SemEval-2019 Task 1
Cross-lingual Semantic Parsing with UCCA
Hershcovich, Daniel; Aizenbud, Zohar; Choshen, Leshem; Sulem, Elior; Rappoport, Ari; Abend, Omri

Published in:
Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019),

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
SemEval-2019 Task 1:
Cross-lingual Semantic Parsing with UCCA

Daniel Hershcovich  
Elior Sulem  
Zohar Aizenbud  
Ari Rappoport  
Leshem Choshen  
Omri Abend

School of Computer Science and Engineering  
Hebrew University of Jerusalem

{danielh,zohara,borgr,eliors,arir,oabend}@cs.huji.ac.il

Abstract

We present the SemEval 2019 shared task on Universal Conceptual Cognitive Annotation (UCCA) parsing in English, German and French, and discuss the participating systems and results. UCCA is a cross-linguistically applicable framework for semantic representation, which builds on extensive typological work and supports rapid annotation. UCCA poses a challenge for existing parsing techniques, as it exhibits reentrancy (resulting in DAG structures), discontinuous structures and non-terminal nodes corresponding to complex semantic units. The shared task has yielded improvements over the state-of-the-art baseline in all languages and settings. Full results can be found in the task’s website https://competitions.codalab.org/competitions/19160.

1 Overview

Semantic representation is receiving growing attention in NLP in the past few years, and many proposals for semantic schemes have recently been put forth. Examples include Abstract Meaning Representation (AMR; Banarescu et al., 2013), Broad-coverage Semantic Dependencies (SDP; Oepen et al., 2016), Universal Decompositional Semantics (UDS; White et al., 2016), Parallel Meaning Bank (Abzianidze et al., 2017), and Universal Conceptual Cognitive Annotation (UCCA; Abend and Rappoport, 2013). These advances in semantic representation, along with corresponding advances in semantic parsing, can potentially benefit essentially all text understanding tasks, and have already demonstrated applicability to a variety of tasks, including summarization (Liu et al., 2015; Dohare and Karnick, 2017), paraphrase detection (Issa et al., 2018), and semantic evaluation (using UCCA; see below). In this shared task, we focus on UCCA parsing in multiple languages.

One of our goals is to benefit semantic parsing in languages with less annotated resources by making use of data from more resource-rich languages. We refer to this approach as cross-lingual parsing, while other works (Zhang et al., 2017, 2018) define cross-lingual parsing as the task of parsing text in one language to meaning representation in another language.

In addition to its potential applicative value, work on semantic parsing poses interesting algorithmic and modeling challenges, which are often different from those tackled in syntactic parsing, including reentrancy (e.g., for sharing arguments across predicates), and the modeling of the interface with lexical semantics.

UCCA is a cross-linguistically applicable semantic representation scheme, building on the established Basic Linguistic Theory typological framework (Dixon, 2010b,a, 2012). It has demonstrated applicability to multiple languages, including English, French and German, and pilot annotation projects were conducted on a few languages more. UCCA structures have been shown to be well-preserved in translation (Sulem et al., 2015), and to support rapid annotation by non-experts, assisted by an accessible annotation interface (Abend et al., 2017).¹ UCCA has already shown applicative value for text simplifica-

¹https://github.com/omriabnd/UCCA-App
Scene Elements

P Process The main relation of a Scene that evolves in time (usually an action or movement).
S State The main relation of a Scene that does not evolve in time.
A Participant Scene participant (including locations, abstract entities and Scenes serving as arguments).
D Adverbial A secondary relation in a Scene.

Elements of Non-Scene Units

C Center Necessary for the conceptualization of the parent unit.
E Elaborator A non-Scene relation applying to a single Center.
N Connector A non-Scene relation applying to two or more Centers, highlighting a common feature.
R Relator All other types of non-Scene relations: (1) Rs that relate a C to some super-ordinate relation, and (2) Rs that relate two Cs pertaining to different aspects of the parent unit.

Inter-Scene Relations

H Parallel Scene A Scene linked to other Scenes by regular linkage (e.g., temporal, logical, purposive).
L Linker A relation between two or more Hs (e.g., “when”, “if”, “in order to”)
G Ground A relation between the speech event and the uttered Scene (e.g., “surprisingly”).

Other

F Function Does not introduce a relation or participant. Required by some structural pattern.

Table 1: The complete set of categories in UCCA’s foundational layer.

UCCA represents the semantics of linguistic utterances as directed acyclic graphs (DAGs), where terminal (childless) nodes correspond to the text tokens, and non-terminal nodes to semantic units that participate in some super-ordinate relation. Edges are labeled, indicating the role of a child in the relation the parent represents. Nodes and edges belong to one of several layers, each corresponding to a “module” of semantic distinctions.

UCCA’s foundational layer covers the predicate-argument structure evoked by predicates of all grammatical categories (verbal, nominal, adjectival and others), the inter-relations between them, and other major linguistic phenomena such as semantic heads and multi-word expressions. It is the only layer for which annotated corpora exist at the moment, and is thus the target of this shared task. The layer’s basic notion is the Scene, describing a state, action, movement or some other relation that evolves in time. Each Scene contains one main relation (marked as either a Process or a State), as well as one or more Participants. For example, the sentence “After graduation, John moved to Paris” (Figure 1) contains two Scenes, whose main relations are “graduation” and “moved”. “John” is a Participant in both Scenes, while “Paris” only in the latter. Further categories account for inter-Scene relations and the internal structure of complex arguments and relations (e.g., coordination and multi-word expressions). Table 1 provides a concise description of the categories used by the UCCA foundational layer.

UCCA distinguishes primary edges, corresponding to explicit relations, from remote edges (appear dashed in Figure 1) that allow for a unit to participate in several super-ordinate relations. Primary edges form a tree in each layer, whereas remote edges enable reentrancy, forming a DAG.

UCCA graphs may contain implicit units with no correspondent in the text. Figure 2 shows the annotation for the sentence “A similar technique is almost impossible to apply to other crops, such as cotton, soybeans and rice.” It includes a single Scene, whose main relation is “apply”, a secondary relation “almost impossible”, as well as two complex arguments: “a similar technique” and the coordinated argument “such as cotton, soybeans, and rice.” In addition, the Scene includes an implicit argument, which represents the agent of the “apply” relation.

While parsing technology is well-established
for syntactic parsing, UCCA has several formal properties that distinguish it from syntactic representations, mostly UCCA’s tendency to abstract away from syntactic detail that do not affect argument structure. For instance, consider the following examples where the concept of a Scene has a different rationale from the syntactic concept of a clause. First, non-verbal predicates in UCCA are represented like verbal ones, such as when they appear in copula clauses or noun phrases. Indeed, in Figure 1, “graduation” and “moved” are considered separate Scenes, despite appearing in the same clause. Second, in the same example, “John” is marked as a (remote) Participant in the graduation Scene, despite not being explicitly mentioned. Third, consider the possessive construction in “John’s trip home”. While in UCCA “trip” evokes a Scene in which “John” is a Participant, a syntactic scheme would analyze this phrase similarly to “John’s shoes”.

The differences in the challenges posed by syntactic parsing and UCCA parsing, and more generally by semantic parsing, motivate the development of targeted parsing technology to tackle it.

3 Data & Resources

All UCCA corpora are freely available. For English, we use v1.2.3 of the Wikipedia UCCA corpus (Wiki), v1.2.2 of the UCCA Twenty Thousand Leagues Under the Sea English-French parallel corpus (20K), which includes UCCA manual annotation for the first five chapters in French and English, and v1.0.1 of the UCCA German Twenty Thousand Leagues Under the Sea corpus, which includes the entire book in German. For consistent annotation, we replace any Time and Quantifier labels with Adverbial and Elaborator in these data sets. The resulting training, development and test sets are publicly available, and the splits are given in Table 2. Statistics on various structural properties are given in Table 3.

The corpora were manually annotated according to v1.2 of the UCCA guidelines, and reviewed by a second annotator. All data was passed through automatic validation and normalization scripts. The goal of validation is to rule out cases that are inconsistent with the UCCA annotation guidelines. For example, a Scene, defined by the presence of a Process or a State, should include at least one Participant.

Due to the small amount of annotated data available for French, we only provided a minimal training set of 15 sentences, in addition to the development and test set. Systems for French were expected to pursue semi-supervised approaches, such as cross-lingual learning or structure projection, leveraging the parallel nature of the corpus, or to rely on datasets for related formalisms, such as Universal Dependencies (Nivre et al., 2016). The full unannotated 20K Leagues corpus in English and French was released as well, in order to facilitate pursuing cross-lingual approaches.

Datasets were released in an XML format, including tokenized text automatically pre-
### Table 2: Data splits of the corpora used for the shared task.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train/Trial</th>
<th>Dev</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sentences</td>
<td>Tokens</td>
<td>Sentences</td>
<td>Tokens</td>
</tr>
<tr>
<td>English-Wiki</td>
<td>4,113</td>
<td>124,935</td>
<td>514</td>
<td>17,784</td>
</tr>
<tr>
<td>English-20K</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>French-20K</td>
<td>15</td>
<td>618</td>
<td>238</td>
<td>6,374</td>
</tr>
<tr>
<td>German-20K</td>
<td>5,211</td>
<td>119,872</td>
<td>651</td>
<td>12,334</td>
</tr>
</tbody>
</table>

### Table 3: Statistics of the corpora used for the shared task.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>En-Wiki</th>
<th>En-20K</th>
<th>Fr-20K</th>
<th>De-20K</th>
</tr>
</thead>
<tbody>
<tr>
<td># passages</td>
<td>367</td>
<td>154</td>
<td>154</td>
<td>367</td>
</tr>
<tr>
<td># sentences</td>
<td>5,141</td>
<td>492</td>
<td>492</td>
<td>6,514</td>
</tr>
<tr>
<td># tokens</td>
<td>158,739</td>
<td>12,638</td>
<td>13,021</td>
<td>144,529</td>
</tr>
<tr>
<td># non-terminals</td>
<td>62,002</td>
<td>4,699</td>
<td>5,110</td>
<td>51,934</td>
</tr>
<tr>
<td>% discontinuous</td>
<td>1.71</td>
<td>3.19</td>
<td>4.64</td>
<td>8.87</td>
</tr>
<tr>
<td>% reentrant</td>
<td>1.84</td>
<td>0.89</td>
<td>0.65</td>
<td>0.31</td>
</tr>
<tr>
<td># edges</td>
<td>208,937</td>
<td>16,803</td>
<td>17,520</td>
<td>187,533</td>
</tr>
<tr>
<td>% primary</td>
<td>97.40</td>
<td>96.79</td>
<td>97.02</td>
<td>97.32</td>
</tr>
<tr>
<td>% remote</td>
<td>2.60</td>
<td>3.21</td>
<td>2.98</td>
<td>2.68</td>
</tr>
<tr>
<td>by category</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Participant</td>
<td>17.17</td>
<td>18.1</td>
<td>17.08</td>
<td>19.86</td>
</tr>
<tr>
<td>% Center</td>
<td>18.74</td>
<td>16.31</td>
<td>18.03</td>
<td>14.32</td>
</tr>
<tr>
<td>% Adverbial</td>
<td>3.65</td>
<td>5.24</td>
<td>4.15</td>
<td>5.67</td>
</tr>
<tr>
<td>% Elaborator</td>
<td>18.98</td>
<td>18.06</td>
<td>18.65</td>
<td>14.88</td>
</tr>
<tr>
<td>% Function</td>
<td>3.38</td>
<td>3.58</td>
<td>2.58</td>
<td>2.98</td>
</tr>
<tr>
<td>% Ground</td>
<td>0.03</td>
<td>0.56</td>
<td>0.37</td>
<td>0.57</td>
</tr>
<tr>
<td>% Parallel Scene</td>
<td>6.02</td>
<td>6.3</td>
<td>6.15</td>
<td>7.54</td>
</tr>
<tr>
<td>% Linker</td>
<td>2.19</td>
<td>2.66</td>
<td>2.57</td>
<td>2.49</td>
</tr>
<tr>
<td>% Connector</td>
<td>1.26</td>
<td>0.93</td>
<td>0.84</td>
<td>0.65</td>
</tr>
<tr>
<td>% Process</td>
<td>7.1</td>
<td>7.51</td>
<td>6.91</td>
<td>7.03</td>
</tr>
<tr>
<td>% Relation</td>
<td>8.58</td>
<td>8.09</td>
<td>9.6</td>
<td>7.54</td>
</tr>
<tr>
<td>% State</td>
<td>1.62</td>
<td>2.1</td>
<td>1.88</td>
<td>3.34</td>
</tr>
<tr>
<td>% Punctuation</td>
<td>11.28</td>
<td>10.55</td>
<td>11.16</td>
<td>13.15</td>
</tr>
</tbody>
</table>

5 Evaluation

**Tracks.** Participants in the task were evaluated in four settings:

1. **English in-domain setting**, using the Wiki corpus.
2. **English out-of-domain setting**, using the Wiki corpus as training and development data, and 20K Leagues as test data.
3. **German in-domain setting**, using the 20K Leagues corpus.
4. **French setting** with no training data, using the 20K Leagues as development and test data.

In order to allow both even ground comparison between systems and using hitherto untried resources, we held both an open and a closed track for submissions in the English and German settings. Closed track submissions were only allowed to use the gold-standard UCCA annotation distributed for the task in the target language, and were limited in their use of additional resources. Concretely, the only additional data they were allowed to use is that used by TUPA, which consists of automatic annotations provided by spaCy: POS tags, syntactic dependency relations, and named entity types and spans. In addition, the closed track only allowed the use of word embeddings provided by fastText (Bojanowski et al., 2017) for all languages.

Systems in the open track, on the other hand, were allowed to use any additional resource, such as UCCA annotation in other languages, dictionaries or datasets for other tasks, provided that they make sure not to use any additional gold standard annotation over the same text used in the UCCA processing using spaCy (see §5), and gold-standard UCCA annotation for the train and development sets. To facilitate the use of existing NLP tools, we also released the data in SDP, AMR, CoNLL-U and plain text formats.

4 TUPA: The Baseline Parser

We use the TUPA parser, the only parser for UCCA at the time the task was announced, as a baseline (Hershcovich et al., 2017, 2018). TUPA is a transition-based DAG parser based on a BiLSTM-based classifier. TUPA in itself has been found superior to a number of conversion-based parsers that use existing parsers for other formalisms to parse UCCA by constructing a two-way conversion protocol between the formalisms. It can thus be regarded as a strong baseline for system submissions to the shared task.

8 https://github.com/UniversalConceptualCognitiveAnnotation/docs/blob/master/FORMAT.md
9 https://github.com/huji-nlp/tupa
10 http://spacy.io
11 http://fasttext.cc
corpora. In both tracks, we required that submitted systems are not trained on the development data. We only held an open track for French, due to the paucity of training data. The four settings and two tracks result in a total of 7 competitions.

**Scoring.** The following scores an output graph $G_1 = (V_1, E_1)$ against a gold one, $G_2 = (V_2, E_2)$, over the same sequence of terminals (tokens) $W$. For a node $v$ in $V_1$ or $V_2$, define $yield(v) \subseteq W$ as its set of terminal descendants. A pair of edges $(v_1, u_1) \in E_1$ and $(v_2, u_2) \in E_2$ with labels (categories) $\ell_1, \ell_2$ is matching if $yield(u_1) = yield(u_2)$ and $\ell_1 = \ell_2$. Labeled precision and recall are defined by dividing the number of matching edges in $G_1$ and $G_2$ by $|E_1|$ and $|E_2|$, respectively. $F_1$ is their harmonic mean:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Unlabeled precision, recall and $F_1$ are the same, but without requiring that $\ell_1 = \ell_2$ for the edges to match. We evaluate these measures for primary and remote edges separately. For a more fine-grained evaluation, we additionally report precision, recall and $F_1$ on edges of each category. \(^13\)

6 **Participating Systems**

We received a total of eight submissions to the different tracks: *MaskParse@Deskīn* (Marzinotto et al., 2019) from Orange Labs and Aix-Marseille University, *HLT@SUDA* (Jiang et al., 2019) from Soochow University, *TūPa* (Pütz and Glocker, 2019) from the University of Tübingen, *UC Davis* (Yu and Sagae, 2019) from the University of California, Davis, *GCN-SEM* (Taslimipoor et al., 2019) from the University of Wolverhampton, *CUNY-PekingU* (Lyu et al., 2019) from the City University of New York and Peking University, *DANGNT@UIT.VNU-HCM* (Nguyen and Tran, 2019) from the University of Information Technology VNU-HCM, and *XLangMo* from Zhejiang University. Some of the teams participated in more than one track and two systems (*HLT@SUDA* and *CUNY-PekingU*) participated in all the tracks.\(^12\)

In terms of parsing approaches, the task was quite varied. *HLT@SUDA* converted UCCA graphs to constituency trees and trained a constituency parser and a recovery mechanism of remote edges in a multi-task framework. *MaskParse@Deskīn* used a bidirectional GRU tagger with a masking mechanism. *TūPa* and *XLangMo* used a transition-based approach. *UC Davis* used an encoder-decoder architecture. *GCN-SEM* uses a BiLSTM model to predict Semantic Dependency Parsing tags, when the syntactic dependency tree is given in the input. *CUNY-PKU* is based on an ensemble that includes different variations of the TUPA parser. *DANGNT@UIT.VNU-HCM* converted syntactic dependency trees to UCCA graphs.

Different systems handled remote edges differently. *DANGNT@UIT.VNU-HCM* and *GCN-SEM* ignored remote edges. *UC Davis* used a different BiLSTM for remote edges. *HLT@SUDA* marked remote edges when converting the graph to a constituency tree and trained a classification model for their recovery. *MaskParse@Deskīn* handles remote edges by detecting arguments that are outside of the parent’s node span using a detection threshold on the output probabilities.

In terms of using the data, all teams but one used the UCCA XML format, two used the CoNLL-U format, which is derived by a lossy conversion process, and only one team found the other data formats helpful. One of the teams (*MaskParse@Deskīn*) built a new training data adapted to their model by repeating each sentence $N$ times, $N$ being the number of non-terminal nodes in the UCCA graphs. Three of the teams adapted the baseline TUPA parser, or parts of it to form their parser, namely *TūPa*, *CUNY-PekingU* and *XLangMo*; *HLT@SUDA* used a constituency parser (Stern et al., 2017) as a component in their model; *DANGNT@UIT.VNU-HCM* is a rule-based system over the Stanford Parser, and the rest are newly constructed parsers.

All teams found it useful to use external resources beyond those provided by the Shared Task. Four submissions used external embeddings, MUSE (Conneau et al., 2017) in the case of *MaskParse@Deskīn* and *XLangMo*, ELMo (Peters et al., 2018) in the case of *TūPa*, \(^14\) and BERT (Devlin et al., 2018) in the case of *HLT@SUDA*.\(^14\) GCN-Sem used ELMo in the closed tracks, training on the available data.

\(^{12}\)We are not aware of any such annotation, but include this restriction for completeness.

\(^{13}\)The official evaluation script providing both coarse-grained and fine-grained scores can be found in https://github.com/huji-nlp/ucca/blob/master/scripts/evaluate_standard.py.

\(^{14}\)GCN-Sem used ELMo in the closed tracks, training on the available data.
Other resources included additional unlabeled data (TüPa and CUNY-PekingU), a list of multi-word expressions (MaskParse@Deskiiñ), and the Stanford parser in the case of DANGNT@UIT.VNU-HCM. Only CUNY-PKU used the 20K unlabeled parallel data in English and French.

A common trend for many of the systems was the use of cross-lingual projection or transfer (MaskParse@Deskiiñ, HLT@SUDA, TüPa, GCN-Sem, CUNY-PKU and XLangMo). This was necessary for French, and was found helpful for German as well (CUNY-PKU).

7 Results

Table 4 shows the labeled and unlabeled F1 for primary and remote edges, for each system in each track. Overall F1 (All) is the F1 calculated over both primary and remote edges. Full results are available online.\(^{15}\)

Figure 3 shows the fine-grained evaluation by labeled F1 per UCCA category, for each system in each track. While Ground edges were uniformly difficult to parse due to their sparsity in the training data, Relators were the easiest for all systems, as they are both common and predictable. The Process/State distinction proved challenging, and most main relations were identified as the more common Process category. The winning system in most tracks (HLT@SUDA) performed better on almost all categories. Its largest advantage was on Parallel Scenes and Linkers, showing was especially successful at identifying constituency boundaries relative to the other systems, which requires a good understanding of syntax.

8 Discussion

The HLT@SUDA system participated in all the tracks, obtaining the first place in the six English and German tracks and the second place in the French open track. The system is based on the conversion of UCCA graphs into constituency trees, marking remote and discontinuous edges for recovery. The classification-based recovery of the remote edges is performed simultaneously with the constituency parsing in a multi-task learning framework. This work, which further connects between semantic and syntactic parsing, proposes a recovery mechanism that can be applied to other grammatical formalisms, enabling the conversion of a given formalism to another one for parsing. The idea of this system is inspired by the pseudo non-projective dependency parsing approach proposed by Nivre and Nilsson (2005).

The MaskParse@Deskiiñ system only participated to the French open track, focusing on cross-lingual parsing. The system uses a semantic tagger, implemented with a bidirectional GRU and a masking mechanism to recursively extract the inner semantic structures in the graph. Multilingual word embeddings are also used. Using the English and German training data as well as the small French trial data for training, the parser ranked fourth in the French open track with a labeled F1 score of 65.4%, suggesting that this new model could be useful for low-resource languages.

The TüPa system takes a transition-based approach, building on the TUPA transition system and oracle, but modifies its feature representations. Specifically, instead of representing the parser configuration using LSTMs over the partially parsed graph, stack and buffer, they use feed-

---

Table 4: Official F1-scores for each system in each track. Prim.: primary edges, Rem.: remote edges.

<table>
<thead>
<tr>
<th># Team All Prim. Rem.</th>
<th># Team All Prim. Rem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Wiki (closed)</td>
<td>English-Wiki (open)</td>
</tr>
<tr>
<td>1 HLT@SUDA 77.4</td>
<td>1 HLT@SUDA 80.5</td>
</tr>
<tr>
<td>2 baseline 77.0</td>
<td>2 baseline 80.0</td>
</tr>
<tr>
<td>3 Davis 72.2</td>
<td>3 baseline 73.5</td>
</tr>
<tr>
<td>4 CUNY-PekingU 71.8</td>
<td>4 CUNY-PekingU 73.5</td>
</tr>
<tr>
<td>5 DANGNT@UIT. VNU-HCM 70.0</td>
<td>5 DANGNT@UIT. VNU-HCM 70.3</td>
</tr>
<tr>
<td>German-20K (closed)</td>
<td>English-20K (open)</td>
</tr>
<tr>
<td>1 HLT@SUDA 72.7</td>
<td>1 HLT@SUDA 76.7</td>
</tr>
<tr>
<td>2 baseline 73.6</td>
<td>2 baseline 67.2</td>
</tr>
<tr>
<td>3 CUNY-PekingU 66.9</td>
<td>3 baseline 73.9</td>
</tr>
<tr>
<td>4 GCN-Sem 62.6</td>
<td>4 TüPa 70.9</td>
</tr>
<tr>
<td>5 baseline 68.4</td>
<td>5 XLangMo 69.5</td>
</tr>
<tr>
<td>German-20K (open)</td>
<td>English-20K (closed)</td>
</tr>
<tr>
<td>1 HLT@SUDA 81.2</td>
<td>1 HLT@SUDA 84.9</td>
</tr>
<tr>
<td>2 baseline 83.8</td>
<td>2 baseline 84.5</td>
</tr>
<tr>
<td>3 CUNY-PekingU 79.7</td>
<td>3 CUNY-PekingU 81.5</td>
</tr>
<tr>
<td>4 GCN-Sem 73.1</td>
<td>4 GCN-Sem 73.6</td>
</tr>
<tr>
<td>5 baseline 71.0</td>
<td>5 baseline 78.4</td>
</tr>
<tr>
<td>German-20K (open)</td>
<td>French-20K (open)</td>
</tr>
<tr>
<td>1 HLT@SUDA 84.9</td>
<td>1 HLT@SUDA 79.6</td>
</tr>
<tr>
<td>2 baseline 85.4</td>
<td>2 baseline 80.0</td>
</tr>
<tr>
<td>3 CUNY-PekingU 84.5</td>
<td>3 CUNY-PekingU 84.3</td>
</tr>
<tr>
<td>4 baseline 79.1</td>
<td>4 baseline 79.6</td>
</tr>
<tr>
<td>5 XLangMo 78.1</td>
<td>5 XLangMo 78.8</td>
</tr>
<tr>
<td>French-20K (open)</td>
<td>English-Wiki (open)</td>
</tr>
<tr>
<td>1 HLT@SUDA 70.0</td>
<td>1 HLT@SUDA 69.0</td>
</tr>
<tr>
<td>2 baseline 70.7</td>
<td>2 baseline 68.2</td>
</tr>
<tr>
<td>3 Davis 70.3</td>
<td>3 Davis 73.5</td>
</tr>
<tr>
<td>4 CUNY-PekingU 70.3</td>
<td>4 XLangMo 69.5</td>
</tr>
<tr>
<td>5 baseline 71.1</td>
<td>5 baseline 69.4</td>
</tr>
</tbody>
</table>

\(^{15}\)http://bit.ly/semeval2019task1results
Figure 3: Each system’s labeled F1 per UCCA category in each track.
forward networks with ELMo contextualized embeddings. The stack and buffer are represented by the top three items on them. For the partially parsed graph, they extract the rightmost and leftmost parents and children of the respective items, and represent them by the ELMo embedding of their form, the embedding of their dependency heads (for terminals, for non-terminals this is replaced with a learned embedding) and the embeddings of all terminal children. Results are generally on-par with the TUPA baseline, and surpass it from the out-of-domain English setting. This suggests that the TUPA architecture may be simplified, without compromising performance.

The UC Davis system participated only in the English closed track, where they achieved the second highest score, on par with TUPA. The proposed parser has an encoder-decoder architecture, where the encoder is a simple BiLSTM encoder for each span of words. The decoder iteratively and greedily traverses the sentence, and attempts to predict span boundaries. The basic algorithm yields an unlabeled contiguous phrase-based tree, but additional modules predict the labels of the spans, discontiguous units (by joining together spans from the contiguous tree under a new node), and remote edges. The work is inspired by Kitaev and Klein (2018), who used similar methods for constituency parsing.

The GCN-SEM system uses a BiLSTM encoder, and predicts bi-lexical semantic dependencies (in the SDP format) using word, token and syntactic dependency parses. The latter is incorporated into the network with a graph convolutional network (GCN). The team participated in the English and German closed tracks, and were not among the highest-ranking teams. However, scores on the UCCA test sets converted to the bi-lexical CoNLL-U format were rather high, implying that the lossy conversion could be much of the reason.

The CUNY-PKU system was based on an ensemble. The ensemble included variations of TUPA parser, namely the MLP and BiLSTM models (Hershovich et al., 2017) and the BiLSTM model with an additional MLP. The system also proposes a way to aggregate the ensemble going through CKY parsing and accounting for remotes and discontinuous spans. The team participated in all tracks, including additional information in the open domain, notably synthetic data based on automatically translating annotated texts. Their system ranks first in the French open track.

The DANGNT@UIT.VNU-HCM system participated only in the English Wiki open and closed tracks. The system is based on graph transformations from dependency trees into UCCA, using heuristics to create non-terminal nodes and map the dependency relations to UCCA categories. The manual rules were developed based on the training and development data. As the system converts trees to trees and does not add reentrancies, it does not produce remote edges. While the results are not among the highest-ranking in the task, the primary labeled F1 score of 71.1% in the English open track shows that a rule-based system on top of a leading dependency parser (the Stanford parser) can obtain reasonable results for this task.

9 Conclusion

The task has yielded substantial improvements to UCCA parsing in all settings. Given that the best reported results were achieved with different parsing and learning approaches than the baseline model TUPA (which has been the only available parser for UCCA), the task opens a variety of paths for future improvement. Cross-lingual transfer, which capitalizes on UCCA’s tendency to be preserved in translation, was employed by a number of systems and has proven remarkably effective. Indeed, the high scores obtained for French parsing in a low-resource setting suggest that high quality UCCA parsing can be straightforwardly extended to additional languages, with only a minimal amount of manual labor.

Moreover, given the conceptual similarity between the different semantic representations (Abend and Rappoport, 2017), it is likely the parsers developed for the shared task will directly contribute to the development of other semantic parsing technology. Such a contribution is facilitated by the available conversion scripts available between UCCA and other formats.

Acknowledgments

We are deeply grateful to Dotan Dvir and the UCCA annotation team for their diligent work on the corpora used in this shared task.

This work was supported by the Israel Science Foundation (grant No. 929/17), and by the HUJI Cyber Security Research Center in conjunction with the Israel National Cyber Bureau in the Prime Minister’s Office.
References


