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

John Giorgi, Gary D. Bader, Bo Wang

October 13, 2022

Introduction

- Novel end-to-end joint learning approach for inter-sentence relation extraction.[1]
- Utilizes sequence to sequence architecture.
- Representation schema for coreferent entities, *n*-ary relations, and disjoint spans in output.

[1] Document-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.
- Competitive results against more complex architectures for datasets with established end-to-end 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 component 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 entites, how to reflect this for each entity's mentions?

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 very tricky.[2]
- Lots of information takes the form of *n*-ary relations, tricky to reconstruct this from binary relations.

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

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@

Linearization Schema Cont.

Schema is:
$< \textit{entity mention}_{1,1} >$; ... ; $< \textit{entity mention}_{1,n} > < \textit{entitytype}_1 > \ldots$
$< \textit{entity mention}_{m,1} >$; ... ; $< \textit{entity mention}_{m,k} > < \textit{entitytype}_m >$
$< \textit{relationtype} >$

Model Structure

Evaluation

Datasets

Conclusion

References I