Parsing Meaning Representations: is Easier Always Better?

Zi Lin, Nianwen Xue
Peking University & Brandeis University
DMR@ACL | 2019
Overview

• Introduction
• MRS v.s. AMR
• Experiment
• Analysis
  • Concept detection
  • Relation detection
Introduction
The boy wants to believe the girl.

Parsing natural language sentences into a formal representation that encodes the meaning of a sentence (usually a graph).
There is no universally accepted standard and existing MRs vary descriptively and theoretically…

- Groningen Meaningbank: Discourse Representation Theory
- Redwoods corpus: Minimal Recursion Semantics
- Prague Dependency Treebank: Functional generative description
- Universal Cognitive Conceptual Annotation: Basic Linguistic Theory
- Abstract Meaning Representation: (Loosely) neo-Davidsonian with some other stuff
Parsing results reported in the literature

| MRs | DRT | MRS | UCCA | AMR |
|-----|-----|-----|------|-----|
| F1  | 77.5\(^1\) | 90.9\(^2\) | 69.9\(^3\) | 74.4\(^4\) |

1. Liu et al. 2018. Discourse representation structure parsing.
2. Chen et al. 2018. Accurate shrg-based semantic parsing.
3. Hershcovich et al. 2019. SemEval 2019 shared task: cross-lingual semantic parsing with UCCA call for participation.
4. Sheng Zhang et al, 2019. AMR parsing as sequence-to-sequence transduction.
Parsing results reported in the literature

| MRs | DRT  | MRS  | UCCA | AMR  |
|-----|------|------|------|------|
| F1  | 77.5¹ | 90.9² | 69.9³ | 76.3⁴ |

1. Liu et al. 2018. Discourse representation structure parsing.
2. Chen et al. 2018. Accurate shrg-based semantic parsing.
3. Hershcovich et al. 2019. SemEval 2019 shared task: cross-lingual semantic parsing with UCCA call for participation.
4. Sheng Zhang et al. 2019. AMR parsing as sequence-to-sequence transduction.
To develop the next generation MRs …

• Which aspects of the MR pose the most challenge to automatic parsing?

• Whether these challenges are “necessary evils”, or they can be simplified without hurting the utility of the MR?
MRS v.s. AMR
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold.
Experiment
Data preparation

• Dataset:
  • SDP2015 for MRS
  • LDC2016E25 for AMR

• Format unification: PENMAN format (using PyDelphin library)

• Parsing model: CAMR (Wang et al., 2015)

• Alignment:
  • Gold for MRS
  • JAMR (Flanigan et al., 2014) for AMR
## Parsing Result

|                          | MRS          | AMR          |
|--------------------------|--------------|--------------|
|                          | Train | Dev | Test | Train | Dev | Test |
| number of graphs/sentences | 35,315  | 1,410 | 1,410 | 36,521  | 1,368 | 1,371 |
| number of tokens per sentence | 22.33 | 22.92 | 23.14 | 17.83 | 21.59 | 22.10 |
| number of nodes per token | 0.96  | 0.97  | 0.93  | 0.68  | 0.70  | 0.70  |

|                          | Node | Edge | $S_{MATCH}$ | Node | Edge | $S_{MATCH}$ |
|--------------------------|------|------|-------------|------|------|-------------|
| CAMR                     | 89.4 | 81.1 | 85.3        | 78.7 | 57.1 | 68.0        |
| Buys and Blunsom (2017)  | 89.1 | 85.0 | 87.0        | -   | -   | -           |
| Chen et al. (2018)       | 94.5 | 87.3 | 90.9        | -   | -   | -           |
| Lyu and Titov (2018)     | -    | -    | -           | 85.9 | 69.8 | 74.4        |

$\triangle S_{MATCH}$

-17.3

-25.8
Analysis
Concept detection

• The first step in constructing a MR graph is determining the nodes.
• The first step in constructing a MR graph is determining the nodes.
  • Word sense disambiguation

**sell-01**: commerce: seller, giving in exchange for money
**sell_out-02**: give in to the man
**sell_out-03**: sell until none is/are left

......
Concept detection

• The first step in constructing a MR graph is determining the nodes.
  • Word sense disambiguation
  • Inferring abstract concepts
Concept detection

• The first step in constructing a MR graph is determining the nodes.
  • Word sense disambiguation
  • Inferring abstract concepts
  • Entity recognition
## Concept detection

| POS | %   | #lemma | #sense | average | score  | WSD   |
|-----|-----|--------|--------|---------|--------|-------|
| n   | 34.46 | 1,420  | 1,434  | 1.01    | 95.35  | 99.76 |
| v   | 20.37 | 838    | 1,010  | 1.21    | 85.56  | 90.58 |
| q   | 13.97 | 25     | 25     | 1.00    | 98.22  | 100.00|
| p   | 12.86 | 96     | 123    | 1.28    | 81.29  | 76.11 |
| a   | 11.45 | 637    | 648    | 1.02    | 90.58  | 99.90 |
| c   | 4.20  | 17     | 19     | 1.12    | 94.46  | 99.61 |
| x   | 2.69  | 80     | 81     | 1.01    | 73.65  | 99.74 |
| total | 100.00 | 3,113  | 3,340  | 1.07    | 90.78  | 97.06 |

| AMR | pred | %   | #lemma | #sense | average | score  | WSD   |
|-----|------|-----|--------|--------|---------|--------|-------|
|     | -    | 1,292 | 1,440  | 1.11   | 77.93  | 94.54  |
Concept detection

| POS | % | #lemma | #sense | average | score | WSD |
|-----|---|--------|--------|---------|-------|-----|
| n   | 34.46 | 1,420 | 1,434 | 1.01 | 95.35 | 99.76 |
| v   | 20.37 | 838   | 1,010 | 1.21 | 85.56 | 90.58 |
| q   | 13.97 | 25    | 25    | 1.00 | 98.22 | 100.00 |
| p   | 12.86 | 96    | 123   | 1.28 | 81.29 | 76.11 |
| a   | 11.45 | 637   | 648   | 1.02 | 90.58 | 99.90 |
| c   | 4.20  | 17    | 19    | 1.12 | 94.46 | 99.61 |
| x   | 2.69  | 80    | 81    | 1.01 | 73.65 | 99.74 |
| total | 100.00 | 3,113 | 3,340 | 1.07 | 90.78 | 97.06 |

AMR

| pred | - | 1,292 | 1,440 | 1.11 | 77.93 | 94.54 |

25
Concept detection

- Word sense disambiguation is not a major contributor to the difficulty in concept detection for AMR

| POS | % | #lemma | #sense | average | score | WSD |
|-----|---|--------|--------|---------|-------|-----|
| n   | 34.46 | 1,420 | 1,434 | 1.01 | 95.35 | 99.76 |
| v   | 20.37 | 838   | 1,010 | 1.21 | 85.56 | 90.58 |
| q   | 13.97 | 25    | 25    | 1.00 | 98.22 | 100.00 |
| p   | 12.86 | 96    | 123   | 1.28 | 81.29 | 76.11 |
| a   | 11.45 | 637   | 648   | 1.02 | 90.58 | 99.90 |
| c   | 4.20  | 17    | 19    | 1.12 | 94.46 | 99.61 |
| x   | 2.69  | 80    | 81    | 1.01 | 73.65 | 99.74 |
| total | 100.00 | 3,113 | 3,340 | 1.07 | **90.78** | **97.06** |

| POS | % | #lemma | #sense | average | score | WSD |
|-----|---|--------|--------|---------|-------|-----|
| pred | - | 1,292 | 1,440 | 1.11 | **77.93** | **94.54** |
We took a closer look at how concept detection fared for lexical categories that are known to have a complex mapping to the concepts they "evoke".

• Phrasal verbs (p.v.)
  e.g. take a bath & bathe -> bathe-01

• Nouns (n.)
  e.g. destruction & destroy -> destroy-01

• Adjectives (adj.)
  e.g. attractive -> attract-01

• Adverbs (adv.)
  e.g. quickly & quick -> quick-01

• Prepositions (prep.)
  e.g. out of mind -> out-06

• Conjunctions (conj.)
  e.g. but -> constrast-01

• Modal verbs (mod)
  e.g. can (modal verbs) & possible -> possible-01
Concept abstraction

- Extract word or word sequences that align with these concepts
- Use a set of heuristics based on morpho-syntactic patterns to determine the type of abstraction in the test set

| type   | n.    | adj.   | adv.  | prep. | conj. | mod.  | p.v. | other | v.    |
|--------|-------|--------|-------|-------|-------|-------|------|-------|-------|
| %      | 35.09 | 10.05  | 1.87  | 1.17  | 1.01  | 2.59  | 0.31 | 0.15  | 47.76 |
| Performance | 83.01 | 84.44  | 80.73 | 73.53 | 96.61 | 66.96 | 83.33 | 44.44 | 74.07 |

Table 3: Individual percentages and scores for different types of AMR predicates
Figure 2: Relative improvement of performance on the test set after correcting each type of POS or construction in AMR
Concept abstraction

Figure 2: Relative improvement of performance on the test set after correcting each type of POS or construction in AMR.
Figure 2: Relative improvement of performance on the test set after correcting each type of POS or construction in AMR
Figure 2: Relative improvement of performance on the test set after correcting each type of POS or construction in AMR

Concept abstraction

can & possible -> possible-01
We next examined how well entities are detected in AMR and MRS parsing.

| Example          | AMR                                           | MRS                                           |
|------------------|-----------------------------------------------|-----------------------------------------------|
| lunar calendar   | (d / date-entity :calendar (m / moon))        | −                                             |
| December (8th)   | (d / date-entity :month 12)                   | (x1 / mofy :carg "Dec")                      |
| Monday           | (d / date-entity :weekday (m / monday))       | (x1 / dofw :carg "Mon")                      |
| (December) 8th   | (d / date-entity :day 8)                      | (x1 / dofm :carg "8")                        |
| night            | (d / date-entity :dayperiod (n / night))      | −                                             |
| New York         | (cl / city                                    | (x1 / named :carg "York")                    |
|                  | :name (n1 / name                              | :ARG1-of (e1 / compound                      |
|                  | :op1 "New" :op2 "York"))                     | :ARG2 (x2 / named :carg "New"))              |
Entity recognition

We next examined how well entities are detected in AMR and MRS parsing.

| dataset          | MRS | AMR |
|------------------|-----|-----|
|                  | #   | #   |
|                  | score | score |
| date entity      | 266  | 273 |
|                  | 92.48 | 66.67 |
| NE detection     | 2,555 | 2,065 |
|                  | 81.96 | 91.09 |
| NE classification| -   | -   |
|                  | -   | 76.46 |

Table 5: Results on entity recognition on the test set
We next examined how well entities are detected in AMR and MRS parsing.

- Date entity detection: AMR << MRS
- e.g. lunar calendar -> (d / date-entity :calendar (m / moon))

| dataset          | MRS # | score  | AMR # | score  |
|------------------|-------|--------|-------|--------|
| date entity      | 266   | 92.48  | 273   | 66.67  |
| NE detection     | 2,555 | 81.96  | 2,065 | 91.09  |
| NE classification| -     | -      | -     | 76.46  |

Table 5: Results on entity recognition on the test set
Entity recognition

We next examined how well entities are detected in AMR and MRS parsing.

- Date entity detection: AMR << MRS
- e.g. lunar calendar -\(\rightarrow\) (d / date-entity :calendar (m / moon))
- Name entity: AMR << MRS
  - detection: AMR > MRS
  - Name entity classification: not needed for MRS

| dataset          | MRS # | score | AMR # | score |
|------------------|-------|-------|-------|-------|
| date entity      | 266   | 92.48 | 273   | 66.67 |
| NE detection     | 2,555 | 81.96 | 2,065 | 91.09 |
| NE classification| -     | -     | -     | 76.46 |

Table 5: Results on entity recognition on the test set
The subtask of relation detection involves identifying and labeling the edges in the MR graph.
The subtask of relation detection involves identifying and labeling the edges in the MR graph.

| dataset   | MRS # | MRS score | AMR # | AMR score |
|-----------|-------|-----------|-------|-----------|
| Overall   | -     | 81.76     | -     | 61.52     |
| All matched | 3,398 | 63.48     | 4,975 | 44.77     |
| ARG0      | 3,087 | 62.00     | 3,680 | 49.43     |
| ARG1      | 2,985 | 68.45     | 5,377 | 53.97     |
| ARG2      | 339   | 35.09     | 1,614 | 37.86     |
| ARG3      | 7     | 57.13     | 123   | 14.63     |
| ARG4      | -     | -         | 39    | 20.51     |
| Reentrancy| 807   | 81.28     | 1,723 | 43.91     |

Table 6: Results on SRL. MRS’s argument number begins at 1 so we just move all the argument to begin at 0 to make them comparable.
• The subtask of relation detection involves identifying and labeling the edges in the MR graph.

| dataset     | MRS # | score | AMR # | score |
|-------------|-------|-------|-------|-------|
| Overall     | -     | 81.76 | -     | 61.52 |
| All matched | 3,398 | 63.48 | 4,975 | 44.77 |
| ARG0        | 3,087 | 62.00 | 3,680 | 49.43 |
| ARG1        | 2,985 | 68.45 | 5,377 | 53.97 |
| ARG2        | 339   | 35.09 | 1,614 | 37.86 |
| ARG3        | 7     | 57.13 | 123   | 14.63 |
| ARG4        | -     | -     | 39    | 20.51 |
| Reentrancy  | 807   | 81.28 | 1,723 | 43.91 |

Table 6: Results on SRL. MRS’s argument number begins at 1 so we just move all the argument to begin at 0 to make them comparable.

(w / want-01
 :ARG0 (b / boy)
 :ARG1 (b2 / believe-01
 :ARG0 (g / girl)
 :ARG1 b))
The subtask of relation detection involves identifying and labeling the edges in the MR graph.

| dataset  | MRS # | MRS score | AMR # | AMR score |
|----------|-------|-----------|-------|-----------|
| Overall  | -     | 81.76     | -     | 61.52     |
| All matched | 3,398 | 63.48     | 4,975 | 44.77     |
| ARG0     | 3,087 | 62.00     | 3,680 | 49.43     |
| ARG1     | 2,985 | 68.45     | 5,377 | 53.97     |
| ARG2     | 339   | 35.09     | 1,614 | 37.86     |
| ARG3     | 7     | 57.13     | 123   | 14.63     |
| ARG4     | -     | -         | 39    | 20.51     |

Table 6: Results on SRL. MRS’s argument number begins at 1 so we just move all the argument to begin at 0 to make them comparable.

• SRL accuracy: AMR << MRS
• Reentrancy: AMR << MRS
• Number of reentrancy: AMR >> MRS
• MRS treats prepositions as predicates, and labels their arguments.
• AMR just drops the preposition when it introduces an oblique argument for a verbal predicate.
Coreference

- AMR resolves sentence-level coreference.
- MRS does not resolve coreference and each instance of the same entity will be a separate concept in the MRS graph.
• AMR concepts show a higher level of abstraction from surface forms
• AMR does a much more fine-grained classification for the named entities than MRS
• Semantic relations are defined differently in AMR and MRS
Summary

• AMR concepts show a higher level of abstraction from surface forms

• AMR does a much more fine-grained classification for the named entities than MRS

• Semantic relations are defined differently in AMR and MRS

These have all contributed to the performance gap between MRS and AMR parsing.

The question is: should AMR be simplified to improve the accuracy of AMR parsing?
Thank you!