Chinese Abstract Meaning Representation Corpus (CAMR) V1.0

Authors: Bin Li, Yuan Wen, Li Song, Rubing Dai, Weiguang Qu, Nianwen Xue

Introduction

The Chinese Abstract Meaning Representation Corpus (CAMR) V1.0 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.

The corpus contains the semantic representation 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 their structures are ill-formed and hard to annotate. The corpus is split into 3 parts by their document IDs as originally released in CTB 8.0. The training set consists of 7,610 sentences from articles 5,061-5,558, the development set has 1,263 sentences from articles 5,000-5,030, and the test set has 1,276 sentences from articles 5,031-5,060. The indices of the unannotated 176 sentences are listed in Table 1.

Indices of Unannotated Sentences
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 Deleted 176 Sentences

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 49 semantic relations in total, with 5 core semantic relations and 44 non-core semantic relations. The details of the relations are shown in Tables 2 and 3.

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 2: Core Semantic Relations in CAMR

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

* are the added relations in CAMR

Table 3: 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 1.0 release contains 10,149 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 text files that correspond to 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:

```
# ::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 sentece. The numerical ID of a concept, prefixed with x, is the index of the word token (or indices of the word tokens). It is aligned with and it is unique with respect to the IDs of other concepts. For example, x7 is the ID of the concept 大家. Where plausible, the functional words also get an ID prefixed with x, but they are generally aligned to relations. For example, "x4/的" is aligned to *:arg0-of*.

The users should refer to the following two papers for further information.

- Bin Li, YuanWen, Lijun Bu,Weiguang Qu, Nianwen Xue. Annotating the Little Prince with Chinese AMRs, *Proceedings of LAW 2016*, Aug 11, 2016. Berlin, Germany.
- Chuan Wang, Bin Li and Nianwen Xue. Transition-based Chinese AMR parsing. *Proceedings of NAACL 2018*, June 1, 2018. New Orleans, Louisiana.

Note that there is a little discrepancy between extraction of sentences in CAMR 1.0 corpus and the pre-release data used in Wang et al (2018). The details are shown in Table 4.

File	Numbers of Sentences	
	CAMR1.0	**Wang's Data (diff from CAMR1.0)**
train	7610	7608 (+No. 4944, 6442, 9232; -No. 6706, 6803, 6804, 7460, 8060)
test	1276	1277(+No. 1600)
dev	1263	1264 (+No. 610)
Totel	10149	10149

Table 4: Discrepancy between Extraction of Sentences in the Two Corpora

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