Georgian Language Extended Representation with Semantic Networks and Processing Using the Wave Principle

Liana Lortkipanidze
Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia
E-mail: liana.lortkipanidze@tsu.ge

Natela Archvadze
Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia
E-mail: natela.archvadze@tsu.ge

Merab Pkhovelishvili
Muskhelishvili Institute of Computational Mathematics, Georgian Technical University
E-mail: m.pkhovelishvili@gtu.ge

Abstract. This paper presents a novel approach to semantic representation and natural language processing (NLP) for the Georgian language. We introduce an extended semantic network (MEN – My Extended Network), adapted from Woods' classical model, which supports transitions between sub-networks and enables context-sensitive processing. Particular focus is given to morphological analysis in Georgian, a language characterized by rich inflectional morphology and complex verb structures. We demonstrate how our system uses 33 base networks and compositional methods to represent over 200,000 active words. A wave-based search algorithm, an extension of breadth-first search with recursive subnetwork transitions, is proposed to efficiently navigate these structures. Implementation is ongoing using Python and Neo4j, and early tests show strong potential for parallelized semantic analysis, machine translation, and information retrieval in Georgian. This approach opens new directions in NLP for low-resource languages..
Keywords: Georgian Language, Extended Semantic Web, Natural Language Processing, Semantic Analysis.

Introduction

With the advancement of information technology, natural language processing (NLP) has become an increasingly important field. NLP encompasses various methods and algorithms for the automatic analysis and generation of text and speech. Due to its unique linguistic features, the Georgian language presents particular challenges for computational processing. However, significant progress has been made in recent years.

Developing software tools for the Georgian language is beneficial for both academic and practical applications. Key areas include text processing (including syntactic and morphological analysis), speech recognition and generation (crucial for voice assistants and translation systems), automatic translation and annotation of Georgian texts, and, finally, the preservation and promotion of the Georgian language online.

This paper focuses on morphological analysis within the broader scope of Georgian language processing while employing extended semantic networks for semantic analysis. Traditionally, computational linguistics has followed a sequential approach: morphological analysis first, then syntactic analysis, followed by semantic analysis. However, in our approach, we justify bypassing syntactic analysis by handling context-dependent languages directly within semantic networks.

Morphological Analysis of the Georgian Language. Morphological analysis is essential for researchers working on the computational processing of the Georgian language. Both theoretical and practical aspects of Georgian morphological analysis are significant. Notably, the widely recognized open-source NLTK library, written in Python, is frequently used as a testbed for syntactic analysis [1].

Several notable works have addressed Georgian morphological analyzers. Among them, [2,6] provides a computational perspective on Georgian morphosyntax, covering morpho-syntactically conditioned long-term dependencies in paradigms and the evaluation of morphological analyzers against a Georgian language corpus.

To address the language-specific focus, we provide a concrete example illustrating the semantic network of a Georgian sentence:

"დედა წავიდა ბაზარში" ("Mother went to the market").

Each word is treated as a vertex in the MEN model, where morphology, syntax, and semantic roles are jointly processed.

The word "წავიდა" is analyzed for root, tense, subject agreement, and is linked to corresponding nodes like <agent>, <action>, and <location>.

This reflects the unique agglutinative nature of Georgian and its capacity for high information density within verbs.

Semantic Representation Methods: Semantic Networks and Frames. Two primary methods exist for semantic representation: semantic networks and frames [3]. Each approach offers distinct advantages depending on the use case. Semantic networks provide a more flexible and generalizable method, making them suitable for various NLP tasks. Frames, on the other hand, offer structured representations ideal for predefined data types.

Extended Semantic Networks (ESN). Extended Semantic Networks (ESN) build upon traditional semantic networks by incorporating deeper semantic relationships. These networks are widely used in knowledge representation, artificial intelligence, the semantic web, and data integration. ESNs facilitate improved information retrieval, automatic inference (knowledge generation), and the unification of disparate databases through shared semantics.

If a system requires modeling knowledge where relationships are complex and dynamic (e.g., NLP and expert systems), network representations are superior. Conversely, frames are more suitable for structured and predefined information.

Our research explores the application of extended semantic networks to the Georgian language and introduces a novel search algorithm within these networks.

Related work includes Helbig (2006) on knowledge representation in semantic networks, as well as studies on morphologically complex languages such as Finnish and Basque, which share structural parallels with Georgian. Works on semantic parsing in under-resourced languages (e.g., Tyers et al., 2020) provide context for our approach.

Extended Semantic Networks for the Georgian Language

We used a tree-based data structure [4] to represent the lexicon of the Georgian language. For morphological analysis, we created 33 networks corresponding to the 33 letters of the Georgian alphabet. When constructing semantic networks, we anticipate having as many extended networks as there are words in the language. Currently, the estimated number of actively used words in Georgian, including dialects and technical terms, is around 200,000.

Sentence Representation Using Extended Semantic Networks. An extended semantic network represents nodes and relationships in a graph structure, capturing objects and their interconnections. Unlike traditional hierarchical models, ESNs allow bidirectional navigation between nodes. We refer to such networks as MEN (My Extended Network).

Woods [5] semantic network formalism provides a foundation for MEN. Our extension introduces subnet transitions: each vertex may point to an entirely new network with its own structure. This enables localized processing (e.g., handling one verb's inflectional complexity) while maintaining global semantic coherence. MEN therefore supports recursive processing and is especially compatible with parallel computing environments.

MEN networks enable bidirectional traversal, including bottom-up navigation. For instance, in Figure 1, the node "information search" links to "global network," and several nodes feature bidirectional connections.

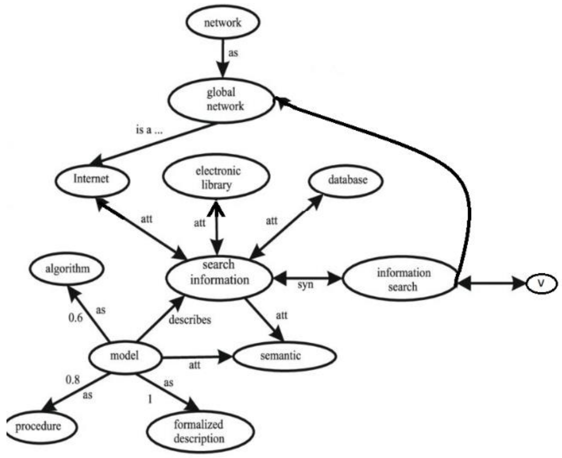

Fig. 1: MEN-Type Network with Transition to a Subnet

In Figure 1, node "information search" connects semantically to the concept of a "global network". The transition vertex V illustrates a subnet jump: the node leads to a specialized subnetwork, for example, for morphological resolution. The result is reintegrated into the main flow, allowing semantically precise parsing without syntactic flattening.

For the representation of knowledge using semantic networks, Woods [5] introduced an additional feature—allowing transitions from any vertex in the network to another network. We apply this idea to MEN networks. If the additional network is of the same type as the main network, this transition would not provide significant value, as the same network could simply continue at that vertex. In this case, the network structure remains intact, and the search methods do not change. However, if the new network specified at the vertex has a different structure, it is processed separately from the main network, and the obtained result is returned to the main network. In Fig.1, vertex V represents a transition to another network. Thus, the purpose of the subgraph is to facilitate transitions between different types of networks and return the results to the main network.

Processing Principle: A sub-network can be processed independently, particularly in multiprocessor systems. The emergence of a new vertex in MEN networks, serving as a transition to a sub-network, primarily depends on the presence of a word in the extended semantic network that is difficult to process and already has a dedicated network for handling it. First, we transition to this new network, process it, and then return to the processed word. The need for an extended network arises from the fact that each vertex itself can represent an entire network. This is the core idea of MEN networks—each word has its own network, and we can transition from any word to another network.

A key feature of MEN is its ability to construct semantic networks compositionally. Starting from lexicon entries, morphological parsing builds the foundational nodes. These are then extended via syntactic roles and contextual dependencies. For Georgian, where verbs encode rich information, the morphological parser first identifies root + affix chains, which then trigger subnet transitions for deeper analysis.

Recent developments in semantic network theory, including Helbig (2006), highlight compositional and frame-based enhancements.

We have contrasted these models with our approach, focusing on the flexibility and recursive structure of MEN networks, particularly their applicability to context-dependent languages like Georgian.

MEN Advanced Network Processing: The Wave Search Algorithm

The **Wave Search Algorithm** is an efficient method for searching within MEN-type extended semantic networks. It operates as follows:

1. The search begins from a marked node.

2. If the target value matches the node value, the search succeeds; otherwise, the process continues in parallel on the first wave of child nodes.
3. If the target is still not found, the search propagates to the second wave (children of child nodes), continuing recursively.
4. The process ends upon locating the target node or reaching terminal nodes.

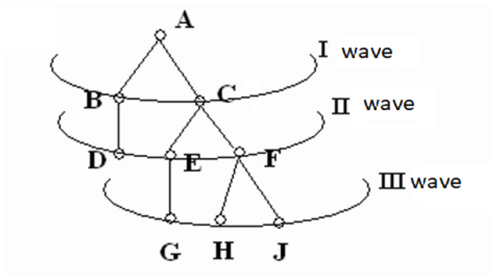

Fig. 2: The Wave Principle in the Semantic Network

In Fig. 2, vertex A represents the network's starting point, vertices B and C belong to the first wave, vertices D, E, and F belong to the second wave, and vertices G, H, and J belong to the third wave.

Since it is possible to move from any vertex of the semantic network to a lower level, there is a chance that a previously visited vertex may be encountered again during the descent. In such cases, if the vertex is already present in the recorded set, it should be traversed only once to prevent looping.

Wave Search Algorithm Considerations

- Level Limitations: A maximum number of waves must be defined to prevent infinite searches.
- Vertex Marking: Nodes must be marked to avoid redundant processing.

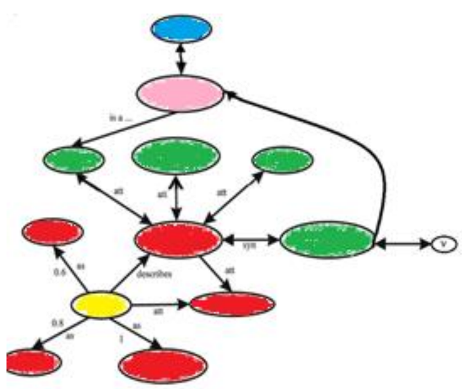

Fig. 3: Wave Propagation in the MEN Extended Network

In this image, the starting vertex is marked in yellow, the vertices of wave I are marked in red, the vertices of wave II in green, the vertices of wave III in purple, and the vertices of wave IV in blue. The lower right section will not be discussed further, as it has already been covered. The vertex of the first wave, named "semantic," also appears in the second wave but is not processed a second time. It is also important to note that vertex V represents a transition to another network. In this case, the second level must first be fully transferred to the other network, processed, and only then can the algorithm proceed with the third and fourth waves.

While the wave search algorithm resembles BFS in layered exploration, it diverges in its capacity to transition into and return from subnets with independent processing logic.

This recursive subnet management enables handling of semantic complexities that BFS does not support, particularly in multi-processor systems.

MEN-Type Networks. When programming on multiprocessor computers, Woods' idea takes on new significance, as processing within a subnet can be performed independently on separate processors, making it efficient for parallel computing. For example, suppose a subject area is described using logical-linguistic methods. The knowledge-based sentences in the main network are represented by semantic networks, while the morphological analysis of words is carried out in subnets. This clearly indicates that the network and subnet have different structures.

In another example, the main network structure represents the alphabet of a natural language, while the subnetwork contains complete, explanatory information about words translated into one or more foreign languages. In this case, the structure of the subnetwork also differs from that of the main network.

The need to combine morphological and semantic analysis directly—without a full syntactic layer—is motivated by Georgian's inflection-heavy nature. For instance, verb roots often contain embedded information about subject, object, and tense. Thus, a MEN node can interpret such information without relying on a traditional syntactic parser. This enables applications such as real-time translation, semantic indexing, and complex question-answering systems.

Conclusion[1]

The discussed methods enable faster and more effective data processing, contributing to reliable performance in modern systems such as Internet search engines. Key takeaways include:

- The Wave Search Algorithm starts from a marked node and requires predefined search conditions.
- Woods' modifications introduce new network transitions.
- Subnet processing is independent and highly efficient in multiprocessor systems.
- Network structures are vital for modern information retrieval systems.

Computational processing of the Georgian language is at a crucial stage. The discussed knowledge representation and retrieval methods, along with algorithmic advancements, will enhance future technological adaptation and digital presence of the Georgian language.

References

1. Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit.* O'Reilly. https://www.nltk.org/book/
2. Irina Lobzhanidze. *Finite-State Computational Morphology: An Analyzer and Generator for Georgian.* Springer, 2022. DOI: https://doi.org/10.1007/978-3-030-90248-3
3. N. Archvadze, M. Pkhovelishvili, L. Shetsiruli. *Morphological Analysis of Words in Clusters.* ISBN 978-5-94777352-1.
4. N. Archvadze, M. Pkhovelishvili. *Georgian Dictionary Representation through Functional Programming.* GESJ: Computer Science and Telecommunications, 2012.
5. William A. Woods. *Foundations for Semantic Networks.* Academic Press, 1975.
6. Lortkipanidze L., Amirezashvili N., Chutkerashvili A., Javashvili N., Samsonadze L., (2017). "Syntax Annotation of the Georgian Literary Corpus", Theoretical Computer Science and General Issues. 11th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2015, Tbilisi, Georgia, 21-26 September, 2015, Revised Selected Papers, pp. 89–97

[1] This article is published with the support of the Shota Rustaveli National Science Foundation of Georgia (SRNSFG) under grant FR-24-11637