The First Shared Task on Discourse Representation Structure Parsing

Lasha Abzianidze Rik van Noord Hessel Haagsma Johan Bos
CLCG, University of Groningen
{l.abzianidze, r.i.k.van.noord, hessel.haagsma, johan.bos}@rug.nl

Abstract

The paper presents the IWCS 2019 shared task on semantic parsing where the goal is to produce Discourse Representation Structures (DRSs) for English sentences. DRSs originate from Discourse Representation Theory and represent scoped meaning representations that capture the semantics of negation, modals, quantification, and presupposition triggers. Additionally, concepts and event-participants in DRSs are described with WordNet synsets and the thematic roles from VerbNet. To measure similarity between two DRSs, they are represented in a clausal form, i.e. as a set of tuples. Participant systems were expected to produce DRSs in this clausal form. Taking into account the rich lexical information, explicit scope marking, a high number of shared variables among clauses, and highly-constrained format of valid DRSs, all these makes the DRS parsing a challenging NLP task. The results of the shared task displayed improvements over the existing state-of-the-art parser.

1 Introduction

Semantic parsing has been gaining in popularity in the last few years. There have been a series of shared tasks in semantic parsing organized, where each task requires to generate meaning representations of specific types: Broad-Coverage Broad-coverage Semantic Dependencies (Oepen et al., 2014, 2015), Abstract Meaning Representation (May, 2016; May and Priyadarshi, 2017), or Universal Conceptual Cognitive Annotation (Hershcovich et al., 2019).

The Discourse Representation Structure (DRS) parsing task extends this development by aiming at producing meaning representations that (i) come with more expressive power than existing ones and (ii) are easily translatable into formal logic, thereby opening the door to applications that require automated forms of inference (Blackburn and Bos, 2005; Dagan et al., 2013). DRSs are meaning representations employed by Discourse Representation Theory (DRT, Kamp and Reyle, 1993). They have been successfully applied for wide-coverage semantic representations (Bos et al., 2004; Bos, 2008), Natural Language Inference (Bos and Markert, 2005; Bjerva et al., 2014), and Natural Language Generation (Basile and Bos, 2013). To the best of our knowledge, there has never been a shared task on scoped meaning representations.

The aim of the task is to compare semantic parsing methods and the performance of systems that take as input an English text and provide as output the scoped meaning representation of that text Since a DRS combines logical (negation, quantification and modals), pragmatic (presuppositions) and lexical (word senses and thematic roles) components of semantics in a single meaning representation, the DRS parsing task shares parts of the following NLP tasks: semantic role labeling, reference resolution, scope detection, named entity tagging, word sense disambiguation, predicate-argument structure prediction, and presupposition projection.

There are only a few previous approaches to DRS parsing. Traditionally, due to the complexity of the task, it has been the domain of symbolic and statistical approaches (Bos, 2008; Le and Zuidema, 2012; Bos, 2015). Recently, however, neural sequence-to-sequence systems achieved impressive performance on the task (Liu et al., 2018; Van Noord et al., 2018), without relying on any external linguistic resources.

SYSTEM INPUT:

Tom isn't afraid of anything.

SYSTEM OUTPUT:

```
b1 REF x1
b1 male "n.02" x1
b1 Name x1 "tom"
b2 REF t1
b2 EQU t1 "now"
b2 time "n.08" t1
b2 NOT b3
b3 REF s1
b3 Time s1 t1
b3 Experiencer s1 x1
b3 afraid "a.01" s1
b3 Stimulus s1 x2
b3 REF x2
b3 entity "n.01" x2
```

BOX FORMAT:

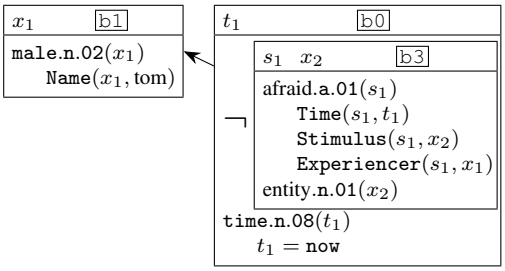

Figure 1: The DRS parsing task: the system input is a short text (the PMB document 99/2308), and the expected output is a DRS in clausal form. Its standard visualisation in box-notation, following DRT, is presented below.

00/3008: He played the piano and she sang.

```
b6 DRS b1                b6 DRS b4
b2 REF x1                b5 REF x3
b2 male "n.02" x1        b5 female "n.02" x3
b1 REF e1                b4 REF e2
b1 play "v.03" e1        b4 sing "v.01" e2
b1 Agent e1 x1           b4 Agent e2 x3
b1 Theme e1 x2           b4 Time e2 t2
b3 REF x2                b4 REF t2
b3 piano "n.01" x2       b4 TPR t2 "now"
b1 REF t1                b4 time "n.08" t2
b1 time "n.08" t1        b6 CONTINUATION b1 b4
b1 TPR t1 "now"          b1 Time e1 t1
```

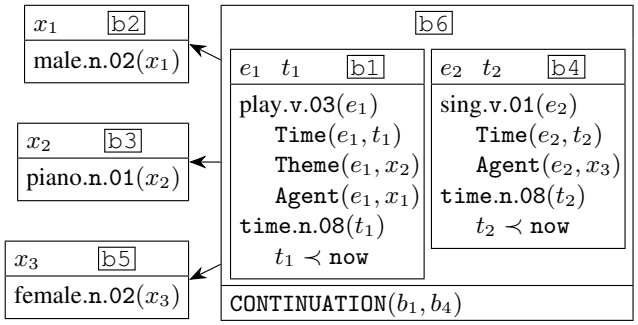

Figure 2: The segmented box b6 consists of a set of labelled boxes, i.e. the discourse segments b1 and b2, and a single discourse condition. In the condition, discourse relation holds between two discourse segments and is formatted in uppercase. The definite noun phrase and the pronouns are presupposition (b2, b3, and b5) triggers.

In the first shared task on DRS parsing, taking into account the information-rich and complex structure of the target meaning representation, we tested participant systems mainly on short, open-domain English texts. In this way, we lowered the threshold for participation to encourage higher results in the shared task and mitigate challenges associated to semantic parsing long texts. In total five systems participated in the shared task. The top-ranked systems outperformed the existing state-of-the-art system in DRS parsing. The shared task was hosted on CodaLab.[1].

2 Task Description

The DRS parsing in a nutshell is presented in Figure 1. Here, the input, a short English sentence, needs to be mapped to the output, a scoped meaning representation in clausal form. Concepts, states and events are represented by the word senses (male.n.02, entity.n.01, afraid.a.01) from WordNet 3.0 (Fellbaum, 1998) and relations are modeled with thematic roles (Name, Experiencer, Stimulus) drawn from an extended version of VerbNet (Bonial et al., 2011).

Each entity needs to introduce a discourse referent, i.e. a variable, in the right scope, form an instance of the right concepts, and be connected to other entities via thematic roles or comparison operators. For example, in Figure 1, *anything* introduces a discourse referent x_2 in the scope b3 with the help of the clause ⟨b3 REF x2⟩. The clause ⟨b3 entity "n.01" x2⟩ makes x_2 an instance of

[1] https://competitions.codalab.org/competitions/20220

entity.n.01. Finally, the clause ⟨b3 Stimulus s1 x2⟩ connects x_2 to the event entity s_1 of *afraid* via the Stimulus thematic role.

The scopes of negation, implication, modal operators or propositional arguments need to be correctly identified. Proper names, pronouns, definite descriptions and possessives are treated as presuppositions and get their own box if they cannot be resolved by the local context. Tense is locally accommodated. For example, Figure 1 shows how the negation operator introduces the scope (b3) and how the named entity *Tom* gives rise to the presupposition (b1). Figure 2 demonstrates how discourse segments get their own scope (b1 and b4) and how definite noun phrases and pronouns trigger presuppositions (b2, b3, and b5). Finally, Figure 3 depicts an implication with two scopes (b3 and b5), modeling semantics of a universal quantifier, and nested presuppositions (b1 and b4) due to a possessive pronoun.

Given the aforementioned nuances of the fine-grained scoped meaning representations, the DRS parsing task represents a challenge for machine learning methods.

3 Discourse Representation Structure

The meaning representations used in this shared task are based on the DRSs put forward in DRT (Kamp and Reyle, 1993) and derived from the Parallel Meaning Bank (Abzianidze et al., 2017). There are some important extensions to the theory, though. First, the DRSs are language-neutral, and all non-logical symbols are disambiguated to WordNet synsets or VerbNet roles. Furthermore, presuppositions are explicitly represented following (Van der Sandt, 1992) and Porjective DRT (Venhuizen et al., 2018). Discourse structure is analysed following by Segmented DRT (Asher and Lascarides, 2003). As in the original DRT, DRSs are displayed in box format for reading convenience (presuppositional DRSs are displayed with outgoing arrows of the boxes that triggered them). DRSs are recursive structures, and for the purpose of evaluation, they are translated into clauses, flattening down the recursion by reification.

A DRS always contains a main labelled box along with an optional set of presupposition DRSs (see Definition 1). For example, the main labelled box in Figure 1 is b0 while b0 is a presupposition. A box can be simple (e.g., the box labelled with b0 in Figure 1) or segmented (e.g., the box labelled with b6 in Figure 2). A simple box consists of a set of discourse referents and a set of conditions. Conditions can be basic or complex. Basic conditions are concept predicates or relations over discourse referents and constants. Indexicals are treated as constants, not as discourse referents Bos (2017), for example, *now* is one of such indexicals (see Figure 1). Complex conditions are those involving labelled boxes. The examples of complex conditions are ¬b3 in Figure 1 and b3 ⇒ b5 in Figure 3. Finally, a segmented box contains a set of labelled boxes (b1 and 4 in Figure 2) and discourse conditions. A discourse condition is a discourse relations over box labels, e.g., CONTINUATION(b_1, b_4) in Figure 2.

Definition 1: A BNF of DRSs: (possibly empty) sets are denoted with curly brackets as {⟨ element⟩}. The string elements for operators and punctuation are in red.

⟨*DRS*⟩ ::= {⟨*DRS*⟩} ⟨*labelled BOX*⟩

⟨*labelled BOX*⟩ ::= ⟨*label*⟩ ⟨*BOX*⟩

⟨*BOX*⟩ ::= ⟨*simple BOX*⟩ | ⟨*segmented BOX*⟩

⟨*simple BOX*⟩ ::= {⟨*discourse referent*⟩} {⟨*condition*⟩}

⟨*condition*⟩ ::= ⟨*basic condition*⟩ | ⟨*complex condition*⟩

⟨*term*⟩ ::= ⟨*discourse referent*⟩ | ⟨*constant*⟩

⟨*basic condition*⟩ ::= ⟨*semantic role*⟩ (⟨*term*⟩, ⟨*term*⟩)
| ⟨*term*⟩ ⟨*comparison operator*⟩ ⟨*term*⟩
| ⟨*concept*⟩.⟨*pos_sense_number*⟩(⟨*term*⟩)

⟨*complex condition*⟩ ::= ¬⟨*labelled BOX*⟩ | ◇⟨*labelled BOX*⟩ | □⟨*labelled BOX*⟩
| ⟨*labelled BOX*⟩⇒⟨*labelled BOX*⟩
| ⟨*discourse referent*⟩:⟨*labelled BOX*⟩

⟨*segmented BOX*⟩ ::= {⟨*labelled BOX*⟩} {⟨*discourse condition*⟩}

⟨*discourse condition*⟩ ::= ⟨*discourse relation*⟩ (⟨*label*⟩, ⟨*label*⟩)

01/2312: He put all his money in the box.

```
b1 REF x1           % He [0...2] his [11...14]
b1 male "n.02" x1   % He [0...2] his [11...14]
b2 REF t1           % put [3...6]
b2 TPR t1 "now"     % put [3...6]
b2 time "n.08" t1   % put [3...6]
b5 REF e1           % put [3...6]
b5 Agent e1 x1      % put [3...6]
b5 Theme e1 x2      % put [3...6]
b5 Time e1 t1       % put [3...6]
b5 put "v.01" e1    % put [3...6]
b2 IMP b3 b5        % all [7...10]
b3 REF x2           % all [7...10]
b3 PartOf x2 x3     % all [7...10]
b3 entity "n.01" x2 % all [7...10]
b4 REF x3           % his [11...14]
b4 Owner x3 x1      % his [11...14]
b4 money "n.01" x3  % money [15...20]
b5 Destination e1 x4 % in [21...23]
b6 REF x4           % the [24...27]
b6 box "n.01" x4    % box [28...31]
```

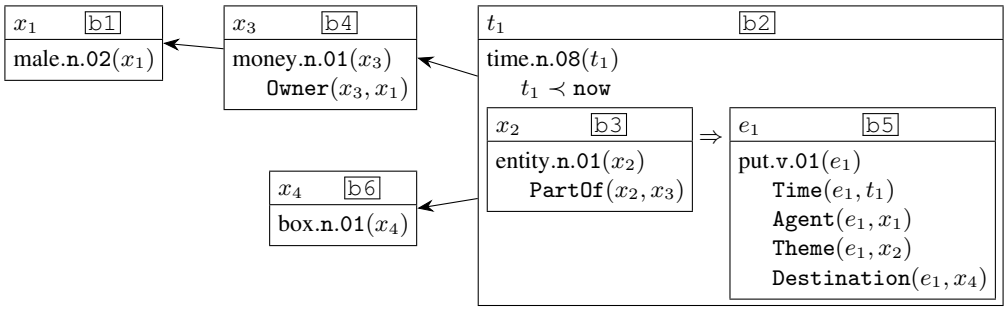

Figure 3: The DRS contains the example of nested presuppositions triggered by the possessive pronoun *his*. The main box `b2` of the DRS presupposes a set of two DRSs. At the same time, one of the presupposed DRSs, namely $\langle\{\texttt{b1}\}, \texttt{b4}\rangle$, itself carries the presupposition `b1`. Note that the presuppositions about a male discourse referent, triggered by *he* and *his* separately, are merged into a single presupposition box `b1`. The clauses are accompanied with aligned tokens.

The clausal form and the box-notation are two different forms of displaying scoped meaning representations van Noord, Abzianidze, Haagsma, and Bos (2018). We consider the clausal form a machine-readable format that is suitable for the evaluation with a continuous score between 0 and 1 (see Section 5). On the other hand, the box-notation is a human-readable format and originates from Discourse Representation Theory. Conversion from the box-notation to the clausal form and vice versa is transparent: each box gets a label, and discourse referents and conditions in the clausal form are preceded by the label of the box they occur in.

4 Data

4.1 Released Data

For the shared task we released the training, development, and test data, taken from the Parallel Meaning Bank (PMB, Abzianidze et al. 2017). The PMB is a parallel corpus annotated with formal meaning representations.[2] These representations capture the most probable interpretation of a sentence; no ambiguities or under-specification techniques are employed. The formal meaning representations are automatically constructed and manually corrected. Completely correct representations are flagged as *gold*. Representations that are partly manually corrected are marked as *silver*, while the rest is marked *bronze*.

The PMB release number used for the shared task is 2.2.0[3], of which some statistics are shown in Table 1. Note that MWE tokens and types are underrepresented in the silver and bronze data compared to the gold data. This is because the gold data contains more manual corrections on the token level than the silver and bronze data. For the example of multi-word expressions see Figure 4. In the shared task, participants were allowed to use the silver and bronze data, this would especially make sense in the case of data-hungry neural models, though there is no guarantee that those representations resemble the gold

[2] A part of the corpus can be viewed online via the PMB explorer: http://pmb.let.rug.nl/explorer

[3] https://pmb.let.rug.nl/data.php

Table 1: Statistics for the PMB release 2.2.0 and the shared task evaluation set.

Data splits	Docs	Tokens	Word types	MWE tokens	MWE types
PMB 2.2.0 gold train	4,597	29,195	4,431	789	494
PMB 2.2.0 gold dev	682	4,067	1,251	71	61
PMB 2.2.0 gold test	650	4,072	1,240	108	100
PMB 2.2.0 silver	67,965	583,835	19,432	4,495	1,902
PMB 2.2.0 bronze	120,662	919,247	23,354	2,054	699
Evaluation set	600	4,066	1,237	92	79

standard.

The data provided to the shared task participants consists of pairs of a raw natural language text and its corresponding scoped meaning representation in clausal form.[4] Whether the meaning representation is of gold, silver or bronze standard is explicitly indicated. To facilitate automatic learning of scoped meaning representations, we also provided automatically induced alignments between clauses and tokens, where token positions are provided with character offsets. The examples of clause-token alignments are give in Figure 2 and Figure 4. The latter represents an exact formatting of the text and clausal form pair provided in the shared task.

```
Nick Leeson was arrested for collapse of Barings Bank PLC.

b1 REF x1                       % Nick~Leeson [0...11]
b1 Name x1 "nick~leeson"        % Nick~Leeson [0...11]
b1 male "n.02" x1               % Nick~Leeson [0...11]
b2 REF t1                       % was [12...15]
b2 TPR t1 "now"                 % was [12...15]
b2 Time e1 t1                   % was [12...15]
b2 time "n.08" t1               % was [12...15]
b2 REF e1                       % arrested [16...24]
b2 Patient e1 x1                % arrested [16...24]
b2 arrest "v.01" e1             % arrested [16...24]
b2 Theme e1 x2                  % for [25...28]
b2 REF x2                       % collapse [29...37]
b2 collapse "n.04" x2           % collapse [29...37]
b2 Patient x2 x3                % of [38...40]
b3 REF x3                       %
b3 Name x3 "barings~bank~plc"   % Barings~Bank~PLC [41...57]
b3 company "n.01" x3            % Barings~Bank~PLC [41...57]
                                % . [57...58]
```

Figure 4: A sample of a training document (57/0762). For each document there is a pair of raw text and the corresponding clausal form. Clausal forms incorporate automatically induced clause-token alignment.

4.2 Evaluation set

The official evaluation set contains 600 instances that were not released previously. They will not be released publicly, but are still available for (blind) scoring via the shared task website.[5] However, during the evaluation phase, we asked the participants to provide DRSs for a set of 12,606 short texts. In addition to the raw texts (600) from the evaluation split, this set contained the train (4,597), development (682),

[4] https://github.com/RikVN/DRS_parsing/tree/master/data/pmb-2.2.0

[5] https://competitions.codalab.org/competitions/20220

and test (650) data from the PMB-2.2.0 release and the sentences (6,077) from the SICK dataset (Marelli et al., 2014). The reason for providing the inflated set of raw texts was three-fold: (i) Disguise the raw texts of the evaluation set to make it hard to tune models on them; (ii) Obtain the complete information about the performance of the systems on the provided training, development and test sets; (iii) Carry out extrinsic evaluation of the participant systems on the natural language inference task.

5 Evaluation Metrics and Baselines

Before comparing a system produced clausal form to the gold one, the produced form is checked on validity—whether it represents a DRS. If the clausal form is invalid, it is replaced by a single non-matching clause. In the shared task, we include three baseline systems. The evaluation and validation scripts and the baselines are publicly available.[6]

5.1 Validation

Not all sets of clauses correspond to a well-formed DRS, e.g., discourse referents found in the conditions should be explicitly introduced in the boxes, or there should exist labelled boxes for the labels used in the discourse conditions. We employ the validator REFEREE (Van Noord et al., 2018) to automatically check a set of clauses on well-formedness. REFEREE does several checks for validity checking. For example, first it scans each clause separately in a clausal form and identifies the types of variables based on the operators. For each discourse referent variable, it checks the existence of the binding discourse referent. During this procedure, REFEREE also detects positions of the boxes in the DRS (i.e., so-called the subordinate relation). Based on this information, it is checked that nested boxes do not create loops and there is a unique main box in the DRS.

All the released clausal forms of the DRSs are valid. We provided the participants with REFEREE in order to help them identify the ill-formed clausal forms produced by their systems.

5.2 Evaluation

The evaluation defines to what degree a system output clausal form is similar to the corresponding gold one. To compare the system output and gold representations, we compute the F1-score over the clauses, following Allen et al. (2008). We use the tool COUNTER (van Noord, Abzianidze, Haagsma, and Bos, 2018), which is specifically designed to evaluate DRSs. It is based on the SMATCH Cai and Knight (2013) tool that is used to evaluate AMR parsers. It is essentially a hill-climbing algorithm that finds the best variable mapping between the produced DRS and the gold standard. To avoid local optima, we restart the procedure 10 times. In order to prevent an inflated F-score, before searching the maximal matching, COUNTER discards those REF-clauses which are deemed redundant. A REF-clause $\langle$b REF x$\rangle$ is redundant if and only if its discourse referent x occurs with a concept predicate in a basic condition of the same box b – in other words, there exists a clause of the form $\langle$b *concept* "*pos.nn*" x$\rangle$.[7]

An example of comparing the clausal forms of two scoped meaning representations is shown in Figure 5. With respect to the optimal mapping, both, the sample system output and gold clauses, include three clauses that could not be matched with each other while four clauses are matched. The optimal mapping gives us a precision and recall of $3/7$, resulting in an F-score of 42.9. Similarly to AMR, we use micro-averaged F-score when evaluating a set of DRSs.

An aspect that is different from the AMR evaluation system is that we generalize over synonyms. In a preprocessing step of the evaluation, all word senses are converted to its WordNet 3.0 synset ID. For example, fox.n.02 and dodger.n.01 both get normalized to dodger.n.01 and are thus able to match.

[6] http://github.com/RikVN/DRS_parsing

[7] In Figure 5 redundant REF-clauses are stricken through.

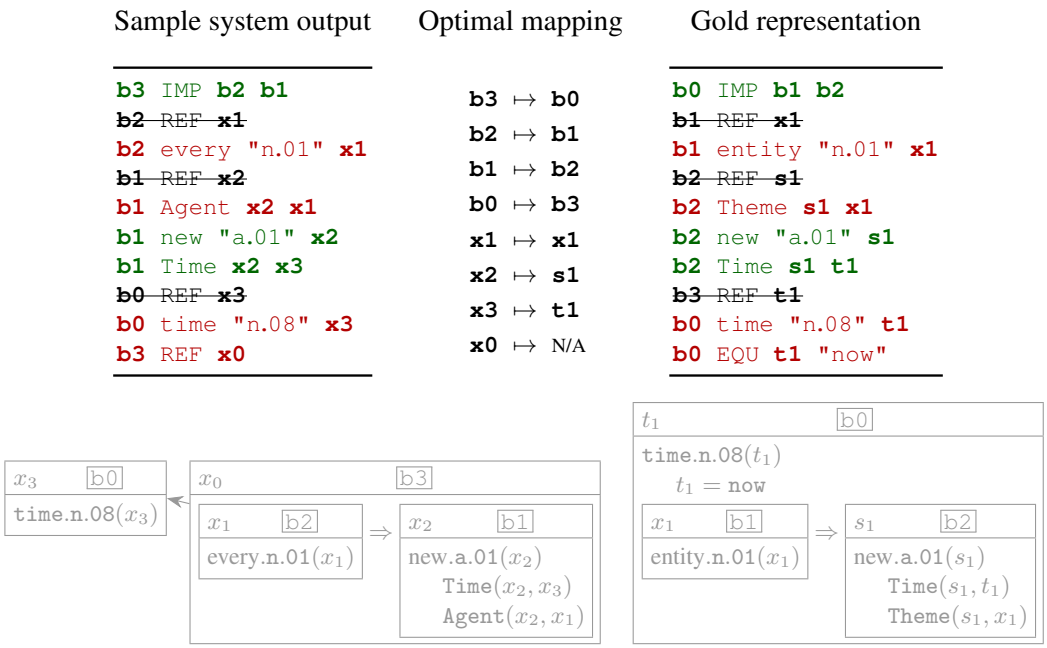

Figure 5: An optimal mapping of variables which maximizes overlap between the system output and gold clausal forms for the sentence (PMB document 00/2302) *Everything is new*. The maximal overlap yields an F-score of 42.9. Matching, non-matching and redundant clauses are in green, red, and stricken through, respectively. The box-notation of scoped meaning representations is not available during the comparison of clausal forms.

To calculate whether two systems differ significantly, we perform approximate randomization Noreen (1989), with $\alpha = 0.05$, $R = 1000$ and $F(model_1) > F(model_2)$ as test statistic for each individual DRS pair.

5.3 Baselines

We provide three baseline parsers: SPAR, SIM-SPAR and AMR2DRS. SPAR simply outputs a default DRS, which is a DRS that is the most similar to the DRSs in our training set.[8] SIM-SPAR outputs the DRS of the most similar sentence in the training set, based on the cosine distance of the average word-embedding vector, calculated using GloVe (Pennington et al., 2014). AMR2DRS is a script that converts the output of an AMR parser to a valid DRS by applying a set of rules, described in Bos (2016) and van Noord, Abzianidze, Haagsma, and Bos (2018). We will provide scores on the development, test and evaluation sets by using the AMR parser of van Noord and Bos (2017).

6 Participating Systems

We received a total of five submissions in the shared task out of 32 registered participants. Three out of five submitted a system paper. The general characteristics of the participating systems are give in Table 2. Following Nissim et al. (2017), we explicitly encouraged the participants to include ablation experiments and negative results (if any). Note that the authors of the systems NOORD ET AL.18 and NOORD ET AL.19 are from the organizers. Below, we provide a short description of each system.

6.1 Van Noord et al. (2018)

NOORD ET AL.18, the parser described in Van Noord et al. (2018), uses a character-level neural sequence-to-sequence model to produce DRSs. They apply a number of methods to improve performance, such as

[8] For PMB release 2.2.0 this is the DRS for *Tom voted for himself.*

Table 2: Overview of the participating systems

	Model	Input	Embeddings	Silver	Bronze
LIU ET AL.	Transformer	char	✖	✔	✔
NOORD ET AL.19	seq2seq	char	✖	✔	✖
NOORD ET AL.18	seq2seq	char	✖	✔	✖
EVANG	stack-LSTMs	word	✔	✖	✖
FANCELLU ET AL.	bi-LSTM	word	✔	✖	✖

rewriting the variables to a more general format, introducing a feature for uppercase letters and *not* using character-level representation for DRS roles and operators. Moreover, they show that performance can be substantially improved by first pre-training on gold and silver data, after which the parser is fine-tuned on only the gold standard data.

6.2 Van Noord (2019)

The system of NOORD ET AL.19 is the parser described in Van Noord et al. (2019) and Van Noord (2019), which follows up on their work previously described in Van Noord et al. (2018). They improve on this work in two ways: (i) by switching their sequence-to-sequence framework from OpenNMT (Klein et al., 2017) to Marian (Junczys-Dowmunt et al., 2018) and (ii) by providing the encoder with linguistic information (lemmas, semantic tags (Abzianidze and Bos, 2017), POS-tags, dependency parses and CCG supertags) that are encoded in a separate encoder.

6.3 Liu et al. (2019)

LIU ET AL. also follow the approach of Van Noord et al. (2018) in terms of pre- and postprocessing the data, but they improve on it by using the Transformer model (Vaswani et al., 2017), instead of a sequence-to-sequence RNN. Also, they show that employing the bronze standard in addition to the gold and silver standard leads to improved performance.

6.4 Evang (2019)

EVANG aim to find a middle-ground between traditional symbolic approaches and the recent neural (sequence-to-sequence) models. They employ a transition-based parser that relies on explicit word-meaning pairs that are found in the training set. Parsing decisions are made based on vector representations of parser states, which are encoded using stack-LSTMs.

6.5 FANCELLU ET AL.[9]

FANCELLU ET AL. propose a graph decoder that given an input sentence encoded via a bidirectional LSTM generates a DAG (Directed Acyclic Graph) as a sequence of fragments from a graph grammar. These fragments are *delexicalized*; predicate names, synset and information on whether the predicate is presupposed or not are predicted in a second step, conditioned on the fragment and the decoding history. Two are the main features of the graph parser: 1) it is agnostic to the underlying semantic formalism and does not need any preprocessing step to deal with variable binding; 2) fragments are aware of the overall graph structure and the graph is built incrementally via a process of non-terminal rewriting. (1) sets this method apart from the graph parser of Groschwitz et al. (2018) where a grammar is extracting via an elaborate pre-processing step, tailored to a specific formalism, whereas (2) allows to leverage neural sequential decoding (stackLSTM, Dyer et al., 2016). The only preprocessing step required is to convert DRSs in clause format into single-rooted, fully instantiated DAGs; we do so by treating both variables

[9]The full list of authors: Federico Fancellu, Sorcha Gilroy, Adam Lopez and Mirella Lapata (University of Edinburgh). Since the authors refrained from submitting a system paper due to the ACL policy for submission, we include a slightly extended summary of their system, generously provided by them.

Table 3: Official results of the shared task for the participating systems

	PMB 2.2.0 (F%)			Evaluation set (%)		
	Train	Dev	Test	Prec.	Rec.	F
AMR2DRS	NA	39.7	40.1	36.7	42.2	38.8
SPAR	NA	40.0	40.8	44.3	35.4	39.4
SIM-SPAR	NA	53.3	57.7	55.7	53.0	54.3
FANCELLU ET AL.	91.1	69.9	73.3	71.9	64.1	67.8
EVANG	84.2	74.4	74.4	71.9	69.9	70.9
NOORD ET AL.18	88.5	81.2	83.3	80.8	78.6	79.7
NOORD ET AL.19	94.9	86.5	86.8	85.5	83.6	84.5
LIU ET AL.	96.9	85.5	87.1	84.8	84.8	84.8

and boxes as nodes and semantic roles, operators and discourse relations as edges between those (where each binary operator or relation gives rise to two edges). Similarly, the only postprocessing step lies in converting the graph back to clause format. This last step can inject errors in the parse and it is the reason why some of the output graphs can be ill-formed.

7 Results

Table 3 shows the official results of the shared task. The system of LIU ET AL. achieved the best performance, though there is no significant difference with the work of NOORD ET AL.19 ($p = 0.23$). The systems of EVANG and FANCELLU ET AL., though clearly outperforming the baselines, are a bit behind the best three systems. However, they can likely improve performance by incorporating silver and bronze standard data. No systems seem to have overfit on the provided dev and test sets. The work of FANCELLU ET AL. is perhaps overfit on the training set, given their high score on train compared to the test sets.

Table 4 shows a more detailed overview of the results. All teams produced a substantial amount of perfect DRSs, but only 31 DRSs of them were perfectly produced by each system. EVANG is the only system with a substantial number of ill-formed DRSs. This hurts their performance, since they get an F-score of 0.0 in evaluation. If we ignore referee and score their ill-formed DRSs as if they were valid, their score increases to 72.1. On the other hand, calculating an F-score for *only* the ill-formed DRSs (without referee) gives us an F-score of 50.6, suggesting that the model would not have scored very well in either way.

Similar as was observed in Van Noord et al. (2018), word sense disambiguation is problematic for the DRS parsers. When assuming oracle sense numbers, all systems obtain a substantially higher F-score (increases of 1.8 to 3.6). NOORD ET AL.19 propose a simple method to improve on this sub-problem by taking the most frequent sense in the training set for a concept, though this only increased their F-scores by 0.2 to 0.4 on the dev and test sets. Nouns are the easiest for all models to correctly produce (possibly due to the frequent `time.n.08`), while adverbs are the hardest, though there are only 12 such clauses in the evaluation set.

8 Analysis

Longer sentences are probably harder to parse, but which systems behave well on longer sentences? Figure 6 shows the performance of the systems plotted over sentence length. As expected, all systems show a clear drop in performance for longer sentences. The work of LIU ET AL. is based on the Transformer model (Vaswani et al., 2017), which claims that performance should not degrade for longer

Table 4: F-scores of fine-grained evaluation of the participating systems on the evaluation set. "Winner out of 5" counts only those instances for which the parser obtained a higher score than all the rest, while "Highest out of 5" allows ties.

	LIU	NOORD19	NOORD18	EVANG	FANCELLU
All clauses	84.8	84.5	79.7	70.9	67.8
DRS Operators	93.9	94.2	91.7	75.2	76.3
VerbNet roles	82.7	83.5	78.1	72.4	66.4
WordNet synsets	83.8	82.3	77.2	67.8	66.5
nouns	89.2	87.5	83.5	75.9	70.3
verbs	69.5	68.9	60.9	44.1	58.3
adjectives	74.8	74.2	66.5	61.5	53.8
adverbs	63.6	45.5	33.3	0.0	31.6
Oracle sense numbers	86.6	87.1	82.6	74.5	69.8
Oracle synsets	90.5	90.7	87.5	80.1	76.5
Oracle roles	88.4	88.5	84.3	74.5	73.7
# of perfect DRSs	214	210	160	95	104
# highest out of 5	383	376	261	171	161
# winner out of 5	100	77	26	18	18
# of ill-formed DRSs	1	0	1	37	5

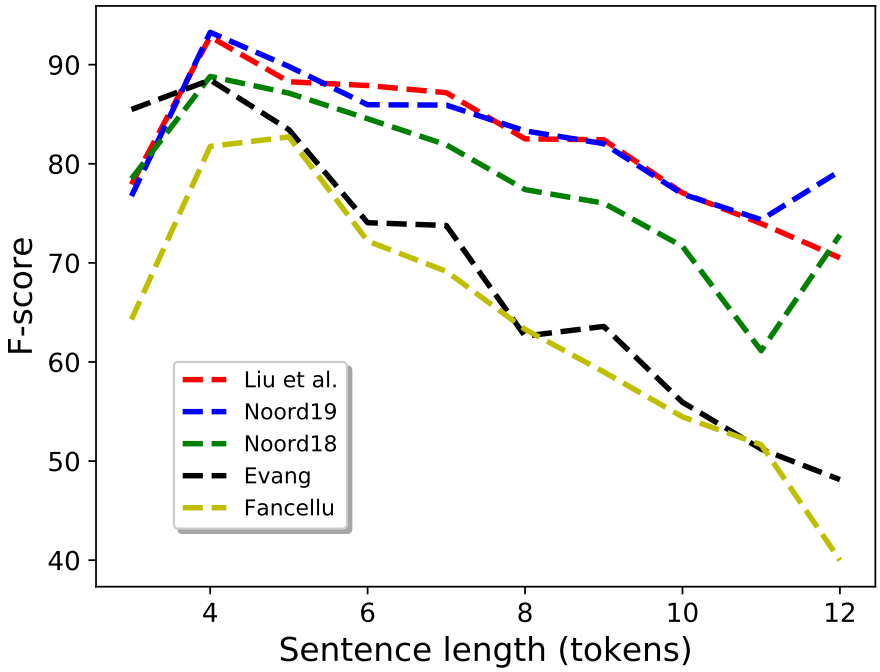

Figure 6: Performance of the systems per sentence length (punctuation are counted as tokens).

sentences. However, for DRS parsing this does not seem to be the case, as LIU ET AL. shows a similar decrease in performance as the neural models of NOORD ET AL.18 and NOORD ET AL.19.

How similar were the outputs of the participating systems to each other? Table 5 shows pairwise comparison of the outputs of the systems on the evaluation set. The only system that has a substantially higher similarity to one of the systems than their official F-score is NOORD ET AL.18 compared to NOORD ET AL.19, which tells us the models make similar mistakes. This makes sense given that they

Table 5: F-scores of all systems compared to each other.

	LIU	NOORD19	NOORD18	EVANG	FANCELLU
LIU		83.4	79.6	70.5	66.6
NOORD19	83.4		83.3	70.9	68.1
NOORD18	79.6	83.3		71.5	67.1
EVANG	70.5	70.9	71.5		61.3
FANCELLU	66.6	68.1	67.1	61.3	

are both character-level sequence-to-sequence models trained on the same data. Additionally, the output of NOORD ET AL.19 comes closest to the output of FANCELLU ET AL. when compared to other systems' outputs. Similarly, EVANG is most similar to NOORD ET AL.18 than to any other systems.

How complementary were the participating systems to each other? If we had an ensemble system, with an oracle component that selected the best DRS for each sentence out of the participants submissions, it would obtain an F-score of 90.5. When only combining the submissions of LIU ET AL. and NOORD ET AL.19, it would already result in an F-score of 89.1. This suggests that the neural models in fact do learn different things, though there is still a significant portion that both methods could not learn.

Finally, are there phenomena that are especially hard for all participating systems? This is not an easy question to answer, and here we show just a first step to such an analysis. Table 6 shows sentences for which systems, on average, performed badly. Some of them show non-standard use of English, others are phenomena that are relatively rare, such as generics, multi-word expressions, and coordination.

Table 6: Sentences for which participating systems, on average, produced the worst DRSs

Sentence	**avg. F**	**Comment**
Thou speakest.	21.4	archaic English
I dinnae ken.	21.8	Scottish
My fault.	24.2	noun phrase
A cat has two ears.	38.1	generic
I look down on liars and cheats.	40.3	coordination, MWE
Get me the number of this young girl.	41.8	imperative
She attends school at night.	44.6	temporal modifier
The union of Scotland and England took place in 1706.	46.4	coordination, MWE
Something I hadn't anticipated happened.	47.0	reduced relative clause
Charles I had his head cut off.	47.2	ordinal, MWE

9 Conclusion

The first shared task on DRS parsing was successful. It improved the state-of-the-art in DRS parsing, and the variety in methods used (models based on recursive neural networks, transformer models, models based on transition-based parsing, graph decoders) gives inspiration for future research. In the future DRS parsing will be made more challenging by moving to longer sentences and texts (where we expect simple seq2seq models to have a harder time), more complex phenomena (ellipsis, comparatives, multi-word expressions), and to languages other than English.

Acknowledgements

We would like to thank the IWCS-2019 organizers, Stergios Chatzikyriakidis, Vera Demberg, Simon Dobnik and Asad Sayeed, for hosting the shared task in Gothenburg. We also would like to thank all participants for their feedback on the task. This work was funded by the NWO-VICI grant "Lost in Translation – Found in Meaning" (288-89-003).

References

Abzianidze, L., J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D.-D. Nguyen, and J. Bos (2017, April). 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*, Valencia, Spain, pp. 242–247. Association for Computational Linguistics.

Abzianidze, L. and J. Bos (2017, September). Towards universal semantic tagging. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, Montpellier, France. Association for Computational Linguistics.

Allen, J. F., M. Swift, and W. de Beaumont (2008). Deep Semantic Analysis of Text. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 343–354. College Publications.

Asher, N. and A. Lascarides (2003). *Logics of conversation*. Studies in natural language processing. Cambridge University Press.

Basile, V. and J. 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*, Sofia, Bulgaria, pp. 1–9.

Bjerva, J., J. Bos, R. Van der Goot, and M. Nissim (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 642–646.

Blackburn, P. and J. Bos (2005). *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.

Bonial, C., W. J. Corvey, M. Palmer, V. Petukhova, and H. Bunt (2011). A hierarchical unification of LIRICS and VerbNet semantic roles. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*, pp. 483–489.

Bos, J. (2008). Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 277–286. College Publications.

Bos, J. (2015). Open-domain semantic parsing with Boxer. In B. Megyesi (Ed.), *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 301–304.

Bos, J. (2016). Expressive power of Abstract Meaning Representations. *Computational Linguistics 42*(3), 527–535.

Bos, J. (2017, September). Indexicals and compositionality: Inside-out or outside-in? In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers*, Montpellier, France. Association for Computational Linguistics.

Bos, J., S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier (2004). Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 1240–1246.

Bos, J. and K. Markert (2005). Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 628–635.

Cai, S. and K. Knight (2013, August). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 748–752. Association for Computational Linguistics.

Dagan, I., D. Roth, M. Sammons, and F. M. Zanzotto (2013). *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Dyer, C., A. Kuncoro, M. Ballesteros, and N. A. Smith (2016). Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 199–209.

Evang, K. (2019). Transition-based DRS parsing using stack-LSTMs. In *Proceedings of the IWCS 2019 Shared Task on Semantic Parsing*.

Fellbaum, C. (Ed.) (1998). *WordNet. An Electronic Lexical Database*. Cambridge, Ma., USA: The MIT Press.

Groschwitz, J., M. Lindemann, M. Fowlie, M. Johnson, and A. Koller (2018). Amr dependency parsing with a typed semantic algebra. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1831–1841.

Hershcovich, D., Z. Aizenbud, L. Choshen, E. Sulem, A. Rappoport, and O. Abend (2019). Semeval 2019 task 1: Cross-lingual semantic parsing with UCCA.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch (2018, July). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, pp. 116–121. Association for Computational Linguistics.

Kamp, H. and U. Reyle (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Dordrecht: Kluwer.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. Rush (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada, pp. 67–72. Association for Computational Linguistics.

Le, P. and W. Zuidema (2012). Learning compositional semantics for open domain semantic parsing. In *Proceedings of COLING 2012*, Mumbai, India, pp. 1535–1552. The COLING 2012 Organizing Committee.

Liu, J., S. Cohen, and M. Lapata (2019). Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS 2019 Shared Task on Semantic Parsing*.

Liu, J., S. B. Cohen, and M. Lapata (2018). Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, Melbourne, Australia, pp. 429–439.

Marelli, M., S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli (2014). A sick cure for the evaluation of compositional distributional semantic models. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

May, J. (2016, June). Semeval-2016 task 8: Meaning representation parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, pp. 1063–1073. Association for Computational Linguistics.

May, J. and J. Priyadarshi (2017, August). Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, pp. 536–545. Association for Computational Linguistics.

Nissim, M., L. Abzianidze, K. Evang, R. van der Goot, H. Haagsma, B. Plank, and M. Wieling (2017). Sharing is caring: The future of shared tasks. *Computational Linguistics 43*(4), 897–904.

Noreen, E. W. (1989). *Computer-intensive Methods for Testing Hypotheses*. Wiley New York.

Oepen, S., M. Kuhlmann, Y. Miyao, D. Zeman, S. Cinkova, D. Flickinger, J. Hajic, and Z. Uresova (2015, June). SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, pp. 915–926. Association for Computational Linguistics.

Oepen, S., M. Kuhlmann, Y. Miyao, D. Zeman, D. Flickinger, J. Hajic, A. Ivanova, and Y. Zhang (2014, August). SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 63–72. Association for Computational Linguistics.

Pennington, J., R. Socher, and C. Manning (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar, pp. 1532–1543.

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

Van Noord, R. (2019). Neural Boxer at the IWCS shared task on DRS parsing. In *Proceedings of the IWCS 2019 Shared Task on Semantic Parsing*.

van Noord, R., L. Abzianidze, H. Haagsma, and J. Bos (2018). Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Van Noord, R., L. Abzianidze, A. Toral, and J. Bos (2018). Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics 6*, 619–633.

van Noord, R. and J. Bos (2017). Neural semantic parsing by character-based translation: Experiments with Abstract Meaning Representations. *Computational Linguistics in the Netherlands Journal 7*, 93–108.

Van Noord, R., A. Toral, and J. Bos (2019). Linguistic information in neural semantic parsing with multiple encoders. In *IWCS 2019-13th International Conference on Computational Semantics-Short papers*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Venhuizen, N. J., J. Bos, P. Hendriks, and H. Brouwer (2018). Discourse semantics with information structure. *Journal of Semantics*.