

Controlled Natural Language Models

Jacinto A. Dávila Q.

CESIMO,
Universidad de Los Andes,
Mérida,
Venezuela

Table of Contents

- 1 Controlled Natural Language
- 2 What is a Model?
- 3 Large Language Models are models
- 4 Grammars as Models of Languages
- 5 Combining Grammars and LLMs: A partial taxonomy of approaches
- 6 Preparing for a comprehensive, hybrid approach: a logic programming grammar for a domain specific controlled natural language

Controlled Natural Language

A controlled natural language (CNL) is a simplified version of a natural language designed to be more precise and unambiguous. It achieves this by restricting the grammar, vocabulary, and syntax of the language. CNLs can be used to represent complex information in a human-readable format.

What is a Model?

Definition

A model is a simplified representation of something. It helps understand, analyze, or predict the behavior of a system or object without dealing with full complexity.

Properties

Models must explain and predict the behavior of a system, even if partially.

Large Language Models are models

Large Language Models (LLMs) are indeed models. They can be seen as complex systems of equations (neural networks) designed to represent and predict language patterns under certain conditions:

- They are simplified representations: LLMs don't capture the full complexity of human language, but they implicitly model key aspects like grammar, semantics, and context.
- They are used for prediction: Given a prompt, an LLM predicts the most likely continuation of the text.
- They are based on data: LLMs are trained on massive amounts of text data to learn language patterns.

Predictability vs. Explain-ability

At some point we used to believe that LLMs were not models, because they are a black box, unaccountable from a symbolic point of view. We changed our opinion after considering that LLMs do represent the grammatical rules of a language, even if they do so in a representation that is not symbolically meaningful. What counts here is that the LLM can actually predict the syntax of expressions in natural language. Thus, with LLMs, predictability takes over from explain-ability, an unusual turn of events.

Grammars as Models of Languages

Definition

A grammar is a set of rules that define the structure of a language.

Formal models

Many grammatical theories, like context-free grammars or dependency grammars, are formal models that use mathematical or symbolic representations to describe language structure.

Generative vs Descriptive

Generative

Grammars aim to generate all and only the well-formed sentences of a language. In this sense, they act as a generative model, even when there is no stochastic process underneath as in the LLMs.

Descriptive

Other grammars are primarily descriptive, aiming to accurately represent the structure of a language as it is used, but they still function as models of that language.

Combining Grammars and LLMs: A partial taxonomy of approaches

1.- Grammar-Guided LLM Generation:

- Constraint-based generation: Use a grammar to define the structural constraints of the output. The LLM then generates text within these constraints. This can improve the grammatical correctness and coherence of the generated text.
- Grammar-informed decoding: Incorporate grammatical knowledge into the decoding process of the LLM. The decoding process produces candidates that are tested for correctness by the grammar.

(Answer Generated by AI tool Gemini)

Combining Grammars and LLMs: A partial taxonomy of approaches

2.- LLM-Enhanced Grammar Induction:

- Grammar extraction: Use LLMs to extract grammatical patterns from large amounts of text data. This can help in automatically constructing or refining grammars.
- Grammar evaluation: LLMs can be used to evaluate the quality of generated text based on a given grammar. This can help in grammar refinement.

(Answer Generated by AI tool Gemini)

Combining Grammars and LLMs: A partial taxonomy of approaches

3.- Hybrid Models:

- Grammar-augmented neural networks: Combine neural networks with formal grammar components. This approach can leverage the strengths of both models.
- Sequence-to-sequence models with grammar injection: Incorporate grammatical information into the input sequence of a sequence-to-sequence model.

(Answer Generated by AI tool Gemini)

a logic programming grammar for a domain specific controlled natural language

we will cover

any amount which you are legally liable to pay
in respect of

a damage which is a

1 bodily injury **or**

2 personal injury **or**

3 property damage **or**

4 nuisance **or**

5 trespass

if the damage occurs during the period of insurance

and the damage occurs in connection with the business.

The translation into Prolog of the rule before

```
we_will_cover(A) :-
    is_a(A, amount),
    you_are_legally_liable_to_pay(A),
    is_a(B, damage),
    is_in_respect_of(A, B),
    (
        is_a(B, bodily_injury)
    ;   is_a(B, personal_injury)
    ;   is_a(B, property_damage)
    ;   is_a(B, nuisance)
    ;   is_a(B, trespass)
    ),
    occurs_during_the_period_of_insurance(B),
    occurs_in_connection_with_the_business(B).
```

A running example:

(Answer Generated by AI tool Gemini begins)

- who affiliated with which entity at which date.
- what entity affiliated with which company at which date.
- when did which entity affiliate with which other entity.
- on what date did which entity affiliate with which company.
- is there an affiliation between which entity and which company.
- when did the affiliation between which entity and which company begin.

(Answer Generated by AI tool Gemini ends)

Demo Online

The Logical English parser has been updated accordingly and the example above can be verified and tested online