Data Mining for Entity Relationship Associations

Gross, Ryan
GWU ID: G47667332
rgross4@gwmail.gwu.edu
CSCI 6443 Data Mining
Fall Semester 2022
Bellaachia, Abdelghani

George Washington University, Department of Computer Science, Washington DC, USA

Data Mining for Entity Relationship Associations

1. Introduction

Creating Artificial Intelligence (AI) that can perform common sense reasoning, sometimes referred to as Artificial General Intelligence (AGI), is still just a dream of many Computer Scientists (CS). An old but promising piece of technology that can enable AGI is the Knowledge Base (KB). KBs are promising because they enable symbolic reasoning. Symbolic reasoning enables the creation of new information from existing information. For example, if Sasha is a mother, then we can infer that Sasha has children. Inference, deductive, and inductive reasoning are tasks that people do everyday, and will be needed for future AGI systems. To date most KBs perform inference from rule sets created by CS professionals. In the future KBs will need to learn these rule sets on their own in an unsupervised way to enable AGI. Chatbots are one use-case that would benefit from unsupervised inference. When people converse, each participant makes inferences to keep the conversation going. For example, if one person mentions they have to stop by daycare after work, the other person might ask how old their children are. If a chatbot meets a new person, it too can infer information about the person by comparing what they know about the person to what they know about similar people. The chatbot can then formulate questions to increase confidence of existing inferences and continue to formulate new inferences. This ability would bring Alexa, Siri, and Google closer to the AI systems depicted in Hollywood, and in a more practical sense, would make interacting with chatbots from banks and IT service desks less awkward.

The problem is unsupervised inference over symbolic data is still a difficult research problem that has yet to make it to commercial systems. In order to achieve this dream, two main steps are required. First unstructured text needs to be converted to triples to put the data in a format that can be stored in the KB. This field of study is Relation Extraction (RE), a sub-field of Natural Language Processing (NLP). Second, association algorithms need to be run on the predicate object pairs commonly found together for subjects. For example, if Sasha, Amy, and Neha are subjects in a KB, and all three subjects have the predicate object pairs, <is_a> <Mother>, and <has> <Children>, then the system can learn that being a mother and having children are associated.

2. Related Work

2.1 Relation Extraction

2.1.1 Open Information Extraction (OpenIE)

OpenIE aims to extract relations through a divide and conquer search algorithm. This algorithm first requires an NLP task called dependency parsing to be performed. Dependency parse is where the root word of the sentence is determined, and then the directly dependent words of the root word are determined. Then words dependent on those words are found, again in a divide and conquer type of approach until all words are linked to their dependent words. OpenIE uses this linked list of words to perform

another divide and conquer algorithm, but this time it tries to reduce the sentence until only a triple is remaining. This method became popular when Stanford implemented this method in their popular open source library and published a paper about it [1].

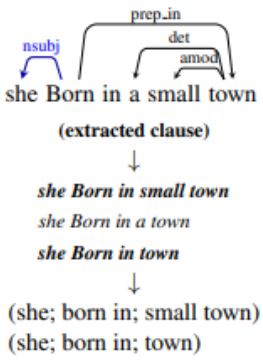

Figure 1. OpenIE example [1].

2.1.2 Graph Convolutional Neural Network (GCN)

GCNs also depend on a dependency parse being run on the sentence prior to processing. GCNs take the dependency parse which is in the shape of a tree and processes them similarly to how images are processed by a Convolution Neural Network (CNN) [2]. By passing the graph representation of the sentence into the CNN, the features of the sentences that are associated with the triple in the training data are found, similarly to how the arrangement of pixels are identified when categorizing an image.

2.1.3 Sequence to Sequence (Seq2Seq)

Seq2Seq has traditionally been useful for translation language, summarizing text, or filling in missing words within a sentence. The task of RE, can be thought of as a language translation, or an extreme version of text summarization. Seq2Seq methods of RE seem to dominate recent research and score the best in ranking on popular training datasets [3]. Seq2Seq methods can be implemented with Recurrent Neural Networks (RNN), Long Short-term Networks (LSTM), Gated Recurrent Unit (GRU), Encoders, Decoders, and Transformers. While all methods have their pros and cons, Transformers [4] currently represent the state-of-the-art text transformation.

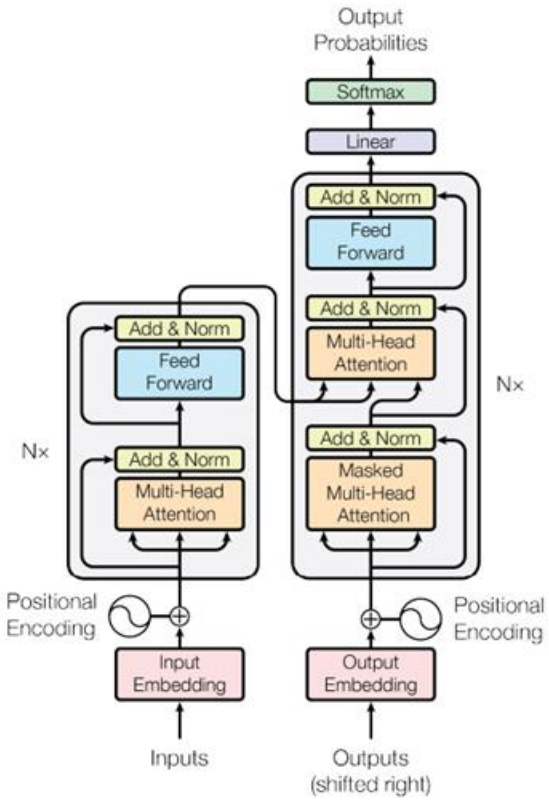

Figure 2. The transformer model [4].

2.2 Association Mining

Association mining determines which entities in a set are commonly found together. The Apriori, Eclat, and FP Growth

represent the three most popular algorithms for association mining [5]. Apriori is a breadth-first search which requires many table scans. The benefit of using Apriori is it requires less memory than the other two methods. Eclat is a depth-first search which reduces the number of table scans, thus making it faster than Apriori, but more memory intensive. FP Growth is known to be the fastest method, as it only requires one table scan. The disadvantage of FP Growth is it requires large amounts of memory.

3. Approach

3.1 Data Set

The dataset used is Wikipedia. Specifically, an already extracted dataset from Kaggle [6] is used, which only extracts the opening text section of each article, and puts this text into JSON arrays. The dataset is 20+ GB once unzipped.

3.2 Algorithms

For RE, the REBEL [7] algorithm is used. The REBEL algorithm is built on the popular BART [8] Transformer algorithm. REBEL comes pre-trained against 200 predicates. The training method used for REBEL is similar to other distantly supervised methods, where triples from DBpedia and text from Wikipedia are used together. However, it is widely understood that this method is noisy, so the REBEL team added a pre-processing step to ensure REBEL only trained on DBpedia triples that are found in the corresponding Wikipedia article.

For association mining, FP Growth was chosen, as it is popular enough to be found in Python libraries and is widely accepted to be one of the fastest association mining algorithms.

3.3 Infrastructure

The entire process was performed on Amazon Web Services (AWS). The decision to use a cloud provider was driven by the fact that the dataset contains millions of articles, and it was taking about twenty seconds to process a single article through REBEL on my computer at home. Using the cloud would enable the process to be parallelized.

3.4 Steps

First the zipped Kaggle dataset needs to be put into AWS's Simple Storage Service (S3). Next an EC2 instance with permissions to download the file needs to download and extract the file. That EC2 instance needs code to parse files and push individual Wikipedia article text to AWS's Simple Queue Service (SQS). This will allow the RE servers to process messages in parallel. Once messages are in the queue, the RE worker EC2 instances running REBEL extract triples and push them back to S3. Using AWS's Semantic Graph service, Neptune, load the triples from S3. Finally use a Neptune workbook to query triples, and run the FP Growth algorithm to find predicate object pairs that are found to be associated with subjects.

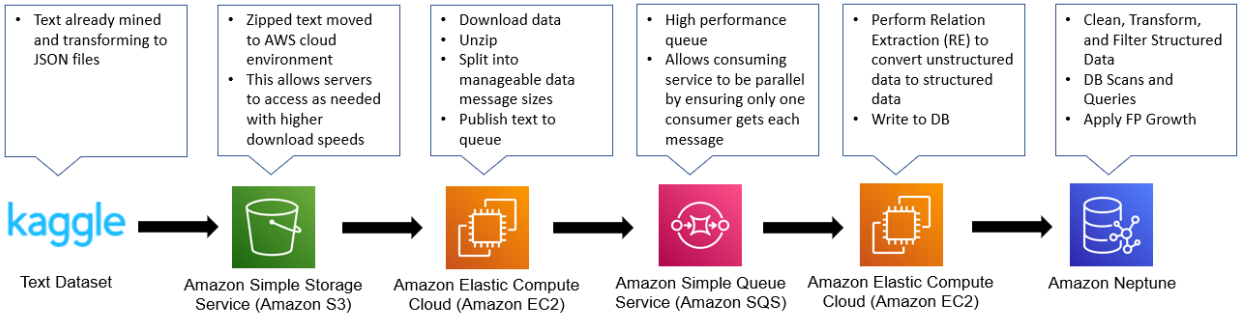

Figure 3. Visual depiction of project approach.

4. Results

The system did find many predicate objects pairs that people would consider common knowledge.

Figure 4. Predicate-object pairs with highest association support

To put the results into perspective, let's take "position held President, residence White House" as an example. If a system using this knowledge base learned that a person was the President, it could then infer they live in the white house with some level of confidence. One can imagine how a chatbot with this inference ability could infer new information when talking to a user, and then use inferred knowledge to continue the conversation. This is similar to how people socialize. As new information is learned, people tie the new information to pre-existing knowledge to ask clarifying or related questions.

While many reasonable predicate object pairs were found, the system also found many more pairs that would be considered non-sense or noise. This comes down to REBEL's extraction and pre-trained model. With more time a new model could be trained that may perform better. Also with REBEL supporting 200 predicates, this limits the variety of results, and misses information from the initial text.

5. Conclusion

The prospect of statistically learning inference rules from plain text remains promising. While the results of this paper show that this task is possible, it also highlights that such a system is not perfect and would require human filtering of learned rules. This unfortunately makes online learning using this method in a chatbot unfeasible. However, as RE methods improve, this concept will become more viable.

References

[1] Angeli, Melvinj, Manning, "Leveraging Linguistic Structure For Open Domain Information Extraction", Department of Computer Science, Stanford University

[2] Guo, Zhijiang, Yan Zhang, and Wei Lu. "Attention guided graph convolutional networks for relation extraction." *arXiv preprint arXiv:1906.07510* (2019).

[3] Papers with Code. Relationship Extraction https://paperswithcode.com/task/relation-extraction

[4] Vaswani, Ashish κ.ά. 'Attention is All You Need'. N.p., 2017. Web.

[5] Khan, Mohammod Akib et al. "Market basket analysis for improving the effectiveness of marketing and sales using Apriori, FP Growth and Eclat Algorithm." (2017).

[6] Kaggle. Plain Text Wikipedia 2020-11 https://www.kaggle.com/datasets/ltcmdrdata/plain-text-wikipedia-202011

[7] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.