Using Large Language Models in the Companion Cognitive Architecture: A Case Study and Future Prospects

Constantine Nakos, Kenneth D. Forbus
Qualitative Reasoning Group, Northwestern University
cnakos@u.northwestern.edu, forbus@northwestern.edu

Abstract

The goal of the Companion cognitive architecture is to understand how to create human-like *software social organisms*. Thus natural language capabilities, both for reading and for conversation, are essential. Recently we have begun experimenting with large language models as a component in the Companion architecture. This paper summarizes a case study indicating why we are currently using BERT with our symbolic natural language understanding system. It also describes some additional ways we are contemplating using large language models with Companions.

Introduction

For a *software social organism* to learn, reason, and interact like a human, it must be able to read and converse in natural language. The Companion cognitive architecture (Forbus & Hinrichs 2017) addresses this challenge with CNLU (Tomai & Forbus 2009), a rule-based semantic parser that produces interpretations in the CycL knowledge representation language that are grounded in the NextKB[1] ontology. CNLU uses FrameNet[2] mappings to OpenCyc in its lexical semantics. It also can use task constraints to help guide abduction, either using domain-specific rules (Tomai & Forbus 2009) or via analogical Q/A training (e.g. Crouse et al. 2018; Wilson et al. 2019). For example, Companions have modeled moral decision-making (Dehghani et al. 2008) and conceptual change (Friedman & Forbus 2009) using stimuli provided using natural language, and language has been used to provide hints to Companions learning complex games (McFate et al. 2014) and performing commonsense reasoning (Blass & Forbus 2017).

One of the advantages of CNLU is that its representations are discrete and inspectable. Unlike neural models, errors it produces can be examined and corrected, a valuable property when precise understanding is essential. It also features a broad-coverage lexicon, and uses capabilities of the Companion architecture for further interpretation and reasoning.

Still, CNLU has its limitations. Disambiguation is a perennial challenge, as choosing between alternative interpretations can depend on subtle commonsense knowledge as well as task constraints that constitute an important component of context. Lexical and grammatical coverage remains an issue due to the sheer breadth of natural language. More generally, we would like CNLU to benefit from recent advances in NLP while retaining its inspectable symbolic representations and ontological grounding.

This paper summarizes our ongoing efforts to integrate large language models (LLMs) within the Companion architecture. The next section presents a case study of how we have already used BERT (Devlin et al. 2018) to assist CNLU with learning by reading, demonstrating that while analogy alone beats BERT alone, analogy plus BERT outperforms analogy alone. The final section outlines three further approaches we are considering for integrating LLMs into the Companion architecture.

Case Study: Learning by Reading

Learning by reading has been explored in Companions in several ways, including learning from instructional analogies (e.g. Barbella & Forbus 2010), learning from multimodal reading (text plus sketches; e.g. Chang 2016), and learning from reading legal cases (e.g. Blass & Forbus 2022, 2023). Until recently, our efforts relied on hand-simplified English text, which does not scale well. Consequently, we explored how to use an LLM (BERT) to augment our analogical Q/A training (AQAT) technique. Prior AQAT work often used ML datasets as the sources of questions and answers, using a connection graph algorithm to induce rule-like constructions (*query cases*) that are then used to help interpret later statements, applied compositionally via analogy. Instead of ML datasets or other hand-built training sets, Ribeiro and Forbus (2021) had a Companion align facts from the NextKB knowledge base with statements from

[1] http://www.qrg.northwestern.edu/nextkb/index.html
[2] https://framenet.icsi.berkeley.edu/

Simple English Wikipedia to learn query cases, a form of distant supervision. It then used these query cases to extract new facts from unseen Simple English Wikipedia text, extracting 66,649 facts with 3,745 distinct relations. Unlike many relation extraction techniques, the facts produced are more than just triples and can contain complex internal structure.

The method uses BERT in two ways. The first is to assist CNLU with disambiguation. The authors fine-tuned BERT on FrameNet data to produce a frame classifier. As most CNLU interpretations are realizations of FrameNet frames, the classifier can be used to choose between them, allowing the system to pick the most likely semantic interpretation. For example, in the sentence "Male deer have antlers", CNLU produces multiples interpretations for *have* corresponding to the FrameNet frames Giving Birth, Ingestion, Opinion, and Possession. The BERT classifier assigns a softmax score of 0.998 to Possession, selecting it as the best interpretation. As AQAT depends on accurate semantics for constructing and applying query cases, this disambiguation step is crucial.

The second use of BERT is to predict fact plausibility. The authors separately fine-tuned BERT to classify facts as plausible or implausible. The resulting classifier is then used to filter the output of the knowledge extraction system, improving the estimated precision of extracted facts from 45.7% (analogy alone) to 71.4% (analogy + BERT). Importantly, BERT alone only managed 20.8% precision, and was limited to simpler forms of facts (i.e. triples). This demonstrates the potential payoffs in integrating LLMs into knowledge-rich cognitive architectures.

Further Prospects for Using LLMs

Here we discuss three additional approaches we are exploring to integrating LLMs into the Companion cognitive architecture. We suspect that, like our learning by reading work, this could benefit any knowledge-rich language using cognitive architecture.

Low-Resource Disambiguation

While the BERT-based disambiguation system described above is useful it suffers from two limitations. First, because the classifier is trained on FrameNet data, its lowest unit of resolution is the FrameNet frame. This is sufficient for making coarse-grained distinctions, such as between the Ingestion and Possession senses of the word *have*. But NextKB is a finer-grained ontology than FrameNet, so some of the interpretations that CNLU produces for a word can correspond to the same FrameNet frame. Making these subtler distinctions will require either finer-grained training data, which is difficult to come by, or a different approach.

Second, as the coverage of CNLU grows, the coverage of the classifier will need to grow with it. Sourcing high-quality semantic annotations to support a constantly growing language system is a daunting prospect. The availability of FrameNet annotations makes them a useful starting point, but the more CNLU's coverage expands, the greater the gap between the available data and that which is needed for successful disambiguation.

These limitations motivate our ongoing work on using the few-shot capabilities of LLMs to help CNLU choose between candidate interpretations. The predictive power of modern LLMs allows them to learn new tasks on the fly, reducing the amount of task-specific training data to a few examples (*few-shot learning*) or none (*zero-shot learning*). By posing the disambiguation task to an LLM in natural language, we hope to tap into the model's knowledge and select the correct interpretation with little annotated data.

We are exploring several schemes for probing the LLM. The challenge lies in converting CNLU's predicate calculus representations into natural language without introducing ambiguities. We are considering using templates and fixed strings to generate English paraphrases of the interpretations for an LLM to choose between, using richer forms of generation to paraphrase the entire sentence according to each of its interpretations, and asking the LLM diagnostic questions to iteratively rule out incorrect interpretations. All of these options seek to frame the task in such a way that the LLM's vast statistical knowledge can be brought to bear.

These approaches have their difficulties. NextKB offers substantial but imperfect support for template-based text generation, making gaps in coverage a risk. More flexible generation might build on CNLU's language-to-meaning mappings, but at the cost of added complexity. Formulating a useful set of diagnostic questions could also be difficult thanks to the open-domain nature of NextKB's ontology. Finally, it remains to be seen whether current LLMs can answer the questions we wish to pose them in a zero- or few-shot setting within a reasonable computational budget.

Simplifying Text Inputs

One weakness of rule-based NLU systems is their brittleness in the face of natural text. Non-standard spelling and grammar, long and complicated sentences, and rare words can all confound even an advanced semantic parser.

LLMs provide an opportunity to shore up this weakness by simplifying the input text into something manageable by the rule-based system. Ideally, this would combine the flexibility of neural language models with the inspectability and high-precision semantics of rule-based systems. Work by Rongali et al. (2022) on regularizing spoken language text for processing by a symbolic language system suggests that this approach has promise.

For example, with Bing Search in Precise mode,

"Please simplify the following sentence into multiple sentences, dividing the description of any physical process into a separate sentence for each fact: "Heat flows from the hot brick to the cold ground because the brick is hotter than the ground.""
yields this simplification:
"The brick is hot. The ground is cold. The brick is hotter than the ground. Because of this temperature difference, heat is flowing from the brick to the ground."

Simplification in this context poses some interesting challenges. First, the simplified text must be faithful to the original. Fidelity is always a goal for text simplification, but it becomes paramount when the text is used for machine reasoning rather than human consumption. One stark example of this comes from a personal experiment with text simplification in the legal domain. Tasked with simplifying a legal case summary, an LLM produced a reasonable simplification but omitted the fact that the police had a search warrant when they broke into a home. Clearly, such omissions can be damaging to downstream reasoning.

Second, the simplified text must actually be easier for the system to parse. It does no good to paraphrase a sentence if it still contains the phrase or grammatical construct that caused CNLU to fail in the first place, and even sentences that appear to be simple can lead to broken or inadequate parses. Several possibilities for addressing this problem exist, such as incorporating CNLU failures into the simplification model's loss function during training or iteratively re-simplifying sentences that still trip up CNLU.

One promising direction is the approach taken by Kirk et al. (2022), who use prompt engineering to encourage an LLM to produce text readable by a Soar agent, finding that terse prompts with a moderate amount of detail produce the best results. In addition, they experiment with interactive generation, where the agent guides the LLM to generate words it is capable of parsing. Prompt engineering and interactive generation should both be helpful for having an LLM produce text simplifications CNLU can understand.

Generating Predicate Calculus from Text

Perhaps the most challenging application of LLMs we are considering is using them to generate predicate calculus from text directly, bypassing CNLU entirely. This approach sacrifices the inspectability of CNLU's internal representations for the power of an LLM. Work such as Wu et al. (2023) suggest that this approach has promise.

What makes using an LLM to generate predicate calculus challenging is the sheer breadth of the NextKB ontology. With over 83,000 collections, 26,000 relations and 5,000 functions to learn, it is not clear how to induce an LLM to produce accurate output without significant amounts of annotated training data and fine-tuning. We are exploring the use of Natural Language Generation to synthesize training data and bootstrap this process, but it remains to be seen how much data is necessary and whether our synthetic examples will generalize to parsing natural text.

Acknowledgments

This research was supported by the US Office of Naval Research.

References

Blass, J. & Forbus, K. 2017. Analogical Chaining with Natural Language Instruction for Commonsense Reasoning. *Proceedings of AAAI 2017*.

Blass, J. & Forbus, K. 2022. The Illinois Intentional Tort Qualitative Dataset. *Proceedings of the 35th International JURIX Conference on Legal Knowledge and Information Systems*.

Blass, J. & Forbus, K. 2023. Analogical Reasoning, Generalization, and Rule Learning for Common Law Reasoning. *Proceedings of ICAIL 2023*.

Barbella, D., and Forbus, K. 2010. Analogical dialogue acts: Supporting learning by reading analogies. Proceedings NAACL HLT 2010: 1st Int. Workshop on Formalisms and Methodology for Learning by Reading.

Blass, J. & Forbus, K. 2022. The Illinois Intentional Tort Qualitative Dataset. *Proceedings of the 35th International JURIX Conference on Legal Knowledge and Information Systems.*

Chang, M.D. 2016. *Capturing Qualitative Science Knowledge with Multimodal Instructional Analogies*. Doctoral dissertation, Northwestern University, Department of Electrical Engineering and Computer Science, Evanston, Illinois.

Crouse, M., McFate, C.J., and Forbus, K.D. 2018. Learning from Unannotated QA Pairs to Analogically Disambiguate and Answer Questions. *Proceedings of AAAI 2018*.

Dehghani, M., Tomai, E., Forbus, K., Klenk, M. 2008. An Integrated Reasoning Approach to Moral Decision-Making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (AAAI). Chicago, IL.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805.

Forbus, K.D. & Hinrichs, T. 2017. Analogy and Qualitative Representations in the Companion Cognitive Architecture. *AI Magazine*.

Friedman, S. and Forbus, K. 2009. Learning Naïve Physics Models and Misconceptions. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

Kirk, J., Wray, R., Lindes, P. & Laird, J. 2022. Improving Language Model Prompting in Support of Semi-autonomous Task Learning. ArXiv.

McFate, C.J., Forbus, K. and Hinrichs, T. 2014. Using Narrative Function to Extract Qualitative Information from Natural Language Texts. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada.

Ribeiro, D. & Forbus, K. 2021. Combining Analogy with Language Models for Knowledge Extraction*, Proceedings of the Third Conference on Automatic Knowledge Base Construction.*

Rongali, S., Arkoudas, K., Rubino, M., Hamza, W. 2022. Training Naturalized Semantic Parsers with Very Little Data, *Proceedings of IJCAI 2022*.

Tomai, E. and Forbus, K. 2009. EA NLU: Practical Language Understanding for Cognitive Modeling. *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference*. Sanibel Island, Florida.

Wilson, J., Chen, K., Crouse, M., C. Nakos, C., Ribeiro, D., Rabkina, I., Forbus, K. D. 2019. Analogical Question Answering in a Multimodal Information Kiosk. In *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems.* Cambridge, MA.

Wu, Y., Jiang, A., Li, W., Rabe, M., Staats, C., Jamnik, M., & Szegedy, C. 2023. Autoformalization with Large Language Models. *Proceedings of NeurIPS.*