DeepStruct: Pretraining of Language Models for Structure Prediction

Chenguang Wang*, Xiao Liu*, Zui Chen*, Haoyun Hong, Jie Tang, Dawn Song

UC Berkeley Tsinghua University

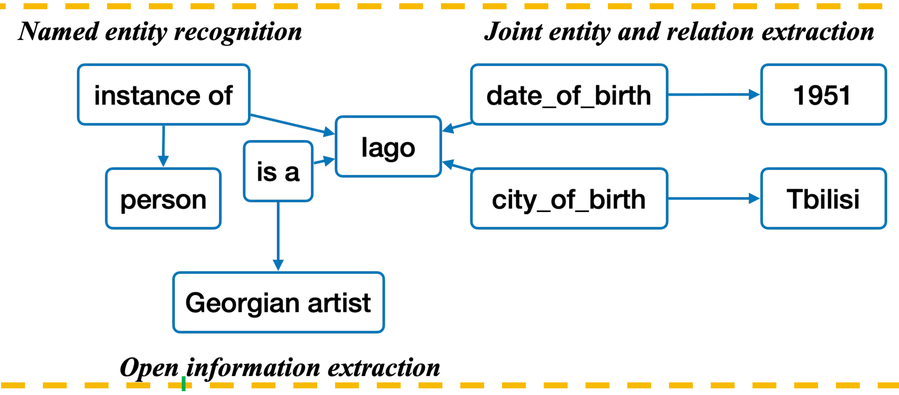
Code

Traditional Understanding

Born in 1951 in Tbilisi, Iago is a Georgian ____. → Language Model → artist

Structure Understanding

Born in 1951 in Tbilisi, Iago is a Georgian artist. → Language Model →



e.g., GPT-3

Predict single words from text

DeepStruct

Predict structures from text

Structure Representation

Joint entity and relation extraction

Born in 1951 in Tbilisi, Iago is a Georgian artist. (Iago; instance of; person) (Tbilisi; instance of; city) (Iago; city_of_birth; Tbilisi)

Named entity recognition

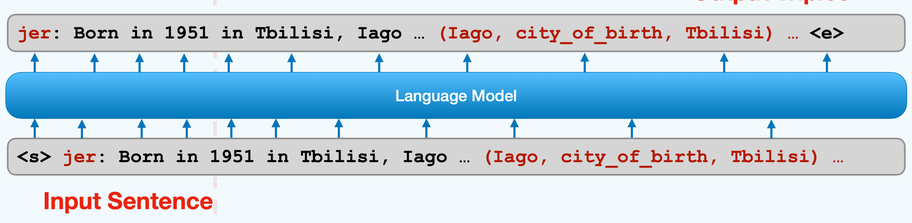
Born in 1951 in Tbilisi, Iago is a Georgian artist. (Iago; instance of; person) (Tbilisi; instance of; city)

Open information extraction

Born in 1951 in Tbilisi, Iago is a Georgian artist. (Iago; is a; Georgian artist)

...

Autoregressive Training



Task-agnostic Training Data

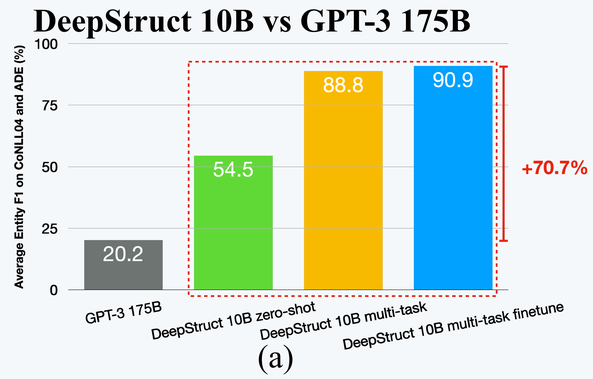
- ~ 51M sentences
- ~ 134M entities
- ~ 114M relations (triples)

Multiple Tasks

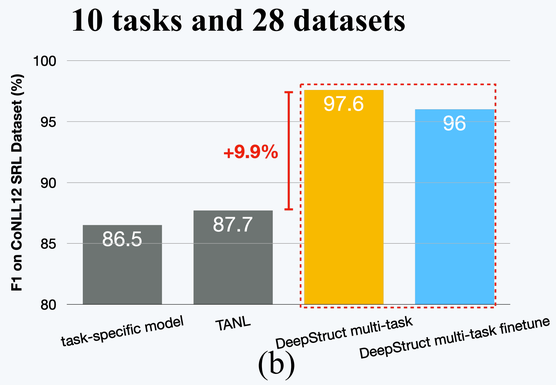
- 10 tasks
- 28 datasets
- ~ 700K sentences

Results:

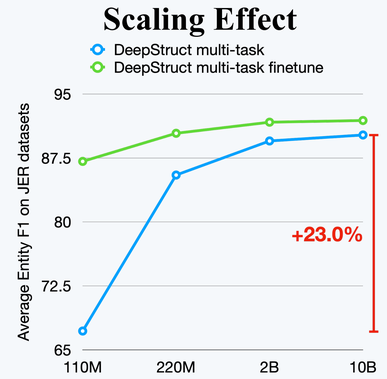
- DeepStruct 10B largely outperforms GPT-3 175B model. (a)
- Achieved state-of-the-art result on 21 of 28 datasets. (b)
- DeepStruct performance increases drastically along with model size. (c)

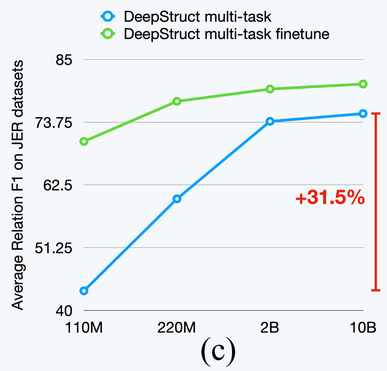


(a)



(b)





(c)