Xuefeng BAI

✉ xfbai.hk@gmail.com · ☎ (+86) 15754604524 · github.com/goodbai-nlp

Education

Zhejiang University, Hangzhou 09/2019 – 03/2023 (expected)
PhD Candidate; Computer Science and Technology
Supervisor: Yue Zhang

Harbin Institute of Technology, Harbin 09/2017 – 07/2019
Master Degree; Computer Science and Technology
Supervisor: Hailong Cao and Tiejun Zhao

Harbin Institute of Technology, Harbin 09/2013 – 07/2017
Undergraduate Degree; Computer Science and Technology

Representative Research

Research Interest: Semantics, Dialogue, Machine Translation, Generation

Semantic-aware Pre-training for Dialogue Understanding

Xuefeng Bai, Linfeng Song and Yue Zhang, accepted by **COLING2022, (CCF-B, THU-B)**

We investigate deep semantic structures in dialogue pre-training. Specifically, we design 3 different semantic-related pre-training tasks to guide the model to focus on the core semantic information in the dialogue text. The experimental results show that the semantically enhanced pre-trained model has achieved significant improvements on both the task-oriented and the chit-chat dialogue understanding tasks.

Graph Pre-training for AMR Parsing and Generation

Xuefeng Bai, Yulong Chen and Yue Zhang, in **ACL2022, (CCF-A, THU-A)**

Python, *Pytorch;* `https://github.com/goodbai-nlp/AMRBART`

We investigate graph self-supervised training to improve the structure awareness of PLMs over AMR graphs. In particular, we introduce two graph auto-encoding strategies for graph-to-graph pre-training and four tasks to integrate text and graph information during pre-training. We further design a unified framework to bridge the gap between pre-training and fine-tuning tasks. Experimental results on both AMR parsing and AMR-to-text generation tasks show the superiority of our model.

Semantic Representation for Dialogue Modeling

Xuefeng Bai, Yulong Chen, Linfeng Song and Yue Zhang, in **ACL2021, (CCF-A, THU-A)**

Python, *Pytorch;* `https://github.com/goodbai-nlp/Sem-Dialogue`

We exploit Abstract Meaning Representation (AMR) as a structured semantic representation to aid dialogue modeling. Compared with textual input, AMR provides external semantic knowledge and reduces data sparsity. We develop an algorithm to automatically construct dialogue-level AMRs from sentence-level AMRs, and we explore two approaches to incorporating AMRs into dialogue systems. Experiments on standard benchmarks show the superiority of our model on both dialogue understanding and response generation tasks.

Online Back-Parsing for AMR-to-Text Generation

Xuefeng Bai, Linfeng Song and Yue Zhang, in **EMNLP2020, (CCF-B, THU-A)**

Python, *Pytorch;* `https://github.com/goodbai-nlp/Gen-Backparsing`

We introduce an online back-parsing mechanism for Graph-to-Text generation. Different from most previous work which designs various graph encoders, we focus on the decoder instead. Sharing the same idea as the back-translation mechanism, our online back-parsing model aims to reconstruct input structures. Extensive experiments on the AMR-to-text show that our model achieves significantly better results than previous work.

Investigating Typed Syntactic Dependencies for Targeted Sentiment Classification Using Graph Attention Neural Network
Xuefeng Bai, Pengbo Liu and Yue Zhang. in **TASLP, (CCF-B, THU-A)**

Python, *Pytorch;* `https://github.com/goodbai-nlp/RGAT-ABSA`

We present a relational graph attentional network for targeted sentiment classification, which can additionally exploit typed relations when encoding a graph compared with the vanilla graph attentional network. Results on standard benchmarks show that our method can effectively leverage syntactic label information for improving targeted sentiment classification performances, obtaining better results than existing systems.

A Bilingual Adversarial Auto-encoder for Unsupervised Bilingual Lexicon Induction
Xuefeng Bai, Hailong Cao, Kehai Chen and Tiejun Zhao. in **TASLP, (CCF-B, THU-A)**

Python, *Pytorch;* `https://github.com/goodbai-nlp/BiAAE`

We propose a novel framework (BiAAE) that aligns words without any supervision. Different from existing works, the proposed model exploits an adversarial auto-encoder which is more powerful and does not rely on the isomorphism hypothesis. Our method significantly outperforms previously unsupervised models and even achieves comparable results with supervised models.

Sentence-State LSTMs For Sequence-to-Sequence Learning
Xuefeng Bai, Yafu Li, Zhirui Zhang, Mingzhou Xu, Boxing Chen, Weihua Luo,
Derek Wong and Yue Zhang. in **NLPCC2021, (CCF-C)**

Python, Pytorch; `https://github.com/goodbai-nlp/SLSTM-nmt`

We propose a graph recurrent neural network-based model for sequence-to-sequence learning. Specifically, the model consists of a S-LSTM encoder and a LSTM decoder. The complexity of S-LSTM encoder is only $\mathcal{O}$(n) as compared to $\mathcal{O}(n^2)$ of Transformer. Experiments on 4 machine translation benchmarks show that our model gives competitive performance while being 1.6 times faster during inference compared with Transformer.

Cross-domain Generalization for AMR Parsing
under review in **EMNLP2022**

We study the cross-domain generalization ability of AMR parsing systems. We systematically evaluate 5 AMR parsers on 5 domains and analyze the key challenges of cross-domain AMR Parsing. Based on our observations, we propose two methods to reduce the domain distribution divergence of text and AMR features, respectively. Experiments on two out-of-domain test sets show the superiority of our method.

PROJECTS

Semantic-guided Pre-training for Dialogue Modeling
Rhino-Bird Focused Research Program with Tencent AI Lab 03/2021 – 04/2022

Python, Pytorch; Research Intern; Supervisor: Linfeng Song

We investigate the feasibility of 1) jointly pre-training semantic graphs and text to model the correspondence between semantic structures and text. 2) employing deep semantic knowledge in dialogue pre-training to guide the model to better understand the core semantic information in the dialogue text. The papers are accepted by ACL2022 and COLING2022, respectively. See the above papers for detailed methods.

Investigating Structured Semantic Representation for Neural Dialogue Modeling
Rhino-Bird Focused Research Program with Tencent AI Lab 03/2020 – 04/2021

Python, Pytorch; Supervisor: Linfeng Song

We 1) propose a better decoding mechanism to generate text which is more faithful to input semantic graphs. 2) explore the use of semantic representations to aid dialogue understanding and dialogue response generation. The papers are accepted in EMNLP2020 and ACL2021, respectively.

Graph Neural Network for Neural Machine Translation
AIR research project with Alibaba DAMO Academy. 12/2019 – 06/2021

Python, Pytorch; Research Intern; Supervisor: Boxing Chen

We slove NMT with a new architecture named Sentence-State LSTM (S-LSTM) which has been proven to be better than Transformer on sentence labeling and sentence classification. Experiments show that our model can give comparable results with Transformer, with a higher inference speed. The paper is accepted in NLPCC.

On Unsupervised Statistical Machine Translation

Intern in Sogou Inc. 07/2018 – 10/2018

Python, Pytorch, Moses; Research Intern; Supervisor: Feifei Zhai

We first align phrases without any supervision via BiAAE, then use the resulting phrase table to initialize a SMT model. Finally, we adopt the interactive back-translation mechanism to refine the translation model. The resulted system obtains significantly better results than word-based unsupervised models.

Publication List

1. **Xuefeng Bai**, Yulong Chen and Yue Zhang. (2022) *Graph Pre-training for AMR Parsing and Generation*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers) (pp. 6001-6015). (CCF-A, THU-A)
2. **Xuefeng Bai**, Yulong Chen, Linfeng Song and Yue Zhang. (2021) *Semantic Representation for Dialogue Modeling*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers) (pp. 4430-4445). (CCF-A, THU-A)
3. **Xuefeng Bai**, Linfeng Song and Yue Zhang. (2021) *Online back-parsing for AMR-to-text generation*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1206-1219). (CCF-B, THU-A)
4. **Xuefeng Bai**, Linfeng Song and Yue Zhang. (2022) *Semantic-based Pre-training for Dialogue Understanding*. In Proceedings of the 29th International Conference on Computational Linguistics (COLING). (CCF-B, THU-B)
5. **Xuefeng Bai**, Pengbo Liu and Yue Zhang. (2021) *Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 503-514. (CCF-B, THU-A)
6. **Xuefeng Bai**, Hailong Cao, Kehai Chen and Tiejun Zhao. (2019) *A bilingual adversarial autoencoder for unsupervised bilingual lexicon induction*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(10), 1639-1648. (CCF-B, THU-A)
7. **Xuefeng Bai**, Yafu Li, Zhirui Zhang, Mingzhou Xu, Boxing Chen, Weihua Luo, Derek Wong and Yue Zhang. (2022) *Sentence-State LSTMs For Sequence-to-Sequence Learning*. In CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC) (pp. 104-115). (CCF-C)
8. **Xuefeng Bai**, Hailong Cao, Kehai Chen and Tiejun Zhao. (2018) *Improving vector space word representations via kernel canonical correlation analysis*. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 17(4), 1-16. (CCF-C)
9. Yulong Chen*, Ming Zhong*, **Xuefeng Bai***, Naihao Deng, Jing Li, Xianchao Zhu, Yue Zhang. (2022) *The Cross-lingual Conversation Summarization Challenge*. In Proceedings of the 15th International Conference on Natural Language Generation (INLG).
10. **Xuefeng Bai**, Yue Zhang, Hailong Cao, Tiejun Zhao. (2019) *Duality regularization for unsupervised bilingual lexicon induction*. ArXiv preprint arXiv:1909.01013.

Awards

National Scholarships for doctoral students	10/2021
National Scholarships for postgraduate students	09/2018
Outstanding postgraduate students in Zhejiang University	10/2021
Top 100 Master Thesis in Harbin Institute of Technology	06/2019
Top 100 graduation projects in Harbin Institute of Technology	06/2017

Others

- **DeepLearning Frameworks:** Pytorch > TensorFlow > Theano
- **Service:** *Area Chair in EMNLP2022*
- **Hobby:** PingPong, Running, Dota