn, **United States**)

- ☑ **[Barack Obama]** was born in the **[United States]**.
- ☒ **[Barack Obama]** is the 44th President of the **[United States]**.



2. Method

2.1 Overview



Goal:

predict relation between entities (e_1, e_2)

Knowledge base $\mathcal{K}$ **:** (e_1, r, e_2)

A bag of sentences: B_{e_1,e_2}

Annotate this bag with the set of relation types:

$$L^{\text{distant}} = \{r \in \mathcal{R} : (e_1, r, e_2) \in \mathcal{K}\}$$

2.2 Model

Bag Encoder Architecture：

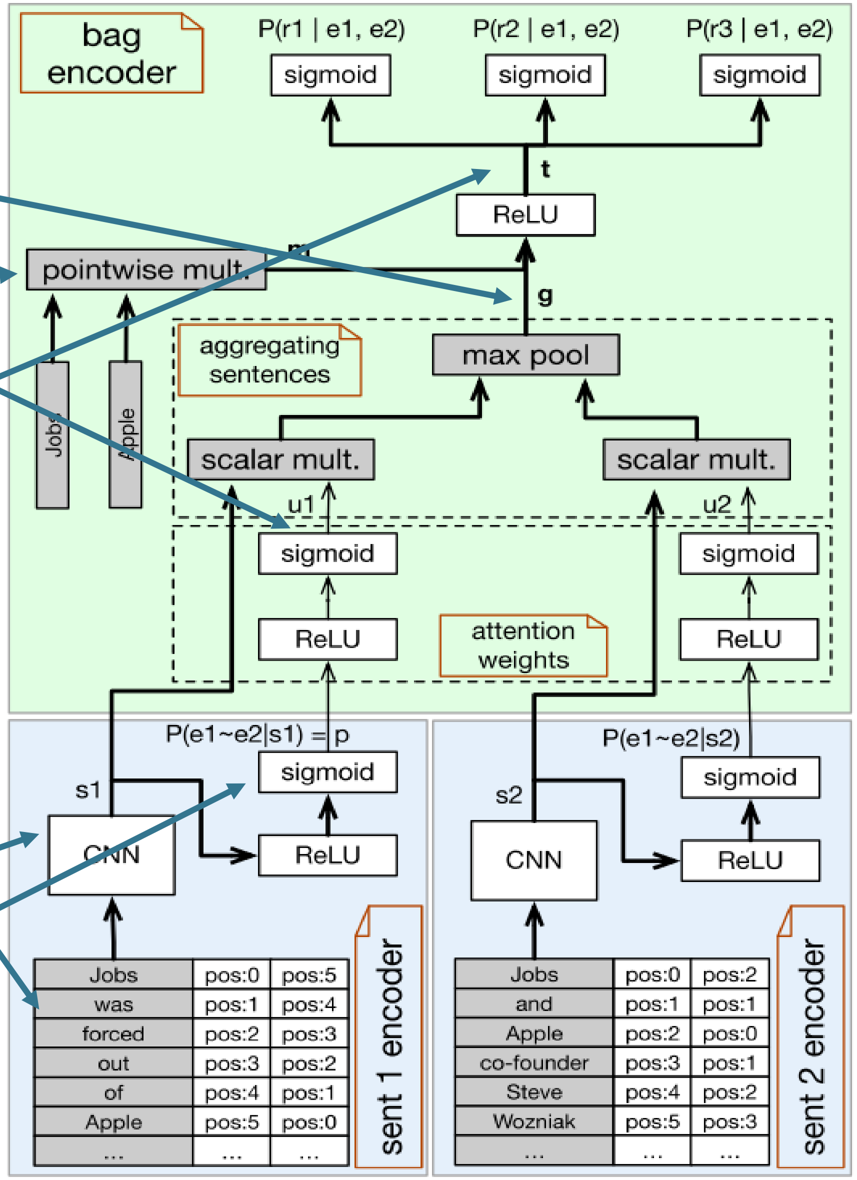
- Attention： $\mathbf{g}_j[k] = \max_{j \in 1,\ldots,n}\{\mathbf{s}_j[k] \times \sigma(u_j)\}$

$$u_j = \mathbf{W}_7\,\mathrm{ReLU}(\mathbf{W}_6\,p + \mathbf{b}_6) + \mathbf{b}_7$$

- Entity Embeddings： $\mathbf{m} = \mathbf{e}_1 \odot \mathbf{e}_2$,
- Output Layer：

$$\mathbf{t} = \mathrm{ReLU}(\mathbf{W}_4[\mathbf{g}; \mathbf{m}] + \mathbf{b}_4)$$

$$P(\mathbf{r} = 1 \mid e_1, e_2) = \sigma(\mathbf{W}_5\mathbf{t} + \mathbf{b}_5),$$



Sentence Encoder Architecture：

- Input Representation： $\mathbf{v}_i = [\mathbf{w}_i; \mathbf{d}_i^{e_1}; \mathbf{d}_i^{e_2}]$, for $i \in 1, \ldots, |s|$
- Word Composition： $\mathbf{c}_x = \mathrm{CNN}_x(\mathbf{v}_1, \ldots, \mathbf{v}_{|s|})$, for $x \in \{2, 3, 4, 5\}$

$$\mathbf{s} = \mathbf{W}_1\,[\mathbf{c}_2; \mathbf{c}_3; \mathbf{c}_4; \mathbf{c}_5] + \mathbf{b}_1,$$

$$P(e_1 \sim e_2 \mid \mathbf{s}) = \qquad (1)$$

$$p = \sigma(\mathbf{W}_3\mathrm{ReLU}(\mathbf{W}_2\mathbf{s} + \mathbf{b}_2) + \mathbf{b}_3)$$

2.3 Model Training

Distant supervision loss:

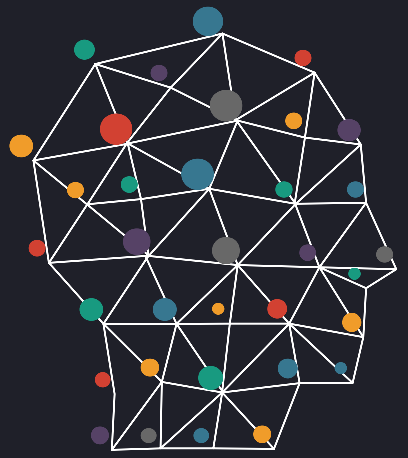
$$\text{DistSupLoss} = \sum_{B_{e_1,e_2}} -\log P(\mathbf{r} = \mathbf{r}^{\text{distant}} \mid e_1, e_2)$$

Direct supervision loss:

$$\text{DirectSupLoss} = \sum_{s,l^{\text{gold}} \in \mathcal{D}} -\log P(l = l^{\text{gold}} \mid \mathbf{s})$$

The model loss is a weighted sum of the direct supervision and distant supervision losses:

$$\text{loss} = \frac{1}{\lambda+1}\text{DistSupLoss} + \frac{\lambda}{\lambda+1}\text{DirectSupLoss}$$



3. Experiment

3.1 Dataset

Distant Supervision Dataset (DistSup):

The FB-NYT dataset. It was generated by aligning Freebase facts with New York Times articles. The dataset has 52 relations with the most common being "location", "nationality", "capital", "place lived" and "neighborhood of". They used all articles for training except those from 2007, which they left for testing.

Direct Supervision Dataset (DirectSup):

The direct supervision dataset was made available by Angeli et al. (2014). The dataset consists of sentences annotated with entities and their relations. It has 22,766 positive examples for 41 relation types in addition to 11,049 negative examples.

3.2 Metrics & Configuration

✔ Metrics:

Use the area under the PR curve (AUC) for early stopping and hyperparameter tuning. Following previous work on this dataset, only keep points on the PR curve with recall below 0.4, focusing on the high-precision low-recall part of the PR curve. As a result, the largest possible value for AUC is 0.4.

✔ Configuration:

Use 90% of the training set for training and keep the other 10% for validation. The main hyperparameter we tune is lambda. The model is trained on machines with P100 GPUs. Each run takes five hours on average. Train for a maximum of 50 epoch. Each dataset is split into minibatches of size 32 and randomly shuffled before every epoch.

3.3 Main Results

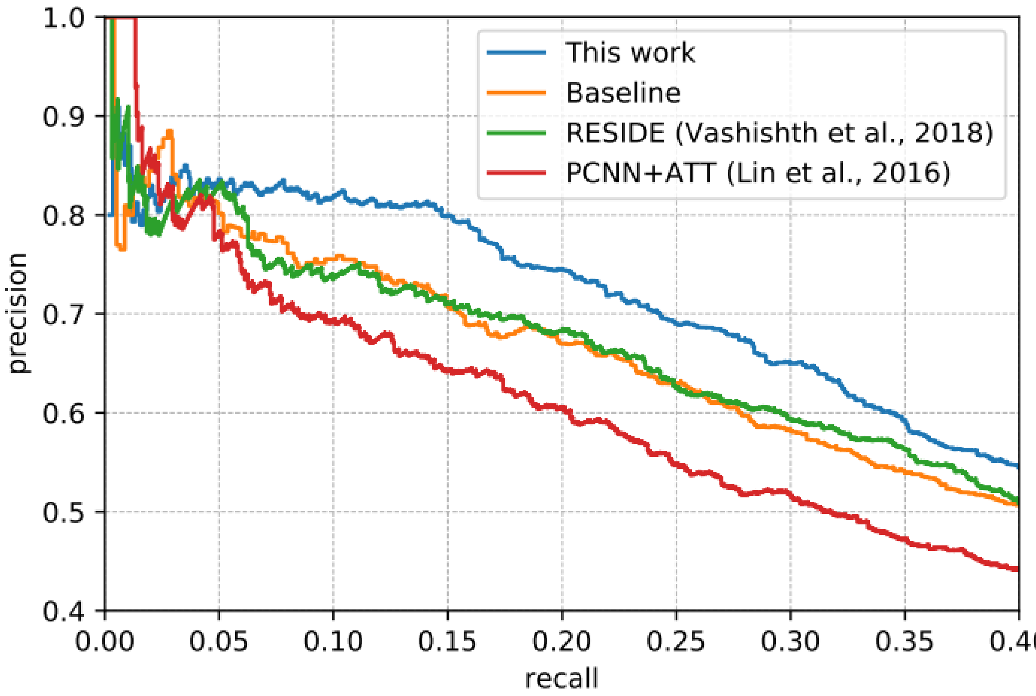
✔ Compared Models:

- **The best model** is trained on the DistSup and DirectSup datasets in our multitask setup and it uses (sigmoid, max pooling) attention.

- **Baseline** is the same model described but trained only on the DistSup dataset and uses the more common (softmax, average pooling) attention.

- **PCNN+ATT** is a model proposed in 2016. It has the same training data and attention form as the **baseline** model. The difference is that it uses PCNN neural network and does not use entity embedding.

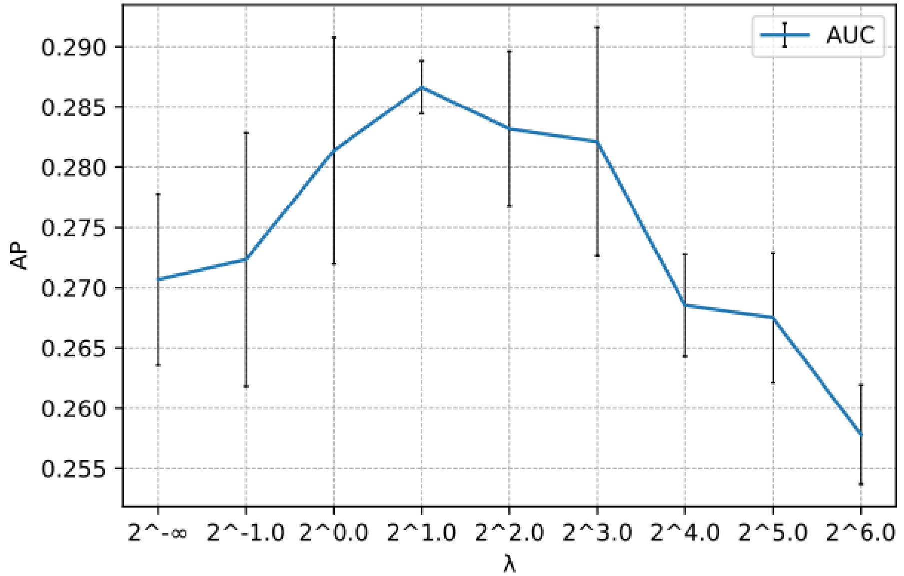
- **Reside** is the best model in the past, which uses graph convolution on the dependency analytic tree.



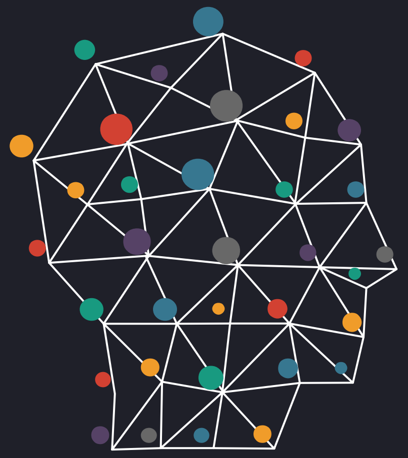
Model	AUC
PCNN+ATT (Lin et al., 2016)	0.247
RESIDE (Vashishth et al., 2018)	0.271
Our baseline	$0.272_{\pm 0.005}$
This work	$\mathbf{0.283}_{\pm 0.007}$

3.4 Controlled experiments

pooling type	supervision signal	attention weight computation		
		uniform	softmax	sigmoid
average pooling	DistSup	0.244± 0.008	0.272± 0.005	0.258± 0.020
	DistSup + DirectSup	0.224± 0.009	0.272± 0.009	0.256± 0.009
	MultiTask (our model)	0.220± 0.012	0.262± 0.014	0.258± 0.015
max pooling	DistSup	0.277± 0.009	0.278± 0.001	0.274± 0.004
	DistSup + DirectSup	0.269± 0.003	0.269± 0.005	0.277± 0.012
	MultiTask (our model)	0.266± 0.007	0.280± 0.004	0.283± 0.007



- **Pooling type:** Use max pooling generally works better than average pooling. This is because max pooling might be better at picking out useful features from each sentence.
- **Supervision signal:** Multitask learning setup leads to considerable improvements (using softmax and sigmoid) because it leads to better attention weights and improves the model's ability to filter noisy sentences.
- **Attention weight computation:** Sigmoidal attention weights give rise to more informative attention weights in cases where all sentences are not useful, or when multiple ones are. This makes the sigmoidal attention weights a better model.
- **Selecting Lambda λ:** It is clear that picking the right value for λ has a big impact on the final result.



4. Conclusion

4 Conclusion

✔ The contributions of this paper are as follows:

1. Improve neural network models for relation extraction by combining distant and direct supervision data. The network uses attention to attend to relevant sentences, and use the direct supervision to improve attention weights, thus improving the model's ability to find sentences that are likely to express a relation.

2. Found that sigmoidal attention weights with max pooling achieves better performance than the commonly used weighted average attention.

3. The model combining both forms of supervision achieves a new state-of-the-art result on the FB-NYT dataset.

Combining Distant and Direct Supervision for Neural Relation Extraction

远程监督与直接监督相结合的神经关系提取

Thanks!