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

We propose 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 two 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 (UCCA; Abend and Rappoport, 2013) 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, 2010/2012), 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 desigining new parsers; and to review existing extensions and possible future directions.

2 Relevance

UCCA resources and applications are valuable for cross-lingual NLP: like Universal Dependencies (UD; Nivre et al., 2019), UCCA's category set can in principle be applied to a wide variety of languages. 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), and text-to-text generation evaluation (Birch et al., 2016; Mareček et al., 2017; Choshen and Abend, 2018; Sulem et al., 2018a; Alva-Manchego et al., 2019). The tuto-

[1] https://github.com/UniversalConceptualCognitiveAnnotation

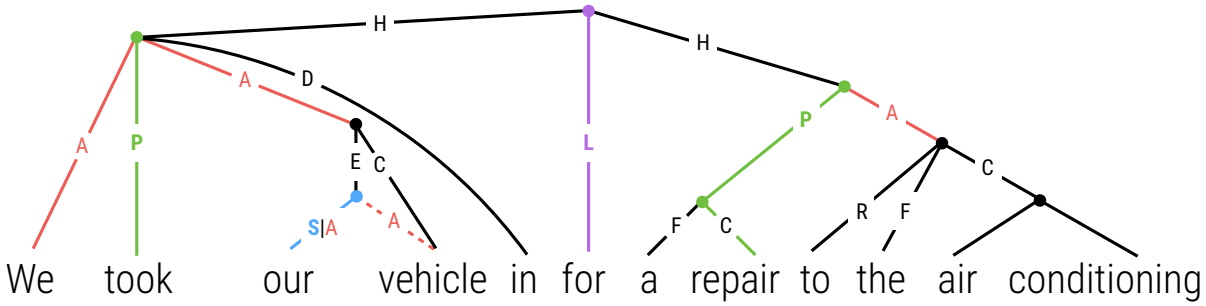

Figure 1: Example sentence from EWT (`reviews-086839-0003`), with its UCCA annotation. Category abbreviations: **H** = parallel scene, **L** = scene linker, **P** = process (dynamic event), **S** = state, **A** = scene participant, **D** = scene adverbial, **E** = non-scene elaborator, **C** = center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

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

Table 1: Data statistics for existing UCCA corpora.

rial 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; Jiang et al., 2019; Zhang et al., 2019), including a SemEval 2019 shared task on cross-lingual UCCA parsing (Hershcovich et al., 2019b), which had 8 participating teams, and a CoNLL 2019 shared task on cross-framework meaning representation parsing (Oepen et al., 2019), where 12 teams 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), and preliminary results from the CoNLL 2019 shared task (Oepen et al., 2019) show that it is also useful as an auxiliary task in itself. 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 (AMR; Banarescu et al., 2013), which relies on PropBank (Palmer et al., 2005), and Universal Decompositional Semantics (Decomp; 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 se-

mantic distinctions. We will give an overview of the recently proposed extensions (to support coreference) and joint parsing experiments (Prange et al., 2019a,b).

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).

COFFEE BREAK

4. **Extensions to UCCA and integration with other schemes (15m).** Semantic roles (Prange et al., 2019a) and coreference (Prange et al., 2019b).

5. **Relation to other representations (15m).** Comparison to other meaning representations (Abend and Rappoport, 2017; Koller et al., 2019) and to UD (Hershcovich et al., 2019a).

6. **Parsing (25m).** TUPA (Hershcovich et al., 2017, 2018), SemEval 2019 Task 1 (Hershcovich et al., 2019b; Jiang et al., 2019), CoNLL 2019 Shared Task (Oepen et al., 2019), and more recent parsers (Zhang et al., 2019).

7. **Monolingual tasks and evaluation (20m).** Sentence simplification (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).

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 constrast 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., 2019), which are also relevant for other structural representations in NLP.

4 Logistics

We are prepared to present this tutorial at ACL 2020 or COLING 2020. ACL would be our preferred venue. We estimate around 200 attendees, based on previous similar tutorials (Schneider et al., 2015; Koller et al., 2019).

The demonstration of the annotation process with UCCAApp (Abend et al., 2017) will require internet access and that participants bring laptops.

We will allow the publication of our slides and video recording of our tutorial in the ACL Anthology. All instructional materials will be openly available.

5 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) 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) is a postdoctoral researcher 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,[2] has teaching experience in an NLP course at the Hebrew University of Jerusalem, and publishes in NLP conferences.

Jakob Prange (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[3] continue to serve as a useful introduction to the scheme, and will serve as a model for the proposed UCCA tutorial.

[2] https://github.com/danielhers/ucca

[3] https://github.com/nschneid/amr-tutorial/

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proc. of ACL*, pages 228–238.

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proc. of ACL*, pages 77–89.

Omri Abend, Shai Yerushalmi, and Ari Rappoport. 2017. UCCAApp: Web-application for syntactic and semantic phrase-based annotation. *Proc. of ACL System Demonstrations*, pages 109–114.

Joakim Nivre et al. 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier Automatic Sentence Simplification Evaluation. In *To Appear in EMNLP-ICJNLP 2019: System Demonstrations*, Hong Kong, China.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *ACL-COLING '98*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the Linguistic Annotation Workshop*.

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. 2016. HUME: Human UCCA-based evaluation of machine translation. In *Proc. of EMNLP*, pages 1264–1274.

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-).

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28(1):41–75.

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. In *Proc. of NAACL-HLT*.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets universal dependencies. In *TLT*, pages 63–75.

Dun Deng and Nianwen Xue. 2017. Translation divergences in chinese–english machine translation: An empirical investigation. *Computational Linguistics*, 43(3):521–565.

R.M.W. Dixon. 2005. *A Semantic Approach to English Grammar*. Oxford University Press, Oxford.

Robert M. W. Dixon. 2010/2012. *Basic Linguistic Theory*. Oxford University Press.

Marianne Djemaa, Marie Candito, Philippe Muller, and Laure Vieu. 2016. Corpus annotation within the French FrameNet: a domain-by-domain methodology. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3794–3801, Portorož, Slovenia. European Language Resources Association (ELRA).

Avi Hayoun and Michael Elhadad. 2016. The Hebrew FrameNet project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4341–4347, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In *Proc. of ACL*, pages 1127–1138.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2018. Multitask parsing across semantic representations. In *Proc. of ACL*, pages 373–385.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2019a. Content differences in syntactic and semantic representation. In *Proc. of NAACL-HLT*.

Daniel Hershcovich, Leshem Choshen, Elior Sulem, Zohar Aizenbud, Ari Rappoport, and Omri Abend. 2019b. SemEval 2019 task 1: Cross-lingual semantic parsing with UCCA. In *Proc. of SemEval*.

Wei Jiang, Zhenghua Li, Yu Zhang, and Min Zhang. 2019. HLT@SUDA at SemEval-2019 Task 1: UCCA graph parsing as constituent tree parsing. In *Proc. of SemEval*, pages 11–15, Minneapolis, Minnesota, USA.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11, Florence, Italy. Association for Computational Linguistics.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

David Mareček, Ondřej Bojar, Ondřej Hübsch, Rudolf Rosa, and Dušan Variš. 2017. CUNI experiments for WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation*, pages 604–611, Copenhagen, Denmark. Association for Computational Linguistics.

Stephan Oepen, Omri Abend, Jan Hajič, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, and Nianwen Xue. 2019. MRP 2019. Cross-framework Meaning Representation Parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1 – 20, Hong Kong, China.

Kyoko Hirose Ohara, Seiko Fujii, Hiroaki Saito, Shun Ishizaki, Toshio Ohori, and Ryoko Suzuki. 2003. The Japanese FrameNet project: A preliminary report. In *Proceedings of pacific association for computational linguistics*, pages 249–254. Citeseer.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Jungyeul Park, Sejin Nam, Youngsik Kim, Younggyun Hahm, Dosam Hwang, and Key-Sun Choi. 2014. Frame-semantic web: a case study for Korean. In *International Semantic Web Conference (Posters & Demos)*, pages 257–260.

Hao Peng, Sam Thomson, and Noah A. Smith. 2017. Deep multitask learning for semantic dependency parsing. In *Proc. of ACL*, pages 2037–2048.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019a. Made for each other: Broad-coverage semantic structures meet preposition supersenses. In *Proc. of CoNLL*. To appear.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019b. Semantically constrained multilayer annotation: The case of coreference. In *Proceedings of the First International Workshop on Designing Meaning Representations (DMR)*, pages 164–176, Florence, Italy. Association for Computational Linguistics.

Nathan Schneider, Jeffrey Flanigan, and Tim O'Gorman. 2015. The logic of AMR: practical, unified, graph-based sentence semantics for NLP. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 4–5, Denver, Colorado, USA.

Elior Sulem, Omri Abend, and Ari Rappoport. 2015. Conceptual annotations preserve structure across translations: A French-English case study. In *Proc. of S2MT*, pages 11–22.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Semantic structural annotation for text simplification. In *Proc. of NAACL*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Simple and effective text simplification using semantic and neural methods. In *Proc. of ACL*.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proc. of EMNLP*, pages 1713–1723.

Liping You and Kaiying Liu. 2005. Building Chinese FrameNet database. *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. Broad-coverage semantic parsing as transduction. In *Proc. of EMNLP-IJCNLP*. To appear.

Sheng Zhang, 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.