THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Mining for Entity Relationship Associations

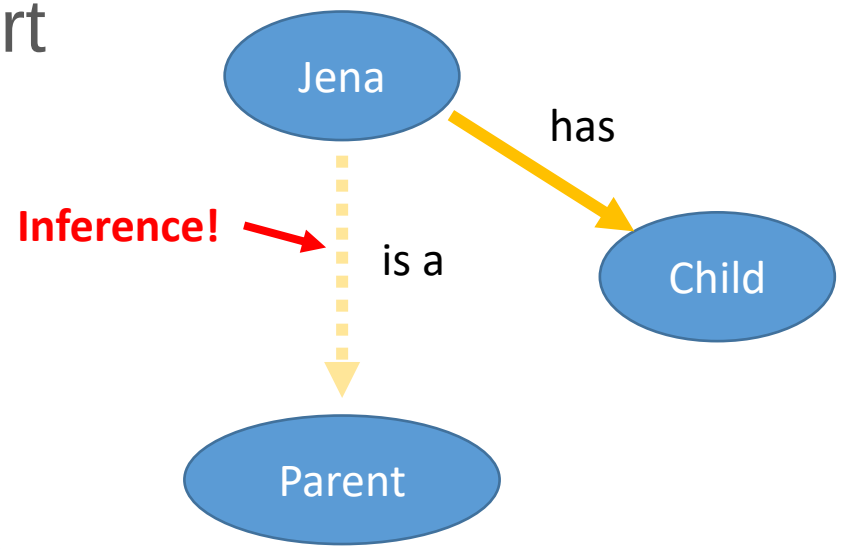
School of Engineering and Applied Science
Department of Computer Science CSCI 6443– Data Mining

Professor: A. Bellaachia

Student: R. Gross (G47667332)

Problem Definition

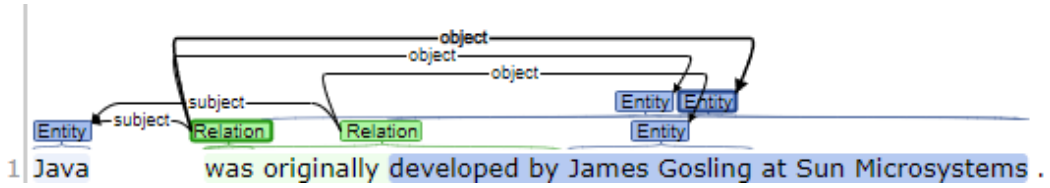
- Many chatbots are a combination of expert systems and machine learning.
- A knowledge base is often used as the “brain” of the chatbot due to its ability to perform inference.
- Traditionally knowledge bases perform inference based on inference rules, which are brittle and don’t scale well.



IF <subject> has Child
THEN <subject> is a Parent

Problem Definition Continued

- Unsupervised learning of entity relationships is difficult and supervised learning datasets are costly to create.
- Performance is subjective and language dependent.
- State-of-the-art NLP algorithms struggle to perform Relationship Extraction (RE) with the precision and recall of a person.



https://corenlp.run/

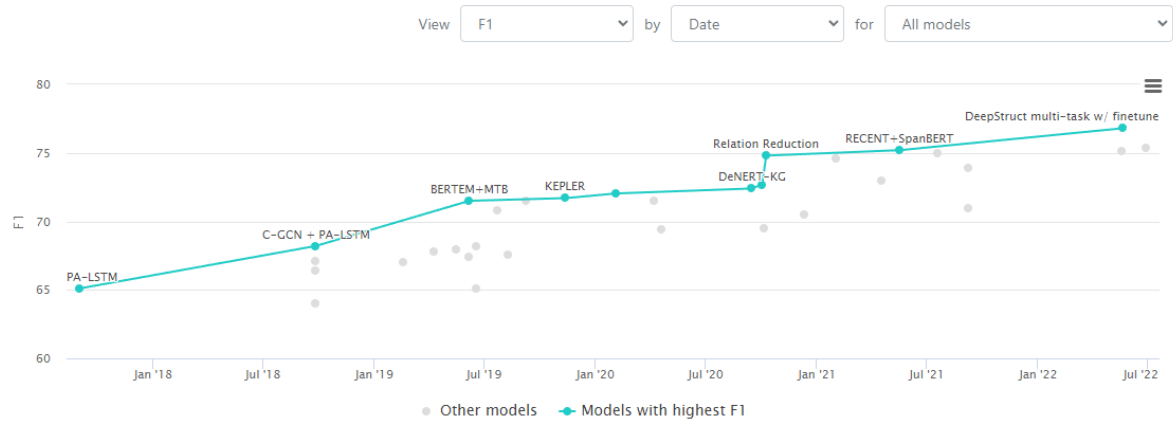
Related Work

- Datasets (not all are free):
  - https://paperswithcode.com/datasets?task=relation-extraction
- Papers:
  - https://paperswithcode.com/task/relation-extraction#papers-list
- Notable Algorithms Types:
  - OpenIE
  - Convolutional Neural Network (CNN)
  - Sequence based networks (RNN, LSTM, GRU, Transformers)

Trend	Dataset	Best Model	Paper	Code	Compare
	DocRED	KD-Rb-l			See all
	TACRED	DeepStruct multi-task w/ finetune			See all
	ACE 2005	PL-Marker			See all
	NYT	DeepStruct multi-task			See all

Relation Extraction on TACRED

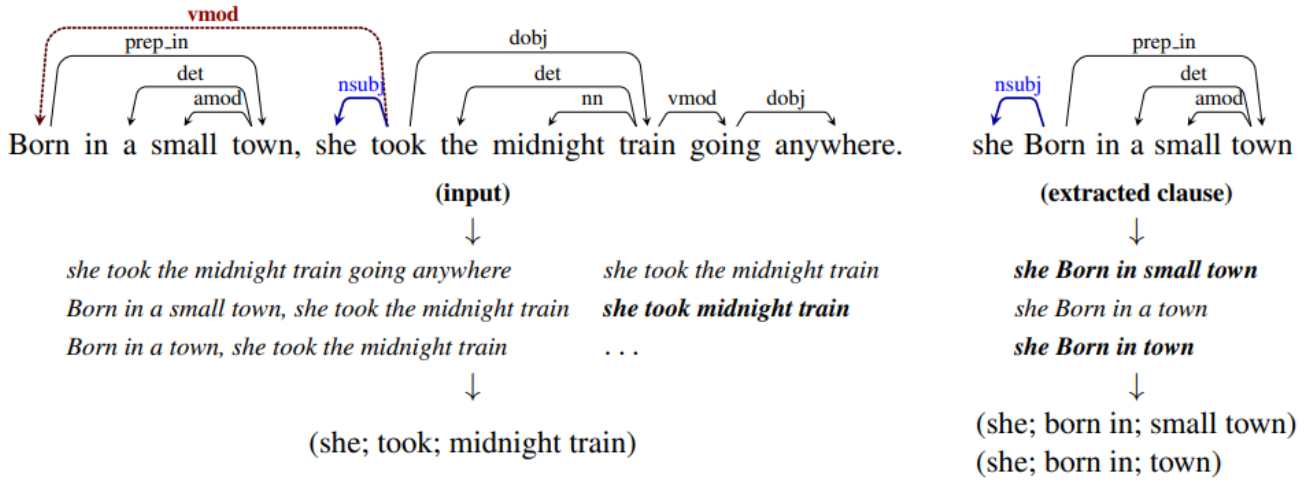
Leaderboard Dataset



https://paperswithcode.com/task/relation-extraction

Related Work Continued

- Methods of doing RE :
  - OpenIE
    - Greedy search on dependency tree
    - Goal is to reduce sentence, to utterance, to triple



https://nlp.stanford.edu/pubs/2015angeli-openie.pdf

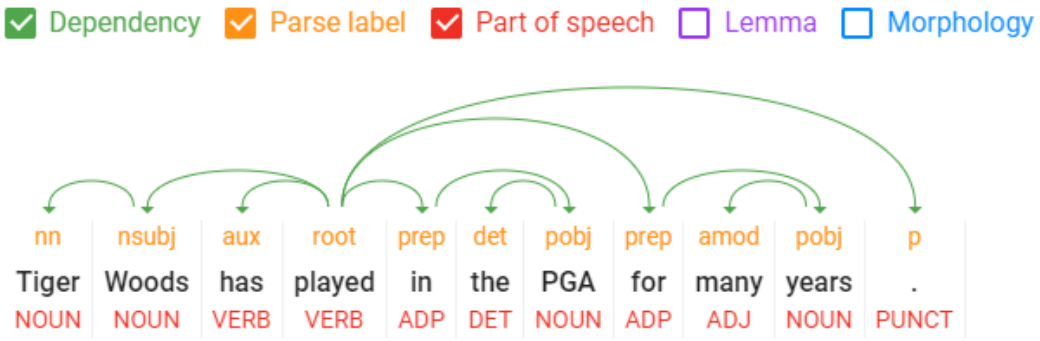
Related Work Continued

- Methods of doing RE:
  - Named Entity Extraction (NER) + Classification of relations between entities using neural networks
    - Text and/or Part of Speech (POS) tags are input features along with entities
  - Graph Convolutional Neural Networks (GCN)
    - Create graph from dependencies
    - GCN learns features and structures in graph that are associated to training data

⟨Tiger Woods⟩$_1$ has played in the ⟨PGA⟩$_2$ for many years.

NER Example

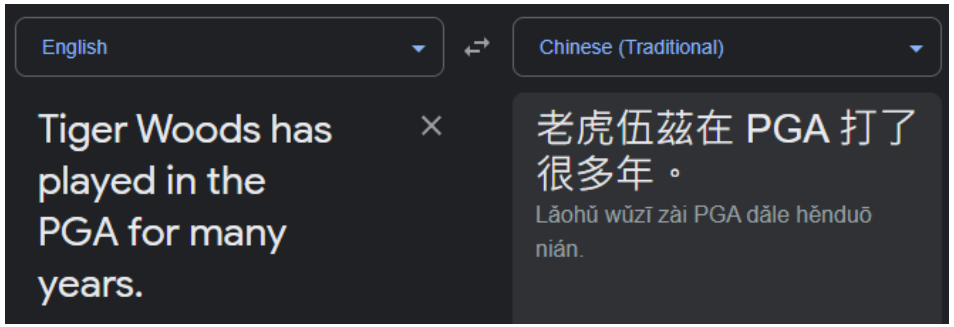
https://cloud.google.com/natural-language



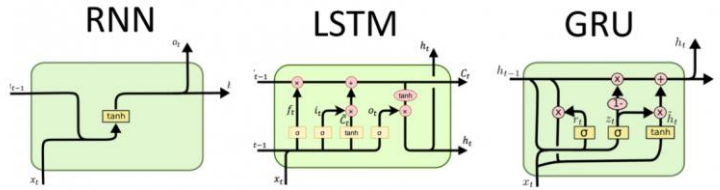
https://cloud.google.com/natural-language

Related Work Continued

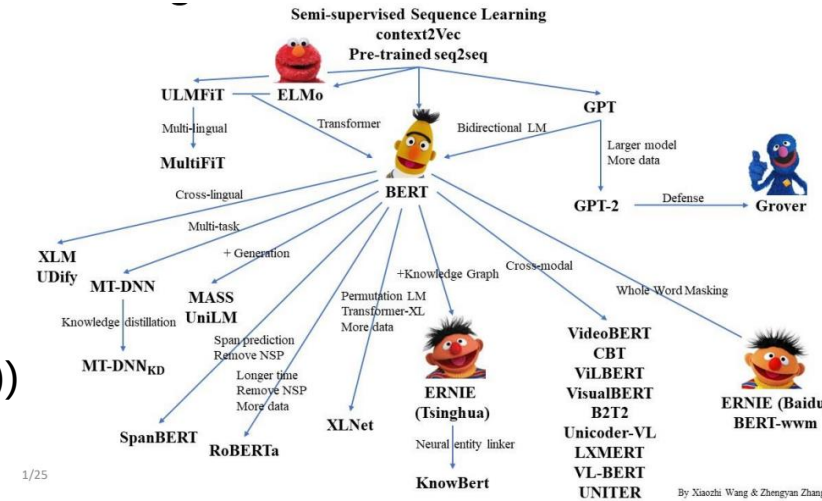
- Methods of doing RE :
  - Applying language generation/translation techniques
    - RNN, LSTM, GRU, Transformers (Seq2Seq (BERT, ERNIE, GPT, BART))



https://translate.google.com/



http://dprogrammer.org/rnn-lstm-gru



https://www.microsoft.com/en-us/research/uploads/prod/2021/06/Pre-training-Models-Xu-Tan.pdf

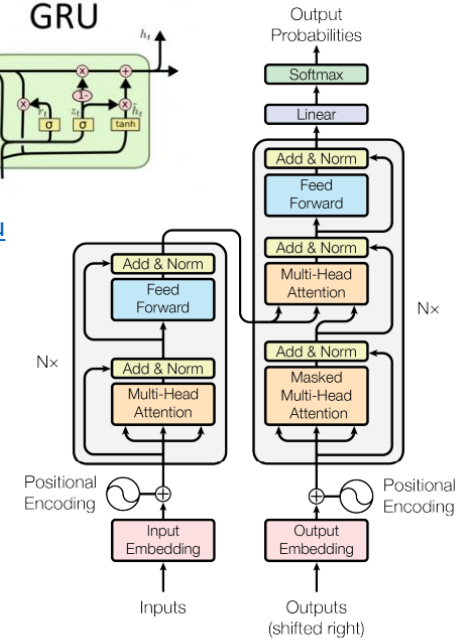
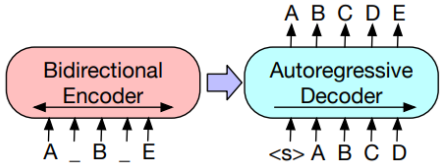


Figure 1: The Transformer - model architecture.

https://arxiv.org/pdf/1706.03762.pdf



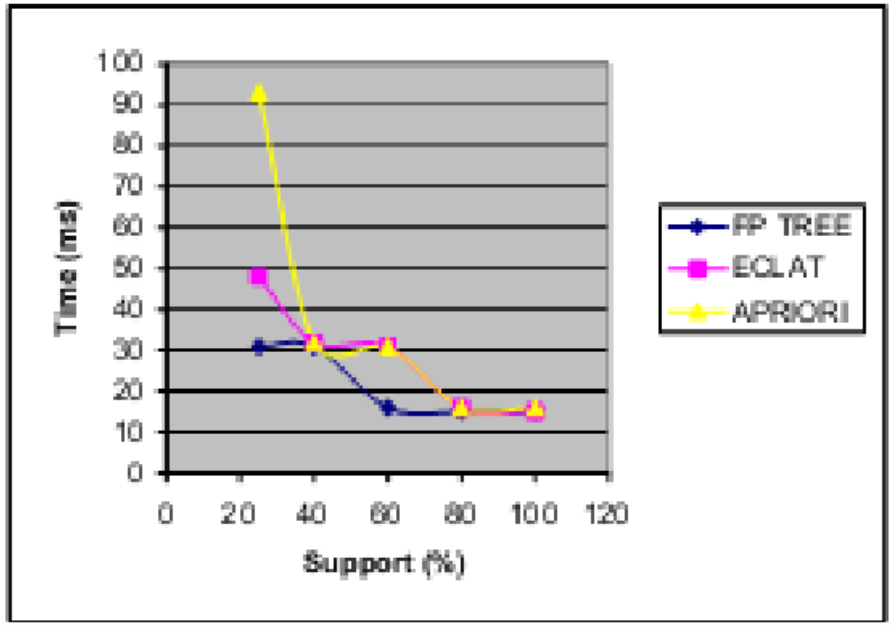
BART

https://arxiv.org/pdf/1906.07510v8.pdf

Related Work Continued

- After RE, associations between relations occurs.
- Leading association algorithms are:
  - Apriori
  - FP Growth
  - Eclat

Figure 1. Comparison of Apriori, Eclat and FP Growth algorithm on artificial dataset.



https://research.ijcaonline.org/volume69/number25/pxc3888502.pdf

Proposed Approach

- Text already mined and transforming to JSON files

Text Dataset

- Zipped text moved to AWS cloud environment
- This allows servers to access as needed with higher download speeds

Amazon Simple Storage Service (Amazon S3)

- Download data
- Unzip
- Split into manageable data message sizes
- Publish text to queue

Amazon Elastic Compute Cloud (Amazon EC2)

- High performance queue
- Allows consuming service to be parallel by ensuring only one consumer gets each message

Amazon Simple Queue Service (Amazon SQS)

- Perform Relation Extraction (RE) to convert unstructured data to structured data
- Write to DB

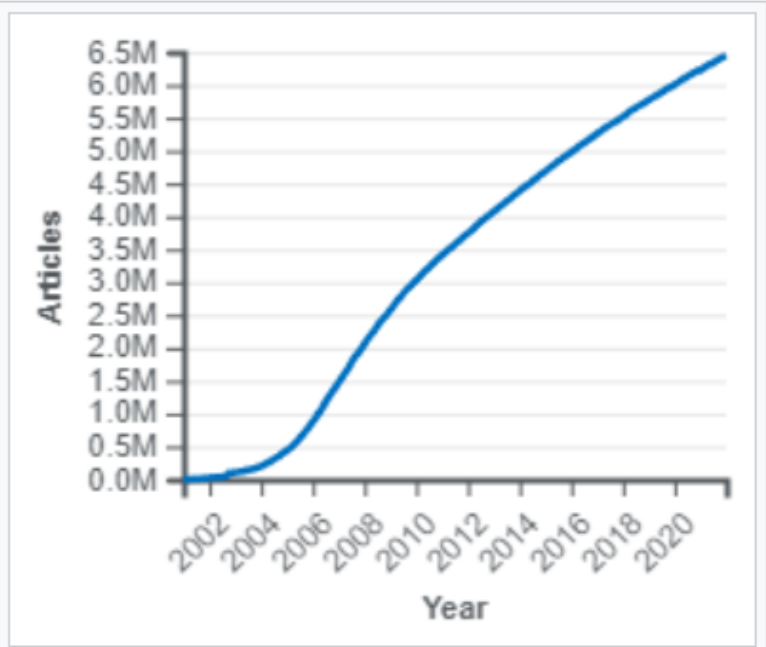
Amazon Elastic Compute Cloud (Amazon EC2)

- Clean, Transform, and Filter Structured Data
- DB Scans and Queries
- Apply FP Growth

Amazon Neptune

Selected Dataset

- Wikipedia
- 6.5M+ English articles as of 2022
- 10TB of data as of 2015
  - https://dumps.wikimedia.org/enwiki/latest/
  - https://www.kaggle.com/datasets/ltcmdrdata/plain-text-wikipedia-202011
  - https://github.com/daveshap/PlainTextWikipedia



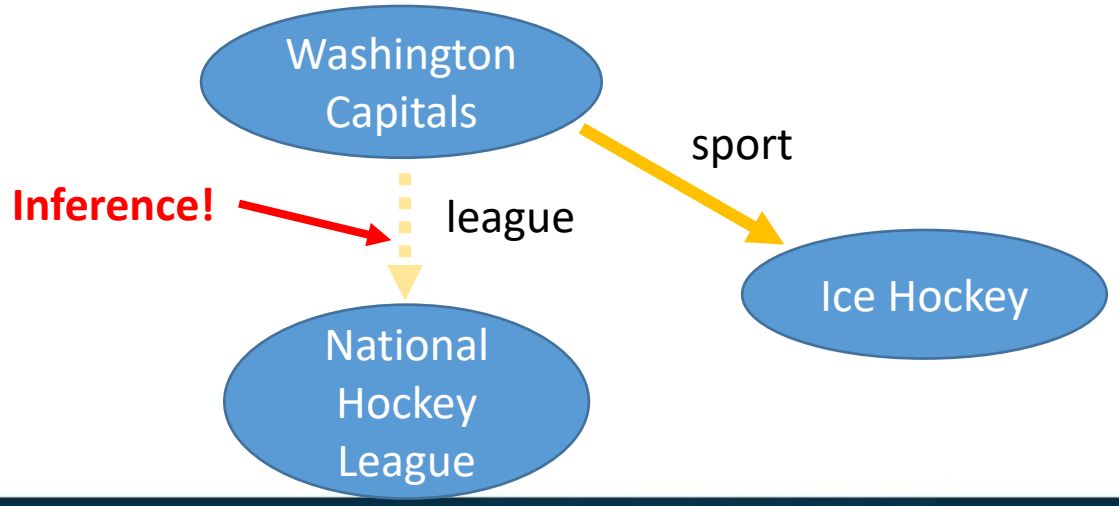
Number of English Wikipedia articles[5]

https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Current Progress

Some early associations

support	itemsets
0.002395	(league Southern Conference, member of SEC)
0.002395	(sport ice hockey, league National Hockey League)



Date: October 2022

Download CSV | Print

Estimated Total $13.50

Name	Type	Created	Messages available
Articles.fifo	FIFO	10/20/2022, 17:04:48 EDT	100158

Name	Instance ID	Instance state	Instance type
Article Queue Worker	i-091ebaa6da955fdaa	Terminated	t2.large
Relation Extraction Worker	i-0ab1850889ef24422	Running	c5.xlarge
Relation Extraction Worker	i-0af5a562cb22ab67e	Running	c5.xlarge
Relation Extraction Worker	i-0d9e48489f40dcf72	Running	c5.xlarge
Relation Extraction Worker	i-0bf6d65705a462b48	Running	c5.xlarge
Relation Extraction Worker	i-0915356135ede0d01	Running	c5.xlarge
Relation Extraction Worker	i-0ca912cea7d0f9f5b	Running	c5.xlarge
Relation Extraction Worker	i-0f543502e49154966	Running	c5.xlarge
Relation Extraction Worker	i-06771b02375ff6f29	Running	c5.xlarge
Relation Extraction Worker	i-0ca674c2d2d2d77f6	Running	c5.xlarge
Relation Extraction Worker	i-09daf15fce53d4b36	Running	c5.xlarge

a. (John_E_Blaha birthDate 1942_08_26) (John_E_Blaha birthPlace San_Antonio) (John_E_Blaha occupation Fighter_pilot)

b. *John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot*

https://webnlg-challenge.loria.fr/

Future Work

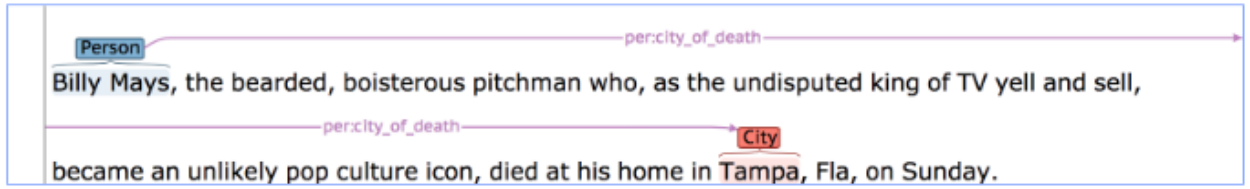
- Generate a new set of training data
- Start with generated triples by determining how often entities and verbs/predicates appear in the same sentence
- Use transformer to generate sentences from generated triple
- This would give the algorithm a more direct triple to sentence training

dbp:yearpro	• 1996 (xsd:integer)

https://dbpedia.org/page/Tiger_Woods

Woods turned professional in 1996

https://en.wikipedia.org/wiki/Tiger_Woods



https://nlp.stanford.edu/projects/tacred/

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC