Boosting a Semantic Parser Using Treebank Trees Automatically Annotated with Unscoped Logical Forms

Miles Frank and Lenhart Schubert
mfrank14@u.rochester.edu
schubert@cs.rochester.edu
Department of Computer Science, University of Rochester, Rochester, NY 14627 USA

Abstract

Deriving structured semantic representations from unrestricted text, in a format suitable for sound, explainable reasoning, is an important goal for achieving AGI. Consequently much effort has been invested in this goal, but the proposed representations fall short in various ways. Unscoped Logical Form (ULF) is a strictly typed, loss-free semantic representation close to surface form and conducive to linguistic inference. ULF can be further resolved into the more precise Episodic Logic. Previous transformer language models have shown promise in the task of parsing English to ULF, but suffered from a lack of a substantial dataset for training. We present a new fine-tuned language model parser for ULF, trained on a greatly expanded dataset of ULFs automatically derived from Brown corpus Treebank parse trees. Additionally, the model uses Parameter Efficient Fine Tuning (PEFT) to leverage a substantially larger base model than its predecessor while maintaining fast training times. We find that training on automatically derived ULFs substantially improves parser performance from the existing smaller dataset (from SEMBLEU score of 0.43 to 0.68), or even the previously used larger, generatively augmented ULF dataset, used with a transition parser (from SEMBLEU score of 0.49 to 0.68).

1 Introduction

Large language models (LLMs) have revolutionized the interactive generation of fluent, coherent text by machines, but their functioning is hidden in their millions or billions of parameters. This blurs the distinction between knowledgeable output and confabulation. Moreover, because they rely on probabilistic mimicry of their vast training data, rather than on rational thought, they do not reason or plan with the kind of reliability and scalability that is required for consequential applications in areas like healthcare, legal matters, police operations, or search and rescue. Ultimately, artificial general intelligence (AGI) requires the ability to reason and plan reliably at scale, and to explain how conclusions or plans were arrived at. For reasoning to be explicit and auditable, the knowledge and rules employed must themselves be made explicit and sufficiently unambiguous. You cannot tell whether *"Alice warned the woman that Bob had left"* plausibly entails *"Bob had left"* or instead, *"Bob had left the woman,"* without clarifying the semantic structure of the premise.[1] Thus effective representation of linguistic content and background knowledge forms the cornerstone of systems designed not only to converse fluently, but also to reason and plan reliably. Such representations should be derivable from language, and enable semantic inference, discourse processing, and explicit, explainable reasoning. Kim and Schubert (2019) describe Unscoped Logical Form (ULF), one such knowledge representation (with a lengthy prior history, e.g., Hwang and Schubert, 1994; Schubert and Hwang, 2000), as an alternative to other popular representations, because it preserves more of the semantic information of natural language while maintaining a strict type system supporting well-founded, natural inference.

Due to their retention of all sentential information and their coherent type structure, ULFs lend themselves to *natural logic*-like inference (Kim et al., 2021c,b), discourse inferences including clause-taking verbs, counterfactuals, questions, requests, and generalizations (Kim et al., 2019), as well as schema-based story representation (Lawley et al., 2019). ULFs, and their subsequent resolu-

[1] As a preview, the alternative VP logical forms are these (hinging on reifier `that` vs. relativizer `that.rel`):

```
((PAST warn.v) (the.d woman.n)
 (that (| Bob| ((PAST have.aux) (PERF leave.v)))))

((PAST warn.v) (the.d (n+preds woman.n (sub that.rel
    (| Bob| ((PAST have.aux) ((PERF leave.v) *h)))))))
```

tion into Episodic Logic, have also proven to be a useful representation for inference within interactive natural language understanding systems (Kane et al., 2020, 2023). Improving the scope and accuracy of ULF parsers will enable generalization of such systems. To provide an initial idea of the form of ULFs and their application to inference, here are three simple examples of the ULFs for the sentences *"Bob pretended to be asleep"*, *"Alice often kids Bob"*, and *"I wish I had turned off the stove"*, along with some inferences derivable by the cited methods:

((| Bob| ((PAST pretend.v) (to (be.v asleep.a)))))
⇒ (| Bob| ((PAST be.v) (not asleep.a)))

(| Alice| frequently.adv-f ((PRES kid.v) | Bob|))
⇒ ((a.d person.n) sometimes.adv-f ((PRES tease.v) (a.d person.n)))

(I.pro ((PRES wish.v) (tht (I.pro ((cf have.aux-s) ((PERF turn_off.v) (the.d stove.n)))))))
⇒ (I.pro ((PAST do.aux-s) not.adv-s (turn_off.v (the.d stove.n))))

(Some syntactic explanations follow later.) Their similarity to surface form should enable the reader to understand the inferences. Unlike inferences by LLMs, such ULF-based inferences are explainable in detail, in this case in terms of the implications of "pretending to," from the plausible assumption that "Bob" and "Alice" are instances of persons, from the entailment "frequently" ⇒ "sometimes," from the approximate synonymy of "kid" and "tease" (as verbs), and (in the last example) from the properties of counterfactual entailment of the subjunctive form. Resolving ULFs into Episodic Logic (EL) involves systematic deindexing, scoping, and reference resolution processes, and this more precise representation enables a superset of FOL inferences as well as uncertain inferences, in conjunction with miscellaneous world and lexical knowledge, and with support from taxonomic, temporal, arithmetic, and other specialist subsystems (e.g., Schubert, 2014). If necessary, ULF can be further converted to Episodic Logic for more granular inference. Resolving ULFs into Episodic Logic (EL) involves systematic deindexing, scoping, and reference resolution processes, and this more precise representation enables a superset of FOL inferences as well as uncertain inferences, in conjunction with miscellaneous world and lexical knowledge, and with support from taxonomic, temporal, arithmetic, and other specialist subsystems (e.g., Schubert, 2014).

The main contributions of this paper are (1) the demonstration that a large corpus of syntactically annotated sentences from a wide spectrum of sources (the Brown corpus) can be rather reliably mapped to ULF – an English-like, highly expressive, coherently typed initial logical form previously shown to be suitable for inference; and (2) the ULF-annotated sentences thus obtained together with a small hand-annotated "gold" training set can be used to fine-tune an LLM for semantic parsing, obtaining a level of accuracy strikingly better than obtained by previous ULF parsers, and comparable to results obtained for other, less comprehensive semantic representations that used much larger hand-annotated training sets than our "gold" corpus.

In the remaining sections, we comment on related representations and prior ULF parsers (Section 2), our rule-based annotation of the Brown corpus Penn Treebank (Marcus et al., 1993) POS tags to obtain a greatly expanded ULF training set (Section 3), our models for fine-tuning and the success metrics (Section 4), and the results with our methods, comparing these to relevant previous semantic parsers (Section 5). We summarize and reiterate our results in the Conclusion (Section 6).

2 Related Work

2.1 Other Knowledge Representations

We briefly discuss the pros and cons of other contemporary knowledge representations including generic First Order Logic (FOL), Discourse Representation Theory (DRT), Abstract Meaning Representation (AMR), and Minimal Recursion Semantics (MRS). Perhaps the most simply formatted representation, FOL is easy to generate inferences from and expressive enough to represent the meaning of most simple, matter-of-fact sentences. Through the use of various syntactic and semantic maneuvers, FOL can also be adapted to sentences that involve more subtle subject matter. However, the required circumlocutions are apt to be awkward and remote from surface form. For example, they may require explicit quantification over possible worlds, or functionalizing of all predicates and quantifiers, and application of a "Holds" or "Is True" predicate to functionalized sentences (Schubert, 2015).

To address some pronoun resolution issues in the conversion of natural language to FOL, Kamp (1981) and Heim (1982) developed Discourse Rep-

resentation Theory. The nested structures in this theory contain free variables to be dynamically interpreted; but because Discourse Representation Theory is convertible to FOL, it shares the expressive limitations of the latter. (An extension of DRT allowing for mental states and attitudes, MS-DRT, seems not to have been deployed as yet in semantic parsing.)

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is less focused on echoing the syntax of sentences, instead striving to represent sentences of similar meaning but different wording as the same AMR graph structure. This is useful in detecting meaning similarity or equivalence, and reduces the need for inferences, such as a "collide" event occurred, given that "Bob was injured in a collision". However, AMR drops important aspects of meaning (such as tense, and the distinction between hypothetical events and real ones), and makes insufficient commitments about the semantic types of its constituents (such as modifiers and quantifiers) to be suitable for reliable inference (again see Schubert, 2015, where other representations are considered as well). The more recent multilingual Uniform Meaning Representation (UMR) (Van Gysel et al., 2021) extends AMR to include temporal and modal dependencies, but due to limited training corpora, the only available parsers use a pipeline approach by first parsing the AMR and then automatically converting to UMR (Chun and Xue, 20240815–20240815).

In view of the considerable attention that AMR has received in the research literature of the last decade, some quick comparisons of AMR and ULF structures can provide an intuitive idea of their characteristics and differences, particularly for readers unfamiliar with ULF. Consider the sentences

> 1. *The broadcast asserted that chemicals were dumped into the river.*
> 2. *The broadcast showed chemicals being dumped into the river.*

The AMR representations of these sentences are identical except for the respective event predicates {assert-02, show-01}:

```
(z0 / {assert-02, show-01}
    :ARG0 (z1 / broadcast
    :ARG1 (z2 / dump-01
              :ARG1 (z3 / chemical)
              :destination (z4 / river)))
```

Note the free variables, generally assumed to be existentially bound at the top level. For version (1), this roughly says that a broadcast z1 asserts an event z2 of dumping a chemical z3 into a river z4. Besides the neglect of tense, one issue is that a dumping event is implicitly assumed to exist, not allowing for a false assertion ("assert" should create an opaque context). Another is that "assert" should take a proposition, not an event, as object argument. (You can assert the Second Amendment, but not the Second World War.) The AMR representation works better for version (2), insofar as it's entirely possible that a broadcast might show a chemical dumping event.

The following are the quite distinct ULF interpretations automatically obtained for (1) and (2) (where the tags ∼1, ∼2, ... indicate positions of corresponding input words, needed for reference resolution and other pragmatic phenomena; they are omitted for ULF evaluations):

```
(((the.d~1 broadcast.n~2)
 ((PAST assert.v~3)
  (that~4
   ((k (plur chemical.n~5))
    ((PAST be.aux~6)
     ((pasv dump.v~7)
    (adv-a (into.p~8 (the.d~9 river.n~10)))))))))
\.)

(((the.d~1 broadcast.n~2)
 ((PAST show.v~3)
  ((k (plur chemical.n~4))
   ((PROG be.aux~5)
    ((pasv dump.v~6)
     (adv-a (into.p~7 (the.d~8 river.n~9))))))))
\.)
```

Some points to note in these examples (as well as the earlier introductory ones) are type/sortal distinctions indicated by dot-suffixes like .d (determiner), .n (nominal predicate), .v (verbal predicate), etc.; and the retention of tense, definite determiners, and plurals. 'plur' shifts a predicate true or false of single entities to a predicate true or false of sets of entities. The operator 'k' type-shifts a monadic predicate P to the abstract *kind* (k P) whose realizations satisfy P.[2] Most notably, the type-shifting operator 'that' in the first ULF maps a sentence meaning to a propositional individual (see Kim and Schubert, 2019). While the proposition exists, it need not be true and the entities it introduces need not exist – this is a matter of inference, for instance for a trustworthy report. In the second ULF, the verbal predicate 'show.v' is treated as taking an object (theme) – namely chemicals, and a predicate – namely, the property of being dumped

[2] But acting on a kind entails acting on an instance of the kind – here, an instance of the kind, chemicals.

into the river, as arguments. (Predicate arguments cannot be quantified over, and the logic remains first-order.)

Minimal Recursion Semantics (MRS) (Flickinger et al., 2012) shares some features with DRT and AMR, though it deals fully with restricted quantification, attitudes, and other phenomena. In its "native form" it uses ordinary predicate + arguments syntax, but assigns names (handles) to predications, using these as placeholders for embedded predications. However, the semantic representations seem under-determined in terms of type structure, and are somewhat hard to understand, because of the indirectness of the structural descriptions – use of handles to flatten the representation, span indices to indicate the scope of handles, and arguments of predicates that include, besides handles (sometimes undefined), various types of unbound variables that are presumably to be closed existentially with some appropriate scope. It is unclear if MRS is intended for reasoning, but we are not aware of recent work in that direction.

2.2 Previous ULF Parsers

Kim et al. (2021a) introduced an LSTM-based transition parser trained on a small, hand-annotated "gold" corpus of English-ULF pairs, achieving accuracy on par with early AMR parsers trained on much larger datasets. Gibson and Lawley (2022) later used a fine-tuned autoregressive language model on the same corpus and reported similar performance, showing that such models can perform well even with limited training data. Their model used the idea from (Mager et al., 2020) and (Bevilacqua et al., 2021) that the parsing task could be performed by seq2seq models similar to previous AMR-to-text models. Building on these, Juvekar et al. (2023) generated a much larger synthetic dataset using the gold data as seed sentences. Their method, grounded in ULF type constraints and linguistic patterns, created up to 116,112 English-ULF pairs, slightly improving upon (Kim et al., 2021a) (see Section 5).

Here, we present a new parser based on a large language model (LLM) trained on ULFs automatically derived from the Brown Treebank, containing about 50,000 sentences (20 words long on average) from many genres. Unlike the original gold corpus, which lacked longer and structurally complex sentences due to annotation costs, the Brown corpus provides broader structural and topical diversity. Whereas Gibson and Lawley used GPT-Too (Mager et al., 2020)[3], we apply Parameter Efficient Fine Tuning (PEFT) to a larger base model for improved performance with minimal training overhead.

3 Expanding the ULF Training Data Using the Penn Treebank Corpus

We now describe how we obtained ULF formulas from Brown corpus Penn Treebank (Marcus et al., 1993) syntax trees, for use in fine-tuning the Gemma-2B model (and also GTP-Too, for comparison). The idea behind use of the Brown corpus was that syntactic constituency trees roughly indicate the compositional semantic structure of sentences, and this should facilitate transduction into ULF. For example, a syntactic VP structure of form

(VP (VBD saw) (NP (DT the) (JJ white) (NN swan)))

(in the Penn Treebank format) can be regarded as indicating that the meaning of the verb phrase is obtained by applying the meaning of the past-tense verb "saw" to the meaning of the object noun phrase (NP). The result is a monadic predicate that can be applied to the meaning of an NP subject such as (NNP Bob) to obtain a sentence meaning. Similarly, the structure of the object NP suggests functional application of the determiner (DT) meaning and the adjective (JJ) meaning to the meaning of the nominal predicate, (NN swan).

3.1 Rule-based adjustments to the Treebank trees

However, there are some immediate adjustments that are needed to obtain a type-coherent structure. First, the past-tense component of (VBD saw) actually has sentence-level significance, placing the seeing-event (with the white swan as its object) in the past relative to the time of assertion. In ULF, (VBD saw) is split into a pair of semantic constituents, (PAST see.v), where "see.v" is an object-taking and subject-taking predicate, and PAST is an unscoped tense operator. Second, the structure of the object NP is insufficient to determine that the adjective should first be applied to the nominal predicate, forming the meaning of "white swan"; this modified nominal predicate is then operated upon by the determiner. In ULF, such determiner phrases are again unscoped semantic constituents. The resulting ULF phrase is thus

[3] "GPT-Too" appears in the title of this paper, referring to small, medium, and large versions of GPT-2 used by the authors for English generation from AMR.

((PAST see.v) (the.d ((MOD-N white.a) swan.n)));

this incorporates a third adjustment, namely conversion of the predicate "white.a" to a nominal-modifier via type-shifting operator MOD-N. This is needed if we take the (natural) view that "white" is lexicalized as a simple predicate (consider "Snow is white"), rather than as a predicate modifier like "fake".[4]

Thus, while syntactic constituency provides a rough indication of semantic structure, a variety of adjustment rules are needed to map Treebank trees to ULF. We use nearly 400 such rules, dealing with issues such as different uses of quotes, punctuation and brackets, inserting silent complementizers, regularizing complex quantifiers (such as "almost all" or "one out of six"), interpreting auxiliaries, distinguishing prepositional phrases used as predicates, predicate modifiers, or argument-suppliers, distinguishing the different semantic functions of participial VPs and subordinate clauses, expanding quantifying pronouns into quantifier-noun combinations (e.g., "nothing," "everybody"), dealing with displaced constituents, interpreting several types of comparatives, and many more.

The writing of these rules was made relatively straightforward by use of our tree transduction language TT, a simpler, more easily used variant of TTT (Purtee and Schubert, 2012). TT match patterns closely mirror the input tree structure, i.e., every sublist in a pattern must correspond to a sublist in the target list structure. The simplest pattern elements can be integers $i = 1, 2, ...$, which will match up to i successive atoms or lists. More often, we make use of TT's regex-like constructs, based on match predicates starting with characters '!', '?', '*', '+' to signal matchability to 1 item, 0 or 1 item, 0 or more items, and 1 or more items respectively; there are over 100 such predicates (separately defined). Some cover extensive data, for example, `!event-noun` covers about 220 event nouns, and a predicate checking for purely intransitive verbs covers over 5,300 verbs. A second class of match predicates, starting with a dot and applicable to atoms only, are interpreted via ISA-hierarchies. For example, `.TIME-PERIOD` checks whether the atom being matched "is a" word like *second, day, summer, pause, ...*, by checking for an ISA-chain of 0 or more links from the word to `.TIME-PERIOD`. (Lexical category can be checked by another ISA-predicate such as `.NN/NNP`, defined to match either `NN` or `NNP`.) Since TT allows for arbitrary nesting of expressions, the match predicates can be used at any structural level. Here is an example of the use of this language to expand a temporal NP such as "last summer," as represented in a constituent tree, into a temporal adverbial "during last summer":

```
(defrule *add-prep-for-definite-embedded-time-np*
 ; E.g., "I know what you did {last summer}"
 ;     parse fragment: (VP (AUX DID)
 ;                         (NP (JJ LAST) (NN SUMMER)))
 '((!atom *expr (!not-prep-or-symb +expr)
       (NP +expr (.NN/NNP .TIME-PERIOD)) *expr)
   (1 2 3 (ADVP (-SYMB- adv-e)
       (PP (-SYMB- {during}.p) 4)) 5)))
```

Every rule consists of a match pattern and an output pattern. Here the match pattern (!atom *expr (!not-prep-or-symb ...) (NP ...) *expr) matches any phrase in parentheses starting with exactly one atomic expression, followed by zero or more arbitrary expressions, followed by two subexpressions of specified forms (the second one being the temporal NP), and possibly additional ones.

When a match succeeds, the matched constituents can be referenced in the output pattern by their position. In the example, position indices 1–5 correspond to the five top-level matched expressions. Non-numeric elements are copied into the output directly, though TT also allows for output elements that are functions of matched input elements. Note the PP adverbial containing during.p (with the time-NP as its complement) in the output. To refer numerically to matched constituents lying within subexpressions of the match pattern, TT uses integers joined by dots. For example, 4.3.2 would refer to whatever piece of the input expression matched `.TIME-PERIOD`.

3.2 From adjusted trees to ULFs

Once transformed, trees are semantically interpreted via a compositional process driven by syntactic types and morphological cues. Lexemes receive type tags via about 50 rules based on word POS – which in many cases has been made semantically more revealing through preprocessing rules, e.g., WDT-REL instead of WDT for *which* or *that* used as a relativizer. Type-shifting operators introduced during preprocessing likewise facilitate function-argument application throughout. The compositional mapping from preprocessed phrases to ULFs is then quite simple, involving a little over

[4] Modified nominals cannot in general be viewed as a conjunction of two predicates, as in "is white and is a swan"; for instance this fails for "white wine," "plastic swan," or "utmost danger".

a page of code.

ULFs derived this way proved effective: in a small evaluation (11 sentences), the raw Brown-derived ULFs scored 0.81 F1 on EL-SMATCH and 0.82 on SEMBLEU, with 952 triples. Our final dataset includes 51,649 English-ULF pairs—substantially larger and more varied than the original gold corpus.

4 Models and Metrics

4.1 Language base models

Our model for deriving ULF from English builds on the training architecture developed by Gibson and Lawley (2022), which in turn built on GPT-Too, an AMR-to-English system (Mager et al., 2020). When run in reverse, Gibson and Lawley's model was shown to also be state-of-the art for the English to ULF parsing task. We apply Gibson and Lawley's architecture, fine-tuning on English-ULF sentence pairs to maximize the joint probabilities of English and ULF tokens. We also use their training process, but instead fine-tune Quantized Low Rank Adapters (QLoRA) (Dettmers et al., 2023) of the pretrained model to perform parameter-efficient fine-tuning (PEFT) to leverage a large base model. The previous LLM model used the 774M parameter version of GPT-Too (i.e., GPT-2L), while we use the 2.5B parameter Google Gemma-2B which would previously have been infeasible to train without parameter-efficient fine-tuning.

4.2 Metrics

We evaluated the model on both a test subset of the previous hand-annotated (gold) dataset ($n = 174$) and a test set of Brown corpus derived ULFs ($n = 174$) using the metrics EL-SMATCH and SEMBLEU. These metrics are borrowed from standard AMR evaluations, but the type-shifting operators of ULF and other differences from AMR require introduction of additional nodes and links to obtain Penman format, after which SMATCH and SEMBLEU can be applied. The SMATCH (Cai and Knight, 2013) score is calculated by (1) extracting all the triples from a hypothesis and reference AMR (e.g., see Figure 1), (2) performing a greedy search to unify variable names between the hypothesis and reference, and finally (3) calculating F1, precision, and recall scores from the matching triples. As noted by Groschwitz et al. (2023), current AMR parsers achieve high SMATCH scores but can still make frequent errors. This is partially because the SMATCH score suffers from two immediate problems: Only taking into account triples (two variables/concepts and a relations) means that larger semantic structure is not captured in the evaluation; and unifying the variables leads to over-counting matching triples where the relation matches but the variables do not map to the same concepts.

```
instance(z0, assert-02)     ARG0(z0, z1)
instance(z1, report-01)     ARG1(z0, z3)
instance(z2, news)          ARG1(z1, z2)
instance(z3, dump-01)       ARG1(z3, z4)
instance(z4, chemical)      destination(z3, z5)
instance(z5, river)
```

Figure 1: Extracted triples for the AMR corresponding to the sentence, "The news report asserted that chemicals were dumped into the river." `z0` through `z5` are variable names, the predicates `instance`, `ARG0`, `ARG1`, and `destination` are the edges of the AMR graph which capture semantic relations between variables. The `instance` predicate maps variables to concepts.

SEMBLEU scores are instead calculated by (1) extracting all n-grams from the hypothesis and reference AMR, where an n-gram includes n concepts connected by $n - 1$ relations (e.g., `assert-01 :ARG1 dump-01 :ARG1 chemical` is a 3-gram roughly corresponding to the meaning "chemicals being dumped is asserted"), (2) calculating an adjusted accuracy of matching n-grams between the hypothesis and reference, (3) multiplying by a brevity penalty. By including longer chains, SEMBLEU captures more complex semantic structures, and not using variables solves the over-counting problem of the SMATCH unification strategy. Because of this and in accordance with previous ULF parsing work, we use SEMBLEU (Song and Gildea, 2019) as a primary evaluation metric and EL-SMATCH for a more detailed breakdown of F1, precision and recall. EL-SMATCH is fully described by Kim and Schubert (2016), but is essentially an adaptation of SMATCH to evaluate ULFs as sets of triples in the same way as AMR.

5 Results

5.1 Results on the gold data in comparison with earlier ULF parsers

Using the 51,649 English-ULF dataset we obtained from the Brown corpus, and employing PEFT, we achieved major gains in all metrics as compared to previous ULF parsers – see Table 1. The results indicate that stronger base models improve evaluation metrics across the board, but have a less substantial effect than the new Brown-based dataset.

Base Model	SEMBLEU	EL-SMATCH		
		F1	Precision	Recall
(Kim et al., 2021a): Transition model	0.47	0.59		
(Gibson and Lawley, 2022): GPT-Too	0.43	0.63		
Trained on Gold + Generated Set				
(Juvekar et al., 2023): Transition model	0.49	0.60		
Trained on Gold + Brown Set (our results)				
GPT-2 124M	0.55	0.60	0.60	0.61
GPT-2 355M	0.66	0.69	0.70	0.68
Google Gemma 2B (PEFT)	**0.68**	**0.72**	**0.73**	**0.71**

Table 1: Results for models tuned on gold training set vs combined gold and Brown-derived training set.

The small gold dataset sufficed to train both Kim et al.'s transition-based and Gibson and Lawley's LLM-based ULF parser to a level of performance comparable with that of early AMR parsers trained on much larger datasets. As noted in Section 2, Juvekar et al. (2023) obtained small improvements over the original transition-based model using up to 116,112 artificially generated, type-consistent English-ULF pairs. The 51,649 English-ULF dataset we obtained from the Brown corpus is not as large as theirs, but we see substantial parsing performance increases over their parser. We suspect that this can be largely attributed to the fact that Brown Treebank sentences are a diverse, naturally occurring set, and that the carefully tuned, rule-based tree-to-ULF parser is almost as accurate as hand annotation of English sentences with ULFs. The substantial gains in SEMBLEU scores show that the model retrieves more individual constituents, and that the overall coherence of the fragments is higher.

5.2 Results on Brown-Derived ULFs

Our model's performance is best described by the results on the hand-annotated gold data. However, since our parser was fine-tuned on a combination of a (small) gold training set and a large set derived from the Brown corpus, it is of interest to look at its performance on Brown data in comparison with its performance on the gold data. Differences are to be expected, in part because the Brown data, though less accurate, clearly impacted performance very significantly, but also because some streamlining of certain syntactic conventions (e.g., the handling of auxiliary verbs and tense/aspect operators) was incorporated into the Brown data which are still in their old form in the gold data. The comparison is provided in Table 2.

As expected, the scores on the Brown-derived test set show substantially better SEMBLEU scores, although surprisingly, the EL-SMATCH scores are scarcely different. In other words, the parser generally matches the overall structure of Brown-derived data better than for gold data, perhaps because of the change in some ULF conventions, but the triple-by-triple match structure is not greatly affected. If we were to create a new gold set abiding by the revised conventions, our parser's performance likely would fall somewhere between the results on the gold and Brown-derived ULFs (i.e., between 0.68 and 0.76 on SEMBLEU). These results are also surprising because the sentence complexity and lengths in the Brown corpus are larger than those in the gold ULF set.

5.3 Comparison to AMR parsers

To relate our work to AMR parsing, we compare our ULF parsing results with results from two AMR parsers in Table 3. Other AMR parsers achieve similar SMATCH scores to (Drozdov et al., 2022) on the AMR 3.0 benchmark dataset. After the proof-of-concept GPT-Too parser (Mager et al., 2020), the first seq2seq parser with benchmark results (Bevilacqua et al., 2021), scored 83.0 on AMR 3.0. More recently Bai et al. (2022) and Vasylenko et al. (2023) build on (Bevilacqua et al., 2021) achieving significant improvements (scores of 84.2 and 84.6 respectively), using novel ideas such as incrementally finding spans to abstract, and inserting the corresponding concepts, treating the transduction between text and AMR as symmetric, and pretraining on AMR graph data rather than (just) text. For parsers of other knowledge representations, the recent English Resource Grammar parser by Lin et al. (2023) (based on Minimal Recursion Semantics) improves performance with a

Model	SEMBLEU	EL-SMATCH		
		F1	Precision	Recall
Gold ULF Test Set	0.68	0.72	**0.73**	0.71
Brown-Derived ULF Test Set	**0.76**	**0.72**	0.72	**0.72**

Table 2: Parser performance on hand-annotated (gold) test set versus performance on a test set of Brown-derived English-ULF pairs.

Parser Model	SEMBLEU	SMATCH/EL-SMATCH
AMR3-structbart-L (Drozdov et al., 2022)	0.56	0.83
AMR2-joint-ontowiki-seed42 (Lee et al., 2022)	0.60	**0.86**
Our Model	**0.68**	0.72

Table 3: Hand annotated test set comparison to AMR parser performance.

neural-symbolic approach, where prior knowledge from the symbolic parser alleviates inaccuracies of the neural model on out-of-distribution evaluation. A recent DRT parser from Yang et al. (2024) similarly proposes a neural-symbolic parser that predicts the scope structure with a rule or dependency based resolver.

As was seen in the discussion of sentences (1) and (2), the greater expressivity of ULF, and its fidelity to the full contents of sentences, results in more variety and complexity in ULF constructions relative to AMR. To re-emphasize this point, sentences such as "Dogs are barking" (thus, presently), "Dogs bark" (thus, generically), and "A dog barked" (thus, in the past) map to distinct ULF representations, while they are assigned the same AMR. This results in higher SMATCH scores for AMR parsers. Other knowledge representations also tend to blur semantic distinctions, or degrade for complex sentences (though apparently not for MRS). For example, DRT parsers score lower on datasets with long and complex sentences (SMATCH score of 87.1 on short example sentences versus 48.7 on longer sentences) (Yang et al., 2024).

Unlike the impressive SMATCH scores of AMR parsers, their SEMBLEU scores are weaker, suggesting that while they are able to adequately generate correct constituents, the arrangement of those constituents is less predictable than for ULF. While the greater expressivity and semantic fidelity of ULF may make it more difficult to generate individually correct constituents, the type coherence of ULF may also help improve the overall structure of the parses. When introducing the SEMBLEU evaluation metric, Song and Gildea (2019) show that SMATCH marks edges as identical regardless of the nodes they attach, leading to inflated scores for parsers that don't accurately capture sentence structure. From our increased SEMBLEU score, we tentatively infer that the ULF type structure is less susceptible to mistakes of this sort.

5.4 Error Analysis

The most common errors we observed in the results for testing on the gold test set were missing implicit references, not generating multi-sentence constructions, and incorrectly identifying proper nouns and quotations. Implicit references (semantic constituents not appearing in the surface text) should show up in ULFs as pronouns or other elements in curly brackets. Errors are possibly due to the Brown-derived ULFs having different proportions of the most common implicit references. The most common form in the gold ULFs is {YOU}.PRO (typically implicit in English imperatives), accounting for over half the implicit references in the gold test set but only 15% of the Brown-derived set. The latter contains more instances of {REF}.N and {FOR}.P (as in "This _ will serve _ to appease him," where the missing items are a nominal and a purposive "for" applied to the action type "to appease him"). Additionally, errors in multi-sentence constructions were expected because the Brown-derived ULFs only contain single sentence examples while the gold set contains examples with multiple punctuation-separated sentences.

The less frequent remaining errors include over-generating special operators and macros, and incorrect bracketing. Specifically, the parser over-generates the N+PREDS macro (typically used for combining a noun with its postmodifiers) which is again over-represented in the Brown-derived ULFs as compared to gold. Also the order in which pre- and post-modifiers are applied to a noun may

be different in gold sentence ULFs and in parser-generated ULFs, though it's sometimes unclear which order is correct. For example, the sentence "Name the disposable razor that 'costs about 19 cents.' " was hand annotated with

```
({you}.pro (name.v (the.d (n+preds
 ((mod-n disposable.a) razor.n)
  (that.rel ((PRES cost.v) (about.adv-s
   (ds currency ``19 cents''))))))))
```

but our model parses it to

```
({you}.pro (name.v (the.d
 ((mod-n disposable.a) (n+preds razor.n
  (that.rel ((PRES cost.v) ((about.mod-a | 19.a|)
   (plur cent.n)))))))))
```

These variant modifier structures have slightly different semantics, but neither is outright mistaken. The other difference between the hand annotation and the parse is the use of the domain-specific representation of currency in the gold ULF, (ds currency "19 cents") and the adv-s vs. mod-a difference. The Brown-derived ULFs do not include domain-specific annotations, so, naturally, the parser handles "19 cents" differently. Now, "19" to be suffixed with .a (the adjectival version of the numeral) and "about" is suffixed with .mod-a, so that it functions as an adjective modifier. In the hand-annotated sentence, the full "19 cents" is annotated in the domain-specific currency context, so there is no adjective 19.a for "about" to modify, and it is instead annotated with suffix .adv-s. Our model parses sentences like this well, but because of similar discrepancies that lead to larger differences from the hand-annotated ULF, their correctness is not reflected in our evaluation metrics.

6 Conclusion

We presented an LLM-based parser that demonstrates significant gains in parsing English to ULF, driven by a new dataset of English-ULF pairs automatically generated from Brown corpus Penn Treebank trees. These gains are evident across all metrics, especially SEMBLEU, which reflect the parser's ability to capture semantic relations and maintain coherence. Our approach outperforms previous ULF parsers and some modern AMR parsers, showing ULF's potential to represent nuanced semantics and complex sentence structures. While evaluation scores on gold test data are lower than on Brown-derived test data, this likely results from updates to ULF annotation principles since the gold data was created, so revising the gold data to align with current standards would be valuable.

With the new Brown ULF dataset, data scarcity is no longer the main challenge in ULF parsing. Future research can instead focus on incorporating learning techniques from AMR parsing, extending the augmentation strategy of Juvekar et al. (2023), or using ULF's type system to constrain generation.

The increased reliability of ULF parsing will make inference and reasoning in AI systems more broadly applicable. An example of a system that relied on rule-based semantic parsing into ULF was the DAVID virtual human (Kane et al., 2020) designed to answer questions in a physical "blocks world". DAVID was answered user questions like *"How many red blocks were to the left of a blue block, before I moved the Nvidia block?"*, based on observing and modeling blocks' spatial relations via cameras, and mapping questions to ULF for spatial model queries. Similarly, the SOPHIE system (Kane et al., 2023), a virtual cancer patient used to help train physicians, makes use of ULF inference in generating dialogue responses. The authors describe a future improvement to their system using a learned ULF parser, to support more logically coherent inferences within the global context.

An intriguing future research direction compatible with our approach to logical form would be to use the type structure of ULF for unsupervised language learning. It appears that the types of ULF and Episodic Logic—names, generalized quantifiers, predicates, predicate and sentence reifying operators, predicate and sentence modifying operators, and a handful more—suffice for human languages in general. We could treat these types as semantically "innate," and take language learning to be learning a mapping from word sequences to structures instantiating these types. The variability of languages, besides different vocabularies, would correspond to different strategies for linearizing and abbreviating internal graph-like structures to facilitate interpretation. Additional learning support besides textual corpora would be needed, such as visual grounding; but it seems that ULF/EL-like presupposed type structure should greatly reduce the demand for data in the learning process.

Acknowledgments

This research was sponsored in part by the University of Rochester's Schwartz Discover Grant. The authors are grateful for the guidance provided by Gene Kim and Lane Lawley. The referees' insights also enabled improvements to the paper.

References

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Jayeol Chun and Nianwen Xue. 20240815–20240815. Uniform meaning representation parsing as a pipelined approach. In *TextGraphs-17*, page 40–52.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv Preprint arXiv:2305.14314 [cs.LG]*.

Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramon Fernandez Astudillo. 2022. Inducing and using alignments for transition-based AMR parsing. *arXiv Preprint arXiv:2205.01464 [cs.CL]*.

Dan Flickinger, Yi Zhang, and Valia Kordoni. 2012. DeepBank: a dynamically annotated treebank of the Wall Street Journal. In *Proc. of the 11th Int. Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 85–96, Lisbon.

Erin Gibson and Lane Lawley. 2022. Language-model-based parsing and english generation for unscoped episodic logical forms. *The International FLAIRS Conference Proceedings*, 35.

Jonas Groschwitz, Shay Cohen, Lucia Donatelli, and Meaghan Fowlie. 2023. AMR parsing is far from solved: GrAPES, the granular AMR parsing evaluation suite. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10728–10752, Singapore. Association for Computational Linguistics.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, UMass Amherst.

Chung Hee Hwang and Lenhart K. Schubert. 1994. Meeting the interlocking needs of LF-computation, deindexing, and inference: An organic approach to general NLU. In *Proc. of the AAAI Fall Symposium, TR FS-94-04*, pages 1297–1302, New Orleans, LA.

Mandar Juvekar, Gene Kim, and Lenhart Schubert. 2023. Semantically informed data augmentation for unscoped episodic logical forms. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 116–133, Nancy, France. Association for Computational Linguistics.

Hans Kamp. 1981. A theory of truth and semantic representation. In P. Portner and B. H. Partee, editors, *Formal Semantics - the Essential Readings*, pages 189–222. Blackwell.

Benjamin Kane, Catherine Giugno, Lenhart Schubert, Kurtis Haut, Caleb Wohn, and Ehsan Hoque. 2023. Managing emotional dialogue for a virtual cancer patient: A schema-guided approach. *IEEE Transactions on Affective Computing, PrePrints*, pages 1–12.

Benjamin Kane, Georgiy Platonov, and Lenhart K. Schubert. 2020. Registering historical context in a spoken dialogue system for spatial question answering in a physical blocks world. In *Proc. of the 23rd Int. Conf. on Text, Speech and Dialogue (TSD 2020)*, pages 487–494, Brno, Czech Republic.

Gene Kim, Viet Duong, Xin Lu, and Lenhart Schubert. 2021a. A transition-based parser for unscoped episodic logical forms. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 184–201, Groningen, The Netherlands (online). Association for Computational Linguistics.

Gene Kim, Mandar Juvekar, Junis Ekmekciu, Viet Duong, and Lenhart Schubert. 2021b. A (mostly) symbolic system for monotonic inference with unscoped episodic logical forms. In *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, pages 71–80, Groningen, the Netherlands (online). Association for Computational Linguistics.

Gene Kim, Mandar Juvekar, and Lenhart Schubert. 2021c. Monotonic inference for underspecified episodic logic. In *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)*, pages 26–40, Groningen, the Netherlands (online). Association for Computational Linguistics.

Gene Kim, Benjamin Kane, Viet Duong, Muskaan Mendiratta, Graeme McGuire, Sophie Sackstein, Georgiy Platonov, and Lenhart Schubert. 2019. Generating discourse inferences from unscoped episodic logical formulas. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 56–65, Florence, Italy. Association for Computational Linguistics.

Gene Kim and Lenhart Schubert. 2016. High-fidelity lexical axiom construction from verb glosses. In

Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, pages 34–44, Berlin, Germany. Association for Computational Linguistics.

Gene Louis Kim and Lenhart Schubert. 2019. A type-coherent, expressive representation as an initial step to language understanding. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 13–30, Gothenburg, Sweden. Association for Computational Linguistics.

Lane Lawley, Gene Louis Kim, and Lenhart Schubert. 2019. Towards natural language story understanding with rich logical schemas. In *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, pages 11–22, Gothenburg, Sweden. Association for Computational Linguistics.

Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. Maximum Bayes Smatch ensemble distillation for AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.

Zi Lin, Jeremiah Liu, and Jingbo Shang. 2023. Neural-symbolic inference for robust autoregressive graph parsing via compositional uncertainty quantification. *Preprint*, arXiv:2301.11459.

Manuel Mager, Ramon Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. *arXiv Preprint arXiv:2005.09123 [cs.CL]*.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the Penn Treebank. *Comput. Linguist.*, 19(2):313–330.

A. Purtee and L.K. Schubert. 2012. TTT: A tree transduction language for syntactic and semantic processing. In *EACL Workshop on Applications of Tree Automata Techniques in Natural Language Processing (ATANLP 2012)*, Avignon, France.

Lenhart Schubert. 2014. NLog-like inference and commonsense reasoning. *Linguistic Issues in Language Technology*, 9.

Lenhart Schubert. 2015. Semantic representation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 4132–4138. AAAI Press.

Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic Logic meets Little Red Riding Hood: A comprehensive natural representation for language understanding. In Lucja M. Iwańska and Stuart C. Shapiro, editors, *Natural Language Processing and Knowledge Representation*, pages 111–174. MIT Press, Cambridge, MA, USA.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4547–4552, Florence, Italy. Association for Computational Linguistics.

Jens E. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35:343–660.

Pavlo Vasylenko, Pere-Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based amr parsing. *Preprint*, arXiv:2306.13467.

Xiulin Yang, Jonas Groschwitz, Alexander Koller, and Johan Bos. 2024. Scope-enhanced compositional semantic parsing for drt. *Preprint*, arXiv:2407.01899.