Universal Discourse Representation Structure Parsing

Jiangming Liu
University of Edinburgh
jiangming.liu@ed.ac.uk

Shay B. Cohen
University of Edinburgh
scohen@inf.ed.ac.uk

Mirella Lapata
University of Edinburgh
mlap@inf.ed.ac.uk

Johan Bos
University of Groningen
johan.bos@rug.nl

We consider the task of cross-lingual 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 𝕌*niversal Discourse Representation Theory (*𝕌*DRT), 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 employ 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 (UCCA; Abend and Rappoport 2013), and Universal Decompositional Semantics (UDS; 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,

Submission received: 8 August 2020; Revised version received: 20 February 2021; Accepted for publication: 7 March 2021

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,a), the Groningen Meaning Bank (GMB; 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 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), while the bottom layer contains conditions over discourse referents. Each referent or condition belongs to a unique box label, showing the referent or the condition which is interpreted in that box (e.g., $b_2 : x_1$ and b_2 : 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.

Despite efforts to create cross-lingual 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, i.e., familiarity with the semantic formalism and language at hand. Since 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). **Cross-lingual** 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 paper, we also aim to develop a cross-lingual 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-

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

English: everyone was killed .
Italian: sono stati uccisi tutti .
they were killed all .

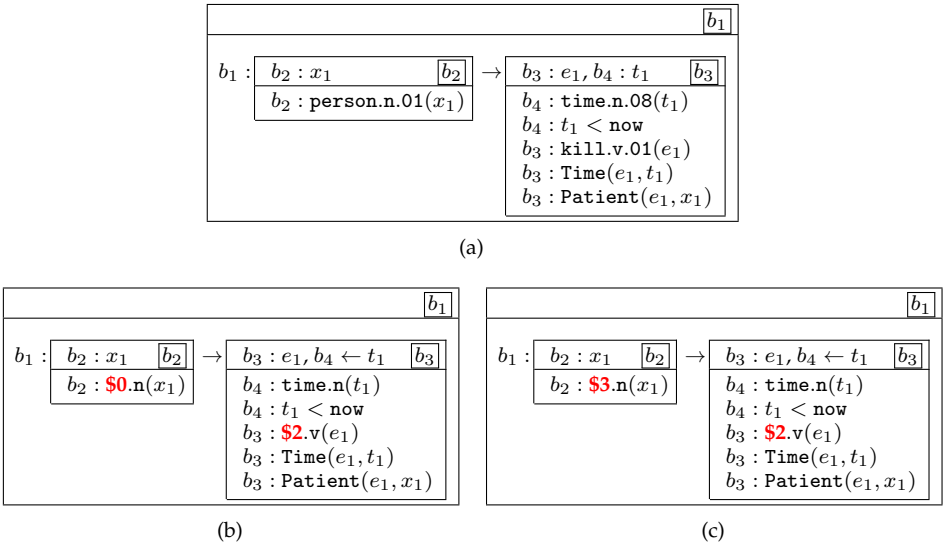

Figure 1: (a) DRS for English sentence *everyone was killed* and its Italian translation *sono stati uccisi tutti*, taken from the Parallel Meaning Bank; (b) 𝕌DRS for English and (c) Italian sentence. The two 𝕌DRSs differ in how they are anchored to the English and Italian sentences (via alignment). 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*.

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 which 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 cross-lingual 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 cross-lingual 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.[2] In this paper, we propose (1) $\mathbb{U}$DRSs to explicitly anchor DRSs to lexical tokens, which we argue is advantageous for both monolingual and cross-lingual parsing; (2) a box-to-tree conversion algorithm which is lossless and reversible; and (3) a general cross-lingual 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 $\mathbb{U}$DRSs based on them. We also highlight the advantages of $\mathbb{U}$DRSs when they are used for multilingual semantic representations.

2.1 Traditional DRSs

Discourse Representation Structures (DRSs), the basic meaning-carrying units of Discourse Representation Theory (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), while the bottom layer contains conditions over the variables. For example, $\texttt{person.n.01}(x_1)$ means that variable x_1 is applied to predicate person.n.01, which in fact is a WordNet synset.

The PMB adopts an extension of DRT which 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 $b_1 : \texttt{time.n.08}(t_1)$ indicate that variable x_1 and condition $\texttt{time.n.08}(t_1)$ should be interpreted within boxes b_2 and b_1, respectively. The boxes are constructed *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 *was killed* gives rise to temporal variable t_1 and condition $\texttt{time.n.08}(t_1)$ and $t_1 < \texttt{now}$; these temporal markers are located in box b_3 together with the predicate "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.

2.2 $\mathbb{U}$niversal 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

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).

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 Guerts 2014). Expressions and constructions carrying presuppositions are called "presupposition triggers", forming a large class including definites and factive verbs.

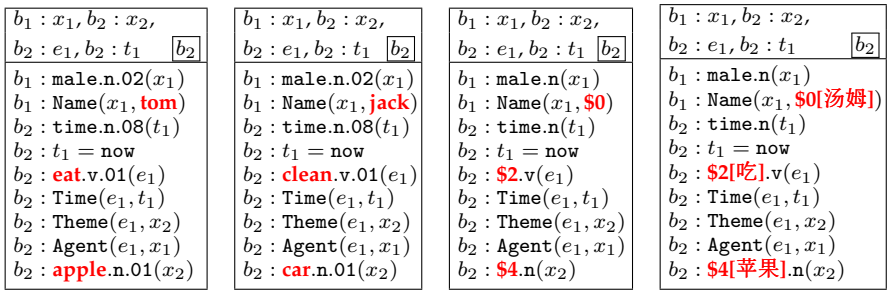

Figure 2: (a) DRS for sentence *Tom is eating an apple*; (b) DRS for sentence *Jack is cleaning a car*; (c) 𝕌DRS for both sentences; (d) 𝕌DRS for sentence 汤姆 正在 吃 一个 苹果 (Tom is eating an apple) constructed via (c) by substituting indices with their corresponding words. Expressions in brackets make the 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.

representations and sense distinctions (aka identical DRSs). For example, the English sentence *everyone was killed* and the Italian sentence *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 *eat/essen* has six senses according to the English WordNet (Fellbaum 1998) but only one in GermaNet (Hamp and Feldweg 1997), and the word *good/好* 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

Since word sense disambiguation is language-specific, we do not consider it part of the proposed cross-lingual meaning representation but assume that DRS operators (e.g., negation) and semantic roles are consistent across languages. We introduce 𝕌niversal Discourse Representation Structures (𝕌DRSs) which replace "DRS tokens" such as constants and predicates of conditions, with *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 : eat.v.01(e_1) is generalized to b_2 : $2.v($e_1$). Here, $2 corresponds to *eating* and v denotes the predicates part of speech (i.e., verb). Note that even though senses are not part of the 𝕌DRS representation, parts of speech (for predicates) are since they can provide cues for sense disambiguation across languages.

𝕌DRS 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, *Tom is eating an apple* and *Jack is washing a car* are represented by the same 𝕌DRS which can be viewed as a template describing an event in past tense with an agent and a theme. 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, 𝕌DRSs can be used to capture basic meaning

4 http://globalwordnet.org/about-gwa/

across languages. The 𝕌DRS in Figure 2(c) can be used to recover the semantics of the 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 which 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[5], and has been used for English DRSs, 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.[6] 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(e_1) 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 since 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; 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 Parallel Meaning Bank (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 REF x_1" indicates that variable x_1 is bound

5 http://globalwordnet.org/resources/wordnets-in-the-world/
6 https://www.wikidata.org/wiki/Wikidata:Main_Page

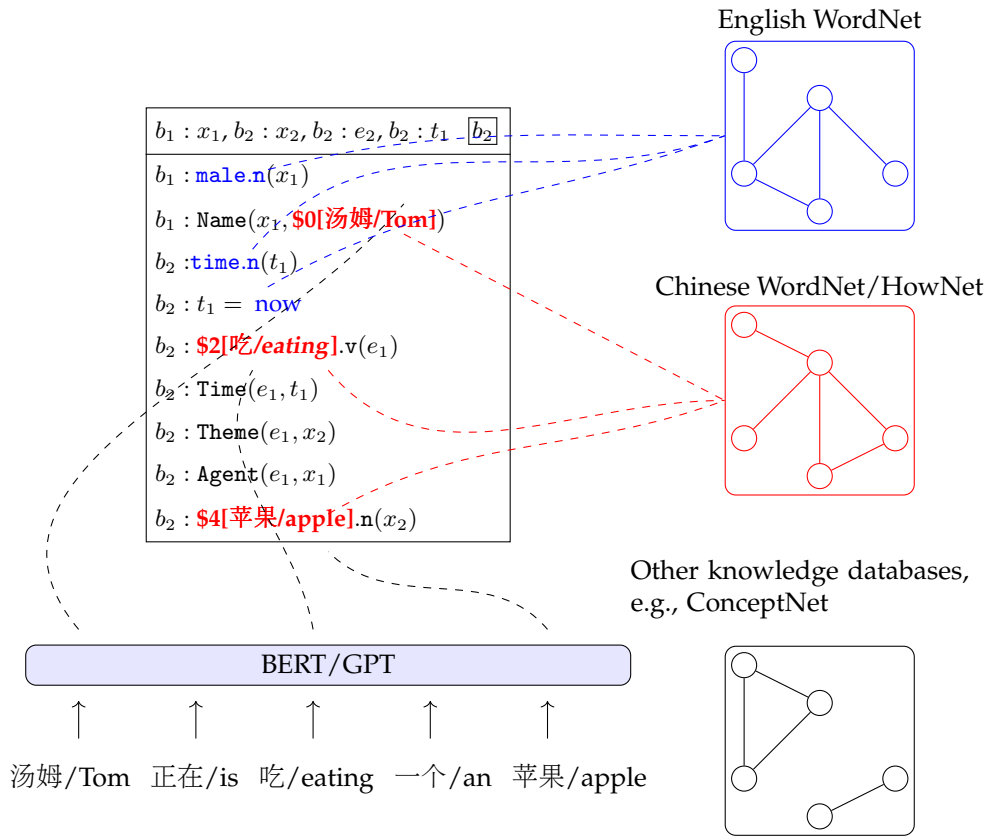

Figure 3: Illustration of how 𝕌DRSs can enrich deep contextual word representations; word senses can be disambiguated as a post-processing step according to the definitions of various language-specific resources.

in box b_2. Analogously, clause "b_3 `kill v.01` e_1" corresponds to condition `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).[7] 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 GETVARIABLEBOUND returns pairs of variables and box labels (indicating where these are bound) by enumerating all nested boxes.[8] The element $P[v]$ represents the label of the box bounding variable v. Basic conditions

7 For the full list of DRS clauses see https://pmb.let.rug.nl/drs.php.
8 Each variable has exactly one bounding box.

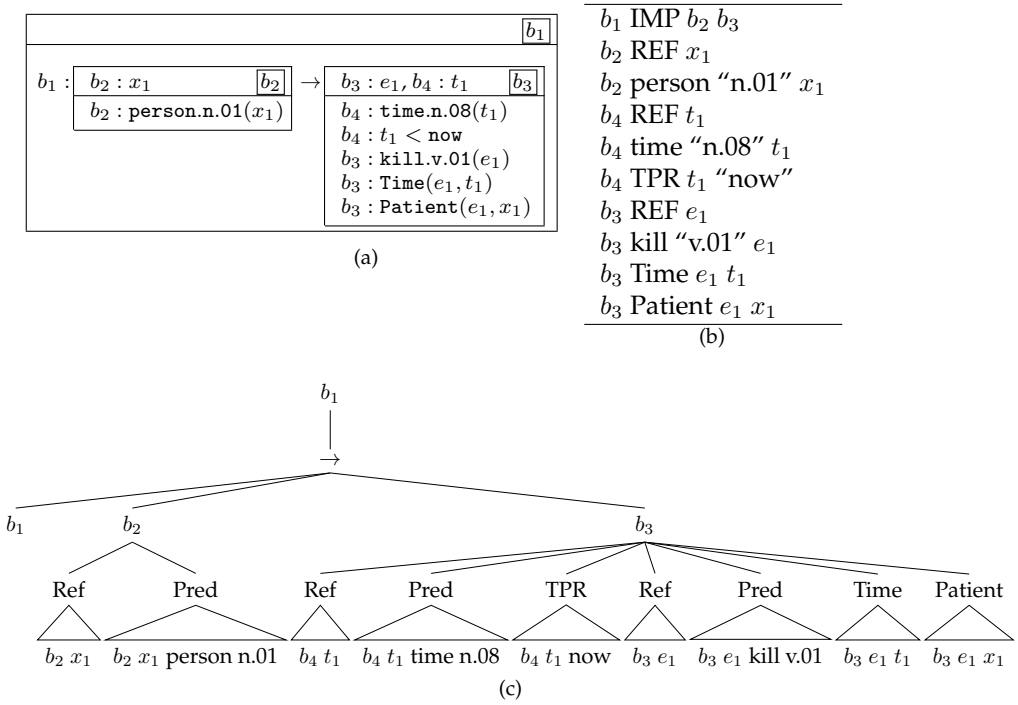

Figure 4: DRS in box format; (b) DRS in clause format (c) DRS in tree format proposed in this paper.

Algorithm 1 Box to Clause

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

1: $P = \text{GETVARIABLEBOUND}(B)$; $V \leftarrow \varnothing$; $C \leftarrow \varnothing$
2: **procedure** TRAVERSAL(b)
3: **for** $cond$ **in** $b.conds$ **do** ▷ each condition
4: **for** v **in** $cond.args$ **do** ▷ 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: $C = C \cup \{(cond.bound\ cond.name\ cond.args)\}$
11: **else if** $cond$ is unary complex **then**
12: $C = C \cup \{(cond.bound\ cond.name\ cond.B)\}$
13: TRAVERSAL($cond.B$)
14: **else if** $cond$ is binary complex **then**
15: $C = C \cup \{(cond.bound\ cond.name\ cond.B1\ cond.B2)\}$
16: TRAVERSAL($cond.B1$); TRAVERSAL($cond.B2$)
17: **end if**
18: **end for**
19: **end procedure**

are converted to a clause in lines 9–10, where $cond.args$ is a list of the arguments of the condition (e.g., predicates and referents). Unary complex conditions (i.e., negation,

Algorithm 2 Box to Tree

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

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

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.[9] An example is shown in Figure 4(b).

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 (DRTS) 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 paper, we propose an improved box-to-tree conversion algorithm, which is reversible and lossless, i.e., 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 $cond.args$ 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.

9 We refer to Bos et al. (2017) for more details on basic and complex conditions in DRS boxes.

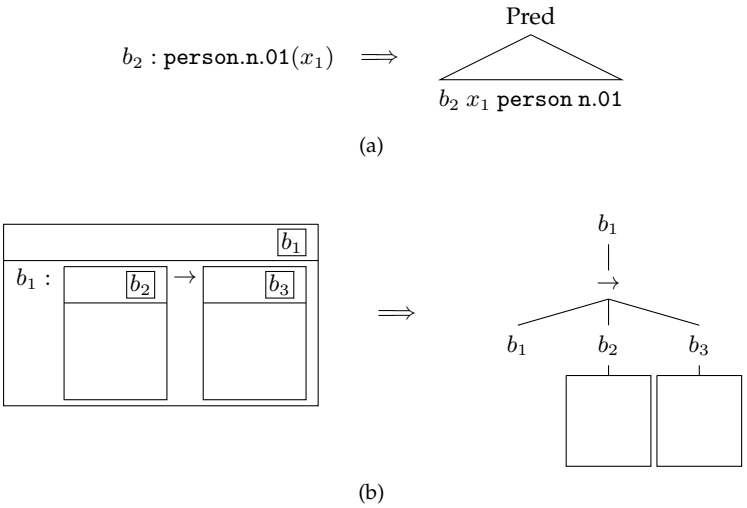

Figure 5: Example of converting a basic condition (a) and a binary complex condition (b) t0 a tree.

An example is shown in Figure 4(c). Basic condition b_2 : 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 (Ref b_2 x_1) (Pred b_2 x_1 person n.01)) (b_3 (Ref b_4 t_1) (Pred b_4 t_1 time n.08) (TPR b_4 t_1 now) (Ref b_3 e_1) (Pred b_3 e_1 kill v.01) (Time b_3 e_1 t_1) (Patient b_3 e_1 x_1)))).

4. Cross-lingual 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 use English WordNet synsets regardless of the source language. The output of our cross-lingual 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 Figure 6) for bootstrapping DRT semantic parsers for languages lacking gold standard training data.

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,

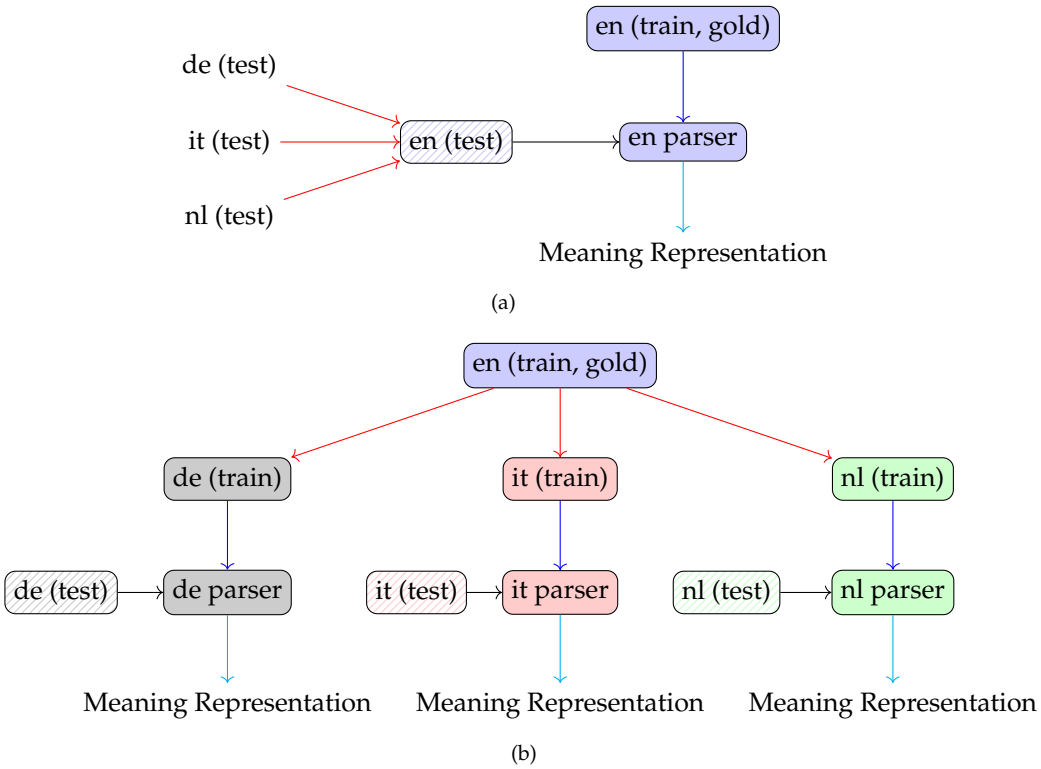

Figure 6: Two approaches for learning cross-lingual DRT parsers. Red arrows denote a machine translation engine; blue arrows denote the training of semantic parsing model, and cyan arrows denote the application of trained parser to test data (drawn in dotted background).

the cross-lingual 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 in 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 employed (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. Since 𝕌DRS indices are anchored to input sentences via word alignments, we would need to employ 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 which 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, $\{s_0, s_1, ..., s_{n-1}\}$ as input and outputs their linearized DRS $\{t_0, t_1, ..., t_{m-1}\}$, 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_{s_k} and position embeddings p_k: $x_k = e_{s_k} + p_k$. The input representations, $x_0, x_1, ..., x_{n-1}$, are fed to the Transformer encoder to obtain their hidden representations, $h_0, h_1, ..., h_{n-1}$:

$$[h_0 : h_{n-1}] = \textsc{LayerNorm}(\textsc{Encoder}([x_0 : x_{n-1}])), \tag{1}$$

where each layer of the ENCODER is:

$$\begin{aligned}
[\bar{x}_0 : \bar{x}_{n-1}] &= \textsc{LayerNorm}([x_0 : x_{n-1}]), \\
[\bar{h}_0 : \bar{h}_{n-1}] &= \textsc{MultiHeadSelfAttn}([\bar{x}_0 : \bar{x}_{n-1}]), \\
[h_0 : h_{n-1}] &= \text{FFN}([\bar{h}_0 : \bar{h}_{n-1}] + [x_0 : x_{n-1}]),
\end{aligned} \tag{2}$$

and LAYERNORM is a layer normalization function (Ba, Kiros, and Hinton 2016); MULTIHEADSEFLATTN is the multi-head self-attention mechanism introduced in Vaswani et al. (2017) which 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:

$$p(t_k|t_{<k}) = \textsc{SoftMax}(g(h_{t_k})), \tag{3}$$

where g is a linear function, and

$$h_{t_k} = \textsc{LayerNorm}(\textsc{Decoder}(y_{<k}, [h_0 : h_{n-1}])), \tag{4}$$

and each layer of the DECODER consists of five components:

$$\begin{aligned} \bar{y}_{<k} &= \textsc{LayerNorm}(y_{<k}), \\ q_{k-1} &= \textsc{MultiHeadAttn}(\bar{y}_{k-1}, \bar{y}_{<k}), \\ \bar{q}_{k-1} &= \textsc{LayerNorm}(q_{k-1} + y_{<k}), \\ \bar{h}_{k-1} &= \textsc{MultiHeadAttn}(\bar{q}_{k-1}, [x_0 : x_{n-1}]), \\ h_{t_k} &= \text{FFN}(\bar{h}_{k-1} + q_{k-1} + y_{k-1}), \end{aligned} \tag{5}$$

and $\textsc{MultiHeadAttn}(y_{k-1}, y_1^{k-1})$ returns the contextual representation for y_{k-1} according to its context information y_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 which 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 $\mathbb{D}_{auto} = D_0, D_1, ..., D_{m-1}$ 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 $\mathbb{D}_{auto}$ and then at each iteration on subset $\mathbb{D}_{auto}/D_1$ which excludes the data with the lowest quality D_i. 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 employs annotations of increasingly better quality. These can be produced

Algorithm 3 Iterative training

Input: M_{init}, the model; $\mathbb{D}$, auto-standard training data
Output: M_{opt}, the optimal model

$M_0 = M_{init}$
for i **in** $1...m$ **do**
    $M_i = \textsc{Train}(\mathbb{D}, M_{i-1})$
    $\mathbb{D} = \mathbb{D}/D_i$
end for
$M_{opt} = M_m$

by other models, human experts or a mixture. Since 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 dataset 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 the Parallel Meaning Bank 2.2.0, which is annotated with DRSs for English (en), German (de), Italian (it), and Dutch (nl). The dataset 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 dataset. We obtained $\mathbb{U}$DRSs 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).

Cross-lingual Setting. All cross-lingual experiments were conducted using Google Translate's API[10], 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.[11] 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.[12]

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.[13]
- **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 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 which incrementally generates DRSs (Evang 2019). It repeatedly selects transition actions within a stack-buffer framework. The stack contains the sequence of generated partial DRS, while the buffer stores incoming words. Transition actions either consume a word in the buffer or

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.'*

DRS	Prec	Rec	F_1
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 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.

merge two partial DRS 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 cross-lingual experiments were carried out on German, Italian, and Dutch. We built four cross-lingual 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 $\mathbb{U}$DRS parsing.

DRS	de			it			nl			all
	Pre	Rec	F_1	Pre	Rec	F_1	Pre	Rec	F_1	avg F_1
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	66.22	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

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

𝕌DRS	Prec	Rec	F_1
Cls-Transformer	94.2	91.1	92.5
Tree-Transformer	**93.9**	**93.6**	**93.8**

Table 3: English 𝕌DRS parsing results (PMB test set); best result per metric is shown in bold.

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 F_1 and 4.4 F_1, respectively) than LSTMs trained on data in the same format. A Transformer trained on trees performs slightly better (by 0.8 F_1) than the same model trained on clauses and is overall best among models employing DRS-based representations.

Our results on the DRS cross-lingual setting are summarized in Table 2. Many-to-One parsers outperform One-to-Many ones, however, the difference is starker for Clauses than for Trees (5.3 vs. 4.6 F_1 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 F_1) compared to the same model trained on clauses. This suggests that 𝕌DRSs parsing indeed benefits more from tree representations.

Our results on the 𝕌DRS cross-lingual 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 datasets 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 due to 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 cross-lingual experiments show that we can indeed bootstrap fairly accurate semantic parsers across languages, without *any* manual annotations on the target language.

𝕌DRS	de Pre	de Rec	de F_1	it Pre	it Rec	it F_1	nl Pre	nl Rec	nl F_1	all avg F_1
Cls	83.0	82.4	82.7	85.2	85.3	85.2	74.9	75.9	75.4	81.1
Cls-o2m	89.0	88.7	88.8	**88.4**	87.9	88.2	86.1	84.6	85.3	87.4
Tree	83.1	82.8	83.0	85.1	85.3	85.2	75.6	76.5	76.1	81.4
Tree-o2m	**89.5**	**88.7**	**89.1**	89.2	**88.2**	**88.7**	**85.7**	**84.8**	**85.2**	**87.7**

Table 4: 𝕌DRS parsing results on German, Italian, and Dutch (PMB test set); best result per metric shown in bold.

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 break-down 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 cross-lingual 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, it only includes global symbols like semantic role names and DRS operators that are shared across languages, thus making the parsing task simpler.

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

DRS	LSTM		Transformer	
	Cls	Tree	Cls	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 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			$\mathbb{U}$DRS	
	Tree	Tree-m2o	Tree-o2m	Tree	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-Npun	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**	—	—

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

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.

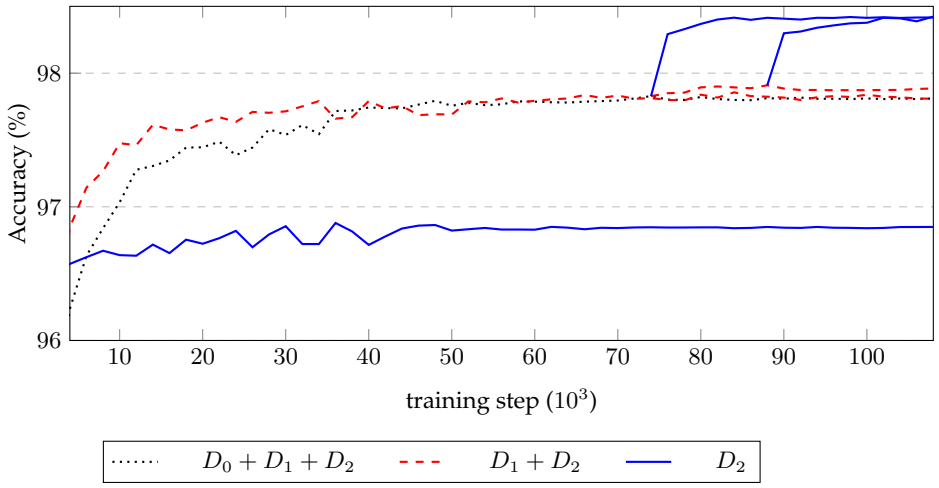

Figure 7: The effect of iterative training on model performance (Tree-Transformer, English development set).

6.4 Scalability Experiments

We further assessed whether the cross-lingual approach advocated in this paper 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 cross-lingual AMR parsers following a *full-cycle* evaluation scheme. The idea is to invert the learning process and bootstrap an English parser from the induced cross-lingual 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 cross-lingual 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.

Table 7 presents our results which are clustered according to language family (we only report 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

Language	F_1	Language	F_1	Language	F_1
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		Kannada	**76.33**
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	Afro-Asiatic		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	Khmer	61.53
Polish	77.48	Sundanese	75.51	Korean	74.14
Portuguese	78.97	Vietnamese	**79.37**	Mongolian	75.47
Punjabi	78.18	Niger-Congo		*Average*	76.47

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.

performance correlates with a higher quality of machine translation and statistical word alignments.[14]

We further investigated the quality of the constructed datasets 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

14 We will make the 𝕌DRS datasets 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.

language	BLEU	F_1
de	65.03	94.21
it	61.22	88.41
nl	69.12	94.06
avg	65.12 ($\pm$3.9)	92.23 ($\pm$1.98)

Table 8: Comparison between gold-standard 𝕌DRSsand constructed 𝕌DRSs by our methods in German, Italian, and Dutch using BLUE and COUNTER; standard deviations are shown in parentheses.

language	Prec	Rec	F_1
ja	62.0	65.6	63.7
zh	57.5	61.7	59.5

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

languages) is 65.12 while the average F_1 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 cross-lingual 𝕌DRS parsing results on the gold-standard test set in German, Italian and Dutch, which are close to English. In order to investigate the cross-lingual 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 4. Our results highlight that translation-based cross-lingual methods will be less accurate for target languages with large typological differences from the source.

6.5 Translation Divergence

Our cross-lingual 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, categorical, 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

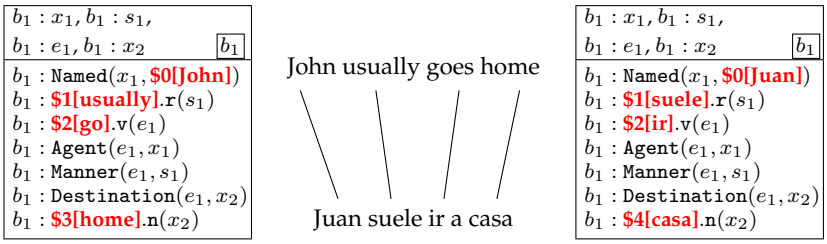

Figure 8: Example of promotional divergence. (a) 𝕌DRS of English sentence *John usually goes home*; (b) word alignments between the two sentences; (c) Incorrect 𝕌DRS of the Spanish translation *Juan suele ir a casa*, constructed via alignments.

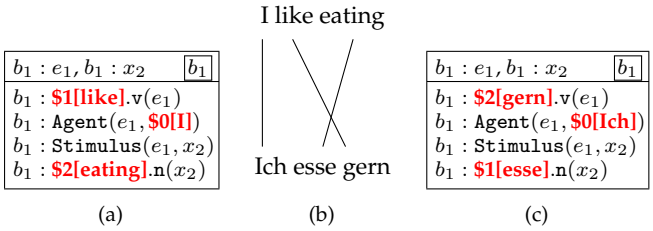

Figure 9: Example of demotional divergence. (a) 𝕌DRS of English sentence *I like eating*; (b) word alignments between two sentences; (c) incorrect 𝕌DRS of German translation *Ich esse gern*, constructed via alignments.

English but as a verb (*sueler*) 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-

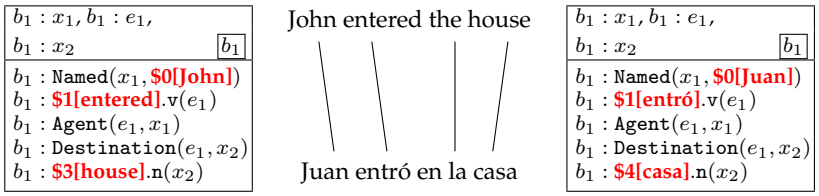

Figure 10: Examples of structural divergence. (a) 𝕌DRS of English sentence *John entered the house*; (b) word alignments between two sentences. (c) correct 𝕌DRS of Spanish translation *Juan entró en la casa*, constructed via alignments.

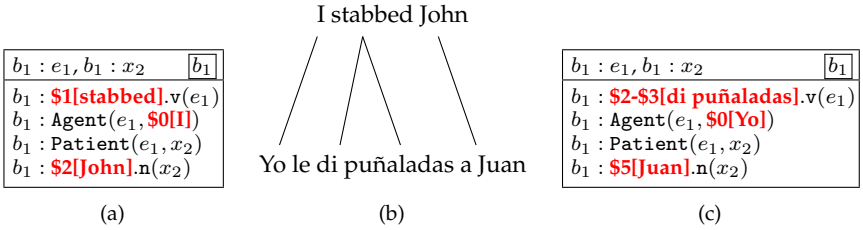

Figure 11: Example of conflational divergence. (a) 𝕌DRS of English sentence *I stabbed John*; (b) word alignments between two sentences; (c) correct 𝕌DRS of Spanish translation *Yo le di puñaladas a Juan*, constructed via alignments.

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.

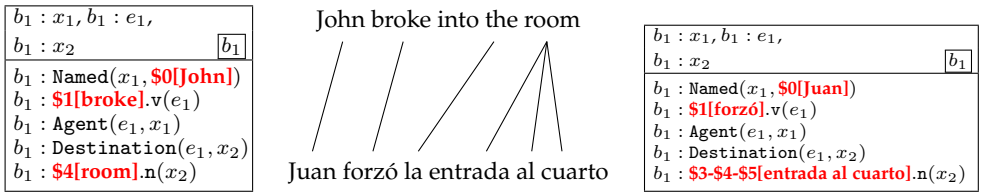

Figure 12: Example of lexical divergence. (a) 𝕌DRS of English sentence *John broke into the room*; (b) word alignments between two sentences. (c) correct 𝕌DRS of Spanish translation *Juan forzó la entrada al cuarto*, constructed via alignments.

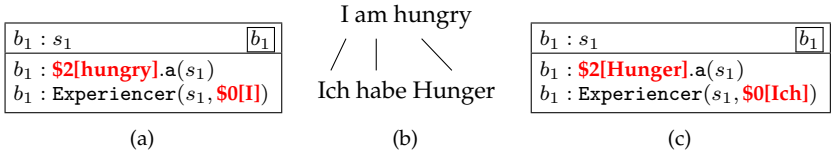

Figure 13: Example of catergorical divergence. (a) 𝕌DRS of English sentence *I am hungry*; (b) word alignments between two sentences; (c) incorrect 𝕌DRS of German translation *Ich habe Hunger*, constructed via alignments.

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.

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

	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

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

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 to compare 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 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 that is 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 which 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 cross-lingual 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 cross-lingual parsers. In addition, we introduce 𝕌DRT, a variant of the DRT formalism that we argue facilitates both monolingual and cross-lingual 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 cross-lingual 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 𝕌DRS representations allows to speed-up meaning bank creation and the annotation process.

8. Conclusion

In this paper, we introduced 𝕌niversal Discourse Representation Structures (𝕌DRSs) as a variant of canonical DRSs; 𝕌DRSs link elements of the DRS structure to tokens in the input sentence and are ideally suited to cross-lingual 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 cross-lingual 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, Bulgaria.

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, Spain.

Akbik, Alan, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 397–407.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baldridge, Jason and Alex Lascarides. 2005a. Annotating discourse structures for robust semantic interpretation. In *Proceedings of th 6th International Workshop on Computational Semantics*, pages 213–239, Tilburg, The Netherlands.

Baldridge, Jason and Alex Lascarides. 2005b. Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 96–103, Ann Arbor, Michigan.

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

Basile, Valerio and Johan Bos. 2013. Aligning formal meaning representations with surface strings for wide-coverage text generation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 1–9, Association for Computational Linguistics, Sofia, Bulgaria.

Basile, Valerio, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 3196–3200, Istanbul, Turkey.

Beaver, David I. and Bart Guerts. 2014. Presupposition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, winter 2014 edition edition.

Bos, Johan. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286.

Bos, Johan. 2015. Open-domain semantic parsing with Boxer. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304, Linköping University Electronic Press, Sweden.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2. Springer, pages 463–496.

Chen, Danqi. 2018. *Neural reading comprehension and beyond*. Ph.D. thesis, Standford University.

Cohen, Shay B., Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, UK.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 2-3(3):281–332.

Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic.

Damonte, Marco and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, Association for Computational Linguistics, New Orleans, Louisiana.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional 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, Minnesota.

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)*, pages 33–43, Berlin, Germany.

Dong, Zhendong, Qiang Dong, and Changling Hao. 2006. Hownet and the

computation of meaning.
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, Association for Computational Linguistics, Gothenburg, Sweden.
Evang, Kilian and Johan Bos. 2016. Cross-lingual 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, Japan.
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.
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.
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, Portugal.
Gangemi, Aldo, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. *Semantic Web*, 8(6):873–893.
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.
Hardalov, Momchil, Ivan Koychev, and Preslav Nakov. 2020. Enriched pre-trained transformers for joint slot filling and intent detection. *arXiv preprint arXiv:2004.14848*.
Hershcovich, 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, Canada.
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, Association for Computational Linguistics, Minneapolis, Minnesota.
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.
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 cross-lingual 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 cross-lingual 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, China.
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.
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, Germany.
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.
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*, volume 1. Mathematisch Centrum, Amsterdam,

pages 277–322.
Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic; An 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, China.
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, Canada.
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, Association for Computational Linguistics.
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.
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.
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, India.
Liu, Jiangming, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne, Australia.
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, Association for Computational Linguistics, Florence, Italy.
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, Sweden.
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, Minnesota.
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, California.
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, Scotland, UK.
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, Association for Computational Linguistics, Baltimore, Maryland.
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, Sweden.
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, Japan.
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.
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, Sweden.

Oepen, Stephan, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. Mrp 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22.

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, Association for Computational Linguistics.

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, British Columbia, Canada.

Padó, Sebastian and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

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 on association for computational linguistics*, pages 311–318, Association for Computational Linguistics.

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.

Van der Sandt, Rob A. 1992. Presupposition projection as anaphora resolution. *Journal of semantics*, 9(4):333–377.

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, Oregon, USA.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Venhuizen, 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, Germany.

Venhuizen, Noortje J., Johan Bos, Petra Hendriks, and Harm Brouwer. 2018. Discourse semantics with information structure. *Journal of Semantics*, 35(1):127–169.

Wada, Hajime and Nicholas Asher. 1986. BUILDRS: An implementation of DR theory and LFG. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, pages 540–545.

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, Association for Computational Linguistics, Online.

Wang, Mengqiu and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

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, Texas.

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, ACM.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 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 cross-lingual adversarial dataset 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, China.

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Yu, Adams Wei, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. QANet: 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.

Zhang, Sheng, Xutai Ma, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2018. Cross-lingual decompositional semantic parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1664–1675, Brussels, Belgium.