Entity Recognition and Linkage for Reference data

Under the guidance of Dr. Zeyd Boukhers

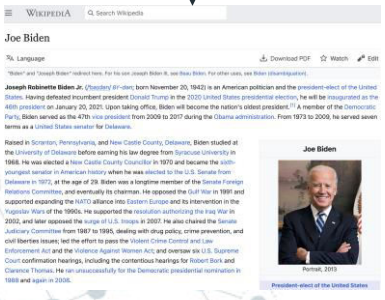
The Beatles:

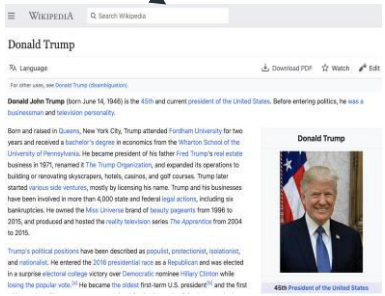
Nagaraj Bahubali Asundi
Sriram Aralappa
Adarsh Anand
Soniya Manchenahalli Gnanendra Prasad
Abinaya Thulsi Chandrasekaran

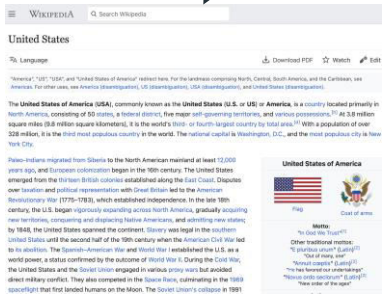
Entity Linking (EL)

- EL - linking entity mention to its unique target entity in the Knowledge Base
- Entity Mentions - terms in the text that refers to real world entities
- Knowledge Base(KB) - a repository where information is stored, organized and shared such as Wikipedia, DBLP etc.

e.g. Joe Biden defeats Trump to become the President-elect of The United States.







KB

Entity Linking for Reference data

- References: a citation string that contains bibliographic information of a scientific paper
- Aim: Link author entities in references to target entities in DBLP

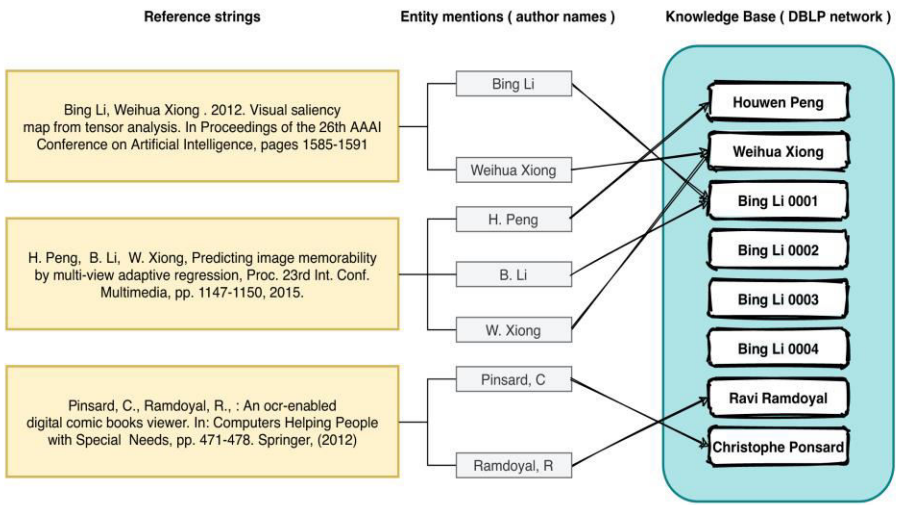
e.g. D. Micciancio, Hao Chen , Adaptive Security of Symbolic Encryption, IEEE Security and Privacy, 2018

KB

Motivation and Challenges

Author name ambiguity

- **synonyms**: same author with different name variations
  e.g. 'Weihua Xiong' and 'W. Xiong'
- **homonyms**: distinct authors sharing the same name
  e.g. 'Bing Li' may refer any of the 4 'Bing Li's



Typos

- Incorrectly written author names
  e.g. Christophe Ponsard written as Christophe Pinsard

Motivation

- Quality of scientific data gathering
- Incorrect identification and credit attribution to authors

Reference2Auth

- Maps author name in reference to its target author in DBLP
- Supervised deep learning model
- Uses citation attributes such as co-authors, title and journal
- Captures co-author patterns and semantic features of title and journal

Data Preparation

Data Collection

- Dump of DBLP bibliographic database (4.4 million records)
- Stored these references in Mongo Database to handle schema less records
- Selected references of top 40 authors – 2K

Feature Generation

- Convert each reference into useful feature
- A reference can contain variable length of authors
- An author can be represented in various styles

Generation of input samples for the model

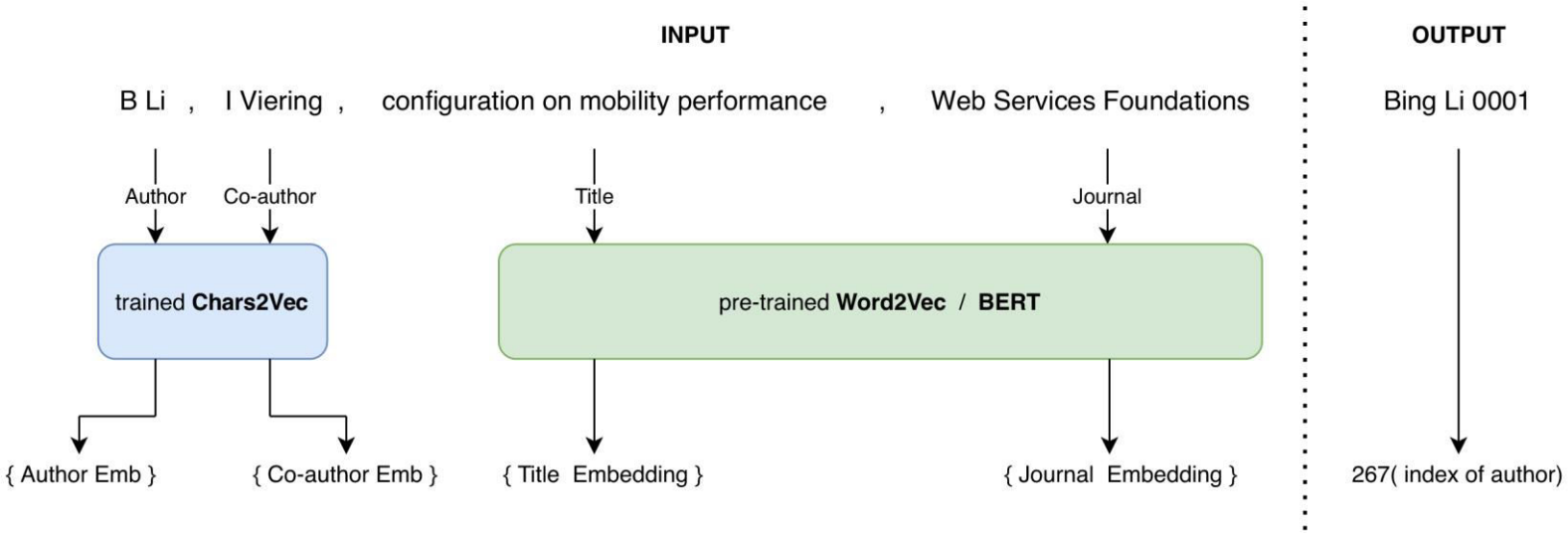
Reference string

{"Author": ["Bing Li 0001", "Ingo Viering", "Meryem Simsek"], "Title": "configuration on mobility performance", "journal": "Web Services Foundations", "Year": "2017"}

Each author name may appear in any of the following forms when cited in a paper.

Input samples	Target author name
[B Li, I Viering, configuration on mobility performance, Web Services Foundations]	Bing Li 0001
[L Bing, Ingo V, configuration on mobility performance, Web Services Foundations]	Bing Li 0001
.	.
.	.
[I Viering, F Berhanu,configuration on mobility performance, Web Services Foundations]	Ingo Viering
[V Ingo, M Simsek,configuration on mobility performance, Web Services Foundations]	Ingo Viering
.	.
.	.
[M Simsek, F Berhanu, configuration on mobility performance, Web Services Foundations]	Meryem Simsek
[S Meryem, I Viering, configuration on mobility performance, Web Services Foundations]	Meryem Simsek

Embeddings for the samples



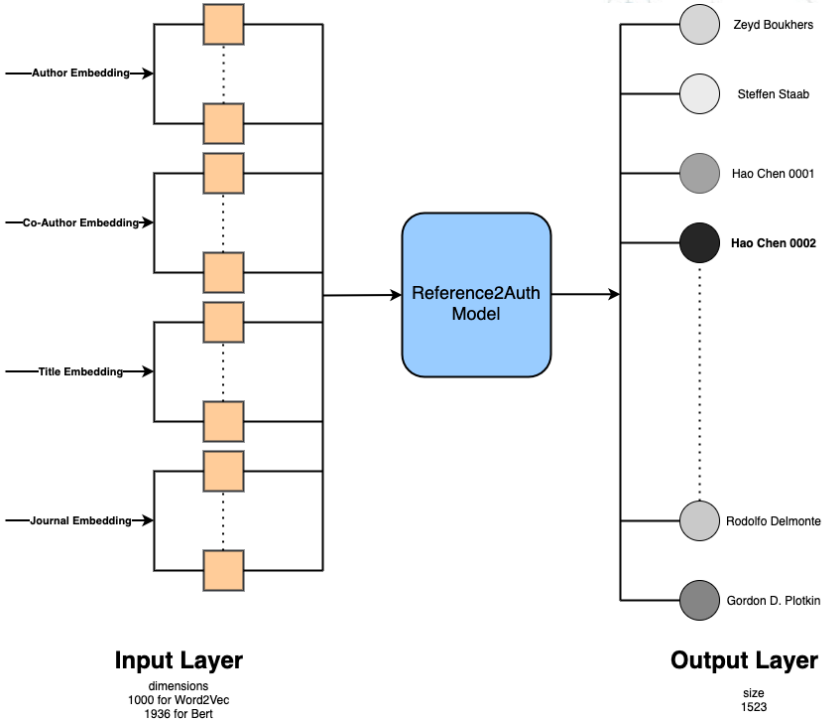
Training of Chars2Vec

- Form positive and negative pairs

  Bing Li and B Li → +ve Bing Li and Hao Chen → -ve

Architecture

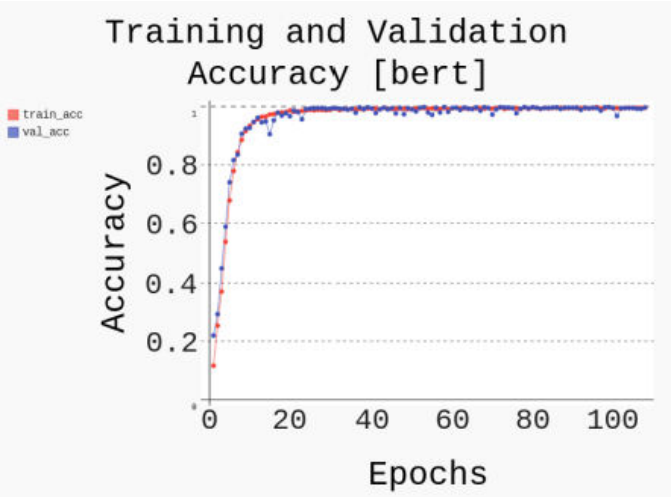
- Sequential, two-layer fully connected deep neural network
- Input - embeddings of author, coauthor, title, journal
- Output – number of unique target author entities
- For each input sample, output is the index value of the target author

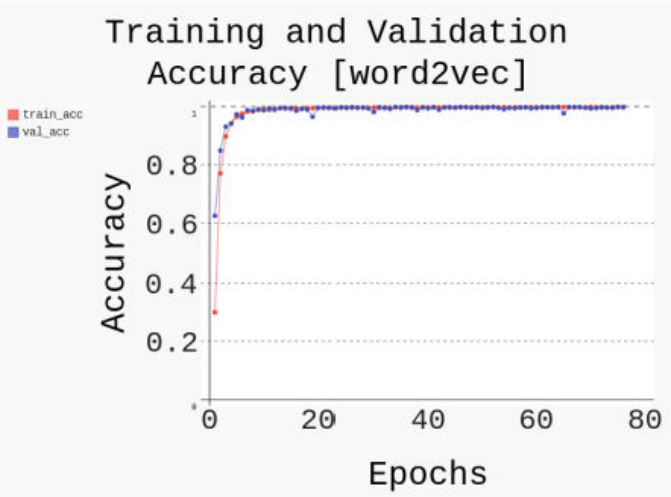


Implementation

- Two versions of embeddings for title and journal – Word2vec and BERT
- Input size is 1000 and 1936 with Word2vec and BERT respectively
- Enabled ‘early stopping’ and ‘model checkpoint’ to avoid overfitting and underfitting

Accuracy Results





	BERT	Word2Vec
Validation	**99.65%**	**99.86%**
Testing	**99.85%**	**99.87%**

Future Enhancements

- Training the model with entire paper instead of title
- Train on entire DBLP dataset
- Address the dynamic nature of bibliographic repositories

Demonstration

- Typos :

  authors : Fahiem Bacchus, Shannon Dalmao, Toniann Pitassi
  title : Value Elimination: Bayesian Inference via Backtracking Search
  journal : UAI

- Synonyms :

  authors : Fahiem Bacchus, Shannon Dalmao, Toniann Pitassi
  title : Value Elimination: Bayesian Inference via Backtracking Search
  journal : UAI

- Homonyms :

  authors : **Bing Li 0001**, Rongrong Ji
  title : Predicting the effectiveness of queries for visual search
  journal : ICASSP

  authors : **Bing Li 0010**, Jin Liu
  title : Research on Semantic-Based Web Services Registry Federation
  journal : GCC

Thank You!

Questions?