Neural ParsCit: A Deep Learning Based Reference String Parser

Animesh Prasad, Manpreet Kaur and Min-Yen Kan
School of Computing, National University of Singapore

❖ Background

- ➢ Extracting information from citation strings is a cornerstone task
- ➢ State of the art systems use eloquent hand made features
- ➢ However the key issue of finding optimal representation able to encompass all such features needs to be addressed

❖ Motivation

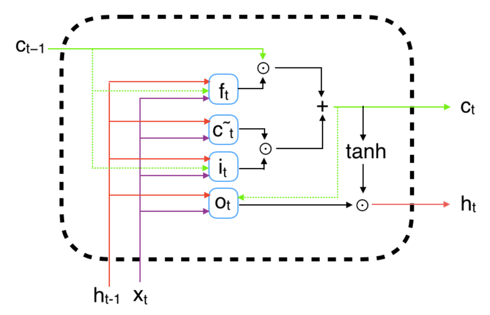
Councill IG, Giles CL, Kan MY (2008) *ParsCit: an open-source CRF reference string parsing package*. *In: LREC*, *vol 8*, *pp 661–667*

Why not have best of both worlds?

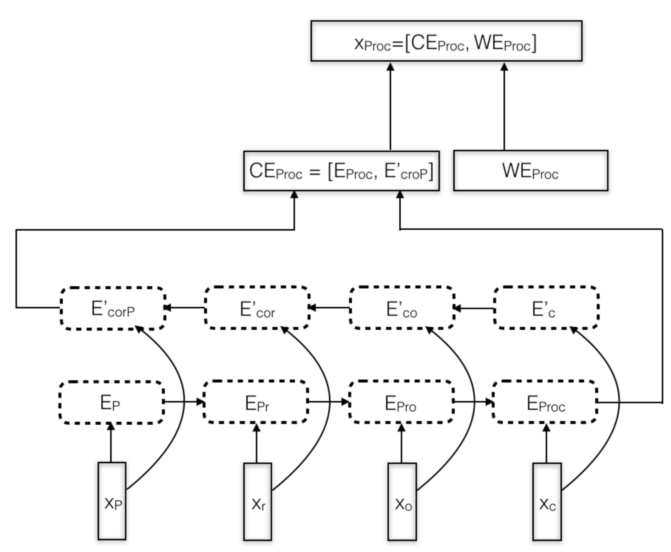
	Linear Chain Conditional Random Field (CRF)	Long Short Term Memory (LSTM)
Pros	Robust, Works with less Gold data	Can incorporate rich semantics (via word embeddings), Can model long range dependencies
Cons	Uses only hand crafted surface level features	Requires lot of Gold data

❖ Proposed Technique

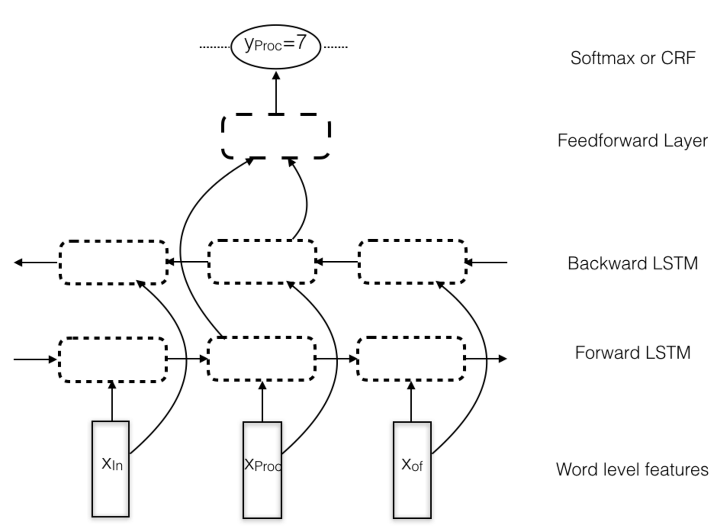
- ➢ **Features**
  - ➢ Word Embeddings and Character Based Word Embeddings
  - ➢ Augmented Word Embeddings
- ➢ **Model**
  - ➢ Bidirectional LSTM output fed to a CRF layer



A simple LSTM cell



An unrolled LSTM across time steps, showing calculation of character based word embedding, similar calculation is performed by LSTM layers on word embeddings as well



Final model with all the layers

❖ Results

- ➢ **Results on Different Model Configurations**
  - ➢ 10 Fold cross validation with 80:10:10 split
  - ➢ Tested on Cora dataset with 13 target labels

Setup	Configuration	Performance (Neural ParsCit) Macro F1	Micro F1
A	LSTM+WE	51.77	80.39
B	A+Augmented WE	60.47	85.11
C	B+Character Based WE	59.60	89.11
D	C-LSTM+Bidirectional LSTM	68.54	88.43
E	D+CRF	90.45	95.68

- ➢ **Results on Different Model Configurations**
  - ➢ 10 Fold cross validation with 90:10 split on best configuration from ablation studies
  - ➢ Tested on Cora + Flux-CiM + ICONIP + cross domain (humanities) + multilingual (Italian + French + German)

	Performance (ParsCit)			Performance (Neural ParsCit)		
Label	P	R	F1	P	R	F1
Author	98.78	98.94	**98.86**	99.32	98.28	98.8
Booktitle	94.19	93.13	93.66	94.41	95.89	**95.15**
Date	97.94	97.92	97.93	98.87	98.23	**98.55**
Editor	90.67	92.86	**91.75**	88.01	92.02	89.97
Institution	78.29	95.77	86.15	91.44	90.97	**91.2**
Journal	91.92	91.45	**91.68**	90.88	91.86	91.37
Location	93.01	92.23	92.62	94.95	92.71	**93.81**
Note	58.69	87.54	70.27	65.62	93.61	**77.15**
Pages	98.59	98.39	98.49	99.27	98.56	**98.91**
Publisher	78.49	92.32	84.85	93.87	92.82	**93.35**
Tech	68.19	92.77	78.61	83.16	86.16	**84.63**
Title	97.85	95.5	96.66	98.34	97.16	**97.74**
Volume	95.24	94.26	94.75	95.49	95.37	**95.43**
Micro			95.48			**96.47****

** p<0.01 on paired student's t-test

❖ Conclusions

- ➢ **Model is able to handle long range dependencies**
  - ➢ Labels farther from beginning show improvement on an average
- ➢ **Features cover both the syntactic and semantic salience**
  - ➢ Similar or Improvement performance on Labels where surface string show elaborate syntactic features
  - ➢ Good performance on semantically close classes (which are otherwise similar on surface level) resulting from incorporating semantics

Sourcecode: https://github.com/opensourceware/Neural-ParsCit
Demo: http://wing.comp.nus.edu.sg/parsCit

Email: animesh@comp.nus.edu.sg
Website: wing.comp.nus.edu.sg