THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Data Mining for Entity Relationship Associations

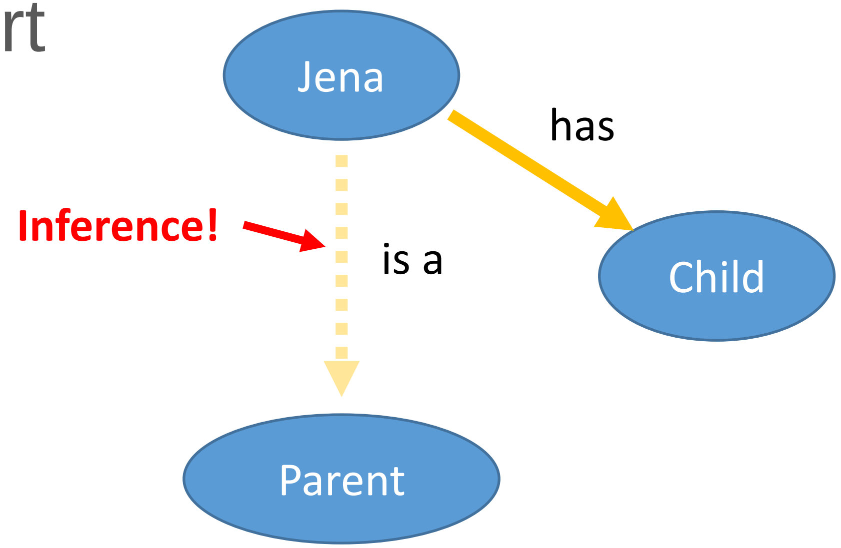
School of Engineering and Applied Science
Department of Computer Science CSCI 6443– Data Mining

Professor: A. Bellaachia

Student: R. Gross (G47667332)

Problem Definition

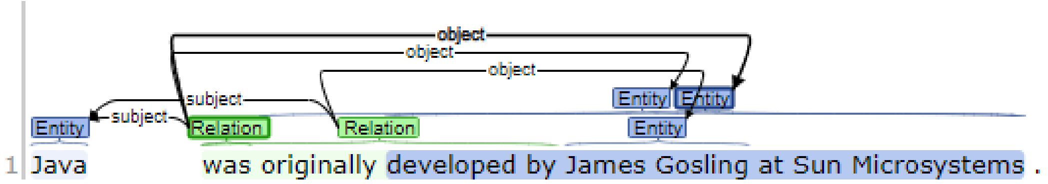
- Many chatbots are a combination of expert systems and machine learning.
- A knowledge base is often used as the “brain” of the chatbot due to its ability to perform inference.
- Traditionally knowledge bases perform inference based on inference rules, which are brittle and don’t scale well.



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.

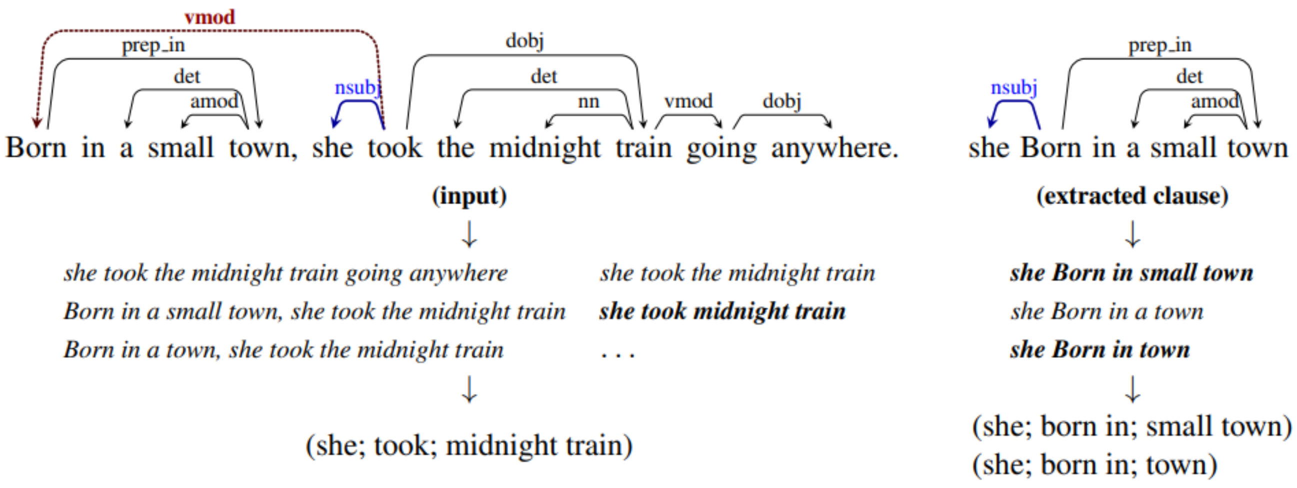


https://corenlp.run/

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Related Work Continued

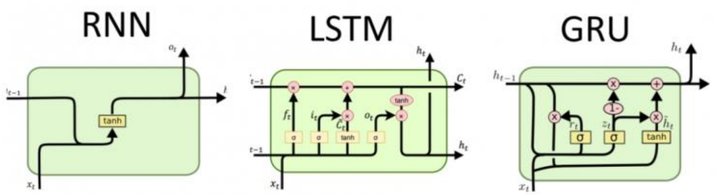
- Methods of doing RE :
  - OpenIE
    - Greedy search on dependency tree
    - Goal is to reduce sentence to utterance, and keep reducing until triple is all that is remaining



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

Related Work Continued

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



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

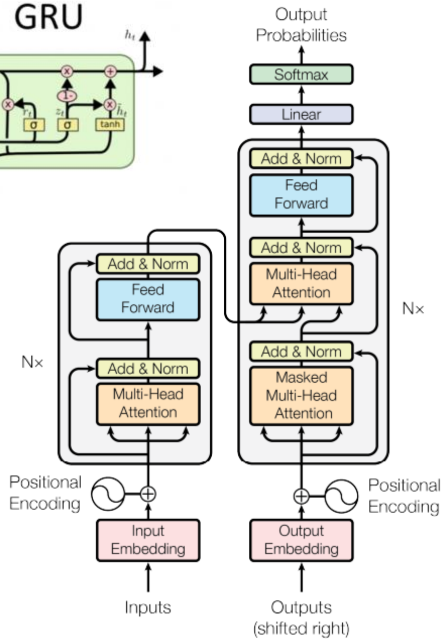
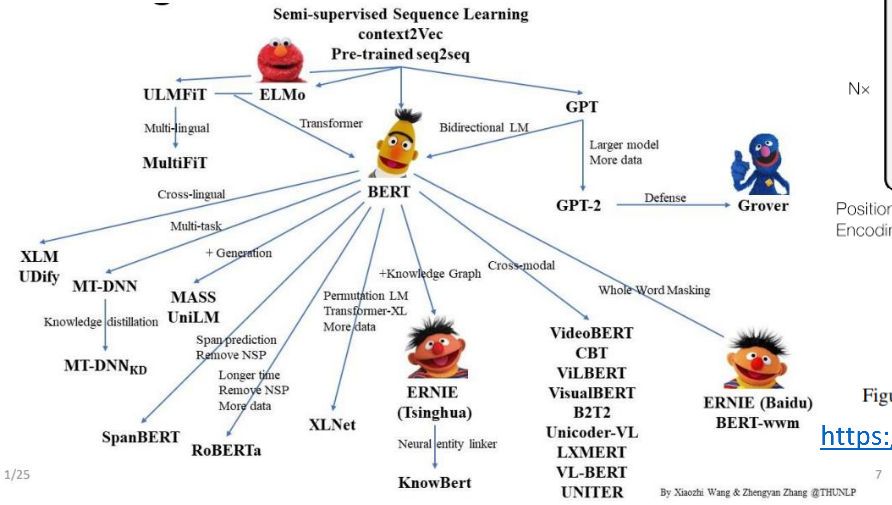
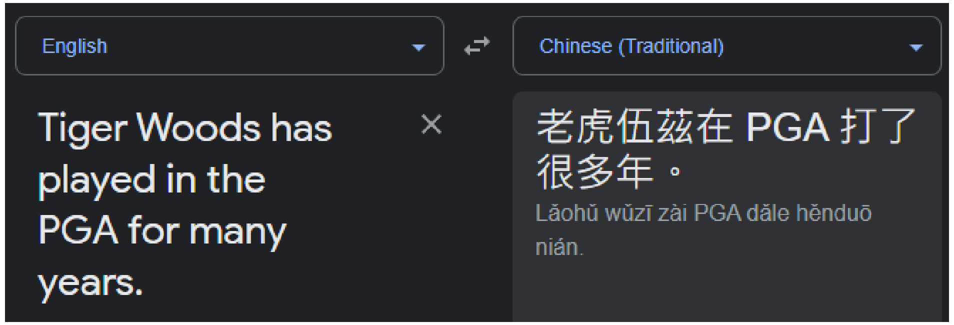


Figure 1: The Transformer - model architecture.

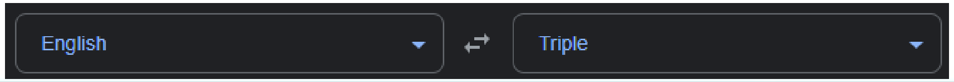
https://arxiv.org/pdf/1706.03762.pdf

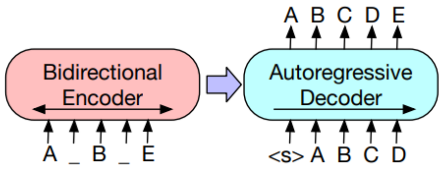


https://www.microsoft.com/en-us/research/uploads/prod/2021/06/Pre-training-Models-Xu-Tan.pdf



https://translate.google.com/





BART

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

Related Work Continued

- After RE, associations between relations occurs.
- Leading association algorithms are:
  - Apriori (Low Mem, Slow Speed)
  - Eclat (Medium Mem, Medium Speed)
  - FP Growth (High Mem, Fast Speed)

Evaluation Criteria	Apriori	FP Growth	Eclat
1. Techniques	Breadth first search	Divide and Conquer	Depth first search and intersection of transaction id.
2. Database Scan	Database is scanned each time a candidate item set is generated	Database Id scanned two times only.	Database is scanned few times.
3. Advantages	-Easy to implement. -Use large item set property.	Database scanned two times only.	No need to scan database each time.
4. Disadvantages	-Require large memory space. -Too many candidate item set	FP tree is expensive to build consumes more memory.	It requires virtual memory to perform the transaction.
5. Data format	Horizontal	Horizontal	Vertical
6. Storage Format	Array	Tree (FP tree)	Array
7. Time	More execution time	Less time as compared to Apriori algorithm	Execution time is less than Apriori algorithm.

https://www.semanticscholar.org/paper/Market-basket-analysis-for-improving-the-of-and-FP-Khan-Solaiman/99115afbb9202eba44c7522dccdf71fec8fd6b21

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

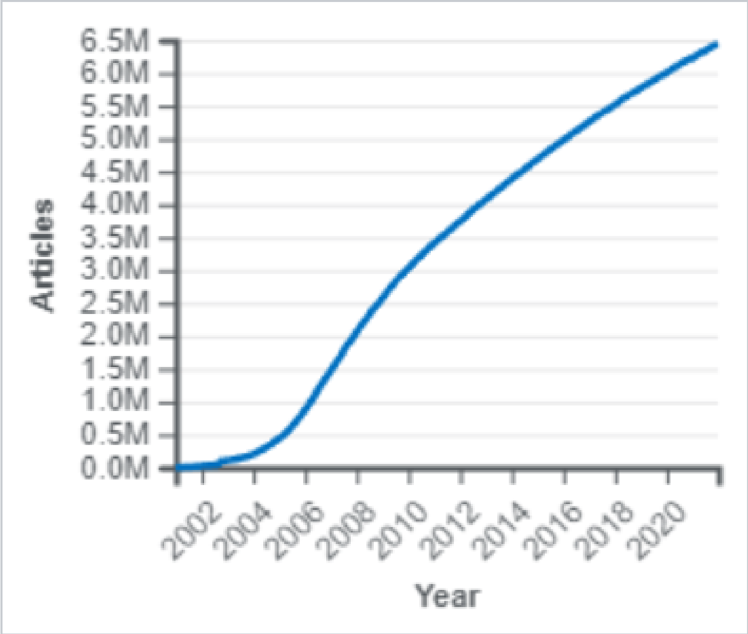
Amazon Elastic Compute Cloud (Amazon EC2)

- 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]

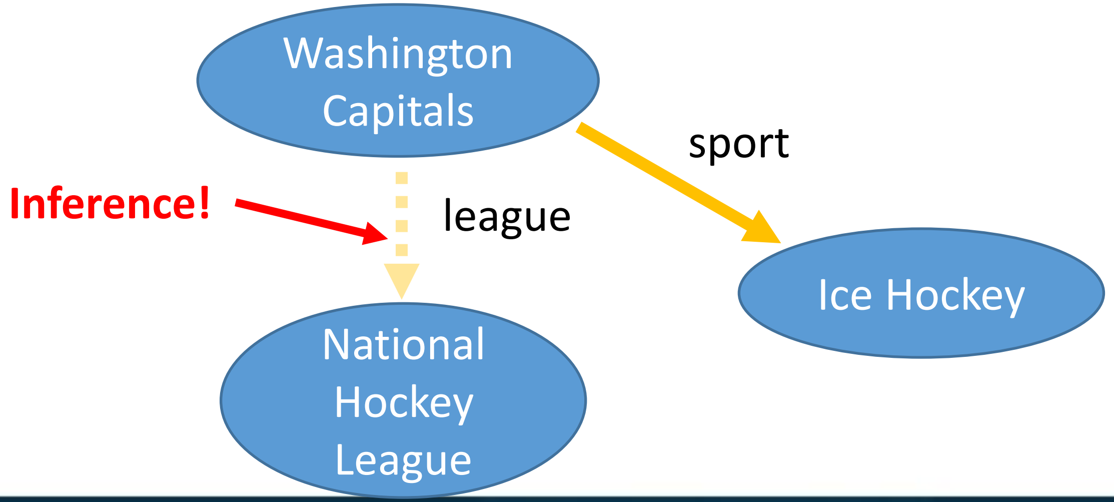
https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

Results

Some early associations

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

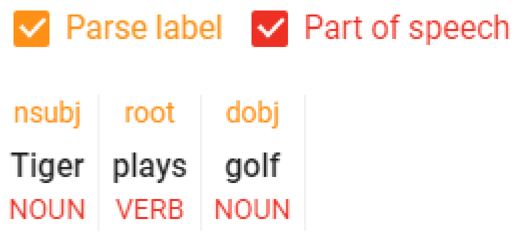


country United States', ' country United States of America'
league NFL', ' sport American football'
position held President', ' residence White House'
member of NATO', ' continent Europe'
diplomatic relation United States', ' member of NATO'
conflict World War I', ' conflict First World War'
league National League', ' league Major League Baseball'
shares border with Pennsylvania', ' instance of state', ' country U
part of Central America', ' part of North America'
instance of color', ' instance of colour'
headquarters location London', ' location London'
sport football', ' sport soccer'
country US', ' country United States', ' country American'
country US', ' country USA', ' country United States'
author Shakespeare', ' author William Shakespeare'
league Segunda División', ' country Spain'
league National Hockey League', ' sport ice hockey'
located in or next to body of water Persian Gulf', ' part of Middle East'
shares border with France', ' participant in World Cup'
member of NFL', ' member of National Football League', ' sport American foot

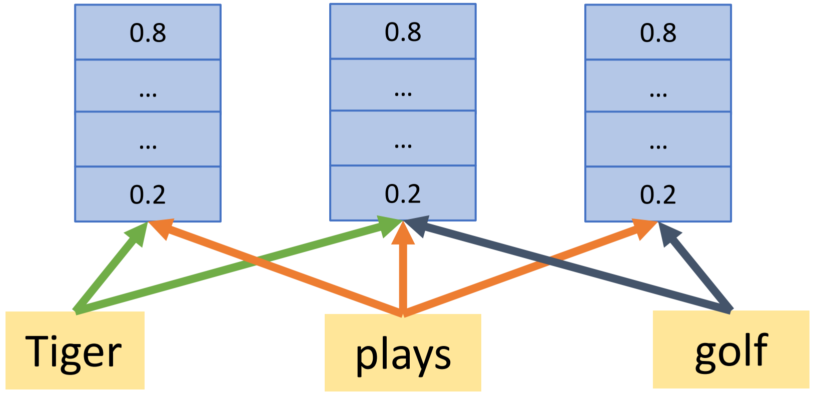
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Future Work

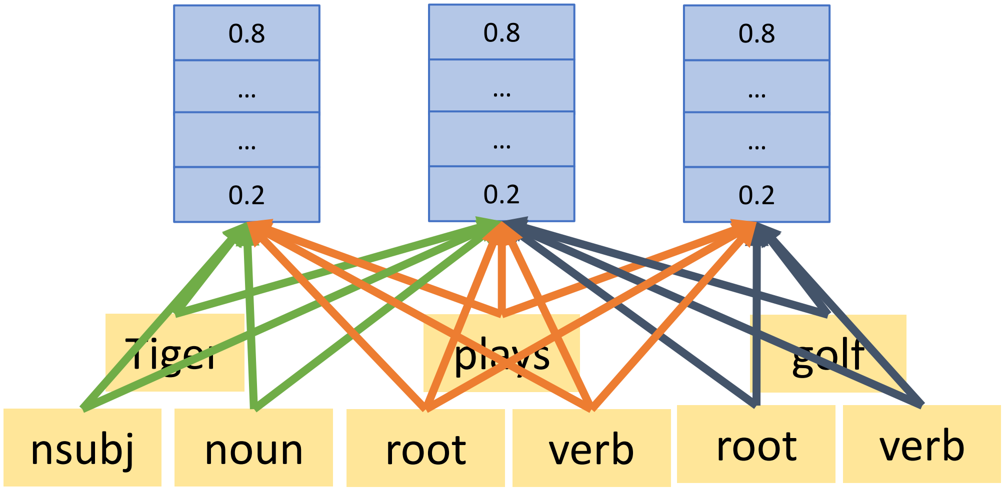
- Investigate methods of factoring in POS tags and parse labels
  - Helps to generalize
  - Due to generalization could reduce training size needed



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



Typical text encoding



Can we do this?

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC