Layered Augmented Transition Networks (LATN): A Modular Approach to Language Parsing and Grounding

Abstract

This paper introduces a novel layered architecture for natural language understanding based on Augmented Transition Networks (ATNs), termed Layered ATN (LATN). Inspired by the layered processing seen in transformer models, LATN decomposes the parsing process into multiple passes. Each pass incrementally builds syntactic and semantic representations of an input sentence. Grounding occurs incrementally within each layer, particularly leveraging scene models in early stages to disambiguate references and guide interpretation. We show that this modular design provides a transparent and interpretable framework for processing language in simple visual domains and offers a foundation for building more complex grounded language models.

1. Introduction and Architecture

Natural language processing systems traditionally face challenges when attempting to parse and interpret complex sentences in a single pass. LATN proposes a modular solution where distinct layers handle lexical analysis, noun phrase extraction, prepositional phrase resolution, verb phrase construction, and anaphora resolution, each building on the output of the previous layer. This approach is inspired by the multiple layers of abstraction observed in transformer architectures, where early layers capture local syntactic structure and later layers support deeper semantic reasoning.

2. Grounding in LATN vs Transformers

A key distinction between the LATN architecture and contemporary transformer-based models lies in the mechanism and timing of grounding. In LATN, grounding is performed explicitly and modularly, with each layer progressively transforming linguistic constructs into references to entities and relationships in an external scene model. Specifically, noun phrase identification in Layer 2 is followed by scene grounding via a SceneObject search, producing one or more referents that represent candidate bindings in the modeled environment. This tightly couples syntactic parsing with semantic resolution in a way that is transparent and interpretable. By contrast, transformer-based architectures encode grounding implicitly through training: scene elements must be embedded into the same high-dimensional space as language tokens, and grounding emerges via the learned

attention weights across multimodal vectors. While this approach enables remarkable flexibility and generalization, it often lacks transparency and requires extensive data-driven training. LATN's approach, in contrast, is more akin to symbolic scene parsing layered on top of vector semantics, allowing for deterministic grounding in simple domains while retaining the extensibility to handle ambiguity and anaphora through deeper layers of interpretation.

3. Conclusion and Future Work

The LATN architecture offers a compelling model for incremental and interpretable language understanding grounded in external scene representations. Future work will explore deeper layers of inference, broader grammatical constructions, and potential integrations with transformer-style attention mechanisms. Ultimately, this approach may serve as a scaffold for hybrid neuro-symbolic systems capable of reasoning in both structured and unstructured domains.