Building a Chinese AMR Corpus with Concept and Relation Alignments

Bin Li, Yuan Wen, Li Song, Weiguang Qu (Nanjing Normal University),
Chuan Wang, Nianwen Xue (Brandeis University)

Abstract Meaning Representation (AMR) is an annotation framework in which the meaning of a full sentence is represented as a rooted, acyclic, directed graph. In this paper, we describe an on-going project in which we built a Chinese AMR (CAMR) corpus, which currently includes 10,040 sentences from the newsgroup and weblog portion of the Chinese TreeBank (CTB). We describe the annotation specifications for the CAMR corpus, which follows the annotation principles of English AMR but make adaptations where needed to accommodate the linguistic facts of Chinese. The CAMR specifications also include a systematic treatment of sentence-internal discourse relations.

One significant change we have made to the AMR annotation methodology is the inclusion of the alignment between word tokens in the sentence and the concepts/relations in the CAMR annotation to make it easier for automatic parsers to model the correspondence between a sentence and its meaning representation. We develop an annotation tool for CAMR that allows an annotator to simply input the offset of a word token in place of a concept during the annotation process. The tool will automatically retrieve the word token based on its offset and generate the concept as well as the concept ID for it. This assumes that the tool does automatic lemmatization, which fortunately is very straightforward for Chinese where there is little inflectional morphology and the concepts are generally the same as their word forms. The tool handles the one-to-one, one-to-zero, zero-to-one, one-to-many and many-to-one alignments between word tokens in a sentence and concepts/relations in its AMR. The tool also allows the annotator to revise the concept, and this is useful when a word does have inflections in a limited number of cases or when the word is misspelled. The annotation tool also keeps track of which words in the sentence have been “covered” by the AMR by highlighting words that the annotator has created concepts for. This is an especially useful feature when annotating long sentences, as it is very easy for the annotator to miss some words. We have annotated 10,040 CTB sentences with the tool, and the inter-agreement as measured by the Smatch score between the two annotators is 0.83, indicating reliable annotation. We plan to publicly release this data for use in linguistic and NLP research.

We also present some quantitative analysis of the CAMR corpus. In CAMR, the AMR of 46.8% of sentences is a graph, 31.92% of the AMRs have non-projective subtrees, and 1.2% of them have cycles. Moreover, the AMR of 89.1% of the sentences have concepts inferred from the context of the sentence but do not correspond to a word or phrase in a sentence, and the average number of such inferred concepts per sentence is 2.84. These statistics will have be taken into account when developing automatic CAMR parsers. As non-projective structures have never been reported in prior work on Chinese dependency annotation or meaning representation annotation, we also analyze the causes of non-projective subtrees and provide a classification of these non-projective subtrees.