


A sequence-to-sequence approach for document-level relation extraction

John Giorgi, Gary D. Bader, Bo Wang

October 18, 2022

Introduction

- Novel end-to-end joint learning approach for inter-sentence relation extraction.¹
- Utilizes sequence to sequence architecture.
- Representation for coreferent entities and n -ary relations (*Linearization schema*).
- Can also handle discontinuous (disjoint) and nested entity spans.

¹Document-level is a stretch, due to encoder limit of 512 tokens they did paragraphs. 

Introduction

- New benchmarks for end-to-end results over some biomedical datasets.
- SOTA for RE with gold entities on two biomedical datasets (DGM, GDA).
- Competitive results against more complex architectures for datasets with established end-to-end and gold entity RE results.

Defining Terms

End-to-end RE:

- Relation extraction depends on entities.
- Pipeline methods (current standard), use one or more models for NER, and one or more models for RE over discovered entities.
- End-to-end approaches use one model (possibly with a classification head) to discover the relations, relying on internal representations to jointly extract and implicitly coordinate entity and relation information.

NB: The authors use *pipeline* to refer to the RE component. In NER/RE practice, pipeline usually refers to the whole system, NER included.

Defining Terms

Coreference:

- The same entity may have one or more mentions in a given text unit (type vs. token).
- If a relation holds between two entities, how to reflect this for each entity's mentions?

Defining Terms

Sequence to sequence (seq2seq):

- Encoder to decoder.
- Encoder maps each input token to a contextual representation.
- Decoder maps each encoder token output and prior context to an output token.
- Sequence cross-entropy loss used in training.

Motivation

- Lots of entity and relation information at the document and cross document level.
- Generalizing sentential pipeline methods (the current standard) for inter-sentential RE is involved.²
- Lots of information takes the form of n -ary relations, not always easy to reconstruct this from binary relations.
- Handling discontinuous/disjoint and nested entity spans is helpful and not entirely solved.

²e.g. our NER/RE system for radiotherapy.

Datasets

- **CDR**
Chemical-induced disease (CID) relations, binary relations.
- **GDA**
Gene-disease associations, binary relations.
- **DGM**
Drug-gene-mutations, ternary relations.
- **DocRED** General domain, binary relations.

Datasets

Table 6: Evaluation datasets used in this paper with details about their annotations. Inter-sentence relations (%) are the fraction of relations in the test set that cross sentence boundaries. We consider a relation intra-sentence if any sentence in the document contains at least one mention of each entity in the relation, and inter-sentence otherwise. *This differs from the estimate in Yao et al. (2019), see Appendix B.

Corpus	Nested Mentions?	Discontinuous Mentions?	Coreferent mentions?	<i>n</i> -ary relations?	Inter-sentence relations (%)
CDR (Li et al., 2016b)	✓	✓	✓	✗	29.8
GDA (Wu et al., 2019)	✓	✗	✓	✗	15.6
DGM (Jia et al., 2019)	✗	✗	✓	✓	63.5
DocRED (Yao et al., 2019)	✗	✗	✓	✗	12.5*

Linearization Schema

X: Variants in the **estrogen receptor alpha (ESR1)** gene and its mRNA contribute to risk for **schizophrenia**.

Y: **estrogen receptor alpha** ; **ESR1** @GENE@
schizophrenia @DISEASE@ @GDA@


Full schema:

$\langle \text{entity mention}_{1,1} \rangle ; \dots ; \langle \text{entity mention}_{1,n} \rangle @ \langle \text{entity type}_1 \rangle @ \dots$
 $\langle \text{entity mention}_{m,1} \rangle ; \dots ; \langle \text{entity mention}_{m,k} \rangle @ \langle \text{entity type}_m \rangle @$
 $@ \langle \text{relation type} \rangle @$

Model Structure

- Seq2seq architectre.
- Encoder: PubMedBERT on DGM, GDA, and CDR. BERT_{BASE} for DocRED.
- Decoder: Single-layer LSTM with randomly initialized weights.
- Generate output token sequence over decoder outputs via beam search at inference time.
- 6 head cross attention mechanism^{3 4} between encoder and decoder.

³<https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture>

⁴https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html#attention_idea 

Model Structure

- Vocabulary restriction: Decoder can only generate special tokens and tokens from input (*copy mechanism*).
- Sorting relations within target strings according to order of appearance in the text⁵.
- Constrained decoding (not used by default): Prevention of syntactically invalid output strings by setting invalid decoder output token scores to near zero.

⁵Authors pick a convention since cross-entropy loss is permutation-sensitive

Training Strategy

- All parameters trained jointly via AdamW optimizer.
- lr is linearly increased for first 10% of training steps, linearly decayed to zero for the rest.
- Top L layers of pre-trained encoder re-initialized before fine-tuning.
- Gradients are scaled to a vector norm of 1.0 before backpropagating.

Training Strategy

- Every forward propagation, hidden state of the LSTM decoder is initialized with the mean of encoder's token embeddings output.
- Decoder uses dropout with probability 0.1 to inputs, and DropConnect with probability 0.5 to the hidden-to-hidden weights.
- Teacher forcing used for decoder at training time⁶.
- Beam search used for output generation at inference time.

⁶<https://cedar.buffalo.edu/~srihari/CSE676/10.2.1%20TeacherForcing.pdf>

RE on Gold Entities with Entity Hinting

X: **estrogen receptor alpha** ; **ESR1** @GENE@
schizophrenia @DISEASE@ @SEP@ Variants in the **estrogen receptor alpha** (**ESR1**) gene and its mRNA contribute to risk for **schizophrenia**.

Full schema:

$\langle \text{entity mention}_{1,1} \rangle ; \dots ; \langle \text{entity mention}_{1,n} \rangle @ \langle \text{entity type}_1 @ \rangle \dots$
 $\langle \text{entity mention}_{m,1} \rangle ; \dots ; \langle \text{entity mention}_{m,k} \rangle @ \langle \text{entity type}_m @ \rangle \dots$
 @SEP@ $\langle \text{input text} \rangle$

Hinting is omitted for end-to-end.

n -ary Relations (DGM)

Method	P	R	F1
Jia et al. (2019) [†]	62.9	76.2	68.9
seq2rel (entity hinting)	84.0	84.8	84.4
seq2rel (entity hinting, relaxed)	84.1	84.9	84.5
seq2rel (end-to-end)	68.9	65.9	67.4
seq2rel (end-to-end, relaxed)	78.3	74.9	76.6

DGM has ternary relations. Baseline, Jia et al. (2019) uses multiscale architecture (uses multiple representations over different sizes of text spans and types of sub-relations). Both use gold entities (entity hinting in seq2rel case).

RE with Gold Entities (CDR, GDA)

Method	CDR			GDA		
	P	R	F1	P	R	F1
Christopoulou et al. (2019)	62.1	65.2	63.6	–	–	81.5
Nan et al. (2020)	–	–	64.8	–	–	82.2
Minh Tran et al. (2020)	–	–	66.1	–	–	82.8
Lai and Lu (2021)	64.9	67.1	66.0	–	–	–
Xu et al. (2021)	–	–	68.7	–	–	83.7
Zhou et al. (2021)	–	–	69.4	–	–	83.9
seq2rel (entity hinting)	68.2	66.2	67.2	84.4	85.3	84.9
seq2rel (entity hinting, relaxed)	68.2	66.2	67.2	84.5	85.4	85.0
seq2rel (end-to-end)	43.5	37.5	40.2	55.0	55.4	55.2
seq2rel (end-to-end, relaxed)	56.6	48.8	52.4	70.3	70.8	70.5

(Not enough room, full breakdown of baseline approaches in paper appendix G)

End-to-end RE (DocRED)

Method	P	R	F1
JEREX (Eberts and Ulges, 2021)	42.8	38.2	40.4
seq2rel (end-to-end)	44.0	33.8	38.2
seq2rel (end-to-end, relaxed)	53.7	41.3	46.7

End-to-end RE (DocRED)

JEREX = BERT with four joint task-specific FFNN-based components in the following order:

- entity mention localization,
- coreference resolution
- entity classification
- relation classification.

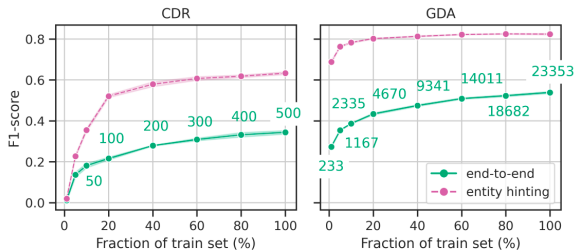
Two versions of the relation classifier, “global relation classifier” (GRC) and “multi-instance relation classifier” (MRC). The authors compare against JEREX-MRC for DocRED end to end.

End-to-end RE Ablation (CDR, DocRED)

	CDR				DocRED			
	P	R	F1	Δ	P	R	F1	Δ
seq2rel (end-to-end)	41.0	35.1	37.8	–	46.9	36.1	40.8	–
- pretraining	9.4	6.9	8.0	-29.8	18.5	7.7	10.8	-30.0
- fine-tuning	24.3	20.5	22.2	-15.6	42.4	15.5	22.7	-18.1
- vocab restriction	39.6	32.2	35.5	-2.3	45.2	35.5	39.7	-1.1
- sorting relations	36.1	29.2	32.3	-5.6	52.9	17.4	26.2	-14.7
+ constrained decoding	40.8	35.6	38.0	+0.2	46.8	35.9	40.6	-0.2

- Fine-tuning here is wrt the encoder.
- Vocab restriction = special tokens + copy mechanism.
- Sorting relations = sorting relations by order of appearance for consistent decoding order.
- Constrained decoding = prevention of syntactically invalid output strings.

Training Set Size vs. Performance (CDR, GDA)



Due to logarithmic shape of e2e f1 against train set size authors believe e2e has potential for better scores (vs. asymptotic curves with entity hinting).

Hyperparameter Tuning

Table 7: Hyperparameter values used for each corpus. Hyperparameters values when using entity hinting, if they differ from the values used without entity hinting, are shown in parentheses. Tuned indicates whether or not the hyperparameters were tuned on the validation sets.

Hyperparameter	Tuned?	CDR	GDA	DGM	DocRED
Batch size	✓	4	4	4	4
Training epochs	✓	130 (70)	30 (25)	30 (45)	50
Encoder learning rate	✗	2e-5	2e-5	2e-5	2e-5
Encoder weight decay	✗	0.01	0.01	0.01	0.01
Encoder re-initialized top L layers	✓	1	1 (2)	1	1
Decoder learning rate	✓	1.21e-4 (1.13e-4)	5e-4 (4e-4)	8e-4 (1.5e-5)	7.8e-5
Decoder input dropout	✗	0.1	0.1	0.1	0.1
Decoder hidden-to-hidden weights dropout	✗	0.5	0.5	0.5	0.5
Target embedding size	✗	256	256	256	256
No. heads in multi-head cross-attention	✗	6	6	6	6
Beam size	✓	3 (2)	4 (1)	3 (2)	8
Length penalty	✓	1.4 (0.2)	0.8 (1.0)	0.2 (0.8)	1.4
Max decoding steps	✗	128	96	96	400

Strengths

Pure RE (with gold entity hinting)

- SOTA on GDA.
- Highly competitive on CDR.
- Markedly simpler architecture than the baselines ⁷.

End-to-end

- Competitive result against SOTA JEREX.
- Simpler architecture than JEREX.

Linearization schema is easy to interpret. Can handle discontinuous/disjoint and nested entities.

⁷Most construct a document-level graph using dependency parsing, heuristics, or structured attention and then update node and edge representations using propagation. Xu et al. 2021 and Zhou et al. 2021 make modifications to transformer architecture and BERT architecture respectively

Limitations

Training strategy:

- Rationale not fully explained.
- Could use variation in lr scheduling⁸ and other hyperparameters.
- Additional LSTM directionality and layering variation.
- Cross-attention mechanism variation.
- Teacher forcing, while common for seq2seq, gives faster convergence but possibly lower performance at inference due to exposure bias.

⁸Some RT NER models rescued by cosine lr schedule

Limitations

Model structure:

- 512 token limitation.
- Bevy of architectures (especially transformer-based) for long document processing⁹

Linearization schema:

- Ideally want order invariant scoring.
- Will this generalize to long documents with frequent correferents?
- $n > 3$ -ary relations?

⁹E.g. Tim and Angus have extensive experience with Hierarchical Transformers <https://arxiv.org/abs/2110.13711>

Conclusion

- Novel application of relatively well understood technique for an increasingly relevant task, esp. for clinical documents.
- Many possible variations of architecture and application.
- Possible to test features/methods from baselines¹⁰ within this architecture.
- Implementation and behavior tracking easier than competition.
- More efficient than a lot of pipeline models¹¹.

Full bibliography in paper.

¹⁰dependency parses, document graphs

¹¹E.g. at least one LPLM for NER, at least one for RE

Other Helpful Links

- Seq2seq:
https://d2l.ai/chapter_recurrent-modern/seq2seq.html#sec-seq2seq
- Beam search:
https://d2l.ai/chapter_recurrent-modern/beam-search.html