组会

曾双

2020.8.30

ACL 2020

SciREX: A Challenge Dataset for Document-Level Information Extraction

Sarthak Jain[2*] **Madeleine van Zuylen**[1] **Hannaneh Hajishirzi**[1,3] **Iz Beltagy**[1]
Allen Institute for AI[1] Northeastern University[2] University of Washington[3]
jain.sar@northeastern.edu
{madeleinev,hannah,beltagy}@allenai.org

EMNLP 2017, ScienceIE SOTA!

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *Proceedings of Empirical Methods in Natural Language Processing*.

EMNLP 2018: SciERC & SciIE

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

NAACL 2019: DyGIE

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of The North American Chapter of the Association for Computational Linguistics (NAACL)*.

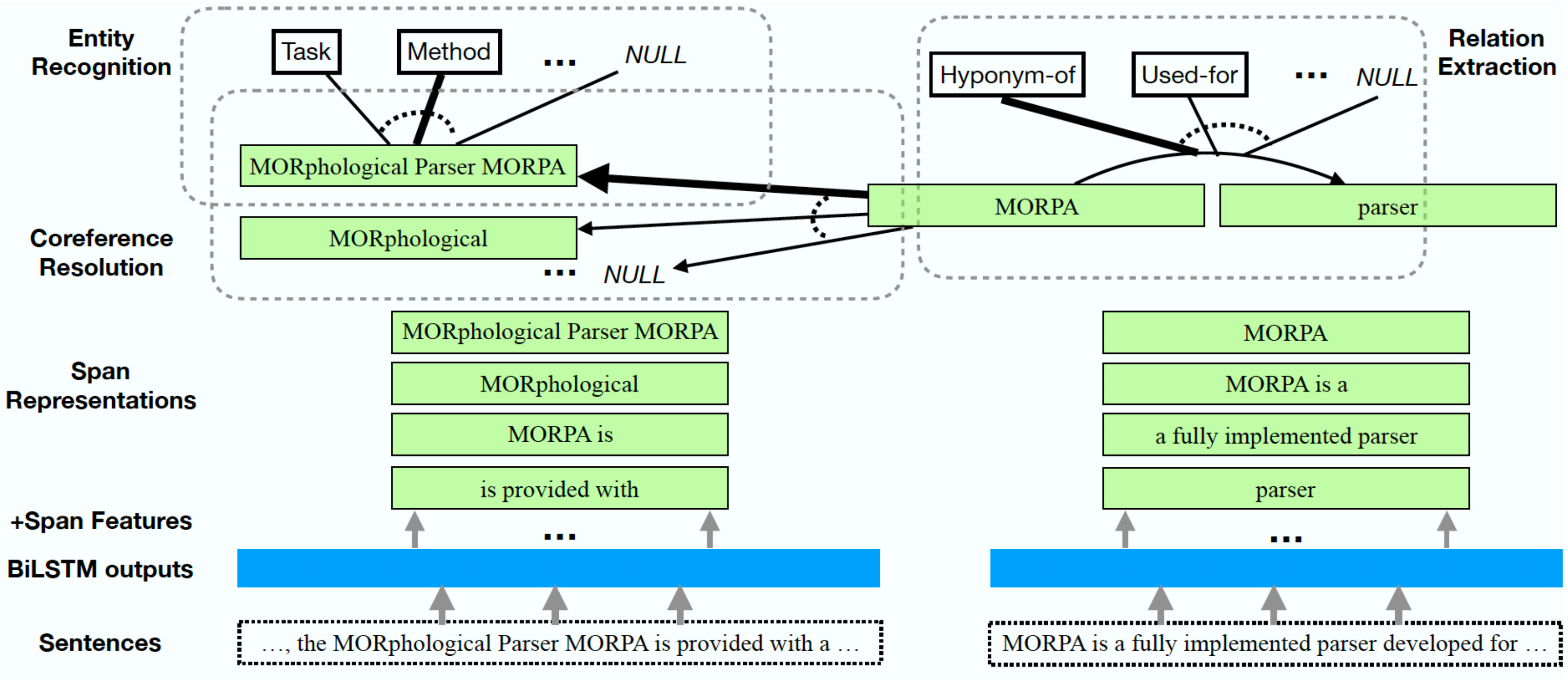
ACE04 ACE05 SOTA !

EMNLP 2019: DyGIE++

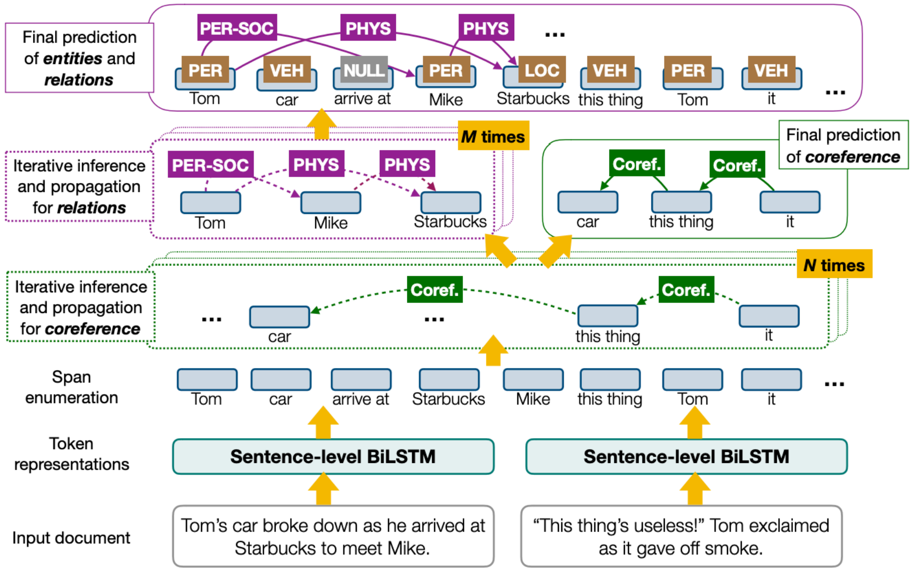
David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

ACE05 SciERC SOTA !

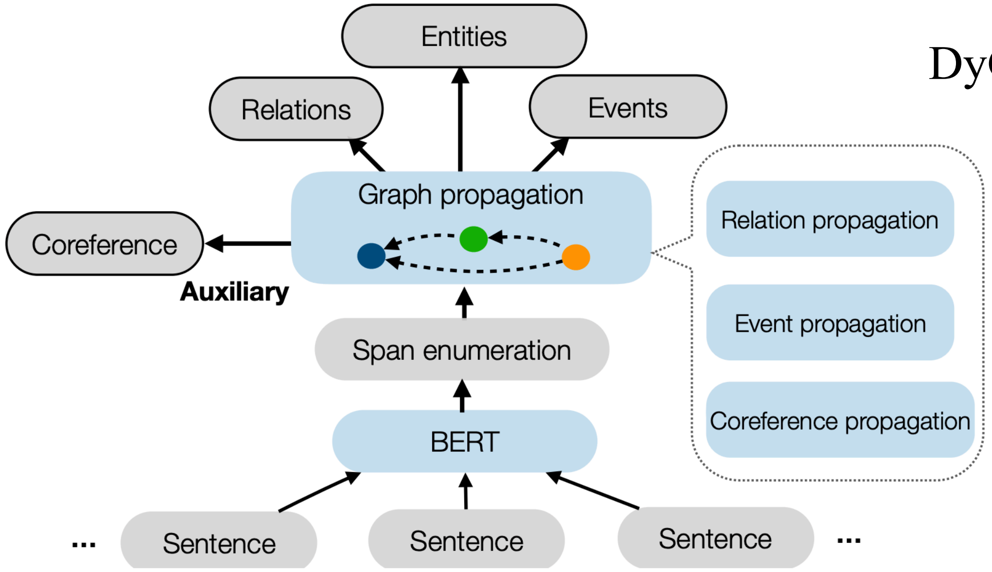
SciIE



DyGIE



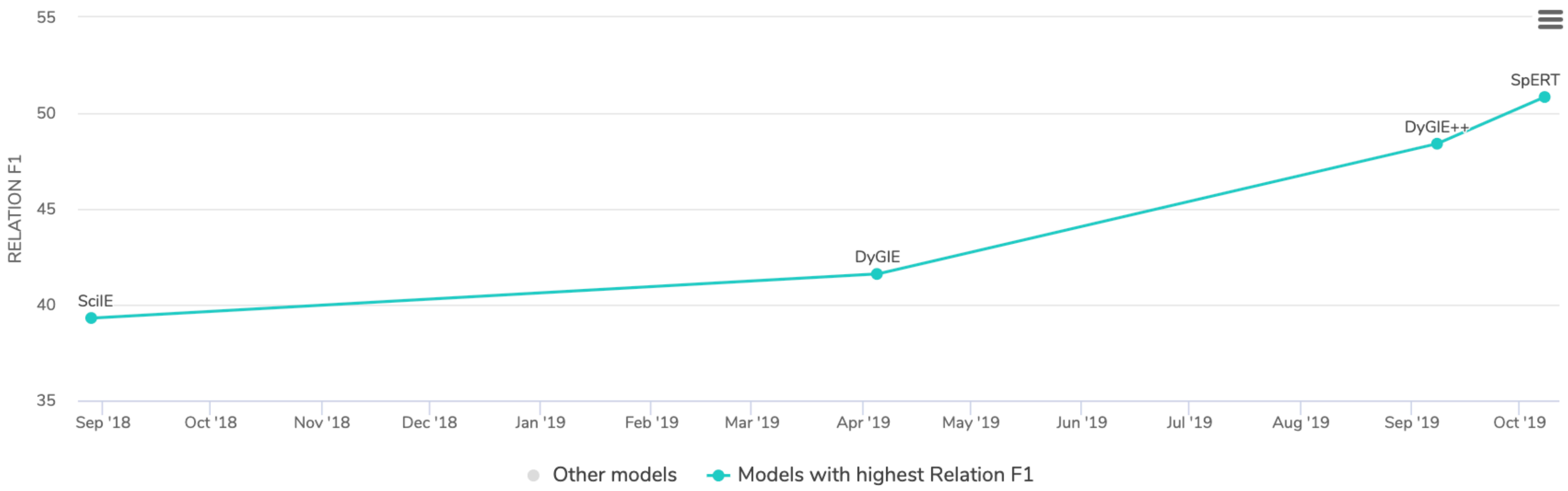
DyGIE ++



Motivation

- 前人工作只关注在句子（TACRED，NYT，WebNLG，ACE04，ACE05，SemEval- 2010 Task 8）或者段落（CDR，GDA，SciERC，DocRED）中抽取关系
- 本文提出一个大规模篇章级信息抽取（doc-level IE）的数据集SciREX。
- 标注了一下信息：
  - Mention span and it type (Dataset, Metric, Task, Method)
  - Salient entity mention
  - Mention resolution
  - Relations between mention clusters (entity)
- 包括了4个IE任务：(1) Entity Recognition; (2) Salient Entity Identification; (3) Mention Resolution; (4) Document-level N-ary Relation Identification (binary, 3-ary, 4-ary; across sentences and sections)
- 第一个做全文IE的工作

Joint Entity and Relation Extraction on SciERC



View Relation F1 Edit

RANK	MODEL	RELATION F1	ENTITY F1	PAPER	CODE	RESULT	YEAR
1	SpERT	50.84	70.33	Span-based Joint Entity and Relation Extraction with Transformer Pre-training			2019
2	DyGIE++	48.40	67.50	Entity, Relation, and Event Extraction with Contextualized Span Representations			2019
3	DyGIE	41.6	65.2	A General Framework for Information Extraction using Dynamic Span Graphs			2019
4	SciIE	39.30	64.20	Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction			2018

https://paperswithcode.com/paper/entity-relation-and-event-extraction-with

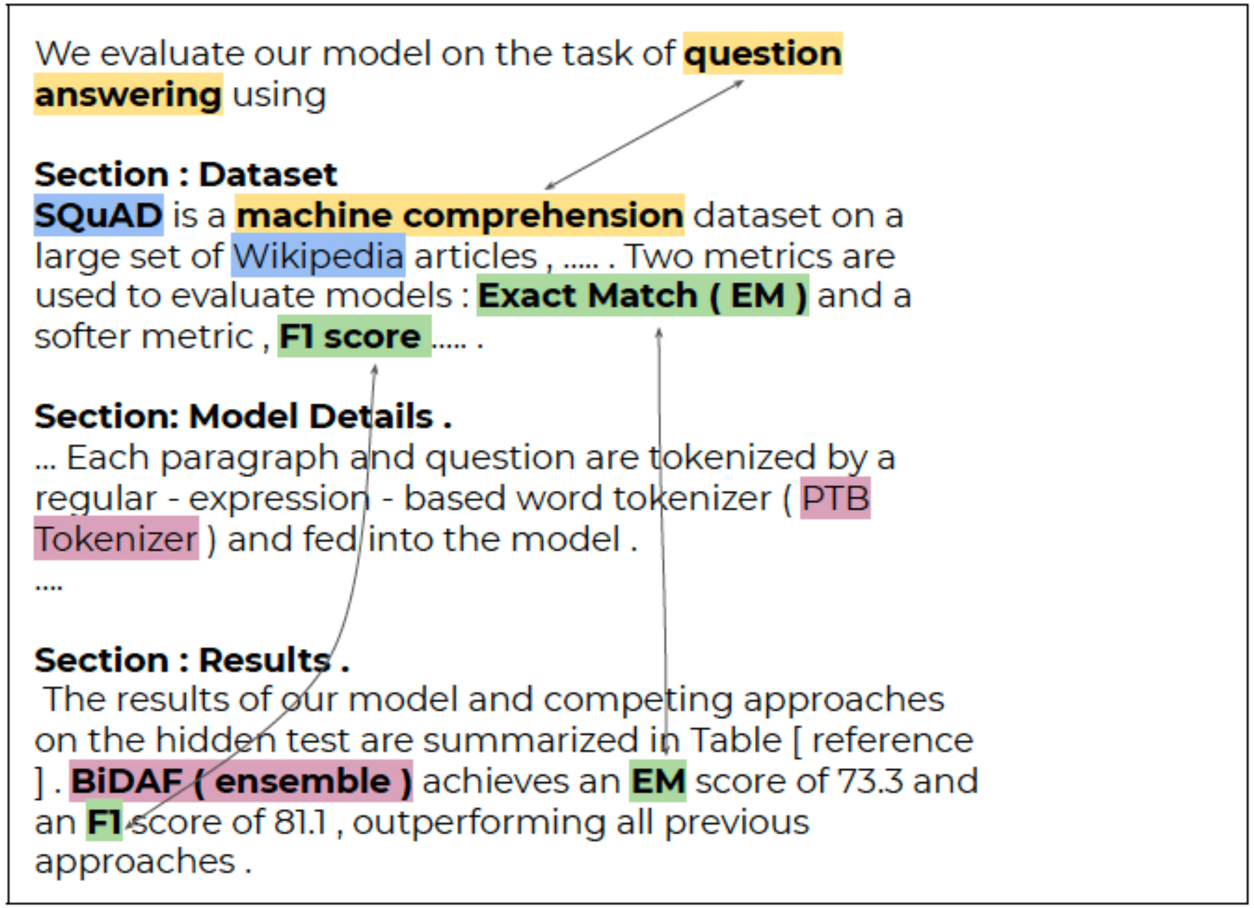


Figure 1: An example showing annotations for entity mentions (Dataset , Metric , Task , Method), coreferences (indicated by arrows), salient entities (bold), and N-ary relation (SQuaD, Machine Comprehension, BiDAF (ensemble), EM/F1) that can only be extracted by aggregating information across sections.

https://allenai.github.io/SciREX/

Method

- 提出了一个端到端的模型

（1）识别mention、mention的显著性和他们的成对共指链接

（2）然后将显著的mention通过识别出来的共指链接进行聚类，变成显著实体。

（3）然后识别这些显著实体之间的篇章级关系

Method

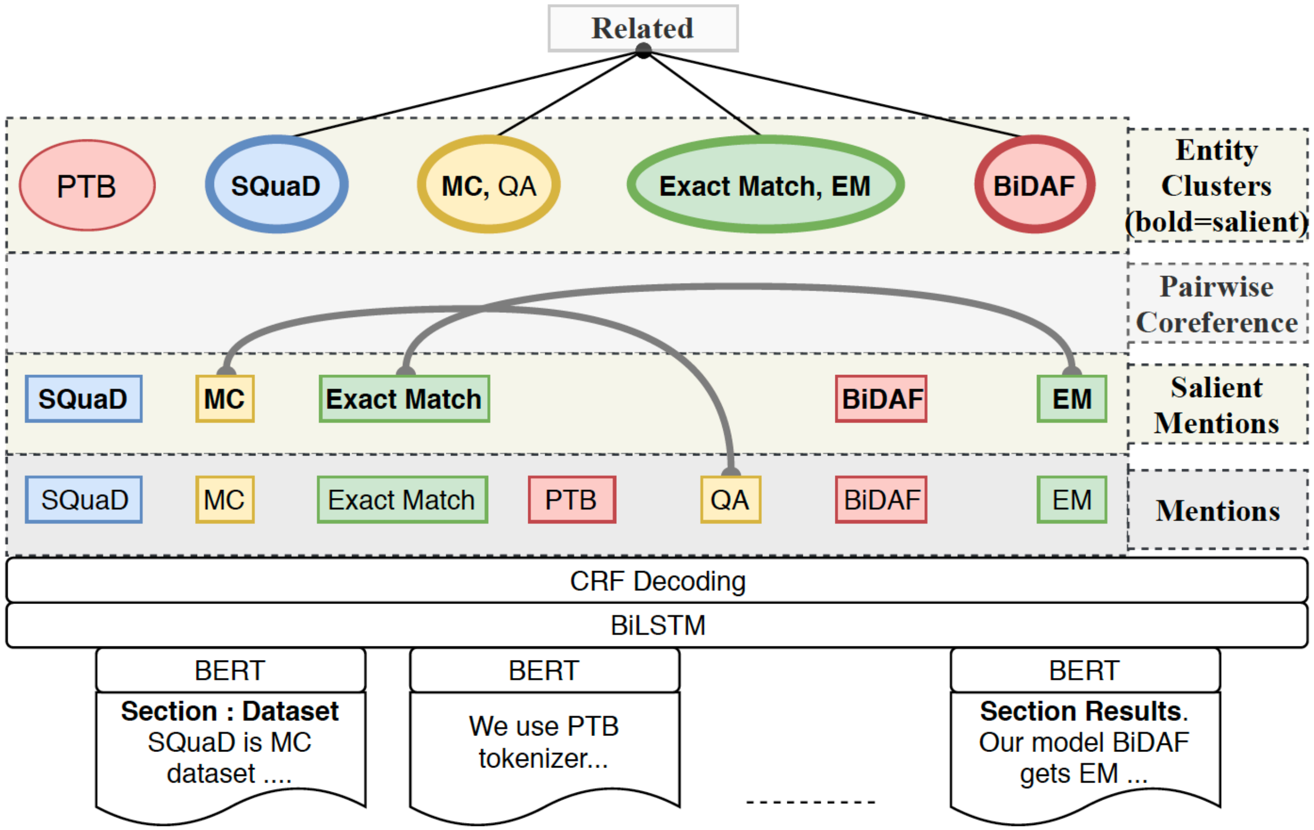


Figure 2: Overview of our model; it uses a two-level BERT+BiLSTM method to get token representations which are passed to a CRF layer to identify mentions. Each mention is classified as being salient or not. A coreference model is trained to cluster these mentions into entities. A final classification layer predicts relationships between 4-tuple of entities (clusters).

Model	P	R	F1
Mention Identification			
DyGIE++	0.703	0.676	0.678
Our Model	0.707	0.717	**0.712**
End-to-end binary relations			
DyGIE++ (Abstracts Only)	0.003	0.001	0.002
DyGIE++ (All sections)	0.000	0.000	0.000
DyGIE++ (SciERC)	0.029	0.128	0.038
Our Model	**0.065**	**0.411**	**0.096**
4-ary relation extraction only			
DocTAET	0.477	**0.885**	0.619
Our Model	**0.531**	0.718	0.611

Table 3: Evaluating state-of-the-art models on subtasks of SciREX dataset because we did not find an existing model that can perform the end-to-end task.

DocTAET (Hou et al., 2019) is a document-level relation classification model that is given a document and a relation tuple to classify if it is expressed in the document. It is formulated as an entailment task with the information encoded as `[CLS] document [SEP] relation` in a BERT style model. This is equivalent to the last step of our model but with gold salient entity clusters as input. Table 3 shows the result on this subtask, and it shows that our relation model gives comparable performance (in terms of positive class F1 score) to that of DocTAET.

Task	Model	P	R	F1
Mention Ident.	DYGIE++	0.676	0.694	0.685
	Our Model	0.637	0.640	0.638
Pairwise Coref. and Clustering	DYGIE++	0.577	0.455	0.476
	Our Model	0.187	0.552	0.255

Table 4: Comparison of DYGIE++ with our model on various subtasks of SCIERC dataset

Task	P	R	F1
Component-wise (gold Input)			
Mention Identification	0.707	0.717	0.712
Pairwise Coreference	0.861	0.852	0.856
Salient Mentions	0.575	0.584	0.579
Salient Entity Clusters	1.000	0.984	0.987
Binary Relations	0.820	0.440	0.570
4-ary Relations	0.531	0.718	0.611
End-to-end (predicted input)			
Salient Entity Clusters	0.223	0.600	0.307
Binary Relations	0.065	0.411	0.096
4-ary Relations	0.007	0.173	0.008
End-to-end (gold salient clustering)			
Salient Entity Clusters	0.776	0.614	0.668
Binary Relations	0.372	0.328	0.334
4-ary Relations	0.310	0.281	0.268

Table 5: Analysis of performance of our model and its subtasks under different evaluation configurations.

Challenges

- This task poses multiple technical and modeling challenges, including
- 1. the use of transformer-based models on long documents and related device memory issues；
- 2. aggregating coreference information from across documents in an end-to-end manner
- 3. identifying salient entities in a document
- 4. performing N-ary relation extraction of these entities.

Challenges

successar commented 23 days ago

Collaborator

Hi

Our model can only be trained on 48Gb GPUs since we apply bert on whole documents (>5000 words on average). You can try to reduce the batch size here

SciREX/scirex/training_config/template_full.libsonnet
Line 98 in eb9f6f3

```
98        batch_size: 50,
```

but I can't say how good the results will be then.

extraction F1 score. All our models were trained using 48Gb Quadro RTX 8000 GPUs. The multitask model takes approximately 3 hrs to train.

Thanks!