Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference

Authors

Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, Xiaofei He

Organization

Zhejiang University, China

What is Natural Language Inference (NLI)?

Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.

Premise → **Hypothesis**

Entailment

What is Natural Language Inference (NLI)?

Premise: An older and younger man smiling.
Hypothesis: Two men are smiling and laughing at the cats playing on the floor.

Premise $\xrightarrow{?}$ **Hypothesis**

Neutral

What is Natural Language Inference (NLI)?

Premise: A black race car starts up in front of a crowd of people
Hypothesis: A man is driving down a lonely road.

Premise $\not\rightarrow$ **Hypothesis**

Contradiction

Applications

- Question Answering
- Machine Translation
- Semantic Search
- Text Summarization

  …

Discourse Marker

- A **discourse marker** is a word or a phrase that plays a role in managing the flow and structure of discourse.

- Examples: *so*, *because*, *and*, *but*, *or*…

Discourse Marker & NLI?

But *Because* *If* *Although* *And* *So*

Entailment Neutral Contradiction

Related Works

- Datasets

  SNLI (Bowman et al., 2015)
  MultiNLI (Williams et al., 2017)

- SOTA Neural Network Models

  CAFE (Tay et al., 2017)
  KIM (Chen et al., 2017)
  DIIN (Gong et al., 2018)

Related Works

- Transfer Learning for NLI

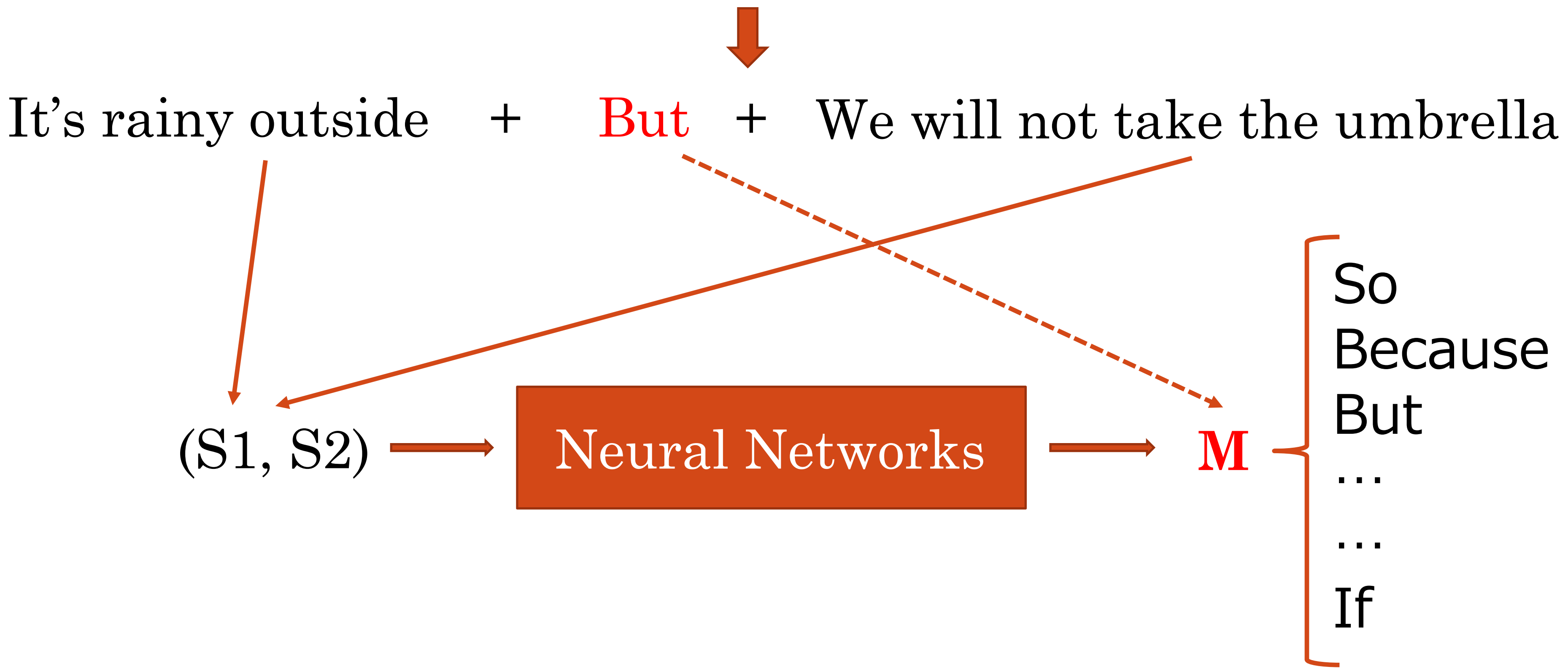
  Skip-thoughts (Vendrov et al., 2016)
  Cove (McCann et al., 2017)

- Discourse Marker Applications

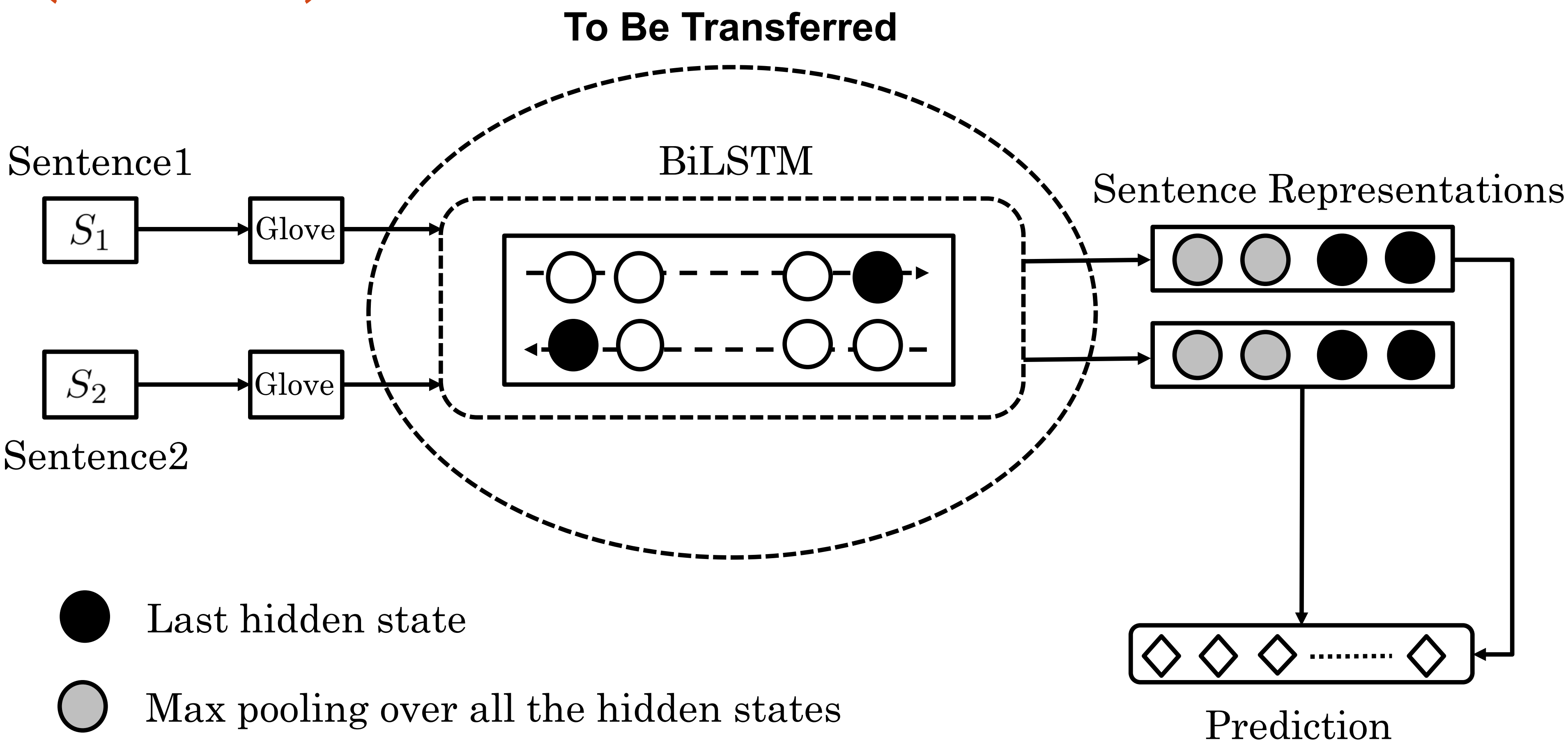
  DisSent (Nie et al., 2017)

Discourse Marker Prediction (DMP)

It's rainy outside **but** we will not take the umbrella

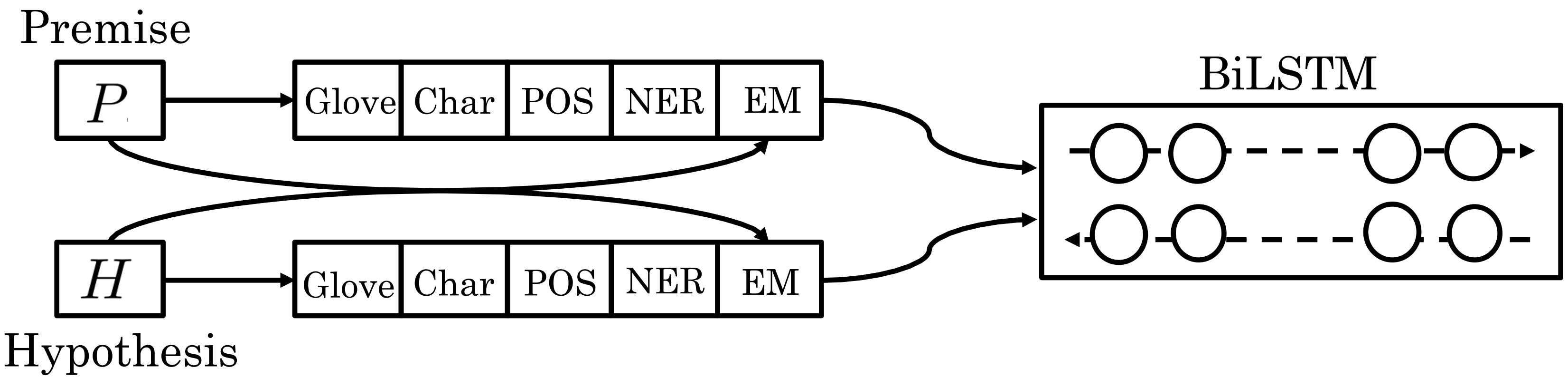


Discourse Marker Prediction (DMP)

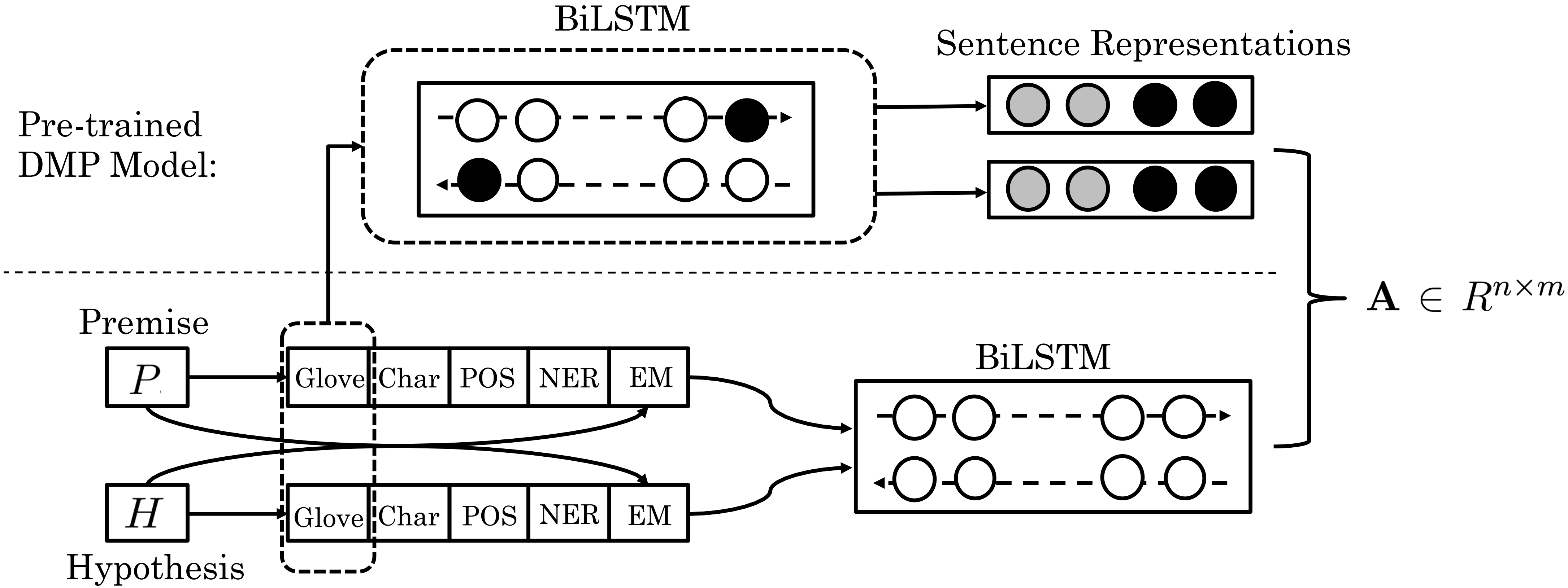


Discourse Marker Augmented Network (NLI Model)

Encoding Layer



Discourse Marker Augmented Network (NLI Model)



Discourse Marker Augmented Network (NLI Model)

Interaction ------ Similarity Matrix

The sentence representation of the **premise**

The sentence representation of the **hypothesis**

$$\mathbf{A}_{ij} = \mathbf{v}_1^\top [\mathbf{p}_i; \mathbf{u}_j; \mathbf{p}_i \circ \mathbf{u}_j; \mathbf{r}_p; \mathbf{r}_h]$$

The i-th word of the **premise**

The j-th word of the **hypothesis**

Discourse Marker Augmented Network (NLI Model)

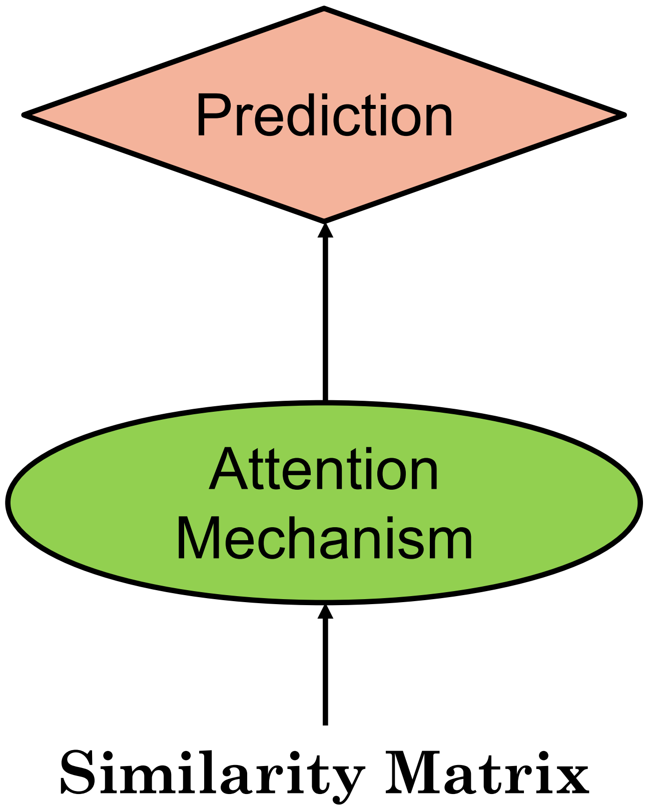
The sentence representation of the **hypothesis**

The sentence representation of the **premise**

$$\mathbf{d} = \mathrm{softmax}(\mathbf{W}[\mathbf{p}^M; \mathbf{u}^M; \mathbf{p}^M \circ \mathbf{u}^M; \mathbf{r}_p \circ \mathbf{r}_h])$$

Modeling vector of the **premise**

Modeling vector of the **hypothesis**



Training

Cross Entropy Loss

$$J_{CE}(\Theta) = -\frac{1}{N}\sum_{k}^{N} log(\mathbf{d}_l^k)$$

- P: "A smiling costumed woman is holding an umbrella."
- H: "A happy woman in a fairy costume holds an umbrella."

Correct Label: *neutral*

Original Labels: *neutral*, *neutral*, *entailment*, *entailment*, *neutral*

Training

$$J_{RL}(\Theta) = -\mathbb{E}_{l\sim\pi(l|P,H)}[R(l,\{l^*\})]$$

Previous action policy that predicts the label given P and H.

$$R(l,\{l^*\}) = \frac{\text{number of } l \text{ in } \{l^*\}}{|\{l^*\}|}$$

$$J(\Theta) = \lambda J_{CE}(\Theta) + (1-\lambda) J_{RL}(\Theta)$$

Experiments (Datasets)

- **Stanford Natural Language Inference (SNLI) (Bowman et al., 2015)**
  570k human annotated sentence pairs

- **Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2017)**
  433k human annotated sentences pairs

- **BookCorpus (Zhu et al., 2015)**
  6.5M pairs of sentences for 8 discourse markers

Discourse Marker	Percentage(%)
but	57.12
because	9.41
if	29.78
when	25.32
so	31.01
although	1.76
before	15.52
still	11.29

Experiments (Results)

Method	SNLI	MultiNLI Matched	MultiNLI Mismatched
300D LSTM encoders(Bowman et al., 2016)	80.6	–	–
300D Tree-based CNN encoders(Mou et al., 2016)	82.1	–	–
4096D BiLSTM with max-pooling(Conneau et al., 2017)	84.5	–	–
600D Gumbel TreeLSTM encoders(Choi et al., 2017)	86.0	–	–
600D Residual stacked encoders(Nie and Bansal, 2017)	86.0	74.6	73.6
Gated-Att BiLSTM(Chen et al., 2017d)	–	73.2	73.6
100D LSTMs with attention(Rocktäschel et al., 2016)	83.5	–	–
300D re-read LSTM(Sha et al., 2016)	87.5	–	–
DIIN(Gong et al., 2018)	88.0	78.8	77.8
Biattentive Classification Network(McCann et al., 2017)	88.1	–	–
300D CAFE(Tay et al., 2017)	88.5	78.7	77.9
KIM(Chen et al., 2017b)	88.6	–	–
600D ESIM + 300D Syntactic TreeLSTM(Chen et al., 2017c)	88.6	–	–
DMAN	**88.8**	**78.9**	**78.2**
BiMPM(Ensemble)(Wang et al., 2017)	88.8	–	–
DIIN(Ensemble)(Gong et al., 2018)	88.9	80.0	78.7
KIM(Ensemble)(Chen et al., 2017b)	89.1	–	–
300D CAFE(Ensemble)(Tay et al., 2017)	89.3	80.2	79.0
DMAN(Ensemble)	**89.6**	**80.3**	**79.4**

Sentence Encoding-Based Models

Other Neural Network Models

Ensemble Models

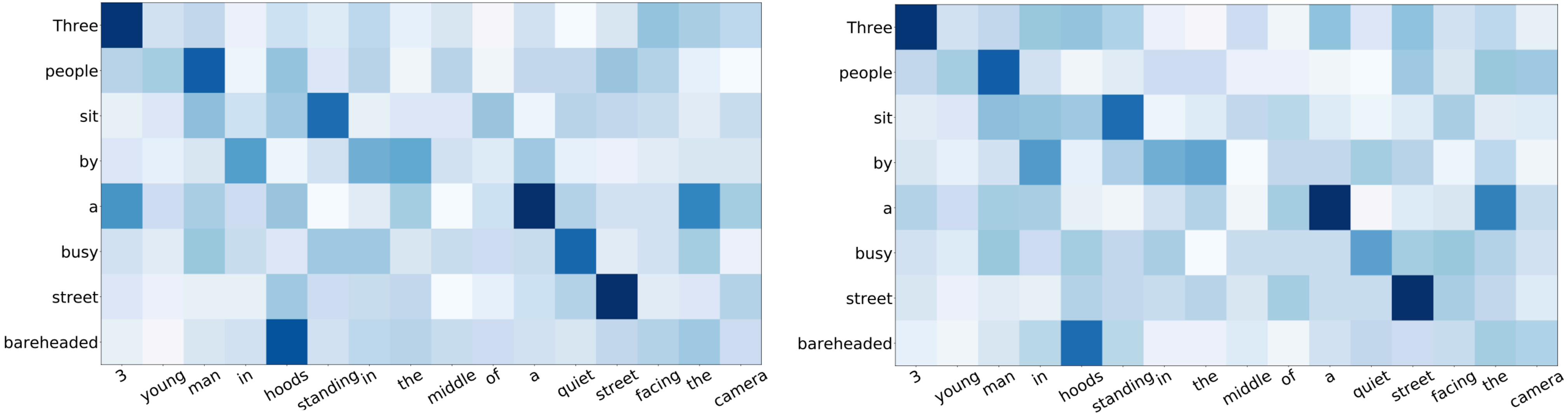
Experiments (Analysis)

Ablation Model	Accuracy
Only Sentence Encoder Model	83.37
No Sentence Encoder Model	87.24
No REINFORCE	88.41
DMAN	88.83

Experiments (Analysis)

(a) Discourse markers augmentation

(b) Without discourse markers augmentation



Premise: *“3 young man in hoods standing in the middle of a quiet street facing the camera.”*
Hypothesis: *“Three people sit by a busy street bare-headed.*

Conclusion

- We solve the task of the natural language inference via transferring knowledge from another supervised task.
- We propose a new objective function to make full use of the labels' information.
- In the future work, we would like to explore some other transfer learning sources.

Thank You !