Chinese UMR annotation: Can LLMs help?

Haibo Sun, Nianwen Xue, Jin Zhao, Liulu Yue, Yao Sun, Keer Xu and Jiawei Wu

Brandeis University

May 15, 2024

Abstract

We explore using LLMs, GPT-4-preview-0125 specifically, to generate draft sentence-level Chinese Uniform Meaning Representations (UMRs) that human annotators can revise to speed up the UMR annotation process.In this study, we use few-shot learning and Think-Aloud prompting to guide GPT-4 to generate UMR sentence-level graphs. Our primary experimental results show that compared with annotating UMRs from scratch, using LLMs as a preprocessing step reduces the annotation time by two thirds on average. This indicates that there is great potential to integrate LLMs into the pipeline for complicated semantic annotation tasks.

Chinese UMR

- Uniform Meaning Representation (UMR) is a recent graph-based cross-lingual semantic representation formalism that includes a sentence-level representation and a document-level representation.
- The sentence-level representation is based on Abstract Meaning Representation (AMR) but has been extended to capture not only predicate-argument structures, word senses, and named entities as AMR does, but also aspectuality of events, person and number attributes of entities, and quantification.
- The document-level representation includes coreference, temporal and modal dependencies that go beyond sentence boundaries.
- Annotating UMR for Chinese requires extra efforts in word segmentation, compound decomposition and determination of multiword expression (Sun et al., 2023; Bonn et al., 2023).

Difficulties in Creating UMR Graphs

Manual annotation:

- Annotators need to have linguistic knowledge to be able to analyze multiple semantic facets;
- Adding abstract features such as modality strengths, aspectuality markers and named entities is very time-consuming and selective on annotators' knowledge and ability.

Parsers:

- Parsers require large amount of data to train or fine-tune;
- Current UMR data or even AMR data is very limited in its scope of language and amount.

Methods for Accelerating Data Creation Process

LLMs:

- Strong abilities in example following;
- Cross-language transferability;
- Probably trained on previous public AMR data.

We tried two In-Context Learning methods serving as a plug-and-play pre-parser to accelerate annotation.

- Few-shot learning with only sentence-graph pairs.
- Think-aloud preambles with verbally expressed process of constructing the semantic graph in the form of chain-of-thought prompting.

Evaluation Scores

Article 1	A-A	A1-G	A2-G	0F-G	7F-G	0T-G	7T-G
CM	78.52	93.72	88.32	79.03	75.93	65.61	81.90
ULRM	53.97	78.92	70.08	46.93	42.28	48.20	47.00
WLRM	53.05	77.64	70.66	41.16	37.62	42.72	42.88
LRM	52.00	78.08	68.66	43.61	38.47	44.44	43.00
SM	60.85	80.08	75.38	55.58	52.69	54.73	53.92
SM++	60.45	79.93	75.06	55.12	52.18	53.99	53.51
Article 2	**A-A**	**A3-G**	**A4-G**	**0F-G**	**7F-G**	**0T-G**	**7T-G**
CM	61.65	97.06	77.86	65.82	64.35	72.02	72.51
ULRM	42.88	85.31	42.44	34.35	34.96	31.37	33.69
WLRM	45.17	90.12	43.84	30.46	32.45	32.39	33.28
LRM	40.77	84.97	40.43	31.23	32.12	28.72	31.30
SM	53.15	87.96	55.00	46.81	46.85	41.62	44.35
SM++	53.33	88.23	54.70	47.15	46.74	41.26	44.12

Figure: Inter-Annotator Agreement (IAA) and Automatic UMR Parsing Accuracy. The gold graph is obtained by merging the independent annotations from two annotators after a discussion between the two annotators. The discrepancy in scores between the gold graphs and those of different annotators reflect the varying levels of proficiency in UMR annotation for the annotators.

- CM (concept match) measures the F1-score of the set of concepts annotated in two graphs;
- ULRM (unlabeled relation match) measures the F1-score of parent-child concept pairs in two graphs;
- LRM (labeled relation match) takes the relation labels into account when measuring the F1 of the parent-child concept pairs;
- WLRM (weighted labeled relation match) is a weighted version of LRM with more weight given to nodes that have more descendants.
- A-A means inter-annotator agreement; A1/3-G and A2/4-G compares the UMRs by two annotators in each article with gold graph;
- 0F-G, 7F-G, 0T-G, 7T-G: the four LLM parses under different setting compared to the gold graphs;
- SM / SM++: Smatch and Smatch++ scores.

Related work

- Preprocessing in annotation is not a new idea and it has been shown to speed up annotation in treebanking (Chiou et al., 2001).
- There is also prior work on using LLMs to generate Abstract Meaning Representations (AMRs) using GPT-4 (Ettinger et al., 2023).

Exemplar Think-Aloud Preambles

Prompt:Parse this sentence: 该周报也曾经短暂的拥有过一个属自己的电视节目。
Prompt translation: The weekly report also once briefly had its own television program.
Exemple response: 这是第 3 句话。
我们需要先判断这句话的句式，然后分析这句话中每个词语之间的依存句法关系。
该句子是一个陈述句，top node 是“拥有”
“拥有”的主语是“周报”，“周报”的修饰是“该”，因此“拥有”:arg0“周报”(s3x2)
“拥有”的宾语是“节目”，而“节目”的定语是“电视”，而这里“电视”是“节目”的媒介，所以“媒介”:medium“电视”
“节目”属于“自己”，而“自己”在这里是一个 pronoun，指代的是“周报”，因此“属”:arg1 s3x2
……
Response translation: This is the third sentence.
We need to first determine the sentence structure of this sentence and then analyze the dependency syntax relationships between each word in the sentence.
The sentence is a declarative sentence, with 'owning' as the top node.
The subject of 'owning' is 'weekly report,' and 'the' modifies 'weekly report,' hence 'owning': arg0 'weekly report' (s3x2).
The object of 'owning' is 'program,' and the attribute of 'program' is 'television,' where 'television' serves as the medium of 'program,' so 'medium': medium 'television.'
'Program' belongs to 'oneself,' and 'oneself' here is a pronoun, referring to 'weekly report,' thus 'belong': arg1 s3x2
...

Abstract Concepts

(1) 这个 关于 他晋升 的秘密
zhege guanyu ta jinsheng de mimi
不能 给 任何 人 讲 !
buneng gei renhe ren jiang
this about he promote DE secret cannot to any person tell !
"You cannot tell anybody the secret that he got promoted!"
(s1x / 讲-01["tell"]
:mode imperative
:modstr NeutNeg
...)

(2) 美国 前 总统 克林顿
meiguo qian zongtong kelindun
US former president Clinton
(s41i2 / individual-person
:name(s41n / name
:op1 " 克林顿"["Clinton"])
:ARG1-of (s41h / have-org-role-91
:ARG2 (s41c / country
:name (s41n2 / name
:op1 " 美国"["US"]))
:ARG3(s41x3 / 总统 ["president"]
:mod (s41x4 / 前 ["former"]))))

Experiment

Two take-aways:

- Evaluation Scores are not significantly lower than IAA;
- Reduced Annotation Time by 66%.

Details:

- We experimented with two Chinese news articles written in 2024.
- Few-shot learning with only sentence-graph pairs can already generate well-formed graphs with highly plausible details, but (1) the extracted concepts are not faithful enough to the original sentences and sometimes word types are expressed in English rather than the target language, and (2) this method is weak in parsing aspect markers of predicates which captures the implicit information of aspectuality of a predicate.
- Think-aloud prompting yields higher accuracy in annotating more implicit semantic features like aspects of predicates and discourse relations. However, errors still exist and are connected to the design of preambles. For example, word segmentation is sometimes not specified in the prompts, thus leading the failure in segmenting compound words of the names of named entities.
- The high temperature we experimented(0.7) does not show obvious improvement on generating semantic parsing for some irregular syntactic structures, and yet it increased the error rate of graph well-formedness.

Speed up

Article	Annotator	From Scratch	Annotator	From Draft Graphs	Ratio
1	A1	8h57min	A3	2h47min	3.19
	A2	9h03min	A4	2h52min	
2	A3	6h49min	A1	2h51min	2.61
	A4	8h47min	A2	3h08min	

Table 2: A comparison between the times needed for annotation from scratch and from draft graphs. The method for calculating the ratio involves computing the average annotation time for each sentence, and then taking the average between the two annotators.

Conclusion

- LLMs, specifically GPT-4, can be used to speed up UMR annotation.
- The accuracy of GPT-generated UMRs is not very far from the IAA from human annotators.

Reference

Julia Bonn, Andrew Cowell, Jan Hajic, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdenka Urešová, Shira Wein, Nianwen Xue, et al. 2023. Umr annotation of multiword expressions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109.

Fu-Dong Chiou, David Chiang, and Martha Palmer. 2001. Facilitating treebank annotation using a statistical parser. In *Proceedings of the first international conference on human language technology research*.

Allyson Ettinger, Jena D Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. " you are an expert linguistic annotator": Limits of llms as analyzers of abstract meaning representation. *arXiv preprint arXiv:2310.17793*.

Haibo Sun, Yifan Zhu, Jin Zhao, and Nianwen Xue. 2023. Umr annotation of chinese verb compounds and related constructions. In *Proceedings of the first international workshop on construction grammars and nlp (cxgs+ nlp, gurt/syntaxfest 2023)*, pages 75–84.

Brandeis