Exploiting the Modelling of Problem Domain Knowledge to Advance Text Analytics: from Computational Linguistics to Large Language Models

Taha Osman

Department of Computer Science
Nottingham Trent University
Nottingham, UK
taha.osman@ntu.ac.uk

***Abstract* - The success of Large Language Models (LLM), most notably ChatGPT, has rejuvenated interest in Text Analytics and its applications. LLM-based systems utilise the new generation transformer neural networks that are trained on colossal amounts of text data (GPT-4 processes 100-trillion parameters) to power NLP/Text Analytics applications such as document summary and conversational agents. Dr Taha Osman has a rich track record in modelling the knowledge embedded in the problem domain into semantic ontologies (knowledge graphs) that have proven effective in improving the classification engines behind information retrieval and sentiment analysis applications. The research talk will explore the fundamentals of his hybrid knowledgebase – Machine Learning NLP methods and their application to different applications areas including digital media, mineral exploration, and social media analytics. The talk will also discuss the recent investigation in introducing domain knowledge to boost the classification accuracy of LLMs (BERT).**

Keywords - Large Lnguage Models, NLP, Domain Knowledge, Semantic Web, Ontologies, Knowledge Graph, BERT

I. INTRODUCTION

There is no doubt that we live in the age of AI and there is world-wide interest about its on Society and the Economy. A tudy commissioned by the UK Government on the impact of AI on Employment in the UK, concluded that by 2030, 30% of the jobs can be automated by AI [1].

The arrival of ChatGPT moved the concern closer to home and academics started worrying about the impact of generative AI on the learning of the students and perhaps about the appeal of the university's offering.

The research community have a glass-half-full view, especially in the field of NLP (Natural Language Processing) where the advent of LLM (Large Language Models) and transformer technology generated enormous interest and rejuvenated research into this area.

ChatGPT, the most popular Generative AI APP, belongs to a family of pre-trained ML systems that use the Transformer deep learning models. Transformer NN were first developed by Google in 2017 to primarily solve the problem of machine translation [2] and very quickly became the engine behind other spectacular AI innovations that invaded main stream use and already found their way to a lot of real life applications from Conversational AI to ART and code generation such as GPT-4; ChatGPT; AlphaCode; GitHub Copilot; Bard; Cohere Generate; Claude; Synthesia; DALL-E 2.

Applied to NLP, the main advantage of transformer neural network architecture is that it can learn the context of the word, i.e. how it relates to other tokens or words in a sentence, over a longer distance and in much more efficient way. More specifically, unlike RNN, transformers are able to process words in parallel while maintaining their order or context. Parallelisation is critical because it means that, with the right hardware (GPUs), transformers can train over very large language models; GPT3 was trained on 45TB of data and used 175 Bn hyper-parameters, which exploded to 170 Tn in GPT4. However, awesome as it is, these Large Language Models or 'transformer systems' are trained over vast volume of generic online text. Therefore, that they can only perform these NLP tasks based on their training data; they have got lots of it, but its general training data, so it can answer general questions and perform general classification based on that training data.

That training data might be old, it might be irrelevant to the problem domain or industry or specific application. This might be problematic when they are used in critical tasks. A study on the accuracy of using Generative AI for a critical service such as legal advice. The study was commissioned by Linklaters global law firm Linklaters global law firm [3]. While a query about the GDPR regulation for retaining customer order information produced a response that surprised the assessors with its accuracy and coverage, the response to a query about sending marketing Emails under PCER guidelines clearly fell short from providing sound advice. Hence, the challenge is how can the awesome power of LLM, with its fantastic ability to understand human conversation and access to extraordinary and often useful general information, be utilised while focusing the response on accurate content of greater relevance to the application area or domain?

Our hypothesis is that for NLP tasks over domain-specific problem domains, we can exploit hand-crafted domain knowledge to improve the Machine Learning classification.

II. PREVIOUS RESEARCH IN KNOWLEDGE-BASED NLP

This section reviews three cases of previous research that contributed to the development of a hybrid methodology integrating linguistic knowledge of the problem domain and the AI classification methods.

(a) In our initial research, we outlined a comprehensive plan for upgrading image search engines to fully leverage advancements in semantic web technologies [4]. The first step involved creating ontologies (domain vocabularies) to establish a consistent understanding of domain entities and potential relationships among them for the search engine. We ensured minimal redundancy in annotations by meticulously organizing the ontology structure, using pragmatic non-semantic techniques where necessary.

Our approach to designing the annotation process was centred around the user experience and considered the dynamic nature of information retrieval (see Figure 1). We devised a template based on sentences to model user queries, especially those involving intricate relationships between query keywords.

The retrieval algorithm employs a modified nearest-neighbour search method to navigate the ontology tree, capable of accommodating complex, relationship-oriented user queries.

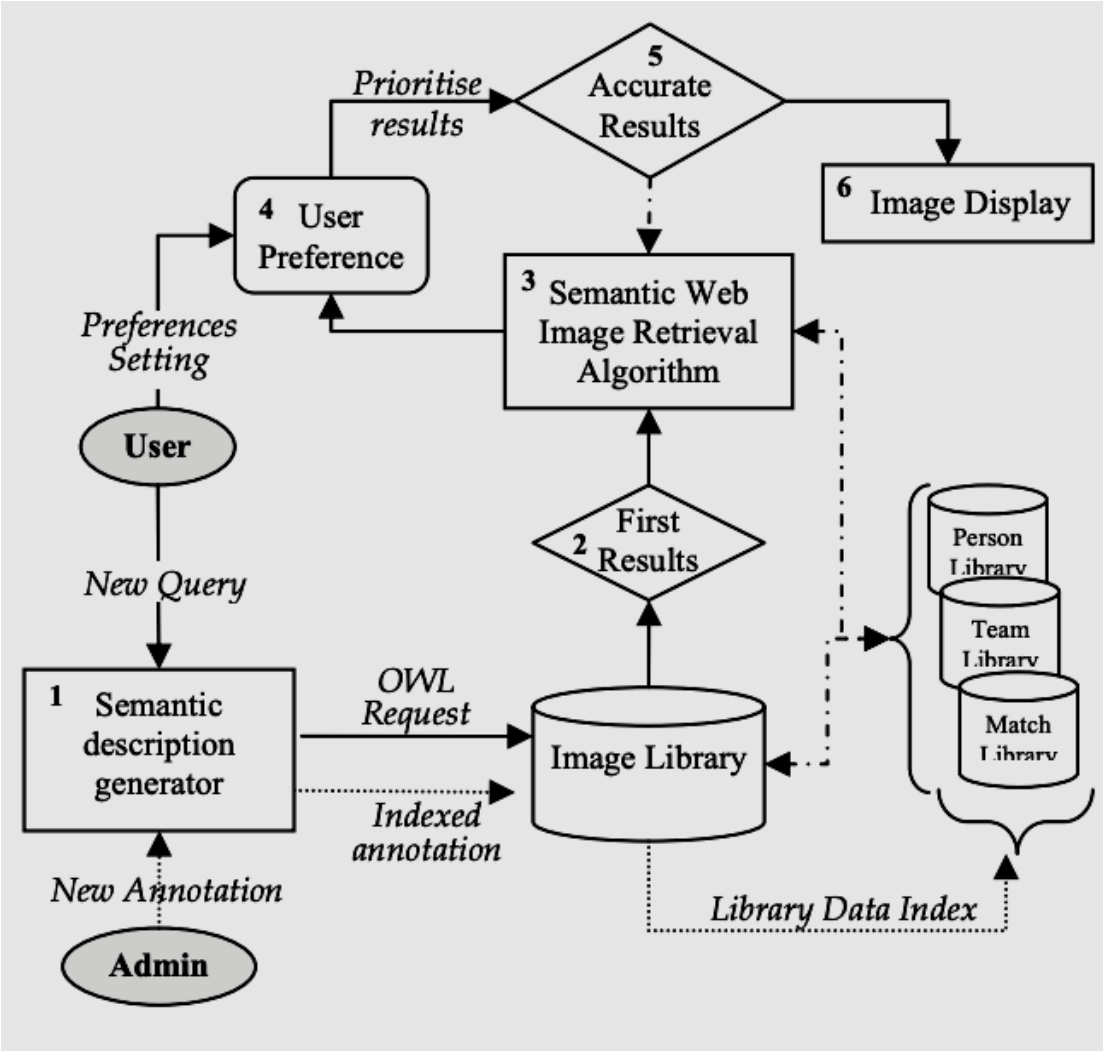

Figure 1: Workflow of the Semantic Web Image Retrieval Mechanism

(b) The work published in [5] focused on relation extraction from unstructured Arabic text is especially challenging due to the Arabic language complex morphology and the variation in word semantics and lexical categories.

The study introduced a hybrid Semantic Knowledge Base - Machine Learning (SKML) methodology for extracting intricate Arabic relations from unstructured Arabic documents. This innovative approach leverages Functional Discourse Grammar (FDG) principles to

underscore the semantic and pragmatic aspects of the language, aiding in identifying relation components. Initially, the FDG-SKML method employs a lexical-based mechanism, utilizing a domain-specific Semantic Knowledge structure to encode semantic associations among the identified relation elements. Evaluation of this initial phase revealed enhanced accuracy in extracting complex Arabic relations.

To further enhance the relation extraction process, the initial mechanism was expanded by incorporating its outcomes into a Machine Learning classifier. This integration facilitated the extraction of particularly complex relations characterized by variations in element presence, order, and correlation. By focusing on Economics as the problem domain, experimental assessment demonstrated the high accuracy of our FDG-SKML approach in complex Arabic relation extraction tasks, with additional improvements observed upon integration with machine learning classifiers.

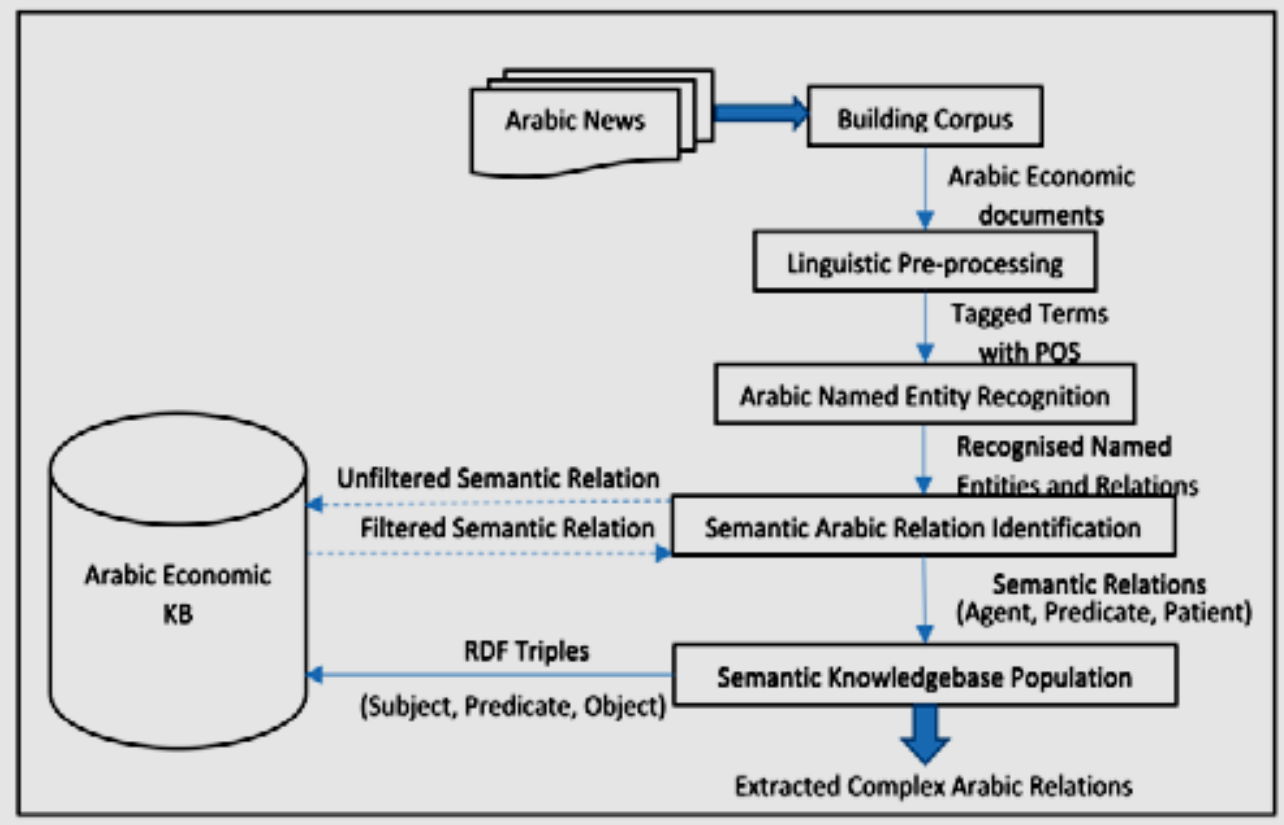

Figure 2: System Architecture of the FDG-SK Approach

(c) The final research effort developed opinion mining tools enable users to efficiently process large number of online reviews to determine the underlying opinions [6]. The study introduced a Hybrid Semantic Knowledgebase-Machine Learning strategy for extracting opinions at the domain feature level and assessing the overall opinion on a multi-point scale. This approach capitalizes on the benefits of a novel Semantic Knowledgebase technique to analyse reviews at the domain feature level, generating structured data linking expressed opinions with specific features (see Figure 3). Additionally, the knowledgebase is augmented with domain-relevant facts sourced from public Semantic datasets, enhancing the semantically-tagged information used to infer valuable insights about the domain and opinions expressed on its features.

By summarizing opinions across multiple reviews and averaging opinions on other cinematic features, this enriched semantic information offers valuable resources for training a machine learning classifier to predict numerical ratings for each review. Experimental evaluation demonstrated that the proposed Hybrid Semantic Knowledgebase-Machine Learning approach enhanced precision and recall in extracting domain features, leading to an enriched dataset of semantic features and improved classification accuracy.

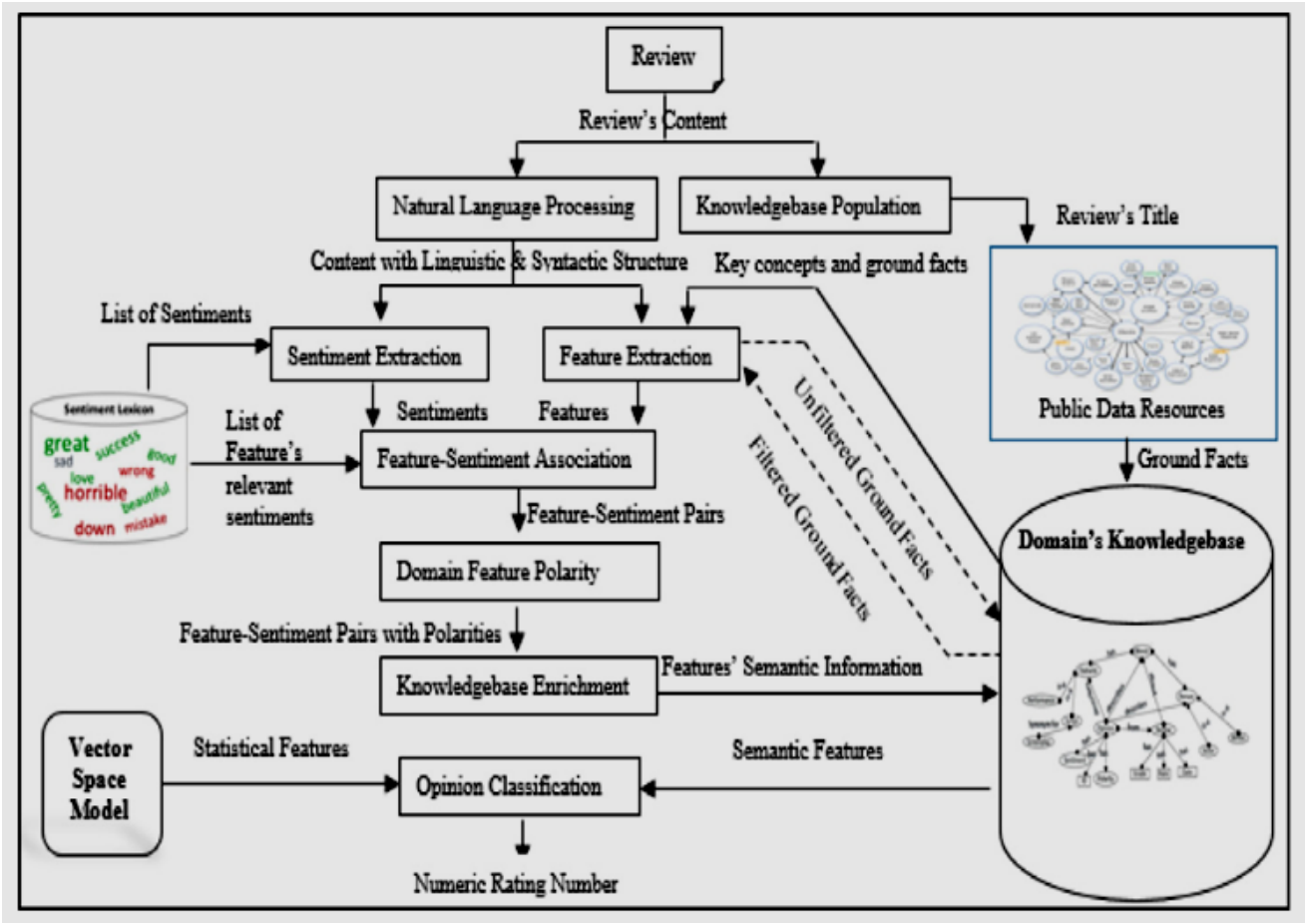

Figure 3. A hybrid semantic knowledgebase-machine learning framework for opinion mining

So, we established reasonable expertise and research track record in fusing human intelligence represented in the understanding of the domain knowledge and the machine intelligence represented by the classification algorithms. We are currently investigating the extension of hybrid methodology to the Transformer-based language processing models.

III. APPLYING THE NLP-AI HYBRID MODEL TO THE TRANSFORMER NETWORKS

The main advantage of transformer networks is its ability to efficiently understand the context of word in text over a longer distance. This ability to deep-learn the contextual meaning of text is critical to the great improvement to language tasks like the text generation in ChatGPT platform.

The transformer model was introduced in 2017 by the Google research team in the prominent paper: Attention is All You Need. It has RNN-like encoder-decoder architecture but with parallel input sequence and embedding generation. The two important concepts in transformer networks are embedding and attention. Embedding is the mapping of input text into a vector that preserves the words' context, this includes the meaning of the word in the language space as well as the position of the word in the sentence. Attention primarily computes the relationship between the words in a sentence, thus complementing the contextual information by deciding which part of the input requires focus.

Our research uses one the most popular Transformer technology, BERT, Bidirectional Encoder Representations from Transformers (BERT) [7], and the selected classification task is sentiment analysis, indicating the conditions of financial health or investment.

Text Classification (Sentiment Analysis) in the Financial Domain using the BERT Transformer model to perform financial opinion mining. It fuses a semantic graph financial domain knowledge with the BERT-Encoded input financial text-to-analyse hence providing mapping of the input financial text to the pre-trained BERT model.

As the transformer model understands natural language by embedding sentence words from the training text, in order to integrate the more complex domain knowledge as opposed to generic text, then we need to embed Knowledge Graphs. We used the Node2vec technology to embed the

knowledge graph into serial text vectors. Node2Vec performs a biased random walk through the nodes to capture local (node role) and global structure (neighbourhood) of the graph.

The integrated embedded objects are processed by the transformer with Multi-headed-attention to generate classifications of raw predictions for each of the three classes (positive, negative, and neutral) indicating the sentiment expressed for the corresponding financial text. The initial experimental results are encouraging and outperformed the traditional machine-learning sentiment analysis methods

IV. CONCLUSIONS

The transformer technology have and Large Language Models have transformed the application and innovation of NLP Systems. Knowledge-based modelling can complement the power of the pre-trained transformer systems by injecting focused human understanding into their ML cycle.

Success in integrating knowledge graphs in BERT input embedding with promising application results in financial text analytics.

The hybrid methodology is applicable to other domains such as the legal domain and digital marketing, and to other NLP tasks such as document classification, information retrieval, and conversational agents.

REFERENCES

[1] GOV.UK, “The impact of AI on UK jobs and training” https://assets.publishing.service.gov.uk/media/656856b8cc1ec500138eef49/Gov.UK_Impact_of_AI_on_UK_Jobs_and_Training.pdf (accessed 29/4/2024)

[2] A. Vaswani A, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Gomez AN, Ł. Kaiser, I. Polosukhin. “Attention is all you need”, Advances in neural information processing systems, 2017, p 30.

[3] linklaters.com, “ChatGPT – 50 questions to road test its legal advice”, https://www.linklaters.com/en/insights/blogs/digilinks/2022/december/chatgpt---50-questions-to-road-test-its-legal-advice (accessed 29/4/20240

[4] T. Osman, D. Thakker and G. Schaefer, “Semantics based intelligent indexing and retrieval of digital images – A case study”, Emergent Web Intelligence: Advanced Information Retrieval, Series on Advanced Information and Knowledge, Springer, 2010, pp. 117-134.

[5] T. Osman, H. Khalil, M. Miltan, K. Shaalan, R.Alfrjani, “Exploiting Functional Discourse Grammar to Enhance Complex Arabic Relation Extraction using a Hybrid Semantic Knowledge Base-Machine Learning Approach”, ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(8), pp.1-30.

[6] R. Alfrjani, T. Osman, G. Cosma, “A Hybrid Semantic Knowledgebase-Machine Learning Approach for Opinion Mining”, Data & Knowledge Engineering, 2019.

[7] A. Gillioz, J. Casas, E. Mugellini and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," 2020 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 2020, pp. 179-183.