Abstract Syntax Networks for Code Generation and Semantic Parsing

Maxim Rabinovich* **Mitchell Stern*** **Dan Klein**
Computer Science Division
University of California, Berkeley
{rabinovich,mitchell,klein}@cs.berkeley.edu

Abstract

Tasks like code generation and semantic parsing require mapping unstructured (or partially structured) inputs to well-formed, executable outputs. We introduce abstract syntax networks, a modeling framework for these problems. The outputs are represented as abstract syntax trees (ASTs) and constructed by a decoder with a dynamically-determined modular structure paralleling the structure of the output tree. On the benchmark HEARTHSTONE dataset for code generation, our model obtains 79.2 BLEU and 22.7% exact match accuracy, compared to previous state-of-the-art values of 67.1 and 6.1%. Furthermore, we perform competitively on the ATIS, JOBS, and GEO semantic parsing datasets with no task-specific engineering.

1 Introduction

Tasks like semantic parsing and code generation are challenging in part because they are *structured* (the output must be well-formed) but not *synchronous* (the output structure diverges from the input structure).

Sequence-to-sequence models have proven effective for both tasks (Dong and Lapata, 2016; Ling et al., 2016), using encoder-decoder frameworks to exploit the sequential structure on both the input and output side. Yet these approaches do not account for much richer structural constraints on outputs—including well-formedness, well-typedness, and executability. The well-formedness case is of particular interest, since it can readily be enforced by representing outputs as abstract syntax trees (ASTs) (Aho et al., 2006), an approach that can be seen as a much lighter weight


```
name: [
  'D', 'i', 'r', 'e', ' ',
  'W', 'o', 'l', 'f', ' ',
  'A', 'l', 'p', 'h', 'a']
cost: ['2']
type: ['Minion']
rarity: ['Common']
race: ['Beast']
class: ['Neutral']
description: [
  'Adjacent', 'minions', 'have',
  '+', '1', 'Attack', '.']
health: ['2']
attack: ['2']
durability: ['-1']
```

```
class DireWolfAlpha(MinionCard):
  def __init__(self):
    super().__init__(
      "Dire Wolf Alpha", 2, CHARACTER_CLASS.ALL,
      CARD_RARITY.COMMON, minion_type=MINION_TYPE.BEAST)
  def create_minion(self, player):
    return Minion(2, 2, auras=[
      Aura(ChangeAttack(1), MinionSelector(Adjacent()))
    ])
```

Figure 1: Example code for the "Dire Wolf Alpha" Hearthstone card.

```
show me the fare from ci0 to ci1

lambda $0 e
  ( exists $1 ( and ( from $1 ci0 )
                    ( to $1 ci1 )
                    ( = ( fare $1 ) $0 ) ) )
```

Figure 2: Example of a query and its logical form from the ATIS dataset. The `ci0` and `ci1` tokens are entity abstractions introduced in preprocessing (Dong and Lapata, 2016).

version of CCG-based semantic parsing (Zettlemoyer and Collins, 2005).

In this work, we introduce *abstract syntax networks* (ASNs), an extension of the standard encoder-decoder framework utilizing a modular decoder whose submodels are composed to natively generate ASTs in a top-down manner. The decoding process for any given input follows a dy-

*Equal contribution.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1139–1149
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1105

namically chosen mutual recursion between the modules, where the structure of the tree being produced mirrors the call graph of the recursion. We implement this process using a decoder model built of many submodels, each associated with a specific construct in the AST grammar and invoked when that construct is needed in the output tree. As is common with neural approaches to structured prediction (Chen and Manning, 2014; Vinyals et al., 2015), our decoder proceeds greedily and accesses not only a fixed encoding but also an attention-based representation of the input (Bahdanau et al., 2014).

Our model significantly outperforms previous architectures for code generation and obtains competitive or state-of-the-art results on a suite of semantic parsing benchmarks. On the HEARTHSTONE dataset for code generation, we achieve a token BLEU score of 79.2 and an exact match accuracy of 22.7%, greatly improving over the previous best results of 67.1 BLEU and 6.1% exact match (Ling et al., 2016).

The flexibility of ASNs makes them readily applicable to other tasks with minimal adaptation. We illustrate this point with a suite of semantic parsing experiments. On the JOBS dataset, we improve on previous state-of-the-art, achieving 92.9% exact match accuracy as compared to the previous record of 90.7%. Likewise, we perform competitively on the ATIS and GEO datasets, matching or exceeding the exact match reported by Dong and Lapata (2016), though not quite reaching the records held by the best previous semantic parsing approaches (Wang et al., 2014).

1.1 Related work

Encoder-decoder architectures, with and without attention, have been applied successfully both to sequence prediction tasks like machine translation and to tree prediction tasks like constituency parsing (Cross and Huang, 2016; Dyer et al., 2016; Vinyals et al., 2015). In the latter case, work has focused on making the task look like sequence-to-sequence prediction, either by flattening the output tree (Vinyals et al., 2015) or by representing it as a sequence of construction decisions (Cross and Huang, 2016; Dyer et al., 2016). Our work differs from both in its use of a recursive top-down generation procedure.

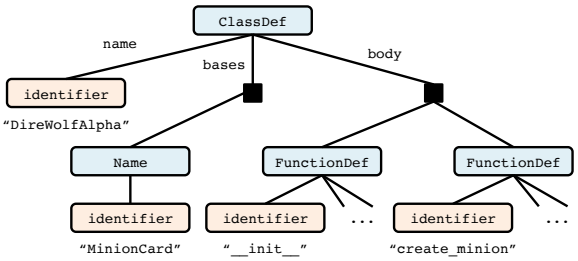
Dong and Lapata (2016) introduced a sequence-to-sequence approach to semantic parsing, including a limited form of top-down recursion, but without the modularity or tight coupling between output grammar and model characteristic of our approach.

Neural (and probabilistic) modeling of code, including for prediction problems, has a longer history. Allamanis et al. (2015) and Maddison and Tarlow (2014) proposed modeling code with a neural language model, generating concrete syntax trees in left-first depth-first order, focusing on metrics like perplexity and applications like code snippet retrieval. More recently, Shin et al. (2017) attacked the same problem using a grammar-based variational autoencoder with top-down generation similar to ours instead. Meanwhile, a separate line of work has focused on the problem of program induction from input-output pairs (Balog et al., 2016; Liang et al., 2010; Menon et al., 2013).

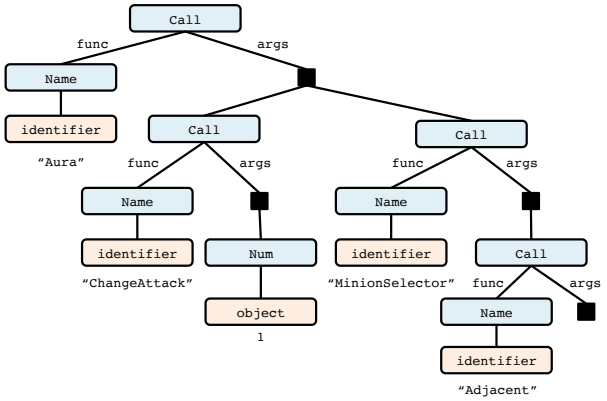
The prediction framework most similar in spirit to ours is the doubly-recurrent decoder network introduced by Alvarez-Melis and Jaakkola (2017), which propagates information down the tree using a vertical LSTM and between siblings using a horizontal LSTM. Our model differs from theirs in using a separate module for each grammar construct and learning separate vertical updates for siblings when the AST labels require all siblings to be jointly present; we do, however, use a horizontal LSTM for nodes with *variable* numbers of children. The differences between our models reflect not only design decisions, but also differences in data—since ASTs have labeled nodes and labeled edges, they come with additional structure that our model exploits.

Apart from ours, the best results on the code-generation task associated with the HEARTHSTONE dataset are based on a sequence-to-sequence approach to the problem (Ling et al., 2016). Abstract syntax networks greatly improve on those results.

Previously, Andreas et al. (2016) introduced neural module networks (NMNs) for visual question answering, with modules corresponding to linguistic substructures within the input query. The primary purpose of the modules in NMNs is to compute deep features of images in the style of convolutional neural networks (CNN). These features are then fed into a final decision layer. In contrast to the modules we describe here, NMN modules do not make decisions about what to generate or which modules to call next, nor do they



(a) The root portion of the AST.



(b) Excerpt from the same AST, corresponding to the code snippet `Aura(ChangeAttack(1),MinionSelector(Adjacent()))`.

Figure 3: Fragments from the abstract syntax tree corresponding to the example code in Figure 1. Blue boxes represent composite nodes, which expand via a constructor with a prescribed set of named children. Orange boxes represent primitive nodes, with their corresponding values written underneath. Solid black squares correspond to constructor fields with `sequential` cardinality, such as the body of a class definition (Figure 3a) or the arguments of a function call (Figure 3b).

maintain recurrent state.

2 Data Representation

2.1 Abstract Syntax Trees

Our model makes use of the Abstract Syntax Description Language (ASDL) framework (Wang et al., 1997), which represents code fragments as trees with typed nodes. *Primitive types* correspond to atomic values, like integers or identifiers. Accordingly, *primitive nodes* are annotated with a primitive type and a value of that type—for instance, in Figure 3a, the `identifier` node storing `"create_minion"` represents a function of the same name.

Composite types correspond to language constructs, like expressions or statements. Each type has a collection of *constructors*, each of which specifies the particular language construct a node of that type represents. Figure 4 shows constructors for the statement (`stmt`) and expression (`expr`) types. The associated language constructs include function and class definitions, return statements, binary operations, and function calls.

Composite types enter syntax trees via *composite nodes*, annotated with a composite type and a choice of constructor specifying how the node expands. The root node in Figure 3a, for example, is

```
primitive types: identifier, object, ...

stmt
 = FunctionDef(
     identifier name, arg* args, stmt* body)
 | ClassDef(
     identifier name, expr* bases, stmt* body)
 | Return(expr? value)
 | ...

expr
 = BinOp(expr left, operator op, expr right)
 | Call(expr func, expr* args)
 | Str(string s)
 | Name(identifier id, expr_context ctx)
 | ...

...
```

Figure 4: A simplified fragment of the Python ASDL grammar.[1]

a composite node of type `stmt` that represents a class definition and therefore uses the `ClassDef` constructor. In Figure 3b, on the other hand, the root uses the `Call` constructor because it represents a function call.

Children are specified by named and typed *fields* of the constructor, which have *cardinalities* of `singular`, `optional`, or `sequential`. By default, fields have `singular` cardinality, meaning they correspond to exactly one child. For instance, the `ClassDef` constructor has a `singular` name field of type `identifier`. Fields of `optional` cardinality are associ-

[1]The full grammar can be found online on the documentation page for the Python `ast` module: https://docs.python.org/3/library/ast.html#abstract-grammar

ated with zero or one children, while fields of `sequential` cardinality are associated with zero or more children—these are designated using `?` and $\star$ suffixes in the grammar, respectively. Fields of `sequential` cardinality are often used to represent statement blocks, as in the `body` field of the `ClassDef` and `FunctionDef` constructors.

The grammars needed for semantic parsing can easily be given ASDL specifications as well, using primitive types to represent variables, predicates, and atoms and composite types for standard logical building blocks like lambdas and counting (among others). Figure 2 shows what the resulting λ-calculus trees look like. The ASDL grammars for both λ-calculus and Prolog-style logical forms are quite compact, as Figures 9 and 10 in the appendix show.

2.2 Input Representation

We represent inputs as collections of named components, each of which consists of a sequence of tokens. In the case of semantic parsing, inputs have a single component containing the query sentence. In the case of HEARTHSTONE, the card's name and description are represented as sequences of characters and tokens, respectively, while categorical attributes are represented as single-token sequences. For HEARTHSTONE, we restrict our input and output vocabularies to values that occur more than once in the training set.

3 Model Architecture

Our model uses an encoder-decoder architecture with hierarchical attention. The key idea behind our approach is to structure the decoder as a collection of mutually recursive modules. The modules correspond to elements of the AST grammar and are composed together in a manner that mirrors the structure of the tree being generated. A vertical LSTM state is passed from module to module to propagate information during the decoding process.

The encoder uses bidirectional LSTMs to embed each component and a feedforward network to combine them. Component- and token-level attention is applied over the input at each step of the decoding process.

We train our model using negative log likelihood as the loss function. The likelihood encompasses terms for all generation decisions made by the decoder.

3.1 Encoder

Each component c of the input is encoded using a component-specific bidirectional LSTM. This results in forward and backward token encodings $(\overrightarrow{\mathbf{h}^c}, \overleftarrow{\mathbf{h}^c})$ that are later used by the attention mechanism. To obtain an encoding of the input as a whole for decoder initialization, we concatenate the final forward and backward encodings of each component into a single vector and apply a linear projection.

3.2 Decoder Modules

The decoder decomposes into several classes of modules, one per construct in the grammar, which we discuss in turn. Throughout, we let $\mathbf{v}$ denote the current vertical LSTM state, and use f to represent a generic feedforward neural network. LSTM updates with hidden state $\mathbf{h}$ and input $\mathbf{x}$ are notated as $\mathrm{LSTM}(\mathbf{h}, \mathbf{x})$.

Composite type modules Each composite type `T` has a corresponding module whose role is to select among the constructors `C` for that type. As Figure 5a exhibits, a composite type module receives a vertical LSTM state $\mathbf{v}$ as input and applies a feedforward network $f_{\mathtt{T}}$ and a softmax output layer to choose a constructor:

$$p\left(\mathtt{C} \mid \mathtt{T}, \mathbf{v}\right) = \left[\mathrm{softmax}\left(f_{\mathtt{T}}\left(\mathbf{v}\right)\right)\right]_{\mathtt{C}}.$$

Control is then passed to the module associated with constructor `C`.

Constructor modules Each constructor `C` has a corresponding module whose role is to compute an intermediate vertical LSTM state $\mathbf{v}_{u,\mathtt{F}}$ for each of its fields `F` whenever `C` is chosen at a composite node u.

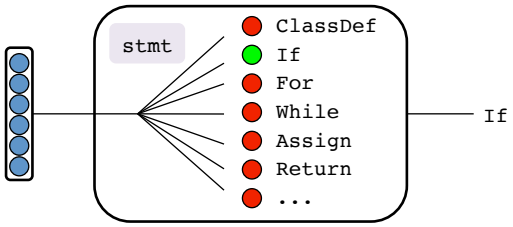
For each field `F` of the constructor, an embedding $\mathbf{e}_{\mathtt{F}}$ is concatenated with an attention-based context vector $\mathbf{c}$ and fed through a feedforward neural network $f_{\mathtt{C}}$ to obtain a context-dependent field embedding:

$$\tilde{\mathbf{e}}_{\mathtt{F}} = f_{\mathtt{C}}\left(\mathbf{e}_{\mathtt{F}},\ \mathbf{c}\right).$$

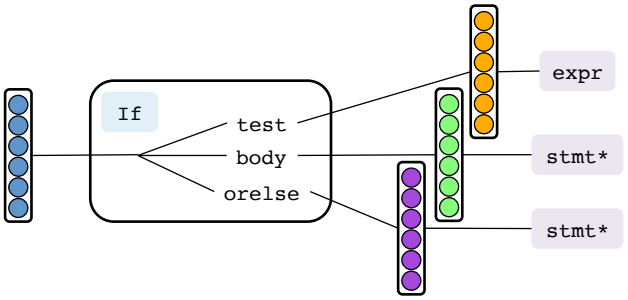
An intermediate vertical state for the field `F` at composite node u is then computed as

$$\mathbf{v}_{u,\mathtt{F}} = \mathrm{LSTM}^{v}\left(\mathbf{v}_{u},\ \tilde{\mathbf{e}}_{\mathtt{F}}\right).$$

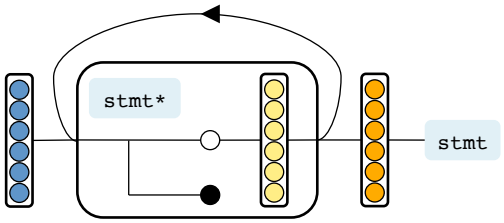
Figure 5b illustrates the process, starting with a single vertical LSTM state and ending with one updated state per field.



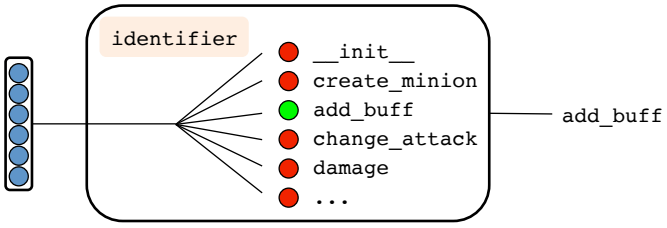
(a) A composite type module choosing a constructor for the corresponding type.



(b) A constructor module computing updated vertical LSTM states.



(c) A constructor field module (`sequential` cardinality) generating children to populate the field. At each step, the module decides whether to generate a child and continue (white circle) or stop (black circle).



(d) A primitive type module choosing a value from a closed list.

Figure 5: The module classes constituting our decoder. For brevity, we omit the cardinality modules for `singular` and `optional` cardinalities.

Constructor field modules Each field `F` of a constructor has a corresponding module whose role is to determine the number of children associated with that field and to propagate an updated vertical LSTM state to them. In the case of fields with `singular` cardinality, the decision and update are both vacuous, as exactly one child is always generated. Hence these modules forward the field vertical LSTM state $\mathbf{v}_{u,\mathtt{F}}$ unchanged to the child w corresponding to `F`:

$$\mathbf{v}_w = \mathbf{v}_{u,\mathtt{F}}. \tag{1}$$

Fields with `optional` cardinality can have either zero or one children; this choice is made using a feedforward network applied to the vertical LSTM state:

$$p(z_\mathtt{F} = 1 \mid \mathbf{v}_{u,\mathtt{F}}) = \text{sigmoid}\left(f_\mathtt{F}^{\text{gen}}(\mathbf{v}_{u,\mathtt{F}})\right). \tag{2}$$

If a child is to be generated, then as in (1), the state is propagated forward without modification.

In the case of `sequential` fields, a horizontal LSTM is employed for both child decisions and state updates. We refer to Figure 5c for an illustration of the recurrent process. After being initialized with a transformation of the vertical state, $\mathbf{s}_{\mathtt{F},0} = \mathbf{W}_\mathtt{F}\mathbf{v}_{u,\mathtt{F}}$, the horizontal LSTM iteratively decides whether to generate another child by applying a modified form of (2):

$$p\left(z_{\mathtt{F},i} = 1 \mid \mathbf{s}_{\mathtt{F},i-1},\ \mathbf{v}_{u,\mathtt{F}}\right) = \text{sigmoid}\left(f_\mathtt{F}^{\text{gen}}\left(\mathbf{s}_{\mathtt{F},i-1},\ \mathbf{v}_{u,\mathtt{F}}\right)\right).$$

If $z_{\mathtt{F},i} = 0$, generation stops and the process terminates, as represented by the solid black circle in Figure 5c. Otherwise, the process continues as represented by the white circle in Figure 5c. In that case, the horizontal state $\mathbf{s}_{u,i-1}$ is combined with the vertical state $\mathbf{v}_{u,\mathtt{F}}$ and an attention-based context vector $\mathbf{c}_{\mathtt{F},i}$ using a feedforward network $f_\mathtt{F}^{\text{update}}$ to obtain a joint context-dependent encoding of the field `F` and the position i:

$$\tilde{\mathbf{e}}_{\mathtt{F},i} = f_\mathtt{F}^{\text{update}}(\mathbf{v}_{u,\mathtt{F}},\ \mathbf{s}_{u,i-1},\ \mathbf{c}_{\mathtt{F},i}).$$

The result is used to perform a vertical LSTM update for the corresponding child w_i:

$$\mathbf{v}_{w_i} = \text{LSTM}^v(\mathbf{v}_{u,\mathtt{F}},\ \tilde{\mathbf{e}}_{\mathtt{F},i}).$$

Finally, the horizontal LSTM state is updated using the same field-position encoding, and the process continues:

$$\mathbf{s}_{u,i} = \text{LSTM}^h(\mathbf{s}_{u,i-1},\ \tilde{\mathbf{e}}_{\mathtt{F},i}).$$

Primitive type modules Each primitive type $\mathtt{T}$ has a corresponding module whose role is to select among the values $\mathtt{y}$ within the domain of that type. Figure 5d presents an example of the simplest form of this selection process, where the value $\mathtt{y}$ is obtained from a closed list via a softmax layer applied to an incoming vertical LSTM state:

$$p\left(\mathtt{y} \mid \mathtt{T}, \mathbf{v}\right) = \left[\mathrm{softmax}\left(f_{\mathtt{T}}\left(\mathbf{v}\right)\right)\right]_{\mathtt{y}}.$$

Some string-valued types are open class, however. To deal with these, we allow generation both from a closed list of previously seen values, as in Figure 5d, and synthesis of new values. Synthesis is delegated to a character-level LSTM language model (Bengio et al., 2003), and part of the role of the primitive module for open class types is to choose whether to synthesize a new value or not. During training, we allow the model to use the character LSTM only for unknown strings but include the log probability of that binary decision in the loss in order to ensure the model learns when to generate from the character LSTM.

3.3 Decoding Process

The decoding process proceeds through mutual recursion between the constituting modules, where the syntactic structure of the output tree mirrors the call graph of the generation procedure. At each step, the active decoder module either makes a generation decision, propagates state down the tree, or both.

To construct a composite node of a given type, the decoder calls the appropriate composite type module to obtain a constructor and its associated module. That module is then invoked to obtain updated vertical LSTM states for each of the constructor's fields, and the corresponding constructor field modules are invoked to advance the process to those children.

This process continues downward, stopping at each primitive node, where a value is generated but no further recursion is carried out.

3.4 Attention

Following standard practice for sequence-to-sequence models, we compute a raw bilinear attention score q_t^{raw} for each token t in the input using the decoder's current state $\mathbf{x}$ and the token's encoding $\mathbf{e}_t$:

$$q_t^{\mathrm{raw}} = \mathbf{e}_t^\top \mathbf{W} \mathbf{x}.$$

The current state $\mathbf{x}$ can be either the vertical LSTM state in isolation or a concatentation of the vertical LSTM state and either a horizontal LSTM state or a character LSTM state (for string generation). Each submodule that computes attention does so using a separate matrix $\mathbf{W}$.

A separate attention score q_c^{comp} is computed for each component of the input, independent of its content:

$$q_c^{\mathrm{comp}} = \mathbf{w}_c^\top \mathbf{x}.$$

The final token-level attention scores are the sums of the raw token-level scores and the corresponding component-level scores:

$$q_t = q_t^{\mathrm{raw}} + q_{c(t)}^{\mathrm{comp}},$$

where $c(t)$ denotes the component in which token t occurs. The attention weight vector $\mathbf{a}$ is then computed using a softmax:

$$\mathbf{a} = \mathrm{softmax}\left(\mathbf{q}\right).$$

Given the weights, the attention-based context is given by:

$$\mathbf{c} = \sum_t a_t \mathbf{e}_t.$$

Certain decision points that require attention have been highlighted in the description above; however, in our final implementation we made attention available to the decoder at all decision points.

Supervised Attention In the datasets we consider, partial or total copying of input tokens into primitive nodes is quite common. Rather than providing an explicit copying mechanism (Ling et al., 2016), we instead generate alignments where possible to define a set of tokens on which the attention at a given primitive node should be concentrated.[2] If no matches are found, the corresponding set of tokens is taken to be the whole input.

The attention supervision enters the loss through a term that encourages the final attention weights to be concentrated on the specified subset. Formally, if the matched subset of component-token pairs is S, the loss term associated with the supervision would be

$$\log \sum_t \exp\left(a_t\right) - \log \sum_{t \in S} \exp\left(a_t\right), \qquad (3)$$

[2] Alignments are generated using an exact string match heuristic that also included some limited normalization, primarily splitting of special characters, undoing camel case, and lemmatization for the semantic parsing datasets.

where a_t is the attention weight associated with token t, and the sum in the first term ranges over all tokens in the input. The loss in (3) can be interpreted as the negative log probability of attending to some token in S.

4 Experimental evaluation

4.1 Semantic parsing

Data We use three semantic parsing datasets: JOBS, GEO, and ATIS. All three consist of natural language queries paired with a logical representation of their denotations. JOBS consists of 640 such pairs, with Prolog-style logical representations, while GEO and ATIS consist of 880 and 5,410 such pairs, respectively, with λ-calculus logical forms. We use the same training-test split as Zettlemoyer and Collins (2005) for JOBS and GEO, and the standard training-development-test split for ATIS. We use the preprocessed versions of these datasets made available by Dong and Lapata (2016), where text in the input has been lowercased and stemmed using NLTK (Bird et al., 2009), and matching entities appearing in the same input-output pair have been replaced by numbered abstract identifiers of the same type.

Evaluation We compute accuracies using tree exact match for evaluation. Following the publicly released code of Dong and Lapata (2016), we canonicalize the order of the children within conjunction and disjunction nodes to avoid spurious errors, but otherwise perform no transformations before comparison.

4.2 Code generation

Data We use the HEARTHSTONE dataset introduced by Ling et al. (2016), which consists of 665 cards paired with their implementations in the open-source Hearthbreaker engine.[3] Our training-development-test split is identical to that of Ling et al. (2016), with split sizes of 533, 66, and 66, respectively.

Cards contain two kinds of components: textual components that contain the card's name and a description of its function, and categorical ones that contain numerical attributes (attack, health, cost, and durability) or enumerated attributes (rarity, type, race, and class). The name of the card is represented as a sequence of characters, while its description consists of a sequence of tokens split on whitespace and punctuation. All categorical components are represented as single-token sequences.

Evaluation For direct comparison to the results of Ling et al. (2016), we evaluate our predicted code based on exact match and token-level BLEU relative to the reference implementations from the library. We additionally compute node-based precision, recall, and F1 scores for our predicted trees compared to the reference code ASTs. Formally, these scores are obtained by defining the intersection of the predicted and gold trees as their largest common tree prefix.

4.3 Settings

For each experiment, all feedforward and LSTM hidden dimensions are set to the same value. We select the dimension from {30, 40, 50, 60, 70} for the smaller JOBS and GEO datasets, or from {50, 75, 100, 125, 150} for the larger ATIS and HEARTHSTONE datasets. The dimensionality used for the inputs to the encoder is set to 100 in all cases. We apply dropout to the non-recurrent connections of the vertical and horizontal LSTMs, selecting the noise ratio from {0.2, 0.3, 0.4, 0.5}. All parameters are randomly initialized using Glorot initialization (Glorot and Bengio, 2010).

We perform 200 passes over the data for the JOBS and GEO experiments, or 400 passes for the ATIS and HEARTHSTONE experiments. Early stopping based on exact match is used for the semantic parsing experiments, where performance is evaluated on the training set for JOBS and GEO or on the development set for ATIS. Parameters for the HEARTHSTONE experiments are selected based on development BLEU scores. In order to promote generalization, ties are broken in all cases with a preference toward higher dropout ratios and lower dimensionalities, in that order.

Our system is implemented in Python using the DyNet neural network library (Neubig et al., 2017). We use the Adam optimizer (Kingma and Ba, 2014) with its default settings for optimization, with a batch size of 20 for the semantic parsing experiments, or a batch size of 10 for the HEARTHSTONE experiments.

4.4 Results

Our results on the semantic parsing datasets are presented in Table 1. Our basic system achieves

[3] Available online at https://github.com/danielyule/hearthbreaker.

ATIS		GEO		JOBS	
System	Accuracy	System	Accuracy	System	Accuracy
ZH15	84.2	ZH15	88.9	ZH15	85.0
ZC07	84.6	KCAZ13	89.0	PEK03	88.0
WKZ14	**91.3**	WKZ14	**90.4**	LJK13	90.7
DL16	84.6	DL16	87.1	DL16	90.0
ASN	85.3	ASN	85.7	ASN	**91.4**
+ SUPATT	85.9	+ SUPATT	87.1	+ SUPATT	**92.9**

Table 1: Accuracies for the semantic parsing tasks. ASN denotes our abstract syntax network framework. SUPATT refers to the supervised attention mentioned in Section 3.4.

System	Accuracy	BLEU	F1
NEAREST	3.0	65.0	65.7
LPN	6.1	67.1	–
ASN	**18.2**	**77.6**	**72.4**
+ SUPATT	**22.7**	**79.2**	**75.6**

Table 2: Results for the HEARTHSTONE task. SUPATT refers to the system with supervised attention mentioned in Section 3.4. LPN refers to the system of Ling et al. (2016). Our nearest neighbor baseline NEAREST follows that of Ling et al. (2016), though it performs somewhat better; its nonzero exact match number stems from spurious repetition in the data.

a new state-of-the-art accuracy of 91.4% on the JOBS dataset, and this number improves to 92.9% when supervised attention is added. On the ATIS and GEO datasets, we respectively exceed and match the results of Dong and Lapata (2016). However, these fall short of the previous best results of 91.3% and 90.4%, respectively, obtained by Wang et al. (2014). This difference may be partially attributable to the use of typing information or rich lexicons in most previous semantic parsing approaches (Zettlemoyer and Collins, 2007; Kwiatkowski et al., 2013; Wang et al., 2014; Zhao and Huang, 2015).

On the HEARTHSTONE dataset, we improve significantly over the initial results of Ling et al. (2016) across all evaluation metrics, as shown in Table 2. On the more stringent exact match metric, we improve from 6.1% to 18.2%, and on token-level BLEU, we improve from 67.1 to 77.6. When supervised attention is added, we obtain an additional increase of several points on each scale, achieving peak results of 22.7% accuracy and 79.2 BLEU.


```
class IronbarkProtector(MinionCard):
  def __init__(self):
    super().__init__(
      'Ironbark Protector', 8,
      CHARACTER_CLASS.DRUID,
      CARD_RARITY.COMMON)
  def create_minion(self, player):
    return Minion(
      8, 8, taunt=True)
```

Figure 6: Cards with minimal descriptions exhibit a uniform structure that our system almost always predicts correctly, as in this instance.


```
class ManaWyrm(MinionCard):
  def __init__(self):
    super().__init__(
      'Mana Wyrm', 1,
      CHARACTER_CLASS.MAGE,
      CARD_RARITY.COMMON)
  def create_minion(self, player):
    return Minion(
      1, 3, effects=[
        Effect(
          SpellCast(),
          ActionTag(
            Give(ChangeAttack(1)),
                 SelfSelector()))
      ])
```

Figure 7: For many cards with moderately complex descriptions, the implementation follows a functional style that seems to suit our modeling strategy, usually leading to correct predictions.

4.5 Error Analysis and Discussion

As the examples in Figures 6-8 show, classes in the HEARTHSTONE dataset share a great deal of common structure. As a result, in the simplest cases, such as in Figure 6, generating the code is simply a matter of matching the overall structure and plugging in the correct values in the initializer and a few other places. In such cases, our system generally predicts the correct code, with the


```
class MultiShot(SpellCard):
  def __init__(self):
    super().__init__(
      'Multi-Shot', 4,
      CHARACTER_CLASS.HUNTER,
      CARD_RARITY.FREE)
  def use(self, player, game):
    super().use(player, game)
    targets = copy.copy(
                  game.other_player.minions)
    for i in range(0, 2):
      target = game.random_choice(targets)
      targets.remove(target)
      target.damage(
        player.effective_spell_damage(3),
        self)
  def can_use(self, player, game):
    return (
      super().can_use(player, game) and
      (len(game.other_player.minions) >= 2))
```

```
class MultiShot(SpellCard):
  def __init__(self):
    super().__init__(
      'Multi-Shot', 4,
      CHARACTER_CLASS.HUNTER,
      CARD_RARITY.FREE)
  def use(self, player, game):
    super().use(player, game)
    minions = copy.copy(
                  game.other_player.minions)
    for i in range(0, 3):
      minion = game.random_choice(minions)
      minions.remove(minion)
  def can_use(self, player, game):
    return (
      super().can_use(player, game) and
      len(game.other_player.minions) >= 3)
```

Figure 8: Cards with nontrivial logic expressed in an imperative style are the most challenging for our system. In this example, our prediction comes close to the gold code, but misses an important statement in addition to making a few other minor errors. *(Left)* gold code; *(right)* predicted code.

exception of instances in which strings are incorrectly transduced. Introducing a dedicated copying mechanism like the one used by Ling et al. (2016) or more specialized machinery for string transduction may alleviate this latter problem.

The next simplest category of card-code pairs consists of those in which the card's logic is mostly implemented via nested function calls. Figure 7 illustrates a typical case, in which the card's effect is triggered by a game event (a spell being cast) and both the trigger and the effect are described by arguments to an `Effect` constructor. Our system usually also performs well on instances like these, apart from idiosyncratic errors that can take the form of under- or overgeneration or simply substitution of incorrect predicates.

Cards whose code includes complex logic expressed in an imperative style, as in Figure 8, pose the greatest challenge for our system. Factors like variable naming, nontrivial control flow, and interleaving of code predictable from the description with code required due to the conventions of the library combine to make the code for these cards difficult to generate. In some instances (as in the figure), our system is nonetheless able to synthesize a close approximation. However, in the most complex cases, the predictions deviate significantly from the correct implementation.

In addition to the specific errors our system makes, some larger issues remain unresolved. Existing evaluation metrics only approximate the actual metric of interest: functional equivalence. Modifications of BLEU, tree F1, and exact match that canonicalize the code—for example, by anonymizing all variables—may prove more meaningful. Direct evaluation of functional equivalence is of course impossible in general (Sipser, 2006), and practically challenging even for the HEARTHSTONE dataset because it requires integrating with the game engine.

Existing work also does not attempt to enforce *semantic* coherence in the output. Long-distance semantic dependencies, between occurrences of a single variable for example, in particular are not modeled. Nor is well-typedness or executability. Overcoming these evaluation and modeling issues remains an important open problem.

5 Conclusion

ASNs provide a modular encoder-decoder architecture that can readily accommodate a variety of tasks with structured output spaces. They are particularly applicable in the presence of recursive decompositions, where they can provide a simple decoding process that closely parallels the inherent structure of the outputs. Our results demonstrate their promise for tree prediction tasks, and we believe their application to more general output structures is an interesting avenue for future work.

Acknowledgments

MR is supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Google Fellowship. MS is supported by an NSF Graduate Research Fellowship.

References

Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. 2006. *Compilers: Principles, Techniques, and Tools (2Nd Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Miltiadis Allamanis, Daniel Tarlow, Andrew D. Gordon, and Yi Wei. 2015. Bimodal modelling of source code and natural language. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. pages 2123–2132.

David Alvarez-Melis and Tommi S. Jaakkola. 2017. Tree-structured decoding with doubly-recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR) 2017*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Oral.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Matej Balog, Alexander L. Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. 2016. Deepcoder: Learning to write programs. *CoRR* abs/1611.01989.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155. http://dl.acm.org/citation.cfm?id=944919.944966.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 740–750. http://aclweb.org/anthology/D/D14/D14-1082.pdf.

James Cross and Liang Huang. 2016. Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 1–11.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. *CoRR* abs/1601.01280.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 199–209.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke S. Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1545–1556.

Percy Liang, Michael I. Jordan, and Dan Klein. 2010. Learning programs: A hierarchical bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. pages 639–646.

Percy Liang, Michael I. Jordan, and Dan Klein. 2013. Learning dependency-based compositional semantics. *Comput. Linguist.* 39(2):389–446. https://doi.org/10.1162/COLI_a_00127.

Wang Ling, Phil Blunsom, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, Fumin Wang, and Andrew Senior. 2016. Latent predictor networks for code generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Chris J. Maddison and Daniel Tarlow. 2014. Structured generative models of natural source code. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. pages 649–657.

Aditya Krishna Menon, Omer Tamuz, Sumit Gulwani, Butler W. Lampson, and Adam Kalai. 2013. A machine learning framework for programming by example. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. pages 187–195.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke

Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, pages 149–157.

Richard Shin, Alexander A. Alemi, Geoffrey Irving, and Oriol Vinyals. 2017. Tree-structured variational autoencoder. In *Proceedings of the International Conference on Learning Representations (ICLR) 2017*.

Michael Sipser. 2006. *Introduction to the Theory of Computation*. Course Technology, second edition.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. pages 2773–2781.

Adrienne Wang, Tom Kwiatkowski, and Luke S Zettlemoyer. 2014. Morpho-syntactic lexical generalization for ccg semantic parsing. In *EMNLP*. pages 1284–1295.

Daniel C. Wang, Andrew W. Appel, Jeff L. Korn, and Christopher S. Serra. 1997. The zephyr abstract syntax description language. In *Proceedings of the Conference on Domain-Specific Languages on Conference on Domain-Specific Languages (DSL), 1997*. USENIX Association, Berkeley, CA, USA, DSL'97, pages 17–17.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*. pages 658–666.

Luke S. Zettlemoyer and Michael Collins. 2007. Online learning of relaxed ccg grammars for parsing to logical form. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007*. pages 678–687.

Kai Zhao and Liang Huang. 2015. Type-driven incremental semantic parsing with polymorphism. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. pages 1416–1421.

A Appendix

```
expr
  = Apply(pred predicate, arg* arguments)
  | Not(expr argument)
  | Or(expr left, expr right)
  | And(expr* arguments)

arg
  = Literal(lit literal)
  | Variable(var variable)
```

Figure 9: The Prolog-style grammar we use for the JOBS task.

```
expr
  = Variable(var variable)
  | Entity(ent entity)
  | Number(num number)
  | Apply(pred predicate, expr* arguments)
  | Argmax(var variable, expr domain, expr body)
  | Argmin(var variable, expr domain, expr body)
  | Count(var variable, expr body)
  | Exists(var variable, expr body)
  | Lambda(var variable, var_type type, expr body)
  | Max(var variable, expr body)
  | Min(var variable, expr body)
  | Sum(var variable, expr domain, expr body)
  | The(var variable, expr body)
  | Not(expr argument)
  | And(expr* arguments)
  | Or(expr* arguments)
  | Compare(cmp_op op, expr left, expr right)

cmp_op = Equal | LessThan | GreaterThan
```

Figure 10: The λ-calculus grammar used by our system.