Chinese Abstract Meaning Representation Corpus (CAMR) V2.0

Authors: Bin Li, Liming Xiao, Yihuan Liu, Yuan Wen, Li Song, Jayeol Chun, Minxuan Feng, Junsheng Zhou, Weiguang Qu, Nianwen Xue

Introduction

The Chinese Abstract Meaning Representation Corpus (CAMR) V2.0 is constructed following the basic principles of Abstract Meaning Representation (AMR), a compact, readable, whole-sentence semantic representation, while making adaptions where necessary to handle Chinese-specific phenomena.

CAMR V1.0 corpus contains the AMR of 10,149 sentences. The raw text is extracted from the weblog and discussion forum portion of CTB 8.0, which totals 10,325 sentences. 176 of the sentences are left unannotated because they are ungrammatical and hard to interpret.

CAMR V2.0 contains the AMR of 20,078 sentences. CAMR V2.0 includes the 10,145 out of the 10,149 sentences from CAMR V1.0 and adds 9,933 more sentences from the news portion of CTB8.0. The news portion of CTB 8.0 has 10,001 sentences, 68 of which are not annotatable. The sentence IDs of the unannotated 176 sentences in CAMR V1.0 and 72 sentences in CAMR V2.0 are listed in Table 1 and Table 2.

Indices of Unannotated Sentences in CAMR V1.0
92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 3092, 3093, 3096, 3097, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3754, 3942, 4627, 4767, 5043, 5044, 5045, 5048, 5117, 5147, 5275, 5418, 5499, 5559, 5560, 5561, 5562, 5634, 5639, 5640, 5800, 5810, 5830, 6019, 6038, 6139, 6150, 6155, 6164, 6169, 6231, 6247, 6250, 6253, 6353, 6373, 6421, 6424, 6681, 6697, 6754, 6756, 6757, 6758, 6759, 6760, 6761, 6762, 6763, 6797, 6802, 7027, 7312, 7321, 7348, 7349, 7350, 7351, 7352, 7353, 7354, 7355, 7356, 7357, 7375, 7377, 7378, 7384, 7389, 7458, 7459, 7468, 7528, 7532, 7533, 7534, 7588, 7618, 7640, 7677, 7690, 7692, 7699, 7978, 8029, 8041, 8052, 8055, 8058, 8059, 8272, 8276, 8431, 8463, 8464, 8465, 8467, 8470, 8572, 8585, 8837, 9042, 9099, 9159,

9463, 9467, 9474, 9538, 9550, 9597, 9606, 9642, 9775, 9815, 9816, 9817, 9818, 9968, 9992, 10005, 10079, 10093, 10135, 10136, 10145, 10153, 10199

Table 1: Indices of the **Unannotated** 176 Sentences in CAMR V1.0

Indices of Unannotated Sentences in CAMR V2.0
6803, 7532, 7533, 7534, 12658, 12667, 12673, 12678, 12685, 12690, 12694, 12705, 12717, 12722, 12734, 12745, 12752, 12762, 12770, 12782, 12792, 12800, 12807, 12811, 12820, 12833, 12846, 12854, 12873, 12884, 12888, 12903, 12918, 12924, 12928, 12931, 12936, 12941, 12945, 12949, 12953, 12969, 15342, 15506, 19082, 19083, 19084, 19085, 19086, 19087, 19088, 19089, 19090, 19091, 19092, 19093, 19094, 19095, 19096, 19097, 19098, 19099, 19100, 19101, 19102, 19103, 19104, 19105, 19106, 19107, 19108, 20327

Table 2: Indices of the **Unannotated** 72 Sentences in CAMR V2.0

The corpus is split into 3 parts by their document IDs as originally released in CTB 8.0. The train, dev and test sets are exactly the same as in MRP2020 shared task, which could be used to improve the parsing models.

Set	**CAMR1.0**	**Article ID**	**New**	**Article ID**	**CAMR2.0**
train	7,606	5,061-5,558	8,970	15-16, 20, 22, 26, 32-34, 52, 81-318, 321-901, 905	16,576
test	1,276	5,000-5,030	437	1-14, 17-19, 21, 23-25, 27-31, 35-40, 902-904	1,713
dev	1,263	5,031-5,060	526	41-51, 53-80, 319-320	1,789
Total	10,145		9,933		20,078

Table 3: the subsets of CAMR V2.0

Like AMR, a Chinese AMR is a single-rooted, directed, acyclic graph, with the nodes labeled with concepts and edges labeled with semantic relations. There are 51 semantic relations in total, with 5 core semantic relations and 46 non-core semantic relations. The details of the relations are shown in Tables 4 and 5.

ID	Label	Explanation
1	arg0	external argument (Proto-Agent)
2	arg1	internal argument (Proto-Patient)
3	arg2	indirect object / beneficiary / instrument / attribute / end state
4	arg3	start point / beneficiary / instrument / attribute
5	arg4	end point

Table 4: Core Semantic Relations in CAMR

ID	Label	ID	Label	ID	Label
1	accompanier	17	extent	33	poss
2	*aspect	18	frequency	34	purpose
3	beneficiary	19	instrument	35	quant
4	cause	20	li	36	range
5	compared-to	21	location	37	*refer
6	consist-of	22	manner	38	source
7	condition	23	medium	39	subevent
8	cost	24	mod	40	subset
9	*cunit	25	mode	41	superset
10	*dcopy	26	name	42	*tense
11	degree	27	ord	43	time
12	destination	28	part-of	44	topic
13	direction	29	path	45	unit
14	domain	30	*perspective	46	value
15	duration	31	polarity		
16	example	32	polite		

* are the added relations in CAMR

Table 5: Non-core Semantic Relations in CAMR

The Chinese Abstract Meaning Representation Corpus project began at the Nanjing Normal University and Brandeis University in 2014. The project goal is to provide a large, concept/relation-to-word aligned Chinese Abstract Meaning Representation Corpus. The CAMR 2.0 release contains 20,078 sentences extracted from CTB 8.0. The Chinese AMR project is on-going and more data will be released in future versions.

Data

This release contains three sample text files that are from the training, development and test set respectively. Each sentence has 4 fields: the sentence ID, the word segmented sentence, the word segmented sentence with word indices, and the AMR graph. The data is provided in the UTF-8 encoding. All files were automatically verified and manually checked.

Example 1:

```
# ::id export_amr.1617 ::2017-01-06 16:12:33
# ::snt 希望 我 惨痛 的 经历 给 大家 一 个 教训 呀
# ::wid x1_希望 x2_我 x3_惨痛 x4_的 x5_经历 x6_给 x7_大家 x8_一 x9_
个 x10_教训 x11_呀 x12_
    (x1 / 希望-01
            :arg1()  (x6 / 给-01
                    :arg0()  (x5 / 经历
                            :poss()  (x2 / 我)
                            :arg0-of(x4/的)  (x3 / 惨痛-01))
                    :arg2()  (x7 / 大家)
                    :arg1()  (x10 / 教训
                            :quant()  (x8 / 1)
                            :cunit()  (x9 / 个)))
            :mode()  (x11 / expressive))
```

The corpus has the manual annotation of concept-to-word and relation-to-word alignments, using the index of each word in a sentence. The numerical ID of a concept, prefixed with x, is the index of the word token (or indices of the word tokens). For example, x7 is the ID of the concept 大家. Where plausible, the functional words also gfet an ID prefixed with x, but they are generally aligned to relations. For example, “x4/的” is aligned to *:arg0-of*.

The CAMR V2.0 data has been used in MRP2020 shared task, with the best F-score of 0.8. Each CAMR data is distributed in three formats:

(a) the original Chinese AMR data with concept-to-word and relation-to-word alignments as shown in Example 1. The data is in the origin folder.

(b) the English AMR format converted from CAMR. The alignments are given in the line starting with # ::align. The data is in the mrp2020 folder.

(b) the Chinese syntactic dependency tree. It’s converted form CTB. It could be used as a syntactic feature for AMR parsing. The data is in the dependency folder.

Example 2:

```
# ::id export_amr.1617 ::2017-01-06 16:12:33
# ::snt 希望 我 惨痛 的 经历 给 大家 一 个 教训 呀
# ::wid x1_希望 x2_我 x3_惨痛 x4_的 x5_经历 x6_给 x7_大家 x8_一 x9_
个 x10_教训 x11_呀
# ::align a0:10 a2:2 a3:7 a4:1 a5:5 a6:9 a7:3 a8:8 a9:6
(a4 / 希望-01
    :arg1 (a9 / 给-01
              :arg0 (a5 / 经历
                        :poss (a2 / 我)
                        :arg0-of (a7 / 惨痛-01))
```

```
                    :arg2 (a3 /  大家)
                    :arg1 (a0 /  教训
                                   :quant (a8 / 1)
                                   :cunit (a6 /  个)))
         :mode (a1 / expressive))
```

The former represents the concept-to-word alignments separately from the CAMR graph. The ID of a concept, prefixed with “a”, is no longer indicated the index of the word token directly. Instead, in the alignment line, the ID of a concept corresponds to the index of the word token. For example, “a0:10” means “a0” is the concept ID of the 10th word token 教训. In addition, the functional words are unannotated in this format, so the relation-to-word alignments are removed.

Example 3:

#export_amr.1617

ID	Form	LEMMA	CPOS	POS	FEATS	HEAD	DEPREL	DEPS	MISC
1	希望	_	VV	VV	_	0	root	_	_
2	我	_	PN	PN	_	3	nsubj	_	_
3	惨痛	_	VA	VA	_	5	rcmod	_	_
4	的	_	DEC	DEC	_	3	cpm	_	_
5	经历	_	NN	NN	_	6	nsubj	_	_
6	给	_	VV	VV	_	1	ccomp	_	_
7	大家	_	PN	PN	_	6	dobj	_	_
8	一	_	CD	CD	_	9	nummod	_	_
9	个	_	M	M	_	10	clf	_	_
10	教训	_	NN	NN	_	6	dobj	_	_
11	呀	_	SP	SP	_	1	dep	_	_

For the latter, a companion Chinese dependency tree, produced with the UD-style CoreNLP 4.0 dpendency parser, provides the word index (ID), word form (FORM), coarse-part-of-speech (CPOS) tag, part-of-speech (POS) tag, head ID (HEAD) of current token, and dependency relation (DEPREL) to the head. For example, ID of the word 经历 is 5 and its CPOS and POS are NN. The head of 经历 is 给, whose ID is 6, and the dependency relation between them is nsubj. The other fields with underscore are lemma (LEMMA) of word form, morphological features (FEATS), secondary dependencies (DEPS) and miscellaneous (MISC) annotations, which are not available in this task.

The users could refer to the following three papers for details.

- Bin Li, Yuan Wen, Li Song, Weiguang Qu, Nianwen Xue. Building a Chinese AMR Bank with Concept and Relation Alignments. Linguistic issues in Language Technology, Vol.18, 2019.
- Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajič, Daniel Hershcovich, Bin Li, Tim O’Gorman, Nianwen Xue, and Daniel Zeman. MRP 2020: The Second Shared Task on Cross-Framework and Cross-Lingual Meaning Representation Parsing, Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, pages 1–22 Online, Nov. 19-20, 2020.
- Chuan Wang, Bin Li and Nianwen Xue. Transition-based Chinese AMR parsing. *Proceedings of NAACL 2018*, June 1, 2018. New Orleans, Louisiana.

Acknowledgement

This work is supported in part by the National Social Science Foundation of China (18BYY127) and National Natural Science Foundation of China (61772278).

Updates

We will continue to release more annotated data of Chinese Abstract Meaning Representation. Please visit our website (http://www.cs.brandeis.edu/~clp/camr/camr.html) for the latest news.

Copyright