We consider the task of crosslingual semantic parsing in the style of Discourse Representation Theory (DRT) where knowledge from annotated corpora in a resource-rich language is transferred via bitext to guide learning in other languages. We introduce Universal Discourse Representation Theory (UDRT), a variant of DRT that explicitly anchors semantic representations to tokens in the linguistic input. We develop a semantic parsing framework based on the Transformer architecture and utilize it to obtain semantic resources in multiple languages following two learning schemes. The many-to-one approach translates non-English text to English, and then runs a relatively accurate English parser on the translated text, while the one-to-many approach translates gold standard English to non-English text and trains multiple parsers (one per language) on the translations. Experimental results on the Parallel Meaning Bank show that our proposal outperforms strong baselines by a wide margin and can be used to construct (silver-standard) meaning banks for 99 languages.
1. Introduction

Recent years have seen a surge of interest in representational frameworks for natural language semantics. These include novel representation schemes such as Abstract Meaning Representation (AMR; Banarescu et al. 2013), Universal Conceptual Cognitive Annotation (Abend and Rappoport 2013), and Universal Decompositional Semantics (White et al. 2016), as well as existing semantic formalisms such as Minimal Recursion Semantics (MRS; Copestake et al. 2005), and Discourse Representation Theory (DRT; Kamp and Reyle 1993). The availability of annotated corpora (Flickinger, Zhang, and Kordoni 2012; May 2016; Hershcovich, Abend, and Rappoport 2017; Abzianidze et al. 2017) has further enabled the development and exploration of various semantic parsing models aiming to map natural language to formal meaning representations.

In this work, we focus on parsing meaning representations in the style of DRT (Kamp 1981; Kamp and Reyle 1993; Asher and Lascarides 2003), a formal semantic theory designed to handle a variety of linguistic phenomena, including anaphora, presuppositions (Van der Sandt 1992; Venhuizen et al. 2018), and temporal expressions within and across sentences. The basic meaning-carrying units in DRT are Discourse Representation Structures (DRSs), which are recursive, have a model-theoretic interpretation, and can be translated into first-order logic (Kamp and Reyle 1993). DRSs are scoped meaning representations—they capture the semantics of negation, modals, quantification, and presupposition triggers.

Although initial attempts at DRT parsing focused on small fragments of English (Johnson and Klein 1986; Wada and Asher 1986), more recent work has taken advantage of the availability of syntactic treebanks and robust parsers trained on them (Hockenmaier and Steedman 2007; Curran, Clark, and Bos 2007; Bos 2015) or corpora specifically annotated with discourse representation structures upon which DRS parsers can be developed more directly. Examples of such resources are the Redwoods Treebank (Oepen et al. 2002; Baldridge and Lascarides 2005b, 2005a), the Groningen Meaning Bank (Basile et al. 2012; Bos et al. 2017), and the Parallel Meaning Bank (PMB; Abzianidze et al. 2017), which contains annotations for English, German, Dutch, and Italian sentences based on a parallel corpus. Aside from larger-scale resources, renewed interest (Oepen et al. 2020)\(^1\) in DRT parsing has been triggered by the realization that document-level semantic analysis is a prerequisite to various applications ranging from machine translation (Kim, Tran, and Ney 2019) to machine reading (Gangemi et al. 2017; Chen 2018) and generation (Basile and Bos 2013; Narayan and Gardent 2014).

Figure 1(a) shows the DRS corresponding to an example sentence taken from the PMB, and its Italian translation. Conventionally, DRSs are depicted as boxes. Each box comes with a unique label (see \(b_1, b_2, b_3\) in the figure) and has two layers. The top layer contains discourse referents (e.g., \(x_1, t_1\)), and the bottom layer contains conditions over discourse referents. Each referent or condition belongs to a unique box label, showing the referent or the condition that is interpreted in that box (e.g., \(b_2 : x_1\) and \(b_2 : \text{person.n.01}(x_1)\)). The predicates are disambiguated with senses (e.g., n.01 and v.01) provided in WordNet (Fellbaum 1998). More details on the DRT formalism are discussed in Section 2.1.

\(^1\) Details on the IWCS 2019 shared task on Discourse Representation Structure parsing can be found at https://sites.google.com/view/iwcs2019/home.
**Figure 1**

(a) DRS for English sentence *everyone was killed* and its Italian translation *sono stati uccisi tutti*, taken from the Parallel Meaning Bank; (b) Universal Discourse Representation Structures (UDRS) for English; and (c) Italian sentence. The two UDRSs differ in how they are anchored to the English and Italian sentences (via alignment). Anchors $0$ and $2$ in (b) refer to English tokens *everyone* and *killed*, respectively, while anchors $3$ and $2$ in (c) refer to Italian tokens *tutti* and *uccisi*.

Despite efforts to create crosslingual DRS annotations (Abzianidze et al. 2017), the amount of gold-standard data for languages other than English is limited to a few hundred sentences that are useful for evaluation but small-scale for model training. In addition, for many languages, semantic analysis cannot be performed at all due to the lack of annotated data. The creation of such data remains an expensive endeavor requiring expert knowledge, namely, familiarity with the semantic formalism and language at hand. Because it is unrealistic to expect that semantic resources will be developed for many low-resource languages in the near future, previous work has resorted to machine translation and bitexts that are more readily available (Evang and Bos 2016; Damonte and Cohen 2018; Zhang et al. 2018; Conneau et al. 2018; Fancellu et al. 2020). Crosslingual semantic parsing leverages an existing parser in a source language (e.g., English) together with a machine translation system to learn a semantic parser for a target language (e.g., Italian or Chinese).

In this article, we also aim to develop a crosslingual DRT parser for languages where no gold-standard training data is available. We first propose a variant of the DRT formalism, which explicitly anchors semantic representations to words. Specifically, we introduce Universal Discourse Representation Structures (UDRSs) where language-dependent symbols are replaced with anchors referring to tokens (or characters, e.g.,
in the case of Chinese) of the input sentence. UDRSs are intended as an alternative representation that abstracts away from decisions regarding concepts in the source language. As shown in Figure 1(b) and Figure 1(c), “person” and “kill” are replaced with anchors $0$ and $2$, corresponding to English tokens everyone and killed, and $3$ and $2$, corresponding to Italian tokens tutti and uccisi. Also, notice that UDRSs omit information about word senses (denoted by WordNet synsets, e.g., person.n.01, time.n.08 in Figure 1(a)) as the latter cannot be assumed to be the same across languages (see Section 2.2 for further discussion).

Like other related broad-coverage semantic representations (e.g., AMR), DRSs are not directly anchored in the sentences whose meaning they purport to represent. The lack of an explicit linking between sentence tokens and semantic structures makes DRSs less usable for downstream processing tasks and less suitable for crosslingual parsing that relies on semantic and structural equivalences between languages. UDRSs omit lexical details pertaining to the input sentence and as such are able to capture similarities in the representation of expressions within the same language and across languages.

Our crosslingual parser takes advantage of UDRSs and state-of-the-art machine translation to develop semantic resources in multiple languages following two learning schemes. The many-to-one approach works by translating non-English text to English, and then running a relatively accurate English DRS parser on the translated text, while the one-to-many approach translates gold-standard English (training data) to non-English text and trains multiple parsers (one per language) on the translations. In this article, we propose (1) UDRSs to explicitly anchor DRSs to lexical tokens, which we argue is advantageous for both monolingual and crosslingual parsing; (2) a box-to-tree conversion algorithm that is lossless and reversible; and (3) a general crosslingual semantic parsing framework based on the Transformer architecture and its evaluation on the PMB following the one-to-many and many-to-one learning paradigms. We showcase the scalability of the approach by creating a large corpus with (silver-standard) discourse representation annotations in 99 languages.

2. Discourse Representation Structures

We first describe the basics of traditional DRSs, and then explain how to obtain UDRSs based on them. We also highlight the advantages of UDRSs when they are used for multilingual semantic representations.

2.1 Traditional DRSs

DRSs, the basic meaning-carrying units of DRT, are typically visualized as one or more boxes, which can be nested to represent the semantics of sentences recursively. An example is given in Figure 1(a). Each box has a label (e.g., $b_1$) and consists of two layers. The top layer contains variables (e.g., $x_1$), and the bottom layer contains conditions over the variables. For example, person.n.01($x_1$) means that variable $x_1$ is applied to predicate person.n.01, which in fact is a WordNet synset.

---

2 We use the term many-to-one to emphasize the fact that a semantic parser is trained only once (e.g., in English). In the one-to-many setting, multiple semantic parsers are trained, one per target language. The terms are equivalent to “translate test” (many-to-one) and “translate train” (one-to-many) used in previous work (Conneau et al. 2018).
The PMB adopts an extension of DRT that treats presupposition\(^3\) with projection pointers (Venhuizen, Bos, and Brouwer 2013), marking how the accommodation site (box) variables and conditions are bounded and interpreted. For example, \(b_2 : x_1\) and \(b_1 : \text{time.n.08}(t_1)\) indicate that variable \(x_1\) and condition \(\text{time.n.08}(t_1)\) should be interpreted within boxes \(b_2\) and \(b_1\), respectively. The boxes are constructed \textit{incrementally} by a set of rules mapping syntactic structures to variables and conditions (Kamp and Reyle 1993). As shown in Figure 1(a), the phrase \textit{was killed} gives rise to temporal variable \(t_1\) and condition \(\text{time.n.08}(t_1)\) and \(t_1 < \text{now}\); these temporal markers are located in box \(b_3\) together with the predicate “\textit{kill},” and are bound by outer box \(b_4\) (not drawn in the figure) which would be created to accommodate any discourse that might continue the current sentence.

\textbf{2.2 Universal DRSs}

How can DRSs be used to represent meaning across languages? An obvious idea would be to assume that English and non-English languages share identical meaning representations and sense distinctions (aka identical DRSs). For example, the English sentence \textit{everyone was killed} and the Italian sentence \textit{sono stati uccisi tutti} would be represented by the same DRS, shown in Figure 1(a). Unfortunately, this assumption is unrealistic, as sense distinctions can vary widely across languages.\(^4\) For instance, the verb \textit{eat/essen} has six senses according to the English WordNet (Fellbaum 1998) but only one in GermaNet (Hamp and Feldweg 1997), and the word \textit{good/\%\%} has 23 senses in the English WordNet but 17 senses in the Chinese WordNet (Huang et al. 2010). In other words, we cannot assume that there will be a one-to-one correspondence in the senses of the same predicate in any two languages.

Because word sense disambiguation is language-specific, we do not consider it part of the proposed crosslingual meaning representation but assume that DRS operators (e.g., negation) and semantic roles are consistent across languages. We introduce \(\hat{\text{UDRSs}},\) which replace “DRS tokens,” such as constants and predicates of conditions, with \textit{alignments} to tokens or spans (e.g., named entities) in the input sentence. An example is shown in Figure 2(b), where condition \(b_2 : \text{eat.v.01}(e_1)\) is generalized to \(b_2 : \$2.v(e_1)\). Here, \$2 corresponds to \textit{eating} and \(v\) denotes the predicates part of speech (i.e., verb). Note that even though senses are not part of the \(\hat{\text{UDRS}}\) representation, parts of speech (for predicates) are because they can provide cues for sense disambiguation across languages.

\(\hat{\text{UDRSs}}\) representations abstract semantic structures within the same language and across languages. Monolingually, they are generalizations of sentences with different semantic content but similar syntax. As shown in Figure 2, \textit{Tom is eating an apple} and \textit{Jack is washing a car} are represented by the same \(\hat{\text{UDRS}},\) which can be viewed as a template describing an event in past tense with an agent and a theme. \(\hat{\text{UDRSs}}\) are more compact representations and advantageous from a modeling perspective; they are easier to generate compared to DRSs since multiple training instances are represented by the same semantic structure. Moreover, \(\hat{\text{UDRSs}}\) can be used to capture basic meaning across languages. The \(\hat{\text{UDRS}}\) in Figure 2(c) can be used to recover the semantics of the

---

\(^3\) Presupposition is the phenomenon whereby speakers mark linguistically the information that is presupposed or taken for granted rather than being part of the main propositional content of an utterance [Beaver and Gueerts 2014]. Expressions and constructions carrying presuppositions are called “presupposition triggers,” forming a large class including definites and factive verbs.

\(^4\) [http://globalwordnet.org/about-gwa/](http://globalwordnet.org/about-gwa/).
UDRS for both sentences; (d) UDRS for sentence 句子 “汤姆 正在 吃 一个 苹果” (Tom is eating an apple) constructed via (c) by substituting indices with their corresponding words. Expressions in brackets make the anchoring explicit (e.g., $4$ is anchored to Chinese characters 苹果) and are only shown for ease of understanding, they are not the part of the UDRS.

sentence “汤姆 正在 吃 一个 苹果” (Tom is eating an apple) by substituting index $0$ with 汤姆, index $2$ with 吃, and index $4$ with 苹果 (see Figure 2(d)).

**Link to Knowledge Bases.** An important distinction between UDRSs and DRSs is that the former do not represent word senses. We view word sense disambiguation as a post-processing step that can enrich UDRSs with more fine-grained semantic information according to specific tasks and knowledge resources. UDRSs are agnostic when it comes to sense distinctions. They are compatible with WordNet that exists in multiple languages, and has been used for English DRs, but also related to resources such as BabelNet (Navigli and Ponzetto 2010), ConceptNet (Speer, Chin, and Havasi 2017), HowNet (Dong, Dong, and Hao 2006), and Wikidata. An example of how UDRSs can be combined with word senses to provide more detailed meaning representations is shown in Figure 3.

**Link to Language Models.** As explained earlier, predicates and constants in UDRSs are anchored (via alignments) to lexical tokens (see 吃,v(e1) in Figure 2(d)). As a result, UDRSs represent multiword expressions as a combination of multiple tokens aiming to assign atomic meanings and avoid redundant lexical semantics. For example, in the sentence Tom picked the graphic card up, graphic card corresponds to entity $3–4.n(x_2)$ and picked up to relation $1–5.v(x_1, x_2)$. The link between elements of the semantic representation and words in sentences is advantageous because it renders UDRSs amenable to further linguistic processing. For example, they could be interfaced with large-scale pretrained models such as BERT (Devlin et al. 2019) and GPT (Radford et al. 2019), thereby fusing together deep contextual representations and rich semantic symbols (see Figure 3). Aside from enriching pretrained models (Wu and He 2019;
Figure 3
Illustration of how UDRSs can enrich deep contextual word representations; word senses can be disambiguated as a post-processing step according to the definitions of various language-specific resources.

Hardalov, Koychev, and Nakov 2020; Kuncoro et al. 2020), such representations could further motivate future research on their interpretability (Wu and He 2019; Hardalov, Koychev, and Nakov 2020; Kuncoro et al. 2020; Hewitt and Manning 2019; Kulmizev et al. 2020).

3. Computational Formats

DRSs are displayed in a box-like format that is intuitive and easy to read but not particularly convenient for modeling purposes. As a result, DRSs are often post-processed in a format that can be straightforwardly handled by modern neural network models (Liu, Cohen, and Lapata 2018; van Noord et al. 2018b; Liu, Cohen, and Lapata 2019a). In this section, we provide an overview of existing computational formats, prior to describing our own proposed format.

3.1 Clause Format

In the PMB (Abzianidze et al. 2017), DRS variables and conditions are converted to clauses. Specifically, variables in the top box layer are converted to clauses by introducing a special condition called “REF.” Figure 4(b) presents the clause format of the DRS in Figure 4(a); here, “$b_2 \text{REF } x_1$” indicates that variable $x_1$ is bound in box $b_2$. Analogously, clause “$b_3 \text{kill v.01 } e_1$” corresponds to condition $\text{kill.v.01}(e_1)$ which
is bound in box $b_3$, and $b_4$ TRP $t_1$ “now” is bound in box $b_4$ (TRP corresponds to temporal). The mapping from boxes to clauses is not reversible; in other words, it is not straightforward to recover the original box from the clause format and restore the syntactic structure of the original sentence. For instance, the clause format discloses that temporal information is bound to box $b_4$, but not which box this information is located in (i.e., $b_3$ in Figure 4(a)). Although PMB is released with clause annotations, Algorithm 1 re-implements the conversion procedure of Abzianidze et al. (2017) to allow for a more direct comparison between clauses and the tree format introduced below.

In Algorithm 1, the function GET_VARIABLE_BOUND returns pairs of variables and box labels (indicating where these are bound) by enumerating all nested boxes. The element $P[v]$ represents the label of the box bounding variable $v$. Basic conditions are converted to a clause in lines 9–10, where $\text{cond.args}$ is a list of the arguments of the condition (e.g., predicates and referents). Unary complex conditions (i.e., negation, possibility, and necessity) are converted to clauses in lines 11–13, while lines 14–16 show how to convert binary complex conditions (i.e., implication, disjunction, and duplication) to clauses. An example is shown in Figure 4(b).

---

7 For the full list of DRS clauses, see https://pmb.let.rug.nl/drs.php.
8 Each variable has exactly one bounding box.
9 We refer to Bos et al. (2017) for more details on basic and complex conditions in DRS boxes.
Algorithm 1 Box to Clause

| Input: B, DRS in box format |
| Output: C, DRS in clause format |

1: \( P = \text{GETVARIABLEBOUND}(B); V \leftarrow \emptyset; C \leftarrow \emptyset \)

2: procedure TRAVERSAL(b)

3: for cond in b.conds do \( \triangleright \) each condition
4:     for v in cond.args do \( \triangleright \) each argument
5:         if v not in V then
6:             \( C = C \cup \{ (P[v] \ REF v) \}; V = V \cup \{ v \} \)
7:         end if
8:     end for
9:     if cond is basic then
10: \( C = C \cup \{ (\text{cond.bound \ cond.name \ cond.args}) \} \)
11: else if cond is unary complex then
12: \( C = C \cup \{ (\text{cond.bound \ cond.name \ cond.B}) \} \)
13: end if
14: else if cond is binary complex then
15: \( C = C \cup \{ (\text{cond.bound \ cond.name \ cond.B1 \ cond.B2}) \} \)
16: TRAVERSAL(cond.B1); TRAVERSAL(cond.B2)
17: end if
18: end for
19: end procedure

3.2 Tree Format

Liu, Cohen, and Lapata (2018) propose an algorithm that converts DRS boxes to trees, where each DRS box is converted to a subtree and conditions within the box are introduced as children of the subtree. In follow-on work, Liu, Cohen, and Lapata (2019a) define Discourse Representation Tree Structure based on this conversion. Problematically, the algorithm of Liu, Cohen, and Lapata (2018) simplifies the semantic representation as it does not handle presuppositions, word categories (e.g., \( n \) for noun), and senses (e.g., \( n.01 \)). In this article, we propose an improved box-to-tree conversion algorithm, which is reversible and lossless—that is, it preserves all the information present in the DRS box, as well as the syntactic structure of the original text. Our conversion procedure is described in Algorithm 2. Similar to the box-to-clause algorithm, basic conditions are converted to a tree in lines 10–11, where \( \text{cond.args} \) is a list of the arguments of the condition (e.g., predicates and referents). Unary complex conditions (i.e., negation, possibility, and necessity) are converted to subtrees in lines 12–15, while lines 16–20 show how to convert binary complex conditions (i.e., implication, disjunction, and duplication) to subtrees.

An example is shown in Figure 4(c). Basic condition \( b_2 : \text{person.n.01}(x_1) \) is converted to the subtree shown in Figure 5(a). Binary complex condition \( \rightarrow \) is converted to the subtree shown in Figure 5(b). Unary complex conditions are converted in a similar way. The final tree can be further linearized to \( (b_1 \ (\rightarrow \ b_1 \ (b_2 \ (\text{Ref} \ b_2 \ x_1) \ (\text{Pred} \ b_2 \ x_1 \ \text{person} \ n.01)) \ (b_3 \ (\text{Ref} \ b_3 \ t_1) \ (\text{Pred} \ b_4 \ t_1 \ \text{time} \ n.08) \ (\text{TPR} \ b_4 \ t_1 \ \text{now}) \ (\text{Ref} \ b_3 \ e_1) \ (\text{Pred} \ b_3 \ e_1 \ \text{kill} \ v.01) \ (\text{Time} \ b_3 \ e_1 \ t_1) \ (\text{Patient} \ b_3 \ e_1 \ x_1))) \).

4. Crosslingual Semantic Parser

As mentioned earlier, PMB (Abzianidze et al. 2017) contains a small number of gold standard DRS annotations in German, Italian, and Dutch. Multilingual DRSs in PMB
Algorithm 2 Box to Tree

Input: B, DRS in box format
Output: T, DRS in tree format

1: \( P = \text{GET VARIABLE BOUND}(B) \); \( V \leftarrow \emptyset \); \( T = \text{TRAVERSE}(B) \)

2: procedure TRAVERSE(b)
3: \( t = \text{TREE}(b.\text{name}, []) \)
4: for \( \text{cond} \in b.\text{conds} \) do
5: for \( v \in \text{cond.args} \) do
6: if \( v \not\in V \) then
7: \( c = \text{TREE}(\text{REF}, [P[v], v]); \text{ADD CHILD}(t, c); V = V \cup \{v\} \)
8: end if
9: end for
10: if \( \text{cond} \) is basic then
11: \( c = \text{TREE}(\text{cond.name}, \text{cond.bound}, \text{cond.args}) \)
12: else if \( \text{cond} \) is unary complex then
13: \( c = \text{TREE}(\text{cond.name}, []) \)
14: \( \text{ADD CHILD}(c, \text{TREE}(\text{cond.bound}, [])) \)
15: \( \text{ADD CHILD}(c, \text{TRAVERSE}(c.B)) \)
16: else if \( \text{cond} \) is binary complex then
17: \( c = \text{TREE}(\text{cond.name}, []) \)
18: \( \text{ADD CHILD}(c, \text{TREE}(\text{cond.bound}, [])) \)
19: \( \text{ADD CHILD}(c, \text{TRAVERSE}(c.B1)) \)
20: \( \text{ADD CHILD}(c, \text{TRAVERSE}(c.B2)) \)
21: end if
22: \( \text{ADD CHILD}(t, c) \)
23: end for
24: end procedure

Figure 5
Example of converting a basic condition (a) and a binary complex condition (b) to a tree.

use English WordNet synsets regardless of the source language. The output of our crosslingual semantic parser is compatible with this assumption, which is also common in related broad-coverage semantic formalisms (Damonte and Cohen 2018; Zhang et al. 2018). In the following, we present two learning schemes (illustrated schematically in
4.1 Many-to-one Method

According to the many-to-one approach, target sentences (e.g., in German) are translated to source sentences (e.g., in English) via a machine translation system; and then a relatively accurate source DRS parser (trained on gold-standard data) is adopted to map the target translations to their semantic representation. Figure 6(a) provides an example for the three PMB languages.

An advantage of this method is that labeled training data in the target language is not required. However, the performance of the semantic parser on the target languages is limited by the performance of the semantic parser in the source language; moreover, the crosslingual parser must be interfaced with a machine translation system at runtime, since it only accepts input in the pivot language (such as English).

4.2 One-to-many Method

The one-to-many approach constructs training data for the target languages (see Figure 6(b)) via machine translation. The translated sentences are paired with gold DRSs (from the source language) and collected as training data for the target languages.
An advantage of this method is that the obtained semantic parsers are sensitive to the linguistic aspects of individual languages (and how these correspond to meaning representations). From a modeling perspective, it is also possible to exploit large amounts of unlabeled data in the target language to improve the performance of the semantic parser. Also, notice that the parser is independent of the machine translation engine utilized (sentences need to be translated only once) and the semantic parser developed for the source language. In theory, different parsing models can be used to cater for language-specific properties.

The learning schemes just described are fairly general and compatible with either clause or tree DRS formats, or indeed meaning representation schemes that are not based on DRT. However, the proposed UDRS representation heavily depends on the order of the tokens in the natural language sentences, and as a result is less suited to the many-to-one method; parallel sentences in different languages might have different word orders and consequently different UDRSs (recall that the latter are obtained via aligning non-English tokens to English ones). Many-to-one adopts the rather strong assumption that meaning representations are invariant across languages. If the Italian sentence *sono stati uccisi tutti* is translated as *everyone was killed* in English, a parser trained on English data would output the UDRS in Figure 1(b), while the correct analysis would be Figure 1(c). The resulting UDRS would have to be post-processed in order for the indices to accurately correspond to tokens in the source language (in this case Italian). Many-to-one would thus involve the extra step of modifying English UDRSs back to the UDRSs of the source language. Because UDRS indices are anchored to input sentences via word alignments, we would need to use word alignment models for every new language seen at test time, which renders many-to-one for UDRS representations slightly impractical.

### 4.3 Semantic Parsing Model

Following previous work on semantic parsing (Dong and Lapata 2016; Jia and Liang 2016; Liu, Cohen, and Lapata 2018; van Noord et al. 2018b), we adopt a neural sequence-to-sequence model which assumes that trees or clauses can be linearized into PTB-style bracketed sequences and sequences of symbols, respectively. Specifically, our encoder-decoder model builds on the Transformer architecture (Vaswani et al. 2017), a highly efficient model that has achieved state-of-the-art performance in machine translation (Vaswani et al. 2017), question answering (Yu et al. 2018), summarization (Liu, Titov, and Lapata 2019), and grounded semantic parsing (Wang et al. 2020).

Our DRS parser takes a sequence of tokens, \( \{s_0, s_1, \ldots, s_{n-1}\} \) as input and outputs their linearized DRS \( \{t_0, t_1, \ldots, t_{m-1}\} \), where \( n \) is the number of input tokens, and \( m \) is the number of the symbols in the output DRS.

**Encoder.** Each input token is represented by a vector \( x_k \), which is the sum of word embeddings \( e_k \) and position embeddings \( p_k \):

\[
x_k = e_k + p_k
\]

The input representations, \( x_0, x_1, \ldots, x_{n-1} \), are fed to the Transformer encoder to obtain their hidden representations, \( h_0, h_1, \ldots, h_{n-1} \):

\[
[h_0 : h_{n-1}] = \text{LayerNorm}(\text{Encoder}([x_0 : x_{n-1}]))
\]
where each layer of the ENCODER is:

\[
[x_0 : x_{n-1}] = \text{LAYERNORM}([x_0 : x_{n-1}])
\]
\[
[h_0 : h_{n-1}] = \text{MULTIHEADSEATTN}([\bar{x}_0 : \bar{x}_{n-1}])
\]
\[
[h_0 : h_{n-1}] = \text{FFN}(\bar{h}_0 : \bar{h}_{n-1} + [x_0 : x_{n-1}])
\]

and LAYERNORM is a layer normalization function (Ba, Kiros, and Hinton 2016); MULTIHEADSEATTN is the multi-head self-attention mechanism introduced in Vaswani et al. (2017) that allows the model to jointly attend to information from different representation subspaces (at different) positions; and FFN is a two-layer feed-forward network with ReLU function.

**Decoder.** The decoder uses the contextual representations of the encoder together with the embeddings \(y_{<k} = e_{t_{<k}} + p_{t_{<k}}\) of the previously predicted tokens to output the next token \(t_k\) with the highest probability:

\[
p(t_k | t_{<k}) = \text{SOFTMAX}(g(h_{t_k}))
\]

where \(g\) is a linear function, and

\[
h_{t_k} = \text{LAYERNORM}([\text{DECODER}(y_{<k}, [h_0 : h_{n-1}])])
\]

and each layer of the DECODER consists of five components:

\[
\bar{y}_{<k} = \text{LAYERNORM}(y_{<k})
\]
\[
q_{k-1} = \text{MULTIHEADATTN}(\bar{y}_{k-1}, \bar{y}_{<k})
\]
\[
\bar{q}_{k-1} = \text{LAYERNORM}(q_{k-1} + y_{<k})
\]
\[
\bar{h}_{k-1} = \text{MULTIHEADATTN}(\bar{q}_{k-1}, [x_0 : x_{n-1}])
\]
\[
h_{t_k} = \text{FFN}(\bar{h}_{k-1} + q_{k-1} + y_{k-1})
\]

and MULTIHEADATTN\((y_{k-1}, y_{k-1}^{k-1})\) returns the contextual representation for \(y_{k-1}\) according to its context information \(y_{k-1}^{k-1}\).

### 4.4 Training

Our models are trained with standard back-propagation that requires a large-scale corpus with gold-standard annotations. The PMB does not contain high-volume annotations for model training in languages other than English (although gold-standard data for development and testing are provided). The situation is unfortunately common when developing multilingual semantic resources that demand linguistic expertise and familiarity with the target meaning representation (discourse representation theory in our case). In such cases, model training can be enhanced by recourse to automatically generated annotations that can be obtained with a trained parser. The quality of these data varies depending on the accuracy of the underlying parser and whether any manual correction has taken place on the output. In this section, we introduce an
Iterative training method that makes use of auto-standard annotations of varying quality and is model-independent.

Let $D_{auto} = D_0, D_1, \ldots, D_{m-1}$ denote different versions of training data generated automatically; indices denote the quality of the auto-standard data, $D_0$ has lowest quality, $D_{m-1}$ has highest quality, and $D_i (0 \leq i < m)$ is auto-standard data with quality $i$. The model is first collectively trained on all available data $D_{auto}$ and then at each iteration on subset $D_{auto} \setminus D_1$ which excludes the data with the lowest quality $D_1$. So, the model takes advantage of large-scale data for more reliable parameter estimation but is progressively optimized on better quality data. Algorithm 3 provides a sketch of this training procedure. Iterative training is related to self-training, where model predictions are refined by training on progressively more accurate data. In the case of self-training, the model is trained on its own predictions, while in iterative training, the model uses annotations of increasingly better quality. These can be produced by other models, human experts, or a mixture. Because we know a priori the quality of annotations, we can ensure that later model iterations make use of better data.

5. Experiments

In this section we describe the data set used in our experiments, as well as details concerning the training and evaluation of our models.

5.1 Data

Our experiments were carried out on PMB 2.2.0, which is annotated with DRSs for English (en), German (de), Italian (it), and Dutch (nl). The data set contains gold standard training data for English only, while development and test gold standard data is available in all four languages. The PMB also provides silver and bronze standard training data in all languages. Silver data is only partially checked for correctness, while bronze data is not manually checked in any way. Both types of data were built using Boxer (Bos 2008), an open-domain semantic parser that produces DRS representations by capitalizing on the syntactic analysis provided by a robust CCG parser (Curran, Clark, and Bos 2007).

5.2 Settings

All models share the same hyperparameters. The dimension of the word embeddings is 300, the Transformer encoder and decoder have 6 layers with a hidden size of 300 and 6 heads; the dimension of position-wise feedforward networks is 4,096. The models were trained to minimize a cross-entropy loss objective with an $l_2$ regularization term. We
used Adam (Kingma and Ba 2014) as the learning rate optimizer; the initial learning rate was set to 0.001 with a 0.7 learning rate decay for every 4,000 updates starting after 30,000 updates. The batch size was 2,048 tokens. Our hyperparameter settings follow previous work (van Noord et al. 2018b; Liu, Cohen, and Lapata 2019b).

**Monolingual Setting.** Our monolingual experiments were conducted on the English portion of the PMB. We used the standard training/test splits provided in the data set. We obtained UDRSs using the manual alignments from DRS tokens to sentence tokens included in the PMB release. Our models were trained with the iterative training scheme introduced in Section 4.4 using the PMB bronze-, silver- and gold-standard data (we use $D_0$ to refer to bronze, $D_1$ denotes silver, and $D_2$ gold).

**Crosslingual Setting.** All crosslingual experiments were conducted using Google Translate’s API, a commercial state-of-the-art system supporting more than a hundred languages (Wu et al. 2016). Bronze- and silver-training data for German, Italian, and Dutch are provided with the PBM release. For experiments on other non-English languages, we only used the one-to-many method to create training data for UDRS parsing (Section 4.2). For this, the original alignments (of meaning constructs to input tokens) in English UDRSs need to be modified to correspond to tokens in the translated target sentences. We used the GIZA++ toolkit to obtain forward alignments from source to target and backward alignments from target to source (Koehn, Och, and Marcu 2003). UDRSs for which no available alignment for tokens was found were excluded from training. For iterative training, we consider bronze- and silver-training data (if these are available) of lower quality (i.e., $D_0$ and $D_1$, respectively) compared to data constructed by the one-to-many method (i.e., $D_2$).

### 5.3 Evaluation

We evaluated the output of our semantic parser using COUNTER (van Noord et al. 2018a), a recently proposed metric suited for scoped meaning representations. COUNTER operates over DRSs in clause format and computes precision and recall on matching clauses. DRSs and UDRSs in tree format can easily revert to boxes, which in turn can be rendered as clauses for evaluation purposes.

### 5.4 Models

We compared 11 models on the English portion of the PMB data:

- **SPAR** is a baseline system that outputs the same DRS for each test instance.
- **SIM-SPAR** is a baseline system that outputs the DRS of the most similar sentence in the training set, based on a simple word embedding metric (van Noord et al. 2018b).
- **Boxer** is a system that outputs the DRSs of sentences according to their supertags and CCG derivations (Bos 2015). Each word is assigned a lexical

---

10 https://translate.google.com/toolkit.
11 Two sentences were discarded for German and 55 for Italian.
12 UDRSs replace language-specific symbols with anchors without, however, changing the structure of the meaning representation in any way; UDRSs can be evaluated with COUNTER in the same way as DRSs.
13 In PMB (release 2.2.0) this is the DRS for the sentence *Tom voted for himself.*
semantic representation according to its supertag category, and the representation of a larger span is obtained by combining the representations of two continuous spans (or words). The semantic representation of the entire sentence is composed based on the underlying CCG derivation.

- **Graph** is a graph neural network that generates DRSs according to a directed acyclic graph grammar (Fancellu et al. 2019). Grammar rules are extracted from the training data, and the model learns how to apply these to obtain DRSs.

- **Transition** is a neural transition-based model that incrementally generates DRSs (Evang 2019). It repeatedly selects transition actions within a stack-buffer framework. The stack contains the sequence of generated partial DRSs, while the buffer stores incoming words. Transition actions either consume a word in the buffer or merge two partial DRSs to a new DRS. The system terminates when all words are consumed, and only one item remains on top of the stack.

- **Neural-Boxer** is an LSTM-based neural sequence-to-sequence model that outputs DRSs in clause format (van Noord et al. 2018b).

- **MultiEnc** (van Noord, Toral, and Bos 2019) extends Neural-Boxer with multiple encoders representing grammatical (e.g., parts of speech) and syntactic information (e.g., dependency parses). It also outputs DRSs in clause format.

- **Cls/Tree-Transformer** is the Transformer model from Section 4.3; it outputs DRSs in clause and tree format using the box-to-tree conversion algorithm introduced in Section 3.2. For the sake of completeness, we also re-implement LSTM models trained on clauses and trees (**Cls/Tree-LSTM**).

Our crosslingual experiments were carried out on German, Italian, and Dutch. We built four crosslingual Transformer-based models:

- **Cls/Tree-m2o** uses the many-to-one method to translate non-English sentences into English and parse them using an English Transformer trained on clauses or trees.

- **Cls/Tree-o2m** applies the one-to-many method to construct training data in the target languages for training clause and tree Transformer models.

6. Results

We first present results on DRS parsing in order to assess the performance of our model on its own and whether differences in the format of the DRS representations make any difference. Having settled the question of which format to use, we next report experiments on UDTRS parsing.

6.1 DRS Parsing

Table 1 summarizes our results on DRS parsing. As can be seen, neural models overwhelmingly outperform comparison baselines. Transformers trained on trees and clauses perform better (by 4.0 F₁ and 4.4 F₁, respectively) than LSTMs trained on data in the same format. A Transformer trained on trees performs slightly better (by 0.8 F₁)
Table 1
English DRS parsing (PMB 2.2.0 test set); results for SPAR, SIM-SPAR, BOXER, Transition, Graph, Neural-BOXER, and MultiEnc are taken from respective papers; best result per metric shown in bold.

| DRS          | Prec | Rec  | F₁  |
|--------------|------|------|-----|
| SPAR         | 44.4 | 37.8 | 40.8|
| SIM-SPAR     | 57.0 | 58.4 | 57.7|
| BOXER        | 72.1 | 72.3 | 72.2|
| Transition   | 75.6 | 74.6 | 75.1|
| Graph        | –    | –    | 76.4|
| Neural-BOXER | 85.0 | 81.4 | 83.2|
| MultiEnc     | 87.6 | 86.3 | 87.0|
| Cls-LSTM     | 82.5 | 83.3 | 83.9|
| Tree-LSTM    | 84.3 | 84.7 | 84.3|
| Cls-Transformer | 88.1 | 87.7 | 87.9|
| Tree-Transformer | **88.6** | **88.9** | **88.7**|
| w/o bronze   | 86.1 | 86.0 | 86.0 (−2.6) |
| w/o bronze & silver | 79.9 | 84.4 | 82.1 (−6.5) |

Table 2
DRS parsing results on German, Italian, and Dutch (PMB test set); best result per metric shown in bold.

| DRS          | de Pre | de Rec | de F₁ | it Pre | it Rec | it F₁ | nl Pre | nl Rec | nl F₁ | all avg F₁ |
|--------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------------|
| Cls          | 72.1   | 72.6   | 72.3   | 74.2   | 74.4   | 74.3   | 64.3   | 65.2   | 64.8   | 70.5        |
| Cls-m2o      | 84.5   | 83.6   | 84.0   | 85.1   | 85.4   | 85.2   | **84.4** | 84.0   | 84.2   | 84.5        |
| Cls-o2m      | 81.1   | 80.2   | 80.6   | 81.0   | 80.6   | 80.8   | 76.0   | 76.4   | 76.2   | 79.2        |
| Tree         | 72.6   | 72.9   | 72.8   | 75.4   | 75.9   | 75.7   | 65.8   | 66.2   | 66.0   | 71.5        |
| Tree-m2o     | **84.6** | **83.9** | **84.2** | **86.0** | **85.6** | **85.9** | 84.3   | **84.4** | **84.4** | **84.9**   |
| Tree-o2m     | 82.7   | 81.1   | 81.9   | 81.4   | 81.0   | 81.2   | 78.4   | 77.5   | 78.0   | 80.3        |

than the same model trained on clauses and is overall best among models using DRS-based representations.

Our results on the DRS crosslingual setting are summarized in Table 2. Many-to-one parsers outperform one-to-many ones, although, the difference is starker for Clauses than for Trees (5.3 vs. 4.6 F₁ points). With the many-to-one strategy, Tree-based representations are overall slightly better on DRS parsing.

6.2 UDRS Parsing

Table 3 summarizes our results on the UDRS monolingual setting. We observe that a Transformer model trained on tree-based representations is better (by 1.3 F₁) compared to the same model trained on clauses. This suggests that UDRS parsing indeed benefits more from tree representations.

Our results on the UDRS crosslingual setting are shown in Table 4. Aside from UDRS parsers trained with the one-to-many strategy (Cls-o2m and Tree-o2m), we also report the performance of monolingual Transformers (clause and tree formats) trained on the
silver and bronze standard data sets provided in PMB. All models were evaluated on the gold standard PMB test data. We only report the performance of one-to-many UDRS parsers because of the post-processing issue discussed in Section 4.2. Compared to models trained on silver and bronze data, the one-to-many strategy improves performance for both clause and tree formats (by 5.7 $F_1$ and 5.7 $F_1$, on average). Overall, the crosslingual experiments show that we can indeed bootstrap fairly accurate semantic parsers across languages, without any manual annotations on the target language.

6.3 Analysis

In this section, we analyze in more detail the output of our parsers in order to determine which components of the semantic representation are modeled best. We also examine the effect of the iterative training on parsing accuracy.

Fine-grained Evaluation. COUNTER (van Noord et al. 2018a) provides detailed breakdown scores for DRS operators (e.g., negation), Roles (e.g., Agent), Concepts (i.e., predicates), and Synsets (e.g., “n.01”). Table 5 compares the output of our English semantic parsers. For DRS representations, Tree models perform better than Clauses on most components except for adjective and adverb synsets. All models are better at predicting noun senses compared to verbs, adjectives, and adverbs. The clause format is better when it comes to predicting the senses of adverbs. Nevertheless, all models perform poorly on adverbs, which are relatively rare in the PMB. In our crosslingual experiments, we observe that Tree models slightly outperform Clauses across languages. For the sake of brevity, Table 6 only reports a break-down of the results for Trees. Interestingly, we see that the bootstrapping strategies proposed here are a better alternative to just training semantic parsers on PMB’s silver and bronze data (see Tree column in Table 6). Moreover, the success of the bootstrapping strategy seems to be consistent among languages, with many-to-one being overwhelmingly better than one-to-many, even though many-to-one fails to predict the adverb synset. Overall, the prediction
of synsets is a harder task and indeed performance improves when the model only focuses on operators and semantic roles (compare Tree and Tree-o2m columns in DRS and UDRS). Without incorporating sense distinctions, UDRSs are relatively easier to predict. The vocabulary of the meaning constructs is smaller, and it only includes global symbols like semantic role names and DRS operators that are shared across languages, thus making the parsing task simpler.

Table 5
Fine-grained evaluation ($F_1$) on the English PMB test set by Cls/Tree-Transformer and Cls/Tree-LSTM. Best result per meaning construct shown in bold.

| DRS     | LSTM Cls | Transformer Cls | Transformer Tree |
|---------|----------|-----------------|-----------------|
| DRS operator | 90.80    | 91.02           | 94.05           | 95.07 |
| Role    | 81.68    | 83.23           | 87.87           | 88.48 |
| Concept | 81.41    | 82.91           | 85.97           | 86.87 |
| Syns-Noun | 87.41    | 87.79           | 91.57           | 91.86 |
| Syns-Verb | 65.27    | 66.04           | 71.58           | 74.20 |
| Syns-Adjective | 71.74    | 73.12           | 74.73           | 76.93 |
| Syns-Adverb | **66.48** | 60.00           | 60.00           | 54.55 |

Table 6
Fine-grained evaluation ($F_1$%) on German, Italian, and Dutch (test set); best result per synset shown in bold.

| DRS     | Tree | DRS Tree-m2o Tree-o2m | UDRS Tree | UDRS Tree-o2m |
|---------|------|------------------------|-----------|---------------|
| de      |      |                        |           |               |
| DRS operator | 84.89  | 91.75                  | 90.94     | 86.26         | 91.87 |
| Role    | 72.47 | 84.37                  | 82.62     | 73.93         | 83.83 |
| Concept | 69.08 | 81.69                  | 78.21     | —             | —    |
| Syns-Noun | 79.67  | 88.71                  | 86.19     | —             | —    |
| Syns-Verb | 41.38  | 63.69                  | 58.82     | —             | —    |
| Syns-Adjective | 50.32  | 69.54                  | 59.64     | —             | —    |
| Syns-Adverb | 14.29  | 0.00                   | 25.00     | —             | —    |
| it      |      |                        |           |               |
| DRS operator | 87.28  | 91.20                  | 89.54     | 87.82         | 91.20 |
| Role    | 76.18 | 88.60                  | 82.02     | 79.41         | 84.82 |
| Concept | 71.35 | 81.63                  | 77.60     | —             | —    |
| Syns-Noun | 81.71  | 87.69                  | 86.70     | —             | —    |
| Syns-Verb | 44.48  | 66.13                  | 53.91     | —             | —    |
| Syns-Adjective | 52.35  | 70.16                  | 62.11     | —             | —    |
| Syns-Adverb | 0.00   | 0.00                   | 0.00      | —             | —    |
| nl      |      |                        |           |               |
| DRS operator | 79.90  | 92.98                  | 89.05     | 82.69         | 93.58 |
| Role    | 64.08 | 83.54                  | 77.72     | 65.57         | 77.75 |
| Concept | 63.34 | 82.43                  | 74.63     | —             | —    |
| Syns-Noun | 74.66  | 88.59                  | 82.68     | —             | —    |
| Syns-Verb | 32.33  | 67.26                  | 54.81     | —             | —    |
| Syns-Adjective | 41.38  | 64.86                  | 50.00     | —             | —    |
| Syns-Adverb | 0.00   | 0.00                   | **50.00** | —             | —    |
Iterative Training. Figure 7 shows how prediction accuracy varies with the quality of the training data. The black dotted curve shows the accuracy of a model trained on the combination of bronze-, silver-, and gold-standard data ($D_0 + D_1 + D_2$), the red dashed curve shows the accuracy of a model trained on the silver- and gold-standard data ($D_1 + D_2$), and the blue curve shows the accuracy of a model trained only on gold-standard data ($D_2$). As can be seen, the use of more data leads to a big performance boost (compare the model trained on $D_0 + D_1 + D_2$ against just $D_2$). We also show what happens after the model converges on $D_0 + D_1 + D_2$: Further iterations on $D_1 + D_2$ slightly improve performance, while a big boost is gained from continually training on gold-standard data ($D_2$). The relatively small gold-standard data is of high quality but has low coverage; parameter optimization on the combination of bronze-, silver-, and gold-standard data enhances model coverage, while fine-grained optimization on gold-standard data increases its accuracy.

Table 7 presents our results, which are clustered according to language family (we report only the performance of Tree UDRS models for the sake of brevity). All models for all languages used the same settings (see Section 5.2). As can be seen, the majority of languages we experimented with are Indo-European. In this family, the highest $F_1$ is 81.92 for Danish and Yiddish. In the Austronesian family, our parser performs best for Indonesian ($F_1$ is 78.20). In the Afro-Asiatic family, Vietnamese achieves the highest $F_1$ of 79.37. In the Niger-Congo family, the highest $F_1$ is 77.21 for Swahili. In the Turkic family, our parser performs best for Turkish ($F_1$ is 74.39). In the Dravidian and Uralic families, Kannada and Estonian obtain the highest $F_1$, respectively. The worst parsing performance is obtained for Khmer ($F_1$ of 61.53), while for the majority of languages our parser is in the 71%–82% ballpark. Perhaps unsurprisingly, better parsing performance correlates with a higher quality of machine translation and statistical word alignments.\footnote{We will make the UDRS data sets for the 96 languages publicly available as a means of benchmarking semantic parsing performance and also in the hope that some of these might be manually corrected.}
## Table 7
UDRS parsing results via Tree-o2m for 96 languages individually and on average; languages are grouped per language family and sorted alphabetically; best results in each family are shown in bold.

| Language F1 | Language F1 | Language F1 | Language F1 |
|-------------|-------------|-------------|-------------|
| Indo-European | Romanian | 79.65 | Chichewa | 73.74 |
| Afrikaans | 80.29 | Russian | 77.87 | Igbo | 73.39 |
| Albanian | 78.41 | Scots Gaelic | 78.08 | Shona | 74.24 |
| Armenian | 76.93 | Serbian | 76.53 | Sesotho | 76.10 |
| Belarusian | 78.24 | Sindhi | 76.46 | Swahili | 77.21 |
| Bengali | 76.59 | Sinhala | 75.99 | Xhosa | 72.44 |
| Bosnian | 76.66 | Slovak | 76.44 | Yoruba | 74.13 |
| Bulgarian | 79.34 | Slovenian | 75.42 | Zulu | 72.34 |
| Catalan | 79.59 | Spanish | 80.79 | Turkic | |
| Corsican | 77.66 | Swedish | 79.18 | Azerbaijani | 72.66 |
| Croatian | 76.94 | Tajik | 74.70 | Kazakh | 72.53 |
| Czech | 77.20 | Ukrainian | 78.18 | Kyrgyz | 73.40 |
| Danish | **81.92** | Urdu | 77.27 | Turkish | **74.39** |
| French | 79.96 | Welsh | 81.38 | Uzbek | 72.84 |
| Frisian | 80.26 | Yiddish | **81.92** | Dravidian | |
| Galician | 78.88 | Austronesian | | |
| Greek | 78.35 | Cebuano | 75.06 | Malayalam | 75.86 |
| Gujarati | 76.50 | Filipino | 76.58 | Tamil | 74.14 |
| Hindi | 79.31 | Hawaiian | 70.45 | Telugu | 75.32 |
| Icelandic | 80.03 | Indonesian | **78.20** | Uralic | |
| Irish | 79.23 | Javanese | 75.07 | Estonian | **78.04** |
| Kurdish | 73.37 | Malagasy | 73.52 | Finnish | 77.51 |
| Latin | 71.01 | Malay | 77.43 | Hungarian | 73.93 |
| Latvian | 79.34 | Maori | 73.81 | Others | |
| Lithuanian | 77.89 | Samoan | 74.29 | Basque | 74.18 |
| Luxembourgish | 77.21 | | | Chinese | 75.64 |
| Macedonian | 78.57 | Amharic | 72.04 | Esperanto | 79.70 |
| Marathi | 76.54 | Arabic | 74.84 | Georgian | 73.07 |
| Nepali | 76.99 | Hausa | 74.46 | Haitian Creole | 80.74 |
| Norwegian | 80.55 | Hebrew | 78.83 | Hmong | 75.52 |
| Pashto | 76.05 | Maltese | 78.89 | Japanese | 76.03 |
| Persian | 76.73 | Somali | 75.06 | Korean | 74.14 |
| Polish | 77.48 | Sundanese | 75.51 | Khmer | 61.53 |
| Portuguese | 78.97 | Vietnamese | **79.37** | Mongolian | 75.47 |
| Punjabi | 78.18 | Niger-Congo | | Average | 76.47 |

### 6.4 Scalability Experiments

We further assessed whether the crosslingual approach advocated in this article scales to multiple languages. We thus obtained UDRS parsers for all languages supported by Google Translate in addition to German, Italian, and Dutch (99 in total). Specifically, we applied the one-to-many bootstrapping method on the English gold-standard PMB annotations to obtain semantic parsers for 96 additional languages.

Unfortunately, there are no gold-standard annotations to evaluate the performance of these parsers and the effort of creating these for 99 languages would be prohibitive. Rather than focusing on a few languages for which annotations could be procured, we adopt a more approximate but larger-scale evaluation methodology. Damonte and Cohen (2018) estimate the accuracy of crosslingual AMR parsers following a full-cycle...
Table 8
Comparison between gold-standard UDRSs and constructed UDRSs by our methods in German, Italian, and Dutch using BLEU and COUNTER; standard deviations are shown in parentheses.

| language | BLEU  | F1     |
|----------|-------|--------|
| de       | 65.03 | 94.21  |
| it       | 61.22 | 88.41  |
| nl       | 69.12 | 94.06  |
| avg      | 65.12 (±3.9) | 92.23 (±1.98) |

evaluation scheme. The idea is to invert the learning process and bootstrap an English parser from the induced crosslingual parser via back-translation. The resulting English parser is then evaluated against the (English) gold-standard under the assumption that the English parser can be used as a proxy to the score of the crosslingual parser. In our case, we applied the one-to-many method to project non-English annotations back to English, and evaluated the parsers on the PMB gold-standard English test set.

We further investigated the quality of the constructed data sets by extrapolating from experiments on German, Italian, and Dutch for which a gold-standard test set is available. Specifically, using the one-to-many method, we constructed silver-standard test sets and compared these with their gold-standard counterparts provided in the PMB. We first assessed translation quality by measuring the BLEU score (Papineni et al. 2002). We also used COUNTER to evaluate the degree to which silver-standard UDRSs deviate from gold-standard ones. As shown in Table 8, the average BLEU (across three languages) is 65.12, while the average F1 given by COUNTER is 92.23. These results indicate that the translation quality is rather good, at least for these three languages, and the PMB sentences. COUNTER scores further show that annotations are transferred relatively accurately, and that silver-standard data is not terribly noisy, where approximately 8% of the annotations deviate from the gold standard.

In Table 4, we show the crosslingual UDRS parsing results on the gold-standard test set in German, Italian, and Dutch, which are close to English. In order to investigate the crosslingual UDRS parsing in languages that are far from English, we performed UDRS parsing experiments in Japanese and Chinese, two languages that are typologically distinct from English, in the way concepts are expressed and combined by grammar to generate meaning. For each language, we manually constructed gold standard UDRS annotations for 50 sentences. Table 9 shows the accuracy of the Chinese and Japanese parsers we obtained following the one-to-many training approach. These two languages have relatively lower scores compared to German, Italian, and Dutch in

Table 9
UDRS parsing results on gold-standard Japanese (ja) and Chinese (zh) test sets.

| language | Prec | Rec  | F1  |
|----------|------|------|-----|
| ja       | 62.0 | 65.6 | 63.7|
| zh       | 57.5 | 61.7 | 59.5|
6.5 Translation Divergence

Our crosslingual methods depend on machine translation and alignments, which can be affected by translation divergences. In this section, we discuss how translation divergences might influence our methods. We focus on the seven types of divergence highlighted in Dorr (1994) (i.e., promotional, demotional, structural, conflational, categorial, lexical, and thematic divergences) and discuss whether the proposed UDRS representation can handle them.

Promotional Divergence. Promotional divergence describes the phenomenon where the logical modifier of a main verb can be changed. For example, consider the English sentence *John usually goes home* and its Spanish translation *Juan suele ir a casa* (John tends to go home), where the modifier (*usually*) is realized as an adverbial phrase in English but as a verb (*suele*) in Spanish. As shown in Figure 8, to obtain the Spanish UDRS, the English words are replaced with aligned words. However, the adverbial *usually* is replaced with the verb *suele*, which together with the thematic relation, Manner qualifies how the action *ir* is carried out. The divergence will raise a Category Inconsistency in the UDRS, which means that the category (or part of speech) of the translation is not consistent with that of the source language.

Demotional Divergence. In demotional divergence, a logical head into an internal argument position can be changed. For example, consider the English sentence *I like eating* and its German translation *Ich esse gern* (I eat likingly). Here, the head (*like*) is realized as a verb in English but as an adverbial satellite in German. Figure 9 shows the alignments between the two sentences and their UDRSs. Similar to promotional divergence, this also leads to Category Inconsistency in the German UDRS, since *gern* should be an adverb, not a verb.

Structural Divergence. Structural divergences are different in that syntactic structure is changed and, as a result, syntactic relations may also become different. For example, for the English sentence *John entered the house*, the Spanish translation is *Juan entró en la casa* (John entered in the house), where the noun phrase *the house* in English...
becomes a prepositional phrase (en la casa) in Spanish. This divergence does not affect the correctness of the Spanish UDRS (see Figure 10). In general, UDRSs display coarse-grained thematic relations (in this case Destination) abstracting away from how these are realized (e.g., as a noun or prepositional phrase).

Conflational and Lexical Divergence. We discuss both types of divergence together. Words or phrases in the source language can be paraphrased using various descriptions in the target language. In conflational divergence, for example, the English sentence I stabbed John is translated into Spanish as Yo le di puñaladas a Juan (I gave knife-wounds to John) using the paraphrase di puñaladas (gave knife-wounds to) to describe the English word stabbed. Analogously, the word broke in the English sentence He broke into the room is aligned to forzó (force) in Spanish (Juan forzó la entrada al cuarto). The two words do not have exactly the same meaning, and yet they convey the breaking event in their respective language. We expect these divergences to be resolved with many-to-many word alignments, and yield correct UDRSs as long as the translations are accurate (see figures 11 and 12).

Categorical Divergence. The lexical categories (or parts of speech) might change due to the translation from the source to the target language. For example, the English sentence I am hungry is translated to German as Ich habe Hunger (I have hunger), where the adjective hungry in English is translated with the noun Hunger in German. As shown in Figure 13, categorical divergences will often lead to incorrect UDRSs in the target language.
Thematic Divergence. Thematic relations are governed by the main verbs of sentences, and it is possible for thematic roles to change in translation. For example, the English sentence *I like Mary* is translated in Spanish as *María me gusta a mí* (Mary pleases me), where the English subject (*I*) is changed to an object (*me*) in Spanish. As shown in Figure 14, although word alignments can capture the semantic correspondence between words in the two sentences, the Spanish UDRS ends up with the wrong thematic relations showing **Thematic Inconsistency**.

In sum, category and thematic inconsistencies will represent the majority of errors in the construction of UDRSs in another language from English (via translation and alignments). Category inconsistencies can be addressed with the help of language-specific knowledge bases by learning a function \( f(s, c) = (s', c') \), where \( s \) and \( c \) are a translated word and an original category, respectively, and \( s' \) and \( c' \) are the corrected word and category. Addressing thematic inconsistencies is difficult, as it requires comparing verbs between languages in order to decide whether thematic relations must be changed.

In order to estimate how many UDRSs are incorrect and quantify what types of errors they display, we randomly sampled 50 German, Italian, and Dutch UDRSs. As
Figure 14
Example of thematic divergence. (a) UDRS of English sentence *I like Mary*; (b) word alignments between two sentences; (c) incorrect UDRS of Spanish translation *María me gusta a mí*, constructed via alignments.

Table 10
Number of correct and incorrect UDRSs on a sample of 50 sentences for German, Italian, Dutch, and Chinese.

|            | de | it | nl | zh |
|------------|----|----|----|----|
| correct    | 44 | 46 | 45 | 32 |
| translation error | 1  | 0  | 0  | 4  |
| translation divergence error | 1  | 0  | 0  | 2  |
| alignment error    | 4  | 4  | 5  | 12 |

shown in Table 10, we found that alignment errors are the main cause of incorrect UDRSs. Translation divergences do not occur very frequently, even though we used machine translation systems. We also sampled and analyzed 50 UDRSs in Chinese, a language typologically very different from English. Again, the number of translation divergences is small, which may be due to the fact that sentences in PMB are short and thus relatively simple to translate.

7. Related Work

Recent years have seen growing interest in the development of DRT parsing models. Early seminal work (Bos 2008) created Boxer, an open-domain DRS semantic parser, which has been instrumental in enabling the development of the Groningen Meaning Bank (Bos et al. 2017) and the Parallel Meaning Bank (Abzianidze et al. 2017). Le and Zuidema (2012) were the first to train a data-driven DRT parser using a graph-based representation leaving anaphora and presupposition aside. The availability of annotated corpora has further allowed the exploration of neural models. Liu, Cohen, and Lapata (2018) conceptualize DRT parsing as a tree structure prediction problem that they model with a series of encoder-decoder architectures (see also the extensions proposed in Liu, Cohen, and Lapata 2019a). van Noord et al. (2018b) adapt sequence-to-sequence models with LSTM units to parse DRSs in clause format, also following a graph-based representation. Fancellu et al. (2019) represent DRSs as direct acyclic graphs and design a DRT parser with an encoder-decoder architecture that takes as input a sentence and outputs a graph using a graph-to-string rewriting system. In addition, their parser exploits various linguistically motivated features based on part-of-speech embeddings, lemmatization, dependency labels, and semantic tags. Our crosslingual strategies can be applied to their work as well. So our own work unifies the proposals of Liu, Cohen, and Lapata (2018) and van Noord et al. (2018b) under a general
modeling framework based on the Transformer architecture, allowing for comparisons between the two, as well as for the development of crosslingual parsers. In addition, we introduce UDRT, a variant of the DRT formalism that we argue facilitates both monolingual and crosslingual learning.

The idea of leveraging existing English annotations to overcome the resource shortage in other languages by exploiting translational equivalences is by no means new. A variety of methods have been proposed in the literature under the general framework of annotation projection (Yarowsky and Ngai 2001; Hwa et al. 2005; Padó and Lapata 2005, 2009; Akbik et al. 2015; Evang and Bos 2016; Damonte and Cohen 2018; Zhang et al. 2018; Conneau et al. 2018), which focuses on projecting existing annotations on source-language text to the target language. While other work focuses on model transfer where model parameters are shared across languages (Cohen, Das, and Smith 2011; McDonald, Petrov, and Hall 2011; Søgaard 2011; Wang and Manning 2014), our crosslingual parsers rely on translation systems following two ways commonly adopted in the literature (Conneau et al. 2018; Yang et al. 2019; Huang et al. 2019): translating the training data into each target language (one-to-many) to provide data to train a semantic parser per language, and using a translation system at test time to translate the input sentences to the training language (many-to-one). Our experiments show that the combination of one-to-many and UDRT representations allows us to speed-up meaning bank creation and the annotation process.

8. Conclusion

In this article, we introduced Universal Discourse Representation Structures as a variant of canonical DRSs; UDRTs link elements of the DRS structure to tokens in the input sentence and are ideally suited to crosslingual learning; they omit details pertaining to the lexical makeup of sentences and as a result disentangle the problems of translating tokens and semantic parsing. We further proposed a general framework for crosslingual learning based on neural networks and state-of-the-art machine translation and demonstrated it can incorporate various DRT formats (e.g., trees vs. clauses) and is scalable. In the future, we would like to improve the annotation quality of automatically created meaning banks by utilizing human-in-the-loop methods (Zanzotto 2019) that leverage machine learning algorithms (e.g., for identifying problematic annotations or automatically correcting obvious mistakes) or crowdsourcing platforms.

Acknowledgments

We thank the anonymous reviewers for their feedback. We thank Alex Lascarides for her comments. We gratefully acknowledge the support of the European Research Council (Lapata, Liu; award number 681760), the EU H2020 project SUMMA (Cohen, Liu; grant agreement 688139) and Bloomberg (Cohen, Liu). This work was partly funded by the NWO-VICI grant “Lost in Translation – Found in Meaning” (288-89-003).

References

Abend, Omri and Ari Rappoport. 2013. Universal conceptual cognitive annotation (UCCA). In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 228–238, Sofia.

Abzianidze, Lasha, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 242–247, Valencia. https://doi.org/10.18653/v1/E17-2039
transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, MN.

Dong, Li and Mirella Lapata. 2016. Language to logical form with neural attention. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 33–43, Berlin. https://doi.org/10.18653/v1/P16-1004

Dong, Zhendong, Qiang Dong, and Changling Hao. 2006. HowNet and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd., Singapore.

Dorr, Bonnie J. 1994. Machine translation divergences: A formal description and proposed solution. Computational Linguistics, 20(4):597–633.

Evang, Kilian. 2019. Transition-based DRS parsing using stack-LSTMs. In Proceedings of the IWCS Shared Task on Semantic Parsing, pages 16–23, Gothenburg. https://doi.org/10.18653/v1/W19-1202.

Evang, Kilian and Johan Bos. 2016. Crosslingual learning of an open-domain semantic parser. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 579–588, Osaka.

Fancellu, Federico, Sorcha Gilroy, Adam Lopez, and Mirella Lapata. 2019. Semantic graph parsing with recurrent neural network DAG grammars. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2769–2778, Hong Kong.

Fancellu, Federico, Ákos Kádár, Ran Zhang, and Afsaneh Fazly. 2020. Accurate polyglot semantic parsing with DAG grammars. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 3567–3580, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.320

Fellbaum, Christiane, editor. 1998. WordNet: An Electronic Database, MIT Press, Cambridge, MA.

Flickinger, Dan, Yi Zhang, and Valia Kordoni. 2012. DeepBank. A dynamically annotated treebank of the Wall Street Journal. In Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories, pages 85–96, Lisbon.

Gangemi, Aldo, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongioví. 2017. Semantic web machine reading with FRED. Semantic Web, 8(6):873–893. https://doi.org/10.3233/SW-160240

Hamp, Birgit and Helmut Feldweg. 1997. GermaNet—a lexical-semantic net for German. In Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pages 9–15. https://www.aclweb.org/anthology/W97-0802

Hardalov, Momchil, Ivan Koychev, and Preslav Nakov. 2020. Enriched pre-trained transformers for joint slot filling and intent detection. arXiv preprint arXiv:2004.14848.

Hershcovitch, Daniel, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1127–1138, Vancouver. https://doi.org/10.18653/v1/P17-1104

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138, Minneapolis, MN. https://www.aclweb.org/anthology/W19-1419

Hockenmaier, Julia and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. Computational Linguistics, 33(3):355–396. https://doi.org/10.1162/coli.2007.33.3.355

Huang, Chu Ren, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese Wordnet: Design, implementation, and application of an infrastructure for crosslingual knowledge processing. Journal of Chinese Information Processing, 24(2):14–23.

Huang, Haoyang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple crosslingual tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2485–2494, Hong Kong. https://doi.org/10.18653/v1/D19-1252
Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325. https://doi.org/10.1017/S1351324905003840

Jia, Robin and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin. https://doi.org/10.18653/v1/P16-1002

Johnson, Mark and Ewan Klein. 1986. Discourse, anaphora and parsing. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, pages 669–675, Bonn. https://doi.org/10.3115/991365.991559

Kamp, Hans. 1981, A theory of truth and semantic representation. In J. A. G. Groenendijk, T. M. V. Janssen, and M. B. J. Stokhof, editors, *Formal Methods in the Study of Language*, 1, Mathematisch Centrum, Amsterdam, pages 277–322.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Kluwer, Dordrecht.

Kim, Yunsu, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong. https://doi.org/10.18653/v1/D19-6503

Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, pages 1–13, Banff.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* Volume 1, pages 48–54, Edmonton. https://doi.org/10.3115/1073445.1073462

Kulmizev, Artur, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. https://doi.org/10.18653/v1/2020.acl-main.375

Kuncoro, Adhiguna, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics*, 8:776–794. https://doi.org/10.1162/tacl_a_00345

Le, Phong and Willem Zuidema. 2012. Learning compositional semantics for open domain semantic parsing. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1535–1552, Mumbai.

Liu, Jiangming, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne. https://doi.org/10.18653/v1/P18-1040

Liu, Jiangming, Shay B. Cohen, and Mirella Lapata. 2019a. Discourse representation parsing for sentences and documents. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6248–6262, Florence. https://doi.org/10.18653/v1/P19-1629

Liu, Jiangming, Shay B. Cohen, and Mirella Lapata. 2019b. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg. https://doi.org/10.18653/v1/W19-1203

Liu, Yang, Ivan Titov, and Mirella Lapata. 2019. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, MN. https://doi.org/10.18653/v1/N19-1173

May, Jonathan. 2016. SemEval-2016 task 8: Meaning representation parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1063–1073, San Diego, CA. https://doi.org/10.18653/v1/S16-1166

McDonald, Ryan, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh.
Narayan, Shashi and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 435–445, Baltimore, MD. https://doi.org/10.3115/v1/P14-1041

Navigli, Roberto and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 216–225, Uppsala.

Oepen, Stephan, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershovich, Bin Li, Tim O’Gorman, Nianwen Xue, and Daniel Zeman. 2020. MRP 2020: The second shared task on cross-framework and crosslingual meaning representation parsing. In Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, pages 1–22. https://doi.org/10.18653/v1/2020.conll-shared.1

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The lingo redwoods treebank motivation and preliminary applications. In Proceedings of the 19th International Conference on Computational Linguistics—Volume 2, pages 1–5, Taiwan. https://doi.org/10.3115/1071884.1071909

Padó, Sebastian and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 859–866, Vancouver. https://doi.org/10.3115/1220575.1220683

Padó, Sebastian and Mirella Lapata. 2009. Crosslinguot annotation projection for semantic roles. Journal of Artificial Intelligence Research, 36:307–340. https://doi.org/10.1613/jair.2863

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of Association for Computational Linguistics, pages 311–318, Philadelphia, PA.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.

Søgaard, Anders. 2011. Data point selection for cross-language adaptation of dependency parsers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 682–686, Portland, OR.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. Thirty-First AAAI Conference on Artificial Intelligence, pages 4444–4451, San Francisco, CA.

Van der Sandt, Rob A. 1992. Presupposition projection as anaphora resolution. Journal of Semantics, 9(4):333–377. https://doi.org/10.1093/jos/9.4.333

van Noord, Rik, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018a. Evaluating scoped meaning representations. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 1685–1693, Miyazaki.

van Noord, Rik, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018b. Exploring neural methods for parsing discourse representation structures. Transactions of the Association for Computational Linguistics, 6:619–633. https://doi.org/10.1162/tacl_a_00241

van Noord, Rik, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In Proceedings of the 13th International Conference on Computational Semantics–Short Papers, pages 24–31, Gothenburg. https://doi.org/10.18653/v1/W19-0504

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Ilya Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008.

Venuhuizen, Noortje J., Johan Bos, and Harm Brouwer. 2013. Parsimonious semantic representations with projection pointers. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pages 252–263, Potsdam.

Venuhuizen, Noortje J., Johan Bos, Petra Hendriks, and Harm Brouwer. 2018. Discourse semantics with information structure. Journal of Semantics, 35(1):127–169. https://doi.org/10.1093/jos/ffx017

Wada, Hajime and Nicholas Asher. 1986. BUILDRS: An implementation of DR
theory and LFG. CoNLL 1986 Volume 1: The 11th International Conference on Computational Linguistics, pages 540–545, Bonn. https://doi.org/10.3115/991365.991523

Wang, Bailin, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7567–7578, Online. https://doi.org/10.18653/v1/2020.acl-main.677

Wang, Mengqiu and Christopher D. Manning. 2014. Crosslingual projected expectation regularization for weakly supervised learning. Transactions of the Association for Computational Linguistics, 2:55–66. https://doi.org/10.1162/tacl_a_00165

White, Aaron Steven, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1713–1723, Austin, TX. https://doi.org/10.18653/v1/D16-1177

Wu, Shanchan and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pages 2361–2364, Beijing. https://doi.org/10.1145/3357384.3358119

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Yang, Yinfei, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-x: A crosslingual adversarial data set for paraphrase identification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3687–3692, Hong Kong. https://doi.org/10.18653/v1/D19-1382

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual POS taggers and NP brackets via robust projection across aligned corpora. In Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA. https://doi.org/10.3115/1073336.1073362

Yu, Adams Wei, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QA-Net: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541.

Zanzotto, Fabio Massimo. 2019. Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research, 64:243–252. https://doi.org/10.1613/jair.1.11345

Zhang, Sheng, Xutai Ma, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2018. Crosslingual decompositional semantic parsing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1664–1675, Brussels. https://doi.org/10.18653/v1/D18-1194