Galaxia - Graph Language Model

Technology overview
2024

smabbler

Introduction

Smabbler started as an R&D team committed to creating a Natural Language Processing (NLP) technology that is lightweight, scalable, explainable, and data-efficient. We began with the understanding that human language is not merely statistical, but can be understood in terms of structure and complex semantic relationships. That's why early on we chose to focus on combining NLP methods with graph-based language and knowledge representation. These technologies form a solid foundation that can facilitate more reliable development and use of Machine Learning (ML) models, like Deep Neural Networks (DNNs). Thanks to its transparency and reliability it can be used to analyze and monitor content produced by less reliable technologies like Large Language Models (LLMs). It can also reduce the technological debt associated with ML models, and in some circumstances replace them altogether.

Machine Learning (ML) relies on statistical models and needs big data sets to learn patterns. Deep Learning (DL), a subset of ML, requires even more data and powerful GPUs (Graphical Processing Units) for computation. Neural Networks (NNs) have been successful in many applications. One of their main drawbacks, limiting their application, is their lack of transparency (the "black box" problem). The knowledge learned during training is spread across the parameters and structure of a model and cannot be extracted and examined. While LLMs can produce a human readable text, this is neither a scalable nor a reliable way to explore what knowledge is contained within a model.

Smabbler technology was inspired by technologies and approaches such as Symbolic AI, Computational Linguistics, Knowledge Graphs, Ontologies, and Evolutionary Algorithms. These technologies are based on structured knowledge representation which handles semantics and provides transparency.

The main challenge preventing the widespread adoption of these technologies is the difficulty of incorporating new knowledge compared to ML. Smabbler addresses these issues by integrating the mentioned technologies, developing an extensive knowledge base, and providing tools to facilitate knowledge extension in an efficient way.

Technology intersection

An overview of the technologies that inspired the Smabbler technology

I. Symbolic AI

Symbolic AI, also known as Symbolic Reasoning, is a branch of Artificial Intelligence (AI) that involves encoding human knowledge and behavioral rules into computer programs using provable mathematical logic.

Key characteristics and advantages of Symbolic AI over Machine Learning (ML) approaches include:

- **Interpretability:** Symbolic AI offers transparency in the reasoning process and the origins of results.
- **Flexibility:** It can be easily adapted to different domains by adjusting a logic and knowledge base.
- **Knowledge representation:** Knowledge is encoded in an explicit, structured, and both human and computer-readable manner.

One of the major benefits of Symbolic AI is that being rule-based, it does not require extensive training and can run on low-cost CPUs. However, as the knowledge base grows, its complexity makes the incorporation of additional knowledge more challenging.
Symbolic AI and Machine Learning are considered as two distinct "camps" in the field of Artificial Intelligence.

II. Computational Linguistics

Computational Linguistics sits at the intersection of AI, computer science, and linguistics. Much like Symbolic AI, it involves mathematical modeling of natural language and leverages rules, lexicons, and semantics to ensure reliability and transparency in reasoning.

III. Evolutionary Algorithms

Evolutionary algorithms are heuristic search methods inspired by biological evolution. They utilize processes such as reproduction, mutation, recombination, and selection to solve complex optimization and constraint satisfaction problems.

IV. Knowledge Graphs (KGs)

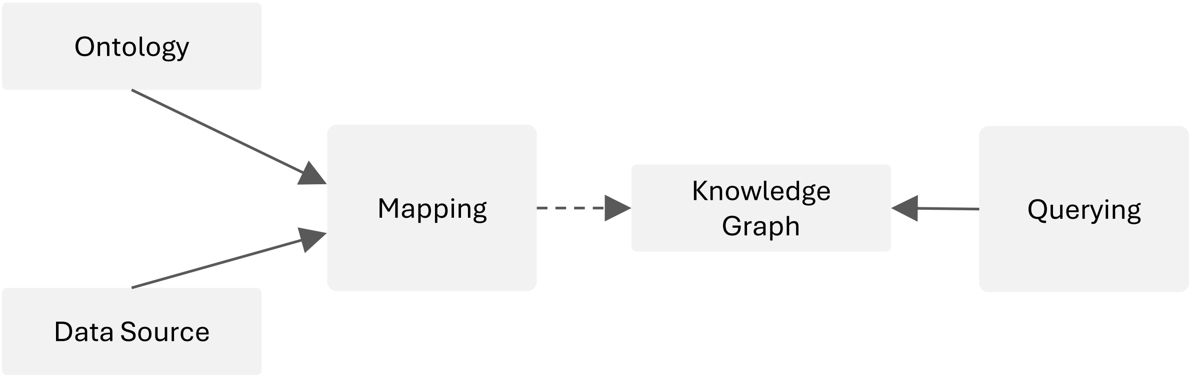
Graphs depict the relationships (edges) between a group of entities (nodes). Knowledge Graphs, also known as Semantic Networks or Semantic Knowledge Graphs, are networks of entities, their semantic types, properties, and relationships. They work with any kind or size of data and can utilize ontologies as their schema. They are a type of 'passive' knowledge representation, serving as a knowledge storage structure.

In comparison to LLMs, Knowledge Graphs do not require large datasets and costly training to work with domain-specific knowledge. Knowledge encoded in graphs is explicit and transparent - it can be examined, validated, and easily corrected. Solutions based on knowledge graphs demonstrate better interpretability and reasoning abilities, compared to LLMs. Graph-based solutions also do not hallucinate. However, the construction and expansion of these systems are difficult and costly. They require a substantial investment to build a comprehensive knowledge base that can reliably support reasoning algorithms. Smabbler solves these challenges by creating a vast knowledge base and offering tools to expand it into new knowledge domains where needed.

Currently, knowledge graphs are a subject of in-depth research, with a significant focus on knowledge completion of the knowledge graph. There is also ongoing research on combining knowledge graphs with LLMs to address the shortcomings of LLMs, such as reasoning abilities and lack of domain knowledge.

V. Ontologies

Ontology is a formal knowledge representation schema that consists of a set of domain concepts (vocabulary) and the relationships between them, organized to link one piece of information to another (e.g., anatomy ontology). Ontology provides organized information to construct a knowledge graph, enabling cross-search of similar or related concepts. The primary challenges for ontology applications are their development and knowledge extension, which heavily depend on human experts.



The technologies and approaches we decided not to use for creating Smabbler technology foundations

I. Machine Learning (ML)
II. Deep Learning (DL)
III. Neural Networks (NNs)
IV. Graph Neural Networks (GNNs)

I. Machine Learning (ML)

Machine Learning is a part of AI that focuses on creating statistical algorithms that can learn from data to perform specific tasks without explicit instructions. ML algorithms develop models based on training datasets (labeled examples) to learn patterns and make predictions. One of the main drawbacks of ML is the lack of transparency and the need for a large amount of data for the system to learn.

II. Deep Learning (DL)

Deep Learning falls within the domain of Machine Learning. It employs neural networks to learn features from data. Deep Learning requires substantial amounts of data and advanced computational capabilities, generally provided by Graphical Processing Units (GPUs), due to its extensive data and depth (number of layers in neural networks). Deep Learning utilizes structured data which can be represented as word sequences or pixel grids.

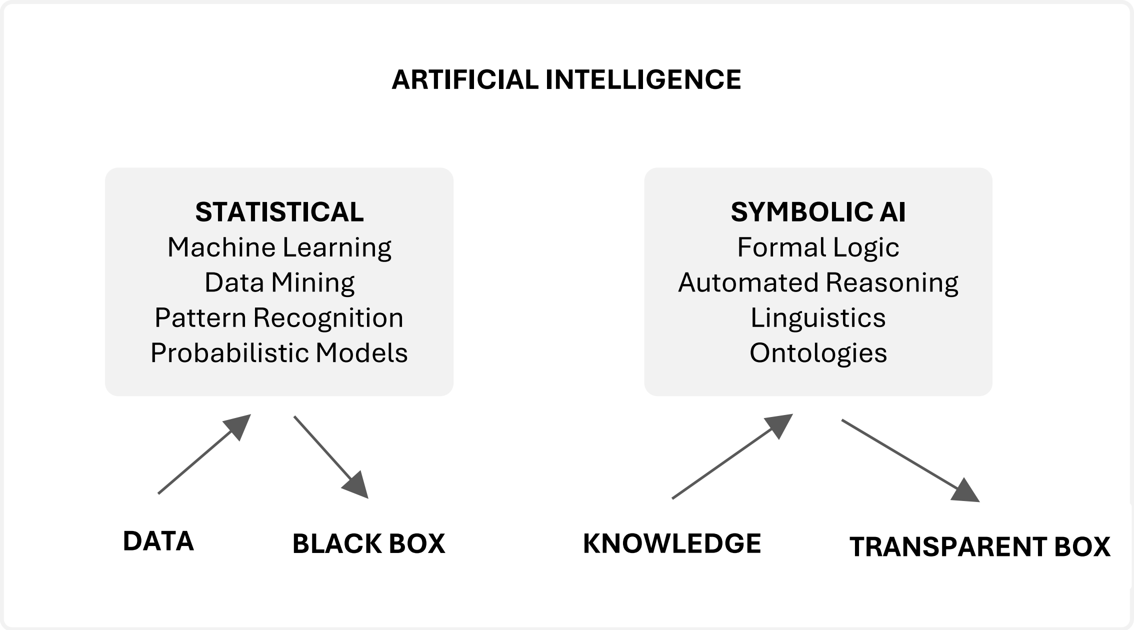
III. Neural Networks (NNs)

Artificial neural networks are a simplified neuron concept organized in a network, heavily based on weights, statistics, and composite functions. Traditional neural networks are unable to model complex relationships.

IV. Graph Neural Networks (GNNs)

This area is a part of deep learning research. GNN is a deep learning architecture that applies deep learning to a graph database comprising edges and nodes and storing information about relationships between data points. GNN arranges graphs so machine learning algorithms can utilize them. The biggest advantage of GNN is its adaptability in modeling complex relationships. However, graphs, particularly Knowledge Graphs, are sensitive to ambiguous and noisy data that are typical of neural networks, and this presents one of the challenges for GNNs.

Statistical (Machine Learning) vs Symbolic AI



Galaxia – foundational technology

Smabbler: Convergence of technologies to build a contextual and scalable Graph Language Model (GLM)

Our main technological goal was to create a flexible, transparent, and scalable natural language processing solution that can work with foundational knowledge and extend it. To achieve this, we combined the Symbolic AI approach, with concepts of Knowledge Graphs and Ontologies.

Galaxia is a graph language model that combines:
- the ability to store and organize information and relationships between them
- easy expansion with new domain knowledge
- mapping relationships between multiple data points
- flexibility to apply to various fields and tasks
- natural language processing capabilities
- easy use of information and deriving features
- transparency
- scalability
- ability to operate without the need for training
- low computational requirements

Language Processing Graph Architecture

We utilize a graph as the fundamental building block for language processing. To overcome the scalability limitations of traditional graphs, we developed a completely new and unique graph architecture. This architecture is inspired by ontology representations of knowledge with concepts (referred to as nodes) and the relationships between them (edges). Moreover, we have significantly expanded on how relationships are built, going beyond the typical 'triples' structure characteristic of ontology.

Knowledge Expansion

Galaxia's structure allows for easy extension with new information and relationships of any size, such as vocabularies, domain taxonomies, and ontologies, due to its similarities to Knowledge Graphs and Ontologies.

GALAXIA's basic entities and functions

Nodes and Edges

Nodes in Galaxia graph represent words, phrases, and definitions. Edges represent direct or indirect relations between them.

Compositionality

One of the core strengths of Galaxia is compositionality - the ability to define new, complex nodes and edges based on pre-existing ones. This is utilized in two main ways. The first one is knowledge graph expansion, where compositionality allows for the easy use of pre-existing knowledge to define new terms and relations. The other one is query building, where compositionality makes retrieving complex information easier.

Natural Language Processing

Compositionality supports all natural language processing tasks that Galaxia performs.

The basic language processing tasks, such as analyzing text structure, dependency parsing, or part of sentence (PoS) tagging are performed automatically without any used involvement. These analyses prepare the text for more advanced tasks.

For tasks such as classification and text or feature extraction, it is enough to specify the type of information or knowledge domain (e.g., health, anatomy, climate emotions) that should be extracted, by providing a simple command that activates Galaxia functions.

Galaxia – first application

smabbler

GALAXIA applications

Galaxia is designed to identify and extract information from text. It serves as a foundational model upon which other models and applications can be developed. Its initial application is automated text labeling.

What is Labeling
In the context of machine learning, data labeling involves adding descriptive labels (features) to provide context, enabling a machine learning model to learn from the data. Data labeling encompasses tasks such as data tagging, annotation, or classification.

Manual effort is typically required for text labeling, which can be costly and time-consuming. While there are methods that utilize the prompting of Large Language Models (LLMs) to obtain labels for training sets, these approaches often suffer from inconsistencies in labels, lack of transparency in results, and the need for manual data cleaning.
In contrast, Smabbler's solution is dedicated to text recognition and extraction, and labeling is a subset of such tasks.

Galaxia for Labeling
Galaxia can be queried to provide labels for unstructured text.
Thanks to its information structure, relationships, and compositionality, Galaxia automates the labeling process without the limitations associated with other approaches, including LLMs. The labels produced are consistent, transparent in origin, and immediately available for use in training machine learning models.

Text labeling: comparison of GALAXIA with LLMs

Galaxia was used to prepare (label) a training set for an ML model built for symptom-based disease identification. To prepare the case study, a publicly available dataset was used. SVC (support vector classifier) was used for the disease classification task.

For the same dataset, Smabbler Galaxia was benchmarked against GPT, Llama, Mistral, and Cohere – with consistently advantageous outcomes.

	Galaxia	**GPT-4**
PROCESSING TIME	5 minutes	2 hours
ACCURACY	96%	86%
EXPLAINABILITY	Yes	No
TECHNOLOGY	Graph	Neural Networks

Large Language Models (LLMs)

Neural network-based LLMs require large amounts of data. To improve their performance, data, storage, and processing capacity are required. To provide reasonable latency, they need large computing resources and expensive hardware, mainly GPUs. When a language model cannot produce a relevant result, it might try to force a response that does not quite fit the input or is factually incorrect, which is termed "hallucinating." LLMs are generative models.

Galaxia Graph Language Model (GLM)

Galaxia GLM utilizes millions of nodes and the relations between them to process natural language. To enhance Galaxia's performance, new domain knowledge can be integrated, and new connections between existing nodes can be created. Galaxia runs on CPUs. When the language model does not have a relevant result, it does not attempt to provide an irrelevant answer. It is designed to identify and extract information from written language. Galaxia is an inference model.

Galaxia – development direction

Next step in technology development: Embeddings

When testing the capabilities of our graph, we confirmed its application in creating a powerful, yet very light, graph embedding system.

What are Word Embeddings

Word embeddings are numeric vector representations used for text analysis. These vectors are created to capture relationships between words, with similar words being represented by similar vectors. The development of word embeddings represents a significant advancement in Natural Language Processing (NLP) and Machine Learning (ML) applications, as it simplifies the generalization process for models.

Graph Embeddings

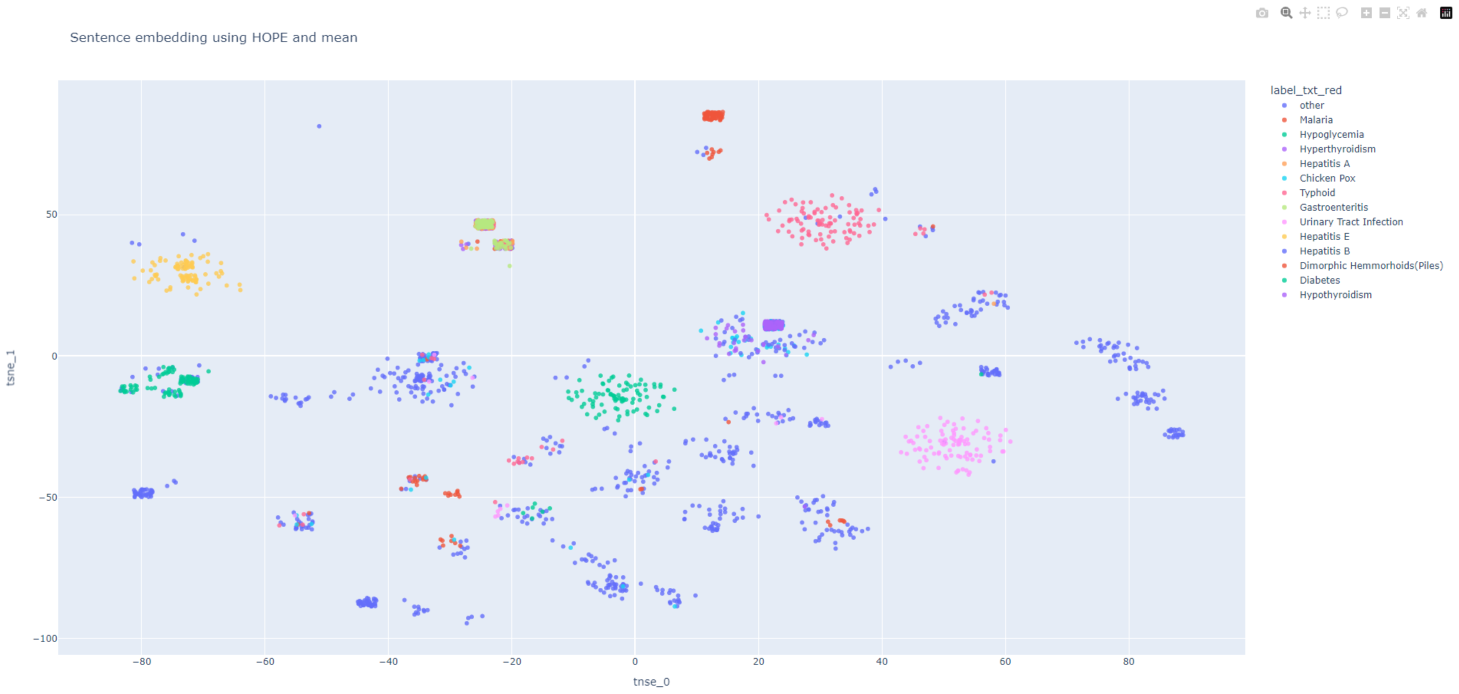
This emergent research area is gaining importance. The intricate relationship structure of semantic graphs presents challenges for use in machine learning applications.

The primary objective of graph embedding is to employ low-dimensional representations in the vector space of a graph and its elements (edges, nodes) while maintaining the graph's structure. These representations have the potential to enrich the input for machine learning models and Natural Language Processing (NLP) applications with semantic information.

Galaxia Graph Embeddings

Thanks to its graph-based language and knowledge representation, Galaxia is a natural fundament for the development of graph-embedding solutions for Natural Language Processing. Instead of using large DNNs, it can produce the embeddings while leveraging advantages of the graph like transparency, computational complexity, and numerical stability.

Galaxia graph embeddings (visualization of results)



Next step in technology development: Graph ML

Graphs differ significantly from typical objects used in Machine Learning due to their complex topology, unlike simple sequences such as strings of text characters. Currently, the primary approach that integrates graphs with machine learning is through Graph Neural Networks (GNNs). This approach aims to utilize graph structure to enhance context and relationships in solutions based on neural networks.

We see potential in taking a different approach by using Machine Learning algorithms to enhance certain features of Galaxia. This concept appears to be free of major challenges. The graph has an organized structure and lacks ambiguous and noisy data, which are common problems when applying graphs in GNNs. By definition, it provides a suitable environment for embedding ML algorithms. Furthermore, the use of ML algorithms could pave the way for expanding Galaxia's capabilities with context-aware natural language generation.

www.smabbler.com

Aga Kopytko
Founder & CTO
akopytko@smabbler.com

smabbler