Abstract

Semantic parses are directed acyclic graphs (DAGs), so semantic parsing should be modeled as graph prediction. But predicting graphs presents difficult technical challenges, so it is simpler and more common to predict the linearized graphs found in semantic parsing datasets using well-understood sequence models. The cost of this simplicity is that the predicted strings may not be well-formed graphs. We present recurrent neural network DAG grammars, a graph-aware sequence model that ensures only well-formed graphs while sidestepping many difficulties in graph prediction. We test our model on the Parallel Meaning Bank—a multilingual semantic graphbank. Our approach yields competitive results in English and establishes the first results for German, Italian and Dutch.

1 Introduction

Semantic parsing is the task of mapping natural language to machine interpretable meaning representations, which in turn can be expressed in many different formalisms, including lambda calculus (Montague, 1973), dependency-based compositional semantics (Liang et al., 2011), frame semantics (Baker et al., 1998), abstract meaning representations (AMR; Banerjee et al. 2013), minimal recursion semantics (MRS; Copestake et al. 2005), and discourse representation theory (DRT; Kamp 1981).

Explicitly or implicitly, a representation in any of these formalisms can be expressed as a directed acyclic graph (DAG). Consider the sentence “Every ship in the dock needs a big anchor”. Its meaning representation, expressed as a Discourse Representation Structure (DRS, Kamp 1981), is shown in Figure 1.¹ A DRS is drawn as a box with two parts: the top part lists variables for discourse referents (e.g. \( x_1, e_1 \)) and the bottom part can contain unary predicates expressing the type of a variable (e.g. ship, need), binary predicates specifying relationships between variables (e.g. PARTOF, TOPIC), logical operators expressing relationships between nested boxes (e.g. \( \Rightarrow, \neg \)), or binary discourse relations (e.g., RESULT, CONTRAST). To express a DRS as a graph, we represent each box as a node labeled \( \square \); each variable as a node labeled by its associated unary predicate; and each binary predicate, logical operator, or discourse relation as an edge from the first argument to the second (Figure 2). To fully realize the representation as a DAG, additional transformations are sometimes necessary: in DRS, when a box represents a presupposition, as box \( b_4 \) does, the label of the node corresponding to the presupposed variable is marked (e.g. \( x_2/\text{dock}^b \)); and edges can be reversed (e.g. \( \text{TOPIC}(s_1, x_3) \) becomes \( \text{TOPICOF}(s_1, x_3) \)).

Since meaning representations are graphs, semantic parsing should be modeled as graph prediction. But how do we predict graphs? A popular approach is to predict the linearized graph—that is, the string representation of the graph found in most semantic graphbanks. Figure 3 illus-

¹For simplicity, our examples do not show time representations, though these are consistently present in our data.
Graph-aware string rewriting

We use a grammar to model the process of rewriting. Formally, our grammar is a graph grammar, specifically a restricted DAG grammar (Björklund...
(et al., 2016), a type of context-free graph grammar designed to model linearized DAGs. Since linearized DAGs are strings, we present it as a string-rewriting system, which can be described more compactly than a graph grammar while making the connection to sequences more explicit. The correspondence between string rewriting and graph grammars is given in the Appendix.

A grammar in our model is defined by a set \( \Sigma \) of terminal symbols consisting of all symbols that can appear in the final string—brackets, variable types, node labels, and edge labels; a set \( N \) of \( n + 1 \) nonterminal symbols denoted by \( \{ L, T_0, \ldots, T_n \} \), for some maximum value \( n \); an unbounded set \( V \) of variable references \( \{ \$1, \$2, \ldots \} \); and a set of productions, which are defined below.

We say that \( T_0 \) is the start symbol, and for each symbol \( T_i \in N \), we say that \( i \) is its rank, and we say that \( L \) has a rank of 0. A nonterminal of rank \( i \) can be written as a function \( i \) of variable references—for example, we can write \( T_2(\$1, \$2) \). By convention, we write the rank-0 nonterminals \( L \) and \( T_0 \) without brackets. Productions in our grammar take the form \( \alpha \rightarrow \beta \), where \( \alpha \) is a function of \( \alpha \) over \( \$1, \ldots, \$i \); and \( \beta \) is a linearized graph in PENMAN format, with each of its sub-trees replaced by either a function or a variable reference. Optionally, the variable name in \( \beta \) may replace one of the variable references in \( \alpha \). All variable references in a production must appear at least twice. Hence every variable reference in \( \alpha \) must contain at least once in \( \beta \), and variables that do not appear in \( \alpha \) must appear at least twice in \( \beta \).

To illustrate, we will use the following grammar, which can generate the string in Figure 3, assuming \( L \) can also rewrite as any node label.

\[
\begin{align*}
T_0 &\rightarrow (b/\Box : \text{IMP}_1 \ T_1(\$1) : \text{IMP}_2 \ T_1(\$1)) \\
T_0 &\rightarrow (x/L) \\
T_0 &\rightarrow (s/L) \\
T_0 &\rightarrow (x/L : \text{TOPICOFT}_0) \\
T_1(\$1) &\rightarrow (b/\Box : \text{DRS} \ T_1(\$1)) \\
T_1(\$1) &\rightarrow (e/L : \text{PIVOT} \ $1 : \text{THEME} \ T_0) \\
T_1(x) &\rightarrow (x/L : \text{PARTOF} \ T_0)
\end{align*}
\]

Our grammar derives strings by first rewriting the start symbol \( T_0 \), and at each subsequent step rewriting the leftmost function in the partially derived string, with special handling for variable references described below. A derivation is complete when no functions remain.

We illustrate the rewriting process in Figure 4. The start symbol \( T_0 \) at step 1 is rewritten by production \( r_1 \) in step 2, and the new \( b \) variable introduced at this step is deterministically renamed to the unique name \( b_1 \). In step 3, the leftmost \( T_1(\$1) \) is rewritten by production \( r_5 \), and the new \( b \) variable is likewise renamed to the unique \( b_2 \). All productions apply in this way, simply replacing a left-hand-side function with a right-hand side expression. These rewrites are coupled with a mechanism to correctly handle multiple references to shared variables, as illustrated in Step 4 when production \( r_7 \) is applied. In this production, the left-hand-side function applies to the \( x \) variable naming the right-hand-side node. When this production applies, \( x \) is renamed to the unique \( x_1 \) as in previous steps, but because it appears in the left-hand-side, the reference \( \$1 \) is bound to this new variable name throughout the partially-derived string. In this way, the reference to \( x_1 \) is passed through the subsequent rewrites, becoming the target of a PIVOT edge at step 10. Derivations of a DAG grammar are context-free (Figure 5).^2

Our model requires an explicit grammar like the one in \( r_1 \ldots r_7 \), which we obtain by converting each DAG in the training data into a sequence of productions. The conversion yields a single, unique sequence of productions via a simple linear-time algorithm that recursively decomposes a DAG into subgraphs (Björklund et al., 2016). Each subgraph consists of single node and its outgoing edges, as exemplified by the PENMAN-formatted right-hand-sides of \( r_1 \) through \( r_7 \). Each outgoing edge points to a nonterminal symbol representing a subgraph. If a subgraph does not share any nodes with its siblings, it is represented by \( T_0 \). But if any subgraphs share a node, then a variable reference must refer for this node in the production associated with the lowest common ancestor of all its incoming edges. For example, in Figure 3, the common ancestor of the two edges targeting \( x_1 \) is the node \( b_1 \), so production \( r_1 \) must contain two copies of variable reference \$1 to account for this.

A more mathematical account can be found along with a proof of correctness in Björklund et al. (2016). Our implementation follows their description unchanged.

^2However, the language of derived strings may not be context-free due to variable references.
derivation, where GEN-FRAG and GEN-LABEL are invoked when rewriting a non-terminal

Figure 4: A partial derivation of the string in Figure 3. The stack operations follow closely each step in the rewriting system of Section 2. Specifically, 

\[ G \text{ outputs a directed acyclic graph } \]

architecture that takes as input a sentence \( w \). We model graph parsing with an encoder-decoder neural network realizer (RNNG, Dyer et al., 2016). As in RNNG, we use inspired by Recurrent Neural Network Grammars (Abzianidze and Bos, 2017) embedding (\( \mathbf{d}_n \)) and dependency label embedding (\( \mathbf{d}_n \)). An input \( \mathbf{x}_i \) is computed as the weighted concatenation of these features followed by a non-linear projection (with vectors and matrices in \textbf{bold}):

\[ \mathbf{x}_i = \tanh(\mathbf{W}^{(1)}[\mathbf{w}_i; \mathbf{w}_i^{p}; \mathbf{l}_i; \mathbf{p}_i; \mathbf{u}_i; \mathbf{d}_i]) \]  

Input encoder

We define the probability of generating graph \( G \) conditioned of input sentence \( w \) as follows:

\[ p(G|w) = p(a|w) = \prod_{i=1}^{|a|} p(a_i|a_{<i}, w) \]  

3 Neural Network Realizer

We model graph parsing with an encoder-decoder architecture that takes as input a sentence \( w \) and outputs a directed acyclic graph \( G \) derived using the rewriting system of Section 2. Specifically, we model its derivation tree in top-down, left-to-right order as a sequence of actions \( a = a_1 \ldots a_{|a|} \), inspired by Recurrent Neural Network Grammars (RNNG, Dyer et al., 2016). As in RNNG, we use a stack to store partial derivations.

We model two types of actions: GEN-FRAG rewrites \( T_i \) nonterminals, while GEN-LABEL rewrites \( L \) nonterminals, always resulting in a leaf of the derivation tree. A third REDUCE action is applied whenever a subtree of the derivation tree is complete, and since the number of subtrees is known in advance, it is applied deterministically. For example, when we predict \( r_1 \), this determines that we must rewrite an \( L \) and then recursively rewrite two copies of \( T_1(\$1) \) and then apply REDUCE. Hence graph generation reduces to predicting rewrites only.

We define the probability of generating graph \( G \) conditioned of input sentence \( w \) as follows:

\[ p(G|w) = p(a|w) = \prod_{i=1}^{|a|} p(a_i|a_{<i}, w) \]  

Figure 5: A derivation tree corresponding to Figure 4. Solid edges rewrite \( T_i \) nonterminals, while dotted rewrite \( L \) nonterminals.

\[ \mathbf{x}_i = \tanh(\mathbf{W}^{(1)}[\mathbf{w}_i; \mathbf{w}_i^{p}; \mathbf{l}_i; \mathbf{p}_i; \mathbf{u}_i; \mathbf{d}_i]) \]  

\[ (2) \]

\[ \text{Universal semantic tags are language neutral tags intended to characterize lexical semantics.} \]  

\[ (2) \]

\[ \text{Universal semantic tags are language neutral tags intended to characterize lexical semantics.} \]  

\[ (2) \]

\[ \text{Universal semantic tags are language neutral tags intended to characterize lexical semantics.} \]  

\[ (2) \]

\[ \text{Universal semantic tags are language neutral tags intended to characterize lexical semantics.} \]  

\[ (2) \]
Graph decoder Since we know in advance whether the next action is GEN-FRAG or GEN-LABEL, we use different models for them.

**GEN-FRAG.** If step \( t \) rewrites a \( T \) nonterminal, we predict the production \( y_t \) that rewrites it using context vector \( c_t \) and incoming edge embedding \( e_t \). To obtain \( c_t \) we use soft attention (Luong et al., 2015) and weight each input hidden representation \( h^c_i \) to decoding hidden state \( h^d_t \):

\[
\text{score}(h^c_i, h^d_t) = h^c_i \mathbf{W}^{(2)} h^d_t
\]

\[
\alpha_{ti} = \frac{\exp(\text{score}(h^c_i, h^d_t))}{\sum_{i'} \exp(\text{score}(h^c_{i'}, h^d_t))}
\]

\[
c_t = \sum_{i=1}^{n} \alpha_{ti} h^c_i
\]

\[
y_t = \mathbf{W}^{(3)} c_t + \mathbf{W}^{(4)} e
\]

The contribution of \( c \) and \( e \) is weighted by matrices \( \mathbf{W}^{(3)} \) and \( \mathbf{W}^{(4)} \), respectively.

We then update the stackLSTM representation using the embedding of the non-terminal fragment \( y_t \) (denoted as \( y^f_t \)), as follows:

\[
h^d_{t+1} = \text{LSTM}(y^f_t, h^d_t) \tag{4}
\]

**GEN-LABEL.** Labels \( L \) can be rewritten to either semantic constants (e.g., ‘speaker’, ‘now’, ‘hearer’) or unary predicates that often correspond to the lemmas of the input words (e.g., ‘love’) or. We predict the former using a model identical to the one for GEN-FRAG. For the latter, we use a selection mechanism to choose an input lemma to copy to output. We model selection following Liu et al. (2018), assigning each input lemma a score \( o_{ji} \) that we then pass through a softmax layer to obtain a distribution:

\[
o_{ji} = h^d_j \mathbf{W}^{(5)} h^c_i
\]

\[
p^\text{copy}_i = \text{SOFTMAX}(o_{ji}) \tag{5}
\]

where \( h_i \) is the encoder hidden state for word \( w_i \).

We allow the model to learn whether to use soft-attention or the selection mechanism through a binary classifier, conditioned on the decoder hidden state at time \( t \), \( h^d_t \). Similar to Equation (4), we update the stackLSTM with the embedding of terminal predicted.\(^5\)

\(^5\)In the PMB, each terminal is annotated for sense (e.g. ‘n.01’, ‘s.01’) and presupposition (e.g. for ‘dock’ in Figure 3) as well. We predict both the sense tag and whether a terminal is presupposed or not independently conditioned on the current stackLSTM state and the embedding of the main terminal labels but are not used to update the state of the stackLSTM.

**REDUCE.** When a reduce action is applied, we use an LSTM to compose the fragments on top of the stack. Using the derivation tree in Figure 5 as reference, let \([c_1, ..., c_n]\) denote the embeddings of one or more sister nodes \( r_i \) and \( p_a \) the embedding of their parent node, which we refer to as children and parent fragments respectively. A reduce operation runs an LSTM over the children fragments and the parent fragment in order and then uses the final state \( u \) to update the stack LSTM as follows:

\[
i = [c_1...c_n, p_a]
\]

\[
\hat{u} = \text{LSTM}(i_t, h^c_t)
\]

\[
h^d_{t+1} = \text{LSTM}(u, h^d_t)
\]

The models are trained to minimize a cross-entropy loss objective \( J \) over the sequence of gold actions \( a_i \) in the derivation:

\[
J = -\sum_{i=1}^{[a_i]} \log p(a_i) \tag{6}
\]

### 4 Experimental Setup

We evaluated our model on the Parallel Meaning Bank (PMB; Abzianidze et al. 2017), a semantic bank where sentences in English, Italian, German, and Dutch have been annotated following Discourse Representation Theory (Kamp and Reyle, 2013). Lexical predicates in PMB are in English, even for non-English languages. Since this is not compatible with our copy mechanism, we revert predicates to their original language by substituting them with the lemmas of the tokens they are aligned to. In our experiments we used both PMB v.2.1.0 and v.2.2.0\(^6\); we included the former release in order to compare against the state-of-the-art seq2seq system of van Noord et al. (2018). Statistics on the data and the grammar extracted from v.2.2.0 are reported in Table 1.

\(^6\)English data is available for both releases at [https://github.com/RikVN/DRS_parsing](https://github.com/RikVN/DRS_parsing); for the other languages, we used the officially released data available at [http://pmb.let.rug.nl/data.php](http://pmb.let.rug.nl/data.php)
as nodes, while conditions, operators, and discourse relations become edge labels between these nodes.

We consider main boxes (see $b_3$, $b_5$, and $b_4$ in Figure 1 separately from presuppositional boxes (see $b_1$), which represent instances that are presupposed in the wider discourse context (e.g., definite expressions). Using Figure 2 as an example, $b_3$, $b_5$, and $b_4$ become nodes in the graph and material implication (IMP stands for $\Rightarrow$) becomes an edge label. If an operator or a relation is binary, as in this case, we number the edge label so as to preserve the order of the operands.

For each node in a main box, we expand the graph by adding all relations and variables belonging to it. We identify the head of the first relation or the first referent mentioned as the head variable. These are ‘ship ($x_1$)’ for $b_2$, and ‘need($e_1$)’ for $b_3$. We attach the head variable as a child of the box-node and follow the relations recursively to expand the subgraph. If while expanding a graph a variable in a condition is part of a presuppositional box, we introduce it as a new node and add to its label the superscript $p$. When expanding the DAG along the edge PartOf, since $x_2$ is also in the presuppositional box in Figure 1, we attach the node ‘dock$p$’. Graphs extracted this way are mostly acyclic except for adjectival phrases and relative clauses where state variables can be themselves root (e.g., big “a.01” $s_1$). We get rid of these extra roots by reversing the direction of the edge involved and adding an ‘-of’ to the edge label to flag this change (see TOPIC-OF).

### 4.2 System Comparison

We compared the performance of our graph parser (seq2graph below) with a sequence-to-sequence model (enhanced with a copy mechanism) which decodes to a string linearization of the graph similar to the one shown in Figure 3 (seq2seq + copy below). We also compare against the recently proposed model of van Noord et al. (2018); they introduce a seq2seq model that generates a DRS as a concatenation of clauses, essentially a flat version of the standard box notation. The decoded string is made aware of the overall graph structure during preprocessing where variables are replaced by indices indicating when they were first introduced and their recency. In contrast, we model the graph structure explicitly. van Noord et al. (2018) experimented with both word and character-based models, as well as with an ensemble of both, using word embedding features. Since all our models are word-based, we compare our results with their best word model, using word embedding features only (trained using 10-fold cross validation).

### 4.3 Model Configurations

In addition to word embeddings$^7$, we also report on experiments which make use of additional features. Specifically, for each word we add information about its universal PoS tag, lemma, universal semantic tag, and dependency label.$^8$

In Section 2, we mentioned that given a production $\alpha \rightarrow \beta$, variable references in $\alpha$ should appear at least once in $\beta$ (i.e., they should have the same rank). In all experiments so far, we did not model this constraint explicitly to investigate whether the model is able by default to predict rank correctly. However, in exploring model configurations we also report on whether adding this constraint leads to better performance.

### 4.4 Cross-lingual Experiments

We conducted two sets of experiments: one monolingual (mono below) where we train and test on the same language and one cross-lingual (cross below), where we train a model on English and test it on one of the other three languages. The goal of our cross-lingual experiments is to examine whether we need data in a target language at all since the semantic representation itself is language agnostic and lexical predicates are dealt with via

<table>
<thead>
<tr>
<th>Gold</th>
<th>train</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>#inst.</td>
<td>#frags</td>
<td>avg. rank</td>
<td>#inst.</td>
</tr>
<tr>
<td>en</td>
<td>4,574</td>
<td>1,105</td>
<td>1.64</td>
</tr>
<tr>
<td>it</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>de</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>nl</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Silver</th>
<th>train</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>#inst.</td>
<td>#frags</td>
<td>avg. rank</td>
<td>#inst.</td>
</tr>
<tr>
<td>en</td>
<td>52,120</td>
<td>11,206</td>
<td>1.90</td>
</tr>
<tr>
<td>it</td>
<td>2,546</td>
<td>1,051</td>
<td>1.52</td>
</tr>
<tr>
<td>de</td>
<td>3,274</td>
<td>1,781</td>
<td>1.54</td>
</tr>
<tr>
<td>nl</td>
<td>874</td>
<td>789</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Table 1: Data statistics of PMB v.2.2.0.

---

$^7$We used word embeddings pre-trained with Glove and available at [https://nlp.stanford.edu/projects/glove/](https://nlp.stanford.edu/projects/glove/).

$^8$Universal PoS tags, lemmas and dependency labels for all languages were obtained using pretrained UDPipe models available at [http://ufal.mff.cuni.cz/udpipe#download](http://ufal.mff.cuni.cz/udpipe#download); gold-standard universal semantic tags were extracted from the PMB release.
the copy mechanism. Most of the features mentioned above are cross-linguistic and therefore fit both mono and cross-lingual settings, with the exception of lemma and word embeddings, where we exclude the former and replaced the latter with multilingual word embeddings.

4.5 System settings

For training, we used the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001 and a decay rate of 0.1 every 10 epochs. Randomly initialized and pre-trained word embeddings have a dimensionality of 128 and 100 respectively, and all other features a dimensionality of 50. In all cross-lingual experiments, the pre-trained word embeddings have a dimensionality of 300. The LSTMs in the encoder and the decoder have a dimensionality of 150 and 100 non-terminal and terminal embeddings during decoding have a dimensionality of 50. The system is trained for 30 epochs, with the best system chosen based on dev set performance.

4.6 Evaluation Metric

We evaluated our system by scoring the similarity between predicted and gold graphs. We used Counter (van Noord et al., 2018), an adaptation of Smatch (Cai and Knight, 2013) to Discourse Representation Structures where graphs are first transformed into a set of ‘source node – edge label – target node’ triples and the best mapping between the variables is found through an iterative hill-climbing strategy. Furthermore, Counter checks whether DRSs are well-formed in that all boxes should be connected, acyclic, with fully instantiated variables, and correctly assigned sense tags.

It is worth mentioning that there can be cases where our parser generates ill-formed graphs according to Counter; this is however not due to the model itself but to the way the graph is converted back in a format accepted by Counter.

All results shown are an averages over 5 runs.

9We experimented with embeddings obtained with iterative procrustes (available at https://github.com/facebookresearch/MUSE) and with Guo et al. (2016)’s ‘robust projection’ method where the embedding of non-English words is computed as the weighted average of English ones. We found the first method to perform better on cross-lingual word similarity tasks and used it in our experiments.

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>illformed</th>
</tr>
</thead>
<tbody>
<tr>
<td>van Noord et al. (2018)</td>
<td>—</td>
<td>—</td>
<td>72.80</td>
<td>20%</td>
</tr>
<tr>
<td>seq2seq + copy</td>
<td>75.57</td>
<td>67.27</td>
<td>71.18</td>
<td>4.12%</td>
</tr>
<tr>
<td>seq2graph</td>
<td>75.51</td>
<td>71.69</td>
<td>73.55</td>
<td>0.40%</td>
</tr>
</tbody>
</table>

Table 2: Model performance (Precision, Recall, F1) on PMB data (v.2.1.0, test set); models were trained on gold standard data.

<table>
<thead>
<tr>
<th></th>
<th>#sents</th>
<th>conversion score</th>
<th>F1</th>
<th>illformed</th>
</tr>
</thead>
<tbody>
<tr>
<td>van Noord et al. (2018)</td>
<td>73.778</td>
<td>100%</td>
<td>82.7</td>
<td>1.10%</td>
</tr>
<tr>
<td>seq2seq+copy</td>
<td>56.694</td>
<td>89.58%</td>
<td>74.67</td>
<td>13.45%</td>
</tr>
<tr>
<td>seq2graph</td>
<td>56.694</td>
<td>89.58%</td>
<td>77.1</td>
<td>0.90%</td>
</tr>
</tbody>
</table>

Table 3: Model performance (Precision, Recall, F1) on PMB data (v.2.1.0, test set); models were trained on gold and silver standard data combined.

5 Results

System comparison Table 2 summarizes our results on the PMB gold data (v.2.1.0, test set). We compare our graph decoder against the system of van Noord et al. (2018) and our implementation of a seq2seq model, enhanced with a copy mechanism. Overall, we see that our graph decoder outperforms both models. Moreover, it reduces the number of illformed representations without any specific constraints or post-processing in order to ensure the well-formedness of the semantics of the output.

The PMB (v.2.1.0) contains a large number of silver standard annotations which have been only partially manually corrected (see Table 1). Following van Noord et al. (2018), we also trained our parser on both silver and gold standard data combined. As shown in Table 3, increasing the training data improves performance but the difference is not as dramatic as in van Noord et al. (2018). We found that this is because our parser requires graphs that are fully instantiated – all unary predicates (e.g. ship(x)) need to be present for the graph to be fully connected, which is often not the case for silver graphs. Our model is at a disadvantage since it could exploit less training data; during grammar extraction we could not process around 20K sentences and in some cases could not reconstruct the whole graph, as shown by the conversion score.10

10The conversion score is computed using Counter where we consider the converted graph representation as a ‘predicted’ DRS and the original DRS structure as gold data. As a comparison, converting the gold DRS structures yields a conversion score of 99.8%.
**Model configurations** Table 4 reports on various ablation experiments investigating which features and combinations thereof perform best. The experiments were conducted on the development set (PMB v2.2.0). We show a basic version of our seq2graph model with word embeddings, to which we add information about rank (+restrict). We also experimented with the full gamut of additional features (+feats) as well as with ablations of individual feature classes. For comparison, we also show the performance of a graph-to-string model (seq2seq+copy).

As can be seen, all linguistic features seem to improve performance. Restricting fragment selection by rank does not seem to improve the overall result showing that our baseline model is already able to predict fragments with the correct rank throughout the derivation. Subsequent experiments report results with this model using all linguistic features, unless otherwise specified.

**Cross-lingual experiments** Results on DRT parsing for languages other than English are reported in Table 5. There is no gold standard training data for non-English languages in the PMB (v.2.2.0). We therefore trained our parser on silver standard data but did use the provided gold standard data for development and testing (see Table 1). We present two versions of our parser, one where we train and test on the same language (s2g mono-silver) and another one where a model is trained on English but tested on the other languages (s2g cross-silver). We also show the results of a sequence-to-sequence model enhanced with a copy mechanism.

In the monolingual setting, our graph parser outperforms the seq2seq baseline by a large margin; we hypothesize this is due to the large percentage of ill-formed semantics, mostly due to training on silver data. The difference in performance between our cross-lingual parser and the monolingual parser for all languages is small, and in Dutch the two parsers perform on par, suggesting that English data and language independent features can be leveraged to build parsers in other languages when data is scarce or event absent. We also conducted various ablation studies to examine the contribution of individual features to cross-linguistic semantic parsing. Our experiments revealed that universal semantic tags are most useful, while the multilingual word embeddings that we have tested with are not. We refer the interested reader to the supplementary material for more detail on these experiments.

**Error Analysis** We further analyzed the output of our parser to gain insight as to what parts of meaning representation are still challenging. Table 6 shows a more detailed break-down of system output as computed by Counter, where operators (e.g., negation, implication), roles (i.e., binary relations, such as ‘Theme’), concepts (i.e., unary predicates like ‘ship’), and synsets (i.e., sense tags like ‘n.01’) are scored separately. Synsets are further broken down into ‘Nouns’, ‘Verbs’, ‘Adverbs’, and ‘Adjectives’. We compare our best seq2graph models (+feats) trained on the PMB v2.2.0, gold and gold+silver data respectively.

Adding silver data helps with semantic elements (operators, roles and concepts), but does not in the case of sense prediction where the only category that benefits from additional data are is nouns. We also found that ignoring the prediction of sense tags altogether helps with the performance of both models.


