Building an Infrastructure for Uniform Meaning Representations (UMRs)

Julia Bonn[1], Matthew Buchholz[1], Jayeol Chun[3], Andrew Cowell[1], William Croft[2], Lukas Denk[2], Sijia Ge[1], Jens E. L. Van Gysel[2], Jan Hajič[4], Kenneth Lai[3], James H. Martin[1], Skatje Myers[1], Alexis Palmer[1], Martha Palmer[1], Benet Post[1], James Pustejovsky[3], Kristine Stenzel[1], Haibo Sun[3], Zdeňka Urešová[4], Rosa Vallejos Yopn[2], Nianwen Xue[3], Jin Zhao[3]

[1]University of Colorado at Boulder, Boulder, CO 80303, USA
[2]University of New Mexico, Albuquerque, NM 87131, USA
[3]Brandeis University, Waltham, MA 02453, USA [4]Charles University, Prague, Czech Republic

University of Colorado Boulder
Brandeis UNIVERSITY

1. Goals of the UMR project

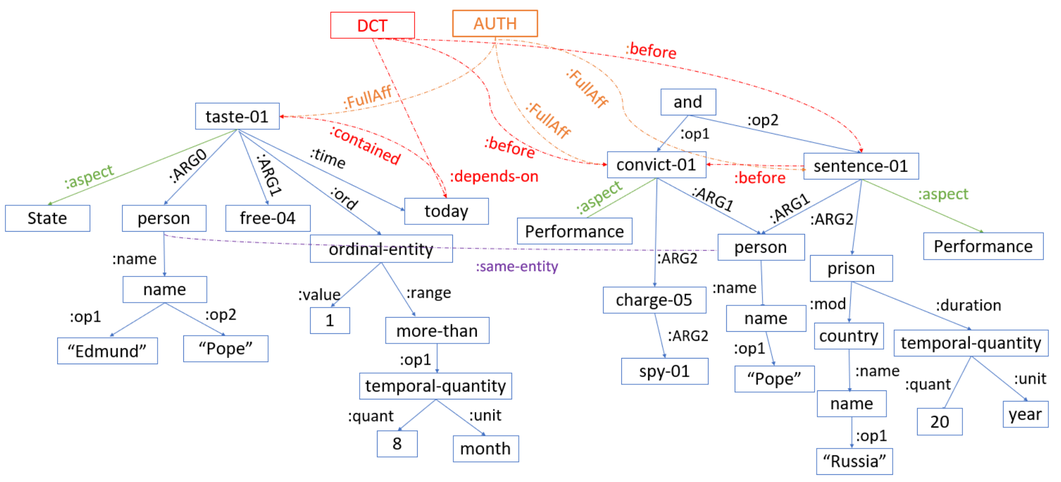
- **UMR-annotated data sets**
  - New data sets for five languages
  - New guidelines for multimodal UMR annotation
  - Multimodal data sets for two languages
  - Semi-automatic conversion of existing English AMR-annotated data sets to UMR
- **Tools to support creation of new data sets**
  - UMR-Writer annotation tool with improved interface
  - Right-to-left language support for UMR-Writer
  - Training materials for UMR annotation and the UMR-Writer tool
  - Platform to host all resources
- **Automated processing**
  - UMR parser to produce meaning representations from textual documents
  - New metrics for interpretable evaluation of UMR parsers
- **Extensions for low-resource languages**
  - Automatic production of partial UMR graphs from language documentation data (IGT)
  - Interface for user configuration of graph production process

2. Motivations

- The motivation of this project is to build the infrastructure for producing data annotated with Uniform Meaning Representations that can be used the NLP community to develop interpretable and controllable AI systems that facilitate information access, human robot communication, for a diverse set of languages, including a few historically under-represented languages.
- This infrastructure of the project, which includes UMR annotation guidelines, data sets, and annotation tools, combined with its outreach efforts, also helps the NLP community to produce UMR-annotated data sets for additional languages.

3. Uniform Meaning Representation (UMR)

- **Graph-based Meaning Representation based on AMR**
  - UMRs are rooted, directed graphs that represent the meaning of text and other modalities (e.g., gesture, videos)
  - UMR concepts are represented as nodes in the graph while UMR relations are represented as edges between them. UMR also represents attributes of concepts.
  - UMR concepts can be *concrete* or *abstract*, with the former being lemmas or sense-disambiguated lemmas and the latter being inferred concepts or types of named entities.
- **Cross-lingual Document-level Representation**
  - UMR represents both predicate-argument structures at the sentence level and coreference, temporal, and modal relations at the document level
  - UMR is designed as a cross-lingual representation based on typological principles and has been tested on languages from diverse language families, including both languages with large number of speakers (Arabic, Chinese, English) and those with small number of speakers (Arapaho, Kukama, Navajo, Quechua, Sanapana)



"Edmund Pope tasted freedom today for the first time in eight months."

"Pope was convicted on spying charges and sentenced to 20 years in a Russian prison."

4. UMR data sets and tools

Data sets

- Planned UMR data sets
  - Text
    - ★ Arabic: 150K words
    - ★ Arapaho: 25K words
    - ★ Chinese: 200K words
    - ★ English: 250K words
    - ★ Quechua: 25K words
  - Multi-modal
    - ★ Arapaho - gesture: 10K
    - ★ English - gesture: 24.5K
- Released UMR data set
  - UMR 1.0 released via LINDAT

Language	sent-level	doc-level
Arapaho	406	109
Chinese	358	358
English	209	202
Kukama	105	86
Navajo	522	168
Sanapaná	602	602

Tools

- Planned UMR tools
  - A baseline UMR parser that parses a textual document into a UMR representation
- Completed UMR tools
  - UMR-Writer: An annotation tool that supports UMR annotation of a diverse set of languages, accepting both keyboard and click-based input methods
  - AnCast: An UMR evaluation metric provides easily interpretable evaluation metrics and supports aligned and unaligned meaning representation graphs

5. Low-resource languages

" Ne'toukutooxebei3i'. "
ne'- toukutooxebei -3i'
then- tie.up.horse -3PL
PREFIX- VAI.INCORP -INFL
"Then they tied up their horses."

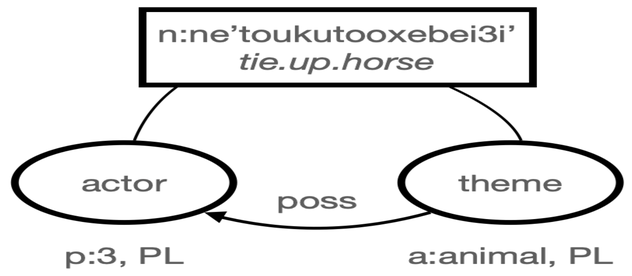


Figure: IGT (left) and partial UMR graph (right) for one Arapaho sentence.

Interlinear glossed text (IGT)

- Widely used data format in linguistics, and especially in endangered language documentation
- Each tier conveys different level of information
  - Original sentence
  - Segmentation of words into meaning-bearing units (morphemes)
  - Linguistic glosses (stem translations, grammatical functions, parts of speech, etc.)
  - Translation into language of wider communication
- Much of this information is redundant with full UMR graph

Automatic subgraph extraction

- With a small amount of input from a language expert, we can produce many parts of the UMR graph automatically
  - Core argument structure: e.g. verb plus subject and object
  - Some grammatical properties of the participants (e.g. plurality, 3rd person)
- Our system produces data format that can be directly imported into the UMR annotation tool; previously, we needed to write new data import functionality for every pre-existing dataset
- System will dramatically reduce annotation time, lowering barrier to entry for many, many languages

6. Outreach efforts

- **Research outreach - completed**
  - 2023 Designing Meaning Representations workshop; co-located with the 15th International Conference on Computational Semantics (IWCS), Nancy, France, June 2023
  - Special event – Amazonian Languages in the Information Age, with presentations on AMR and UMR; Informal session at the 9th International Colloquium on Amazonian Languages, Bogota, Colombia, June 2023
- **Research outreach - coming up**
  - 2024 Designing Meaning Representations workshop proposal submitted
- **Educational outreach - completed**
  - Tutorial – Uniform Meaning Representation, a Cross-lingual Annotation Framework for Document-level Semantics; held at the 13th International Language Resources and Evaluation Conference (LREC 2022), Marseilles, France, June 2022
  - Tutorial – Meaning Representations for Natural Languages: Design, Models, and Applications; held at IJCAI 2023, the 32nd International Joint Conference on Artificial Intelligence, Macao, August 2023
- **Educational outreach - coming up**
  - 1-week UMR Summer School at the University of Colorado Boulder, June 2024

7. Planned platform for community contributions

Data set submission and hosting

- UMR data sets hosted and distributed via LINDAT, part of CLARIN (Common Language Resources and Infrastructure)
- New data sets are welcome!

Support for new contributions

- UMR-Writer for efficiently producing annotations
- Bootstrapping system (work in progress) for partial automation of graph production
- Educational materials
  - Existing: tutorials, annotation guidelines, tool documentation
  - Planned: video tutorials, course materials

Inspiration

- Universal Dependencies (UD) project focuses on treebanks (collections of sentences with syntactic annotation)
- Hundreds of NLP community contributors have added UD treebanks for more than 100 languages

8. Acknowledgements

This work is supported by grants from the CNS Division of National Science Foundation (Awards no: NSF_2213805, NSF_2213804, NSF_IIS 1764048, NSF_1763926 RI) entitled "Building a Broad Infrastructure for Uniform Meaning Representation" and "Developing a Uniform Meaning Representation for Natural Language Processing", respectively. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, or the U.S. government.