Eleventh Annual Conference on Advances in Cognitive Systems Poster Collection (2024) 1-15

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

Miles Frank MFRANK14@U.ROCHESTER.EDU
Lenhart Schubert 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 the reasoning to be made explicit and auditable, the knowledge and rules employed must themselves be made explicit and sufficiently unambiguous. You can't 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. 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 & Schubert, 1994; Schubert & 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.

ULFs, and their further resolution into Episodic Logic, have already proved to be a useful representation for inference within natural language understanding systems (Kane et al., 2020, 2023). Improving the scope and accuracy of ULF parsers will enable generalization of such systems. Because of their retention of all sentential information, and their coherent type structure, ULFs lend themselves to monotonic inference (Kim et al., 2021c,b), discourse inferences including for clause-taking verbs, counterfactuals, questions, requests, and generalizations (Kim et al., 2019), as well as schema-based story representation (Lawley et al., 2019). 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 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 convertible to the more precise Episodic Logic representation); 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 Treebank corpus 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 Comclusion (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), and Abstract Meaning Representation (AMR). 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 use of various syntactic and semantic maneuvers, FOL can also be adapted to sentences involving more subtle subject matter such as beliefs, plans, and imprecise knowledge. 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). Most of all, deriving such circumlocutions automatically is likely to be very hard.

To address some pronoun resolution issues in converting natural language to FOL, Kamp (1981) and Heim (1982) developed Discourse Representation 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.

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 some 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, such as Montague intensional logic, description logics, and conceptual representations).

In view of the disproportionate 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 says, roughly, 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, automatically obtained ULF interpretations of (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))))))))
\.)
```

(ULF formulas are case-insensitive except for names such as | New York|, but we have used upper case for tense/aspect operators for clarity.) 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.[1] 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 into the river, as arguments. (Predicate arguments cannot be quantified over, and the logic remains first-order.)

1. But acting on a kind entails acting on an instance of the kind – here, an instance of the kind, chemicals.

2.2 Previous ULF Parsers

Kim et al. (2021a) developed an LSTM-based transition parser trained on a limited hand-annotated "gold" corpus of English-ULF pairs, achieving accuracies comparable with those obtained by early AMR parsers trained on much larger datasets. Gibson and Lawley (2022) subsequently introduced another English-to-ULF parser, leveraging a large language model fine-tuned on the same gold corpus, and obtained very similar results. Their work demonstrated the suitability of pre-trained autoregressive language models for the English-to-ULF parsing task, even with a very limited training dataset. In slightly later work, Juvekar et al. (2023) used the gold data as a source of seed sentences to randomly generate a greatly expanded dataset of samples consistent with ULF type constraints, and favoring conformity with statistical patterns of language. The generated ULFs were paired with automatic translations into English, thus providing a training corpus of up to 116,112 English-ULF pairs. This method provided small improvements to the accuracy of Kim et al.'s transition parser (see Section 5).

In this paper, we present a new English-to-ULF parser based on a large language model (LLM) and trained on a dataset of ULFs automatically derived from Brown corpus Treebank parses. While the original hand-annotated ULF dataset contained sentences from diverse domains, the cost of hand annotation kept the proportion of longer, more complex sentences rather low. Using the tree-annotations of the Brown corpus, we were able to significantly increase not only the size of the training dataset, but also the structural and topical diversity and lengths of ULF-annotated sentences. Additionally, while Gibson and Lawley's (2022) LLM-based parser used GPT-Too (Mager et al., 2020)[2] as a base model, with the use of Parameter Efficient Fine Tuning (PEFT), we can employ a larger base model to boost performance while conserving low cost training.

3. Expanding the ULF Training Data Using the Brown Treebank Corpus

We now describe how we obtained ULF formulas from Brown corpus 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, given the compositional semantic types and surface-like form of ULF. For example, a syntactic VP structure of form

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

(which is in the standard 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).

2. "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. Gibson and Lawley used the large, 774M parameter version, also called GPT-2L.

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 and an agent such as Bob as its subject) 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 that would be used, in conversion to Episodic Logic, to explicitly relate the seeing event to a NOW point (assertion event), placing it in the relative past. Second, the structure of the object NP fails to indicate in what order the determiner and the adjective do their semantic work. 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, forming a determiner phrase. In ULF, such determiner phrases are again unscoped semantic constituents. In the conversion to Episodic Logic, they are shifted to sentence-initial position, with introduction of a variable bound by the determiner and inserted as argument of the nominal (viewed as a restrictor predicate) as well as of the VP predicate.[3] The resulting ULF phrase is thus

((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" or "erstwhile").[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, structuring unduly "flat" phrases to properly reflect their functional structure, 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, inserting missing trace constituents, 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. 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}/{this morning}"; sample
;        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)))
```

3. This approach keeps Episodic Logic first-order, unlike Richard Montague's approach of treating determiners as second-order predicates taking the restrictor predicate and VP predicate as arguments.
4. Contrary to a common, convenient assumption, 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", or "plastic swan".

Every rule consists of a match pattern and an output pattern. The TT syntax has regex-like constructs, but allows for arbitrary nesting of expressions and separately defined match predicates. 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. Boolean predicates prefixed with !, ?, *, or + match exactly one, zero or one, zero or more, or at least one matching expression respectively. (These are implemented as simple Lisp functions.) Predicates prefixed with a dot refer to an "ISA" hierarchy of features of atomic expressions. For example, NN/NNP is a feature of just the symbols NN and NNP, while TIME-PERIOD is stored as a feature of many lexical items such as *time, past, minute, week, May,* and many more. Features can form arbitrary acyclic directed graphs.

When a match succeeds, the matched constituents can be referred to in the output pattern by their location in the match. In the example, 1, 2, 3, 4, 5 refer to the expressions matched by the five top-level expressions of the match patterns. The non-numeric elements of the output pattern are generated as-is (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, since this is the 2nd element of the 3rd element of the 4th element of the match pattern. An important consideration in the design of TT was ease of specifying rules, and their legibility, which is ensured by the way match-pattern bracket structure directly echoes the input structure to be matched.

3.2 From adjusted trees to ULFs

When all applicable adjustments have been made to a Treebank tree, semantic interpretation of the resulting structure becomes a quite systematic compositional process, starting with interpretation of lexical items using their syntactic type and inflectional morphology. Type-shifting operators will already have been introduced in the adjustment process, so that composition is a matter of identifying function-argument roles for the immediate constituents of each subphrase.

The ULFs derived from the Brown tree structures in this way turned out to be accurate enough to have a very positive impact on parser training. A small random sample of 11 sentences from the Brown-derived ULFs was hand-corrected, and the F1-score on EL-SMATCH for the uncorrected ULFs was 0.81, and the SEMBLEU score was 0.82. The number of triples involved in the tests was 952. Our new Brown-derived ULF dataset contains 51,649 English-ULF sentence pairs.

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 (as noted earlier) 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. Low Rank Adaptation (LoRA) (Hu et al., 2021), and its derivative QLoRA, freezes most layers of the pretrained model and instead trains smaller rank decomposition matrices of each layer, greatly reducing the number of trainable parameters while preserving the accuracy gains from using a larger 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 PEFT. The best results for our model were achieved with the parameters, $rank = 8$, $alpha = 32$, and $dropout = 0.1$, which are standard for similar tasks.

4.2 Metrics

We evaluate the model on both a test subset of the previous hand-annotated (gold) dataset ($n = 174$) as well as a test set of Brown corpus derived ULFs ($n = 174$) using the EL-SMATCH and SEMBLEU metrics. 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 & 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. This 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 (e.g., any `ARG0(a, b)` could match `ARG0(z0, z1)` even if `a` and `z0` represent completely different 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 of concepts, SEMBLEU captures more complex semantic structures, and not using variables solves the over-counting problem engendered by the SMATCH unification strategy. Because of this, and in accordance with previous ULF parsing work, we use SEMBLEU (Song & Gildea, 2019) as a primary evaluation metric and EL-SMATCH for a more detailed F1, precision, and recall breakdown. 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

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

Base Model	SEMBLEU	EL-SMATCH		
		F1	Precision	Recall
Trained on Gold Set				
(Kim et al., 2021a): Transition model	0.47	0.59[5]		
(Gibson & 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**

Using the 51,649 English-ULF dataset we obtained from the Brown corpus, and employing PEFT, we obtained a substantial increase in all metrics as compared to previous ULF parsers, as shown in Table 1. This lists the performance metrics for all ULF parsers to date, starting with Kim et al.'s original LSTM-based transition parser, which provided the set of 1,378 gold data in all cases. 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.

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,

5. using PyTorch as in (Juvekar et al., 2023), for comparison purposes

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 not only that the model retrieves more individual constituents, but that the overall coherence of the parses 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.

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

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**

As one might expect, the scores on the Brown-derived test set show substantially better SEMBLEU scores, though 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

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

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

To relate our work to AMR parsing, we compare our ULF parsing results with results from two AMR parsers in Table 3. 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. As a further example, 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 the higher SMATCH scores for AMR parsers. On the other hand, AMR parsers perform more poorly on SEMBLEU, 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 significantly different proportions of the most common implicit references. The most common form in the gold ULFs is {YOU}.PRO (typically left 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 instead 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"). Similarly, 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 these domain specific annotations, so it is expected that the parser would handle "19 cents" differently, which in turn causes "19" to be suffixed with .a (the adjectival version of the numeral) and so "about" is suffixed with .mod-a, i.e., it functions as an adjective modifier. In the hand annotated sentence, because the full "19 cents" is annotated in the domain specific currency context, there is no adjective 19.a for "about" to modify, so it is instead annotated with suffix .adv-s. Our model actually parses certain 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 the trees of the Brown Treebank corpus. The improved performance is evident across all evaluation metrics, particularly in SEMBLEU scores, highlighting the parser's ability to correctly derive semantic relations between constituents and maintain overall coherence. The results also indicate that our approach offers better parsing scores compared to previous ULF parsers and even some contemporary AMR parsers, showcasing the potential of ULF in representing nuanced semantics and complex sentence structures. While evaluation scores for gold test data are lower than those for artificial Brown-derived test data, this discrepancy can be attributed to some changes to ULF annotation principles since the creation of the guidelines that the gold data adhered to. Thus a worthwhile future effort would be revision of the gold data to conform with the updated standards.

From the results we obtained by training on the Brown ULF dataset, lack of training data no longer seems to be a primary research concern for ULF parsing development. Instead, future ULF parsing work could involve implementing more sophisticated learning techniques used in state of the art AMR parsers, extending the data augmentation technique described by Juvekar et al., or otherwise leveraging the underlying ULF type structure 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 for answering questions in a physical "blocks world". DAVID was able to answer user questions such as *"How many red blocks were to the left of a blue block, before I moved the Nvidia block?"*, based on observing and modeling block movements and spatial relations via Kinect cameras, and mapping questions to ULF and hence into queries to the spatial and historical models. The limited domain enabled very accurate parsing into ULF, but generalizing to miscellaneous indoor and outdoor situations will require broader coverage. Similarly, the SOPHIE system (**?**), a virtual cancer patient used to help train doctors, makes limited use of ULF inference in generating appropriate responses in dialogues with users. The authors

describe a future improvement to their semantic understanding system using a learned ULF parser, to allow for inferences that are logically coherent within the global dialogue context. With an accurate logical form parser, other such systems that rely on natural language understanding will be able to maintain a high level of logical consistency, which is especially important in sensitive contexts such as medical domains.

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 the goal of language learning to be learning a mapping from word sequences to structures instantiating these (latent) types. The variability of languages, besides their different vocabularies, would be the result of different strategies for linearizing and abbreviating internal graph-like structures (whether more AMR-like or ULF/EL-like) in ways that facilitate interpretation, such as, by keeping lexical items that belong to the same phrase close together. Of course additional learning support besides textual corpora would be needed, such as visual grounding or predetermined allowable lexical types for a substantial sub-vocabulary of the language to be learned; but it seems that ULF/EL-like presupposed type structure should greatly reduce the demand for data in the learning process.

Acknowledgements

This research was sponsored in part by the University of Rochester's Schwartz Discover Grant Program for undergraduate researchers. The authors are grateful for the expert help and guidance provided by Gene Kim and Lane Lawley. The referees' comments also led to improvements in the paper.

References

Banarescu, L., et al. (2013). Abstract Meaning Representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 178–186). Sofia, Bulgaria: Association for Computational Linguistics.

Cai, S., & Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 748–752). Sofia, Bulgaria: Association for Computational Linguistics. From `https://aclanthology.org/P13-2131`.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *arXiv Preprint arXiv:2305.14314 [cs.LG]*.

Drozdov, A., Zhou, J., Florian, R., McCallum, A., Naseem, T., Kim, Y., & Astudillo, R. F. (2022). Inducing and using alignments for transition-based AMR parsing. *arXiv Preprint arXiv:2205.01464 [cs.CL]*.

Gibson, E., & Lawley, L. (2022). Language-model-based parsing and english generation for unscoped episodic logical forms. *The International FLAIRS Conference Proceedings*, *35*. From

https://journals.flvc.org/FLAIRS/article/view/130703.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. Doctoral dissertation, UMass Amherst.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv Preprint arXiv:2106.09685 [cs.CL]*.

Hwang, C. H., & Schubert, L. K. (1994). Meeting the interlocking needs of LF-computation, deindexing, and inference: An organic approach to general NLU. *Proc. of the AAAI Fall Symposium, TR FS-94-04* (pp. 1297–1302). New Orleans, LA.

Juvekar, M., Kim, G., & Schubert, L. (2023). Semantically informed data augmentation for unscoped episodic logical forms. *Proceedings of the 15th International Conference on Computational Semantics* (pp. 116–133). Nancy, France: Association for Computational Linguistics. From https://aclanthology.org/2023.iwcs-1.14.

Kamp, H. (1981). A theory of truth and semantic representation. In P. Portner & B. H. Partee (Eds.), *Formal semantics - the essential readings*, 189–222. Blackwell.

Kane, B., Giugno, C., Schubert, L., Haut, K., Wohn, C., & Hoque, E. (2023). Managing emotional dialogue for a virtual cancer patient: A schema-guided approach. *IEEE Transactions on Affective Computing, PrePrints*, (pp. 1–12).

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

Kim, G., Duong, V., Lu, X., & Schubert, L. (2021a). A transition-based parser for unscoped episodic logical forms. *Proceedings of the 14th International Conference on Computational Semantics (IWCS)* (pp. 184–201). Groningen, The Netherlands (online): Association for Computational Linguistics. From https://aclanthology.org/2021.iwcs-1.18.

Kim, G., Juvekar, M., Ekmekciu, J., Duong, V., & Schubert, L. (2021b). A (mostly) symbolic system for monotonic inference with unscoped episodic logical forms. *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)* (pp. 71–80). Groningen, the Netherlands (online): Association for Computational Linguistics. From https://aclanthology.org/2021.naloma-1.9.

Kim, G., Juvekar, M., & Schubert, L. (2021c). Monotonic inference for underspecified episodic logic. *Proceedings of the 1st and 2nd Workshops on Natural Logic Meets Machine Learning (NALOMA)* (pp. 26–40). Groningen, the Netherlands (online): Association for Computational Linguistics. From https://aclanthology.org/2021.naloma-1.5.

Kim, G., Kane, B., Duong, V., Mendiratta, M., McGuire, G., Sackstein, S., Platonov, G., & Schubert, L. (2019). Generating discourse inferences from unscoped episodic logical formulas. *Proceedings of the First International Workshop on Designing Meaning Representations* (pp. 56–65). Florence, Italy: Association for Computational Linguistics. From https://aclanthology.org/W19-3306.

Kim, G., & Schubert, L. (2016). High-fidelity lexical axiom construction from verb glosses. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 34–44). Berlin, Germany: Association for Computational Linguistics. From https://aclanthology.org/S16-2004.

Kim, G. L., & Schubert, L. (2019). A type-coherent, expressive representation as an initial step to language understanding. *Proceedings of the 13th International Conference on Computational Semantics - Long Papers* (pp. 13–30). Gothenburg, Sweden: Association for Computational Linguistics. From https://aclanthology.org/W19-0402.

Lawley, L., Kim, G. L., & Schubert, L. (2019). Towards natural language story understanding with rich logical schemas. *Proceedings of the Sixth Workshop on Natural Language and Computer Science* (pp. 11–22). Gothenburg, Sweden: Association for Computational Linguistics. From https://aclanthology.org/W19-1102.

Lee, Y.-S., Astudillo, R., Thanh Lam, H., Naseem, T., Florian, R., & Roukos, S. (2022). Maximum Bayes Smatch ensemble distillation for AMR parsing. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5379–5392). Seattle, United States: Association for Computational Linguistics. From https://aclanthology.org/2022.naacl-main.393.

Mager, M., Astudillo, R. F., Naseem, T., Sultan, M. A., Lee, Y.-S., Florian, R., & Roukos, S. (2020). GPT-too: A language-model-first approach for AMR-to-text generation. *arXiv Preprint arXiv:2005.09123 [cs.CL]*.

Schubert, L. (2014). NLog-like inference and commonsense reasoning. *Linguistic Issues in Language Technology*, *9*. From https://aclanthology.org/2014.lilt-9.9.

Schubert, L. (2015). Semantic representation. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (p. 4132–4138). AAAI Press.

Schubert, L. K., & Hwang, C. H. (2000). Episodic Logic meets Little Red Riding Hood: A comprehensive natural representation for language understanding. In L. M. Iwańska & S. C. Shapiro (Eds.), *Natural language processing and knowledge representation*, 111–174. Cambridge, MA, USA: MIT Press.

Song, L., & Gildea, D. (2019). SemBleu: A robust metric for AMR parsing evaluation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4547–4552). Florence, Italy: Association for Computational Linguistics. From https://aclanthology.org/P19-1446.