Cross-lingual Semantic Representation for NLP with UCCA

Abend, Omri; Dvir, Dotan; Hershcovich, Daniel; Prange, Jakob; Schneider, Nathan

Published in:
Proceedings of the 28th International Conference on Computational Linguistic

Publication date:
2020

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

Document license:
CC BY

Citation for published version (APA):
Tutorial at COLING 2020
Cross-lingual Semantic Representation for NLP with UCCA

Omri Abend*, Dotan Dvir*, Daniel Hershcovich**, Jakob Prange***, and Nathan Schneider***

*Hebrew University of Jerusalem  
**University of Copenhagen  
***Georgetown University

oabend@cs.huji.ac.il, dotan.dvir@mail.huji.ac.il, dh@di.ku.dk  
jakob@cs.georgetown.edu, nathan.schneider@georgetown.edu

Abstract

This is an introductory tutorial to UCCA (Universal Conceptual Cognitive Annotation), a cross-linguistically applicable framework for semantic representation, with corpora annotated in English, German and French, and ongoing annotation in Russian and Hebrew. UCCA builds on extensive typological work and supports rapid annotation. The tutorial will provide a detailed introduction to the UCCA annotation guidelines, design philosophy and the available resources; and a comparison to other meaning representations. It will also survey the existing parsing work, including the findings of three recent shared tasks, in SemEval and CoNLL, that addressed UCCA parsing. Finally, the tutorial will present recent applications and extensions to the scheme, demonstrating its value for natural language processing in a range of languages and domains.

1 Introduction

Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013), abbreviated as “UCCA”, is a symbolic meaning representation (MR) that supports human annotation of text with broad coverage. While several meaning representation schemes share this goal (Abend and Rappoport, 2017), UCCA targets a level of semantic granularity that abstracts away from syntactic paraphrases in a typologically-motivated, cross-linguistic fashion, building on Basic Linguistic Theory (Dixon, 20102012), an influential framework for linguistic description. The scheme does not rely on language-specific resources, and sets a low threshold for annotator training.

UCCA has been annotated on several corpora of different genres and languages[1] as summarized in table [1]. Pilot studies have been conducted in additional languages. A web-based annotation system is available (Abend et al., 2017).

In UCCA, an analysis of a text passage is a directed acyclic graph over semantic elements called units. The principal kind of unit is a scene, which describes an action, movement or state, and is similar to FrameNet’s notion of a frame. Figure [1] contains three scenes, evoked, respectively, by the verb took, the noun phrase a repair, and the possessive our. Several elements are exemplified, including participants, secondary relations, and scene linkage. The graph is anchored in the text tokens (the leaves generally correspond to one or more tokens), and relations between units are indicated by the categories assigned to the edges connecting them.

The goals of this tutorial are: to describe the UCCA representation as a linguistic scheme and how it is being used computationally, especially for cross-lingual and multilingual NLP; to familiarize participants with existing UCCA parsers and equip them with the conceptual tools required for designing new parsers; and to review existing extensions and possible future directions [2].

2 Relevance

UCCA resources and applications are valuable for cross-lingual NLP: like Universal Dependencies (Nivre et al., 2020), UCCA’s category set can in principle be applied to a wide variety of languages.

Table 1: Data statistics for existing UCCA corpora.

<table>
<thead>
<tr>
<th>Genre</th>
<th># Tokens</th>
<th># Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiki Encyclopedia</td>
<td>158,573</td>
<td>5,141</td>
</tr>
<tr>
<td>20k Literary Online</td>
<td>12,574</td>
<td>492</td>
</tr>
<tr>
<td>EWT Online reviews</td>
<td>55,590</td>
<td>3,813</td>
</tr>
<tr>
<td>WSJ Financial news</td>
<td>2,273</td>
<td>100</td>
</tr>
<tr>
<td>LPP Literary Online</td>
<td>1,322</td>
<td>100</td>
</tr>
<tr>
<td><strong>German</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20k Literary Online</td>
<td>144,531</td>
<td>6,510</td>
</tr>
<tr>
<td>LPP Literary Online</td>
<td>18,653</td>
<td>1,042</td>
</tr>
<tr>
<td><strong>French</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20k Literary Online</td>
<td>12,954</td>
<td>492</td>
</tr>
<tr>
<td><strong>Hebrew, Russian</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPP Literary Online</td>
<td></td>
<td>annotation underway</td>
</tr>
</tbody>
</table>

Figure 1: Example sentence from EWT (reviews-086839-0003), with its UCCA annotation. Category abbreviations: H = parallel scene, L = scene linker, P = process (dynamic event), S = state, A = scene participant, D = scene adverbial, E = non-scene elaborator, C = center (non-scene head), R = relator, F = functional element.

It is also cross-linguistically stable, and reflects a level of semantic structure that is usually preserved in translations (Sulem et al., 2015). UCCA has been applied in NLP to text simplification (Sulem et al., 2018b, 2020), and text-to-text generation evaluation (Birch et al., 2016, Mareček et al., 2017, Choshen and Abend, 2018, Sulem et al., 2018b, Alva-Manchego et al., 2019, Xu et al., 2020). The tutorial will describe the guidelines and rationale behind UCCA, helping potential application designers understand what abstractions it makes.

Significant effort has been devoted to building UCCA parsers (Hershcovich et al., 2017, Hershcovich et al., 2018, Jiang et al., 2019, Lyu et al., 2019, Tuan Nguyen and Tran, 2019, Taslimipoor et al., 2019, Marzinotto et al., 2019, Pütz and Glocker, 2019, Yu and Sagae, 2019, Zhang et al., 2019a, Hershcovich and Arviv, 2019, Li et al., 2019, Donatelli et al., 2019, Che et al., 2019, Bai and Zhao, 2019, Lai et al., 2019, Koreeda et al., 2019, Straka and Straková, 2019, Cao et al., 2019, Zhang et al., 2019b, Droganova et al., 2019, Chen et al., 2019, Arviv et al., 2020, Samuel and Straka, 2020, Dou et al., 2020). including a SemEval 2019 shared task on cross-lingual UCCA parsing (Hershcovich et al., 2019b), which had 8 participating teams, as well as CoNLL 2019 and CoNLL 2020 shared tasks on cross-framework and cross-lingual meaning representation parsing (Oepen et al., 2019, Oepen et al., 2020), where 12 and 4 teams, respectively, submitted parsed UCCA graphs. This tutorial will allow researchers interested in UCCA parsing, and more generally graph parsing, deepen their understanding of the framework, and what properties make it unique. The tutorial will include a brief survey of the various approaches taken by existing parsers, and prepare attendees to work on UCCA parsing themselves.

Furthermore, UCCA parsing has been shown to benefit from multi-task learning (Caruana, 1997) with
other meaning representations (Hershcovich et al., 2018), although results from the CoNLL 2019 and CoNLL 2020 shared tasks (Oepen et al., 2019; Oepen et al., 2020) show that multi-task meaning representation parsing is difficult. The tutorial will compare and contrast UCCA and other meaning representations, and will thereby inform participants of the potential advantages and difficulties in employing multi-task learning across semantic schemes.

UCCA defines a small inventory of coarse-grained categories so as not to rely on language-specific lexical resources, and can thus in principle be applied to a great variety of languages. This distinguishes UCCA from finer-grained sentence-structural representations like FrameNet (Baker et al., 1998), the Abstract Meaning Representation (Banarescu et al., 2013), which relies on PropBank (Palmer et al., 2005), and Universal Decompositional Semantics (White et al., 2016). For example, FrameNet requires a different ontology for each new language addressed (Ohara et al., 2003; You and Liu, 2005; Borin et al., 2013; Park et al., 2014; Hayoun and Elhadad, 2016; Djemaa et al., 2016), and AMR underwent significant customization to be applicable to Chinese (Li et al., 2016). Decomp takes a different approach to multilinguality, where the parser is required to parse sentences in other languages to their corresponding English semantic forms (Zhang et al., 2018). The tutorial will address contemporary issues in the field, such as the question of how to represent semantic structure multilingually with broad coverage, which is actively being explored from many angles.

While UCCA structures and categories are intentionally coarse, the scheme has a multi-layered architecture, which allows for refinement using additional layers, which serve as “modules” of semantic distinctions. We will give an overview of the recently proposed extensions (to support coreference) and joint parsing experiments (Prange et al., 2019a; Prange et al., 2019b).

3 Agenda

The planned division of time is as follows:

1. Bird’s eye view (45m). Design philosophy, notion of scenes, basic explanation of categories, simple examples.
2. Annotation guidelines (35m). Linguistic details, interesting constructions in several languages.
3. Data and annotation (10m). Overview of annotated data (see §1) and the annotation process and software (Abend et al., 2017).
4. Extensions to UCCA and integration with other schemes (15m). Semantic roles (Prange et al., 2019a), coreference (Prange et al., 2019b), and fine-grained implicit arguments (Cui and Hershcovich, 2020).
5. Relation to other representations (15m). Comparison to other meaning representations (Abend and Rappoport, 2017; Koller et al., 2019; Hershcovich et al., 2020) and to UD (Hershcovich et al., 2019a).
6. Parsing (25m). TUPA (Hershcovich et al., 2017; Hershcovich et al., 2018; Hershcovich and Arviv, 2019; Arviv et al., 2020), SemEval 2019 Task 1 (Hershcovich et al., 2019b; Jiang et al., 2019), CoNLL 2019 and CoNLL 2020 Shared Tasks (Oepen et al., 2019; Oepen et al., 2020), and more recent parsers (Zhang et al., 2019a).
7. Monolingual tasks and evaluation (20m). Sentence simplification (Sulem et al., 2018a; Sulem et al., 2018b), evaluation of sentence simplification (Sulem et al., 2018a; Alva-Manchego et al., 2019) and grammatical error correction (Choshen and Abend, 2018).
8. Cross-linguistic studies and applications (15m). Analysis of cross-linguistic stability (Sulem et al., 2015), machine translation evaluation (Birch et al., 2016; Mareček et al., 2017; Xu et al., 2020) and simplification for machine translation (Sulem et al., 2020).
3.1 Prerequisites

No prior knowledge is assumed about linguistics and typology. The necessary background will be provided as part of the tutorial. However, participants are expected to know about basic data structures such as trees and graphs. For the parsing section, prior knowledge is assumed about common machine learning techniques, including supervised learning and neural networks.

3.2 Reading list

The following are recommended to read before the tutorial, as they provide background and frame the context in which the tutorial materials lie:

1. Chapter 3 of Dixon (2005) contains an introduction to some basic concepts in semantics on which UCCA is based.

2. Kiperwasser and Goldberg (2016) present a transition-based parser using an architecture on which TUPA, the first UCCA parser, is based (Hershcovich et al., 2017).

3. Peng et al. (2017) performed multi-task learning for meaning representation parsing, inspiring work on cross-framework parsing for UCCA (Hershcovich et al., 2018).

4. Abend and Rappoport (2017) compare and contrast several meaning representations according to various aspects.

5. Deng and Xue (2017) investigate translation divergences using a hierarchical alignment, and discuss bridging them with cross-lingual semantic representations.

6. Croft et al. (2017) list typologically-informed design criteria for Universal Dependencies (Nivre et al., 2020), which are also relevant for other structural representations in NLP.

4 Presenters

The instruction in this tutorial involves organizers at various career stages, exhibiting geographic diversity, and diversity in terms of gender.

Omri Abend ([https://www.cse.huji.ac.il/~oabend](https://www.cse.huji.ac.il/~oabend)) is a Senior Lecturer (Assistant Professor) of Computer Science and Cognitive Science at the Hebrew University of Jerusalem. Research interests: computational semantics and specifically, cross-linguistically applicable semantic and grammatical representation, semantic parsing, corpus annotation and evaluation. Relevant experience: co-developer of the UCCA scheme, partner in all annotation and application efforts related to UCCA, and in some of the parsing efforts. Publishes regularly in NLP conferences (ACL, NAACL, EMNLP etc.).

Dotan Dvir has been managing the UCCA manual annotation project at the Hebrew University of Jerusalem since 2017. She was involved in writing version 2 of the UCCA guidelines. She has in-depth knowledge of the UCCA guidelines and is experienced in instructing annotators about them. Before joining the UCCA project, she had been working as a text analyst in IBM’s Project Debater (2014-2017).

Daniel Hershcovich ([https://danielhers.github.io](https://danielhers.github.io)) is a Tenure-Track Assistant Professor at the University of Copenhagen, Denmark. Daniel pioneered the work on UCCA parsing, and is interested in semantic parsing and meaning representations. Daniel develops and maintains the UCCA toolkit Python codebase ([https://github.com/danielhers/ucca](https://github.com/danielhers/ucca)) has teaching experience in NLP and ML courses, and publishes in NLP conferences.

Jakob Prange ([https://prange.jakob.georgetown.domains](https://prange.jakob.georgetown.domains)) is pursuing his Ph.D. at Georgetown University, investigating design, annotation, and parsing strategies for various meaning representations. Among other formalisms (SNACS, frame semantics, STAG, CCG), he has studied and worked with UCCA over the past two years, which recently resulted in two published proposals of novel
UCCA extensions, for coreference and semantic roles. He has experience with teaching in multicultural classroom settings and presenting research at international conferences.

Nathan Schneider (http://nathan.cl) leads an interdisciplinary computational linguistics research group at Georgetown University. He has worked on the design and parsing of a range of broad-coverage representations for different aspects and granularities of meaning, including multiword expressions, supersenses, frame semantics, AMR, and UCCA (as a multiyear collaboration with the copresenters). He has experience teaching meaning representations in classroom settings as well as conference tutorials—notably, a tutorial on AMR [Schneider et al., 2015] whose materials continue to serve as a useful introduction to the scheme, and will serve as a model for the proposed UCCA tutorial.

References


Lars Borin, Markus Forsberg, and Benjamin Lyngfelt. 2013. 2) close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. Veredas-Revista de Estudos Linguísticos, 17(1-).


Yue Zhang, Wei Jiang, Qingrong Xia, Junjie Cao, Rui Wang, Zhenghua Li, and Min Zhang. 2019b. SUDA–Alibaba at MRP 2019: Graph-based models with BERT. In Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Computational Natural Language Learning, pages 149 – 157, Hong Kong, China.