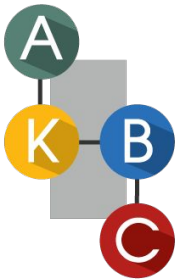
Using BibTeX to Automatically Generate Labeled Data for Citation Field Extraction

Dung Thai, Zhiyang Xu, Nicholas Monath, Boris Veytsman, Andrew McCallum



AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Citation Field Extraction

Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30, Provided by the SAO/NASA Astrophysics Data System.

Krucker, S.; Benz, A. O. In Bastian, T. S., Gopalswamy, N., Shibasaki, K., Eds., *Proceedings of the Nobeyama Symposium*, Vol. 479, pages 25–30, 1999.

S. Krucker, A.O. Benz, in *Proceedings of the Nobeyama Symposium*, vol. 479, ed. by T.S. Bastian, N. Gopalswamy, K. Shibasaki (1999), vol. 479, pp. 25–30. URL `http://esoads.eso.org/abs/1999spro.proc.25K`. Provided by the SAO/NASA Astrophysics Data System

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Citation Field Extraction

Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30,

Krucker, S AUTHOR and Benz, A. O. AUTHOR , " Heating events in the quiet solar corona TITLE ," Proceedings of the Nobeyama Symposium JOURNAL , edited by T. S. Bastian EDITOR , N. Gopalswamy EDITOR , and K. Shibasakii EDITOR , Vol. 479 VOLUME , December MONTH 1999 YEAR , pp. 25-30 PAGES ,

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Citation Field Extraction

Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30,

Krucker, S AUTHOR and Benz, A. O. AUTHOR , " Heating events in the quiet solar corona TITLE ,"
Proceedings of the Nobeyama Symposium JOURNAL , edited by T. S. Bastian EDITOR ,
N. Gopalswamy EDITOR , and K. Shibasakii EDITOR , Vol. 479 VOLUME , December MONTH
1999 YEAR , pp. 25-30 PAGES ,

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Citation Field Extraction

Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30, Provided by the SAO/NASA Astrophysics Data System.

Krucker, S.; Benz, A. O. In Bastian, T. S., Gopalswamy, N., Shibasaki, K., Eds., *Proceedings of the Nobeyama Symposium*, Vol. 479, pages 25–30, 1999.

S. Krucker, A.O. Benz, in *Proceedings of the Nobeyama Symposium*, vol. 479, ed. by T.S. Bastian, N. Gopalswamy, K. Shibasaki (1999), vol. 479, pp. 25–30. URL `http://esoads.eso.org/abs/1999spro.proc.25K`. Provided by the SAO/NASA Astrophysics Data System

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Motivation

Scientific Impact

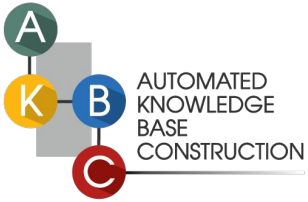
- CFE is a useful test bed for sequence labeling problems
- Dense, complex labeling space

Practical Applications

- Citation knowledge graph (CORA, WebKB)
- Document classification (CiteSeer, PubMed Diabetes)
- Entity resolution (CiteSeer, Arxiv High-Energy Physics)

Lack of Dataset

- Human annotated dataset is costly
- Coverage of scarce citation field, citation style, etc.
- Available dataset - UMass Citation Field Extraction is fairly small



Motivation

Scientifically Interesting

- Clear sequence-to-sequence learning setting
- Dense, complex labeling space

Practical Applications

- Citation knowledge graph (CORA, WebKB)
- Document classification (CiteSeer, PubMed Diabetes)
- Entity resolution (CiteSeer, Arxiv High-Energy Physics)

Lack of Dataset

- Human annotated dataset is costly
- Coverage of scarce citation field, citation style, etc.
- Available dataset - UMass Citation Field Extraction is fairly small

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Motivation

Scientifically Interesting

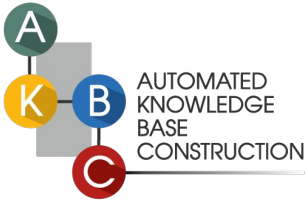
- Clear sequence-to-sequence learning setting
- Dense, complex labeling space

Practical Applications

- Citation knowledge graph (CORA, WebKB)
- Document classification (CiteSeer, PubMed Diabetes)
- Entity resolution (CiteSeer, Arxiv High-Energy Physics)

Lack of Dataset

- Human annotated dataset is costly
- Coverage of scarce citation field, citation style, etc.
- Available dataset - UMass Citation Field Extraction is fairly small



Contributions

- Most previous work focused on complicated sequence modeling
- Would straightforward Deep Neural Networks training on noisy, large-scale data work better?

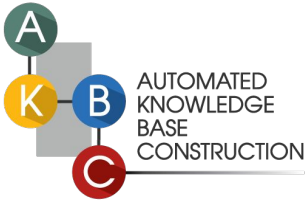
Research Question

- Automatically generate labeled citations from BibTeX
- Simple, reliable way to extract citation field labels

Data Generation Process

- Achieve 24.48% relative error reduction on UMass CFE, results in span level F1 96.3%
- A pre-trained MLM for citations
- New benchmarks for the Citation Field Extraction task

Experimental Results



Contributions

- Most previous work focused on complicated sequence modeling
- Would straightforward Deep Neural Networks training on noisy, large-scale data work better?

Research Question

- Automatically generate labeled citations from BibTeX
- Simple, reliable way to extract citation field labels

Data Generation Process

- Achieve 24.48% relative error reduction on UMass CFE, results in span level F1 96.3%
- A pre-trained MLM for citations
- New benchmarks for the Citation Field Extraction task

Experimental Results

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Contributions

- Most previous work focused on complicated sequence modeling
- Would straightforward Deep Neural Networks training on noisy, large-scale data work better?

Research Question

- Automatically generate labeled citations from BibTeX
- Simple, reliable way to extract citation field labels

Data Generation Process

- Achieve 24.48% relative error reduction on UMass CFE, results in span level F1 96.3%
- A pre-trained MLM for citations
- New benchmarks for the Citation Field Extraction task

Experimental Results

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Generate Citation from BibTeX

```
@inproceedings{kingma:vae,
  title={Auto-encoding variational {Bayes}},
  author={Kingma, Diederik P and Welling, Max},
  booktitle={ Int. Conf. on Learning Representations },
  year={2014}
}
```

BibTeX Entry

BibTeX Style

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Int. Conf. on Learning Representations.*

Citation PDF

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Citation Field Extraction

Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30, Provided by the SAO/NASA Astrophysics Data System.

Krucker, S.; Benz, A. O. In Bastian, T. S., Gopalswamy, N., Shibasaki, K., Eds., *Proceedings of the Nobeyama Symposium*, Vol. 479, pages 25–30, 1999.

S. Krucker, A.O. Benz, in *Proceedings of the Nobeyama Symposium*, vol. 479, ed. by T.S. Bastian, N. Gopalswamy, K. Shibasaki (1999), vol. 479, pp. 25–30. URL `http://esoads.eso.org/abs/1999spro.proc.25K`. Provided by the SAO/NASA Astrophysics Data System

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Annotated Generate Citation from BibTeX

```
@inproceedings{kingma:vae,
  title={[T] Auto-encoding variational {Bayes} [T]},
  author={Kingma, Diederik P and Welling, Max},
  booktitle={[B] Int. Conf. on Learning Representations [B]},
  year={[Y] 2014 [Y]}
}
```

Annotated BibTeX Entry

BibTeX Style

Diederik P Kingma and Max Welling. [Y] 2014 [Y]. [T] Auto-encoding variational Bayes [T]. In [B] *Int. Conf. on Learning Representations* [B].

Annotated Citation PDF

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

BibTeX CFE Dataset

Parameter	BibTeX dataset
Number of annotated references	41,572,904
Average reference length (in tokens)	33.09
Number of segment labels	59
Number of segments	298,013,391
Average segment length (in tokens)	3.26
Vocabulary size	2,823,254
Number of styles	26
Number of BibTeX sources	6023

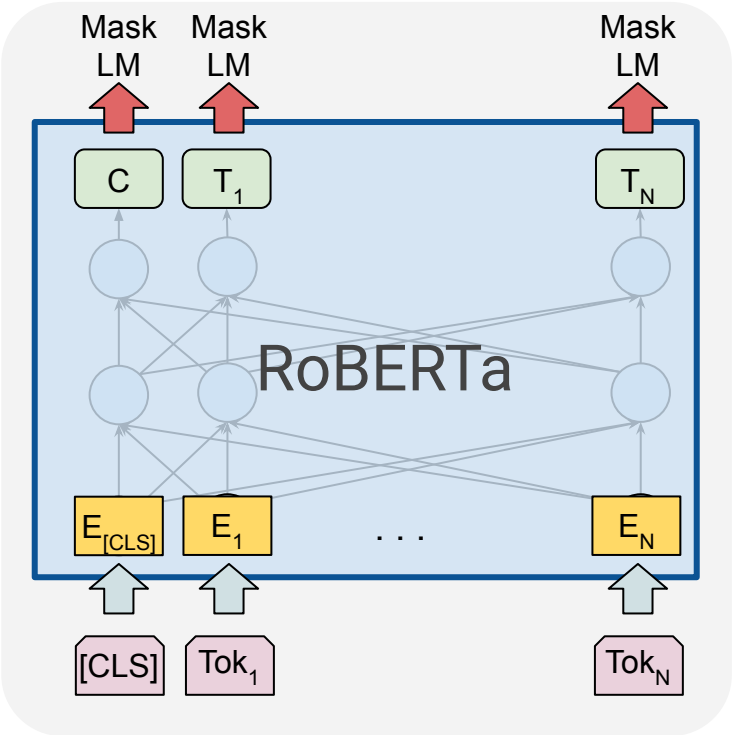
Table 1: Summary of our BibTeX CFE dataset.

Label	Number of segments
author	91,324,094
year	52,946,966
title	42,846,934
journal	20,620,003
publisher	9,777,982
editor	3,481,227
location	3,125
category	219

Table 2: Segment counts for some labels of interest.

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Pre-trained MLM on CFE Dataset



+

S. Krucker, A.O. Benz, in *Proceedings of the Nobeyama Symposium*, vol. 479, ed.

Krucker, S.; Benz, A. O. In Bastian, T. S., Gopalswamy, N., Shibasaki, K., Eds.,

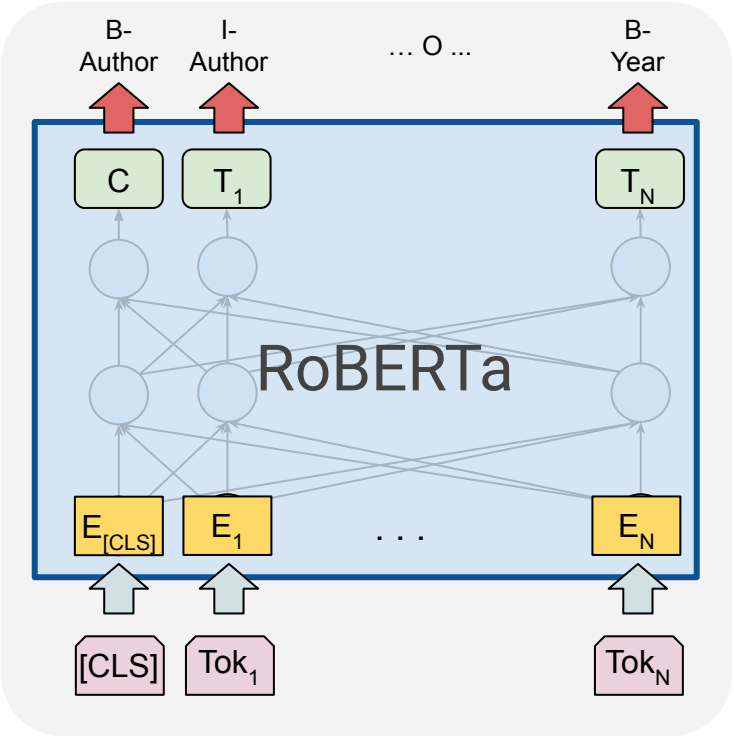
Krucker, S. and Benz, A. O., "Heating Events in the Quiet Solar Corona," *Proceedings of the Nobeyama Symposium*, edited by T. S. Bastian, N. Gopalswamy, and K. Shibasaki, Vol. 479, December 1999, pp. 25–30, Provided by the SAO/NASA Astrophysics Data System.

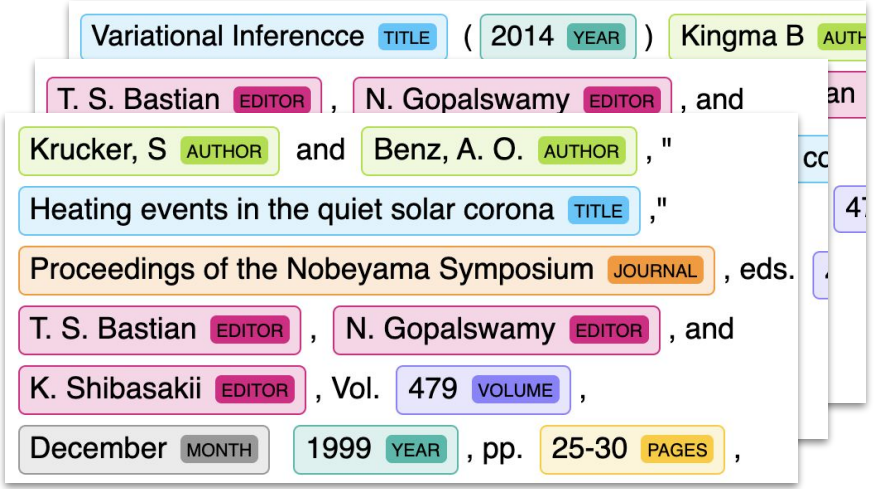
Krucker, S.; Benz, A. O. In Bastian, T. S., Gopalswamy, N., Shibasaki, K., Eds., *Proceedings of the Nobeyama Symposium*, Vol. 479, pages 25–30, 1999.

S. Krucker, A.O. Benz, in *Proceedings of the Nobeyama Symposium*, vol. 479, ed. by T.S. Bastian, N. Gopalswamy, K. Shibasaki (1999), vol. 479, pp. 25–30. URL `http://esoads.eso.org/abs/1999spro.proc.25K`. Provided by the SAO/NASA Astrophysics Data System

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Fine-tune Sequence Labeling





AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Performance on UMass CFE

(a.k.a., performance on human labeled dataset)

Model	UMass Dev			UMass Test		
	P	R	F1	P	R	F1
Thai et al. [2018]	–	–	–	–	–	0.951
GloVe	**0.982**	0.923	0.925	0.940	0.934	0.937
ELMo	0.954	0.947	0.950	0.955	0.946	0.951
BERT	0.941	0.932	0.936	0.932	0.925	0.928
RoBERTa	0.932	0.944	0.938	0.925	0.940	0.933
RoBERTa (+LM)	0.940	0.948	0.944	0.934	0.948	0.940
RoBERTa (+BIBTEX)	0.956	0.960	0.958	0.959	0.963	0.961
RoBERTa (+BIBTEX+LM)	0.954	**0.964**	**0.959**	**0.960**	**0.967**	**0.963**

Table 3: Span level results on UMass CFE dataset.

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Performance on UMass CFE

(a.k.a., performance on human labeled dataset)

	SoTa	Our	Δ
title	0.9258	**0.9661**	+0.0403
publisher	0.8525	**0.9180**	+0.0655
booktitle	0.4416	**0.6769**	+0.2353
institution	0.5455	**0.9091**	+0.3636
school	0.5000	**0.8000**	+0.3000
year	**0.9944**	0.9929	-0.0015
journal	**0.9583**	0.9409	-0.0174

Table 4: Per label F1 of **RoBERTa (+BibTeX+LM)** compared to SoTA.

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

BibTeX CFE Benchmark

Labels	Precision	Recall	F1	Count
author	0.981	0.988	0.984	119,003
title	0.937	0.951	0.944	564,813
year	0.998	0.964	0.981	555955
pages	0.997	0.989	0.993	376960
journal	0.970	0.997	0.983	307,135
volume	0.994	0.986	0.990	232883
institution	0.889	0.832	0.860	22,558
school	0.893	0.873	0.883	12,271
organization	0.905	0.952	0.928	8,040
edition	0.876	0.551	0.677	1,538
chapter	0.960	0.582	0.725	1,278
overall	**0.972**	**0.968**	**0.970**	**3,760,465**

Table 5: Performance of **RoBERTa (+BibTeX+LM)** on a subset of citation field labels.

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

BibTeX CFE Benchmark

Models	Math			Physics			Econs			CompSci		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RoBERTa	0.832	0.809	0.820	0.860	0.803	0.831	0.832	0.784	0.807	0.858	0.810	0.833
RoBERTa (+LM-BibTeX)	**0.846**	**0.819**	**0.832**	**0.874**	**0.811**	**0.841**	**0.850**	**0.796**	**0.822**	**0.872**	**0.820**	**0.845**

Table 6: Sequence tagger performances on selected domain.

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Conclusion

- We confirm that standard Transformer-based model training on noisy, large-scale data works better
- Achieve new SoTA UMass CFE (span level F1 96.3%)

Research Findings

- We release the code and the BibTeX entries for generating the dataset as well as the evaluation dataset
- The pre-trained MLM model will be provided

Data & Pre-trained

- A more effective training procedure on noisy large-scale dataset
- Further improve the new CFE benchmark

Future Work

AUTOMATED KNOWLEDGE BASE CONSTRUCTION

Thanks!

Paper Link