LiLT volume 18, issue (1) June 2019

Building a Chinese AMR Bank with Concept and Relation Alignments

Bin Li, Yuan Wen, Li Song, Weiguang Qu, *Nanjing Normal University* Nianwen Xue ✉, *Brandeis University*

Abstract

Abstract Meaning Representation (AMR) is a meaning representation framework in which the meaning of a full sentence is represented as a single-rooted, acyclic, directed graph. In this article, we describe an on-going project to build a Chinese AMR (CAMR) corpus, which currently includes 10,149 sentences from the newsgroup and weblog portion of the Chinese TreeBank (CTB). We describe the annotation specifications for the CAMR corpus, which follow 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, and the inter-agreement as measured by the Smatch score between the two annotators is 0.83, indicating reliable annotation. We also present some quantitative analysis of the CAMR corpus. 46.71% of the AMRs of the sentences are non-tree graphs. Moreover, the AMR of 88.95% of the sentences has concepts inferred from the context of the sentence but do not correspond to a specific word

or phrase in a sentence, and the average number of such inferred concepts per sentence is 2.88. These statistics will have to be taken into account when developing automatic Chinese AMR parsers.

1 Introduction

Abstract Meaning Representation (AMR) is a novel annotation framework to represent the "meaning" of a sentence with a single rooted, acyclic[1], directed graph (Banarescu et al., 2013), departing from previous practices of performing partial semantic annotation that focuses on certain aspect of meaning. Some well-known examples of partial semantic annotation efforts include the annotation of predicate-argument structure of verbs (Palmer et al., 2005, Xue and Palmer, 2009) and predicative or relational nouns (Meyers et al., 2004), the annotation of entities and relations along the lines of Automatic Content Extraction project (Doddington et al., 2004), the annotation of discourse relations (e.g., the Penn Discourse TreeBank (Prasad et al., 2008)), as well as the annotation of temporal relations (Pustejovsky et al., 2003) and factuality (Saurí and Pustejovsky, 2009). The choice to annotate aspects of meaning instead of "whole-sentence" meaning is predicated on the assumption that focusing on a single aspect of meaning is more likely to lead to consistent annotation and consistently annotated data in turn lead to more accurate machine learning based automatic systems. This is a reasonable assumption when the meaning components of a sentence are not well-understood. However, such a fragmented approach to meaning annotation also leads to redundancies or even conflicts between the different meaning components, thus diminishing the value of these annotated resources when they have to be used in conjunction. There will also inevitably be gaps that are not covered by any of the meaning components which will be problematic for applications that need to reason over the semantics of entire sentences. The "holistic" approach of AMR to annotating the meaning of entire sentences attempts to address this issue by modeling the meaning of a sentence with a single rooted, directed acyclic graph. In general this is considered to be a welcome development in spite of the fact that there are still aspects of sentence meaning that AMR leaves out in exchange for expedience in annotation. For ex-

[1]Banarescu et al. (2015) reports that about 0.3% sentences are cyclic in AMR Sembank.

ample, AMR currently does not annotate tense, aspect, nor does it annotate the phenomena of quantification. However, these linguistic phenomena can be added without substantially modifying the AMR formalism.

Compared with other semantic annotation efforts such as the Semantic Dependency annotation (Oepen et al., 2014) that is largely based on Minimum Recursion Semantics (MRS) (Copestake et al., 2005), the tectogrammatical layer of the Prague Dependency TreeBank (Böhmová et al., 2003), as well as the Groningen Meaning Bank (Bos et al., 2017) which is largely based on the Discourse Representation Theory (Kamp and Reyle, 1993), one salient characteristic of AMR annotation is the relaxation of the strict correspondence between the meaning representation and its underlying morpho-syntactic representation. This has a number of consequences for AMR annotation in practice. First of all, AMR can be annotated independently of morpho-syntactic structures and does not have to be linked to syntactic units such as words and phrases in the annotation process. The practical benefit of this is that it makes annotation scalable, eliminating the time needed to first build morpho-syntactic structures before any semantic annotation can start. Second, the relaxation of the strict correspondence between syntactic representation and semantic representation allows more freedom in handling syntax-semantic mismatches. This includes cases where function words that are crucial building blocks of the syntactic structure can be left out of the meaning representation because they do not contribute to the meaning of the sentence (e.g., infinitive "to" in English). Conversely, there are also cases where constructs (i.e., concepts or relations in AMR) in the meaning representation are inferred from the context and do not necessarily correspond to any words (e.g., "person" can be inferred from "the young"). A third type of syntax-semantic mismatch is reflected in cases where there is a complicated correspondence between the meaning representation and the surface syntactic structure. For example, a single concept or relation in AMR can be posited to represent meaning conveyed in discontinuous constructions such as "as $\cdots$ as $\cdots$" which can be collapsed into a single relation *:compared-to*. Third, since AMR abstracts away from elements of surface syntactic structure such as word order and morpho-syntactic markers, which account for much of the cross-linguistic variations,

it makes a more portable semantic annotation framework across languages, as the preliminary AMR annotation on Chinese and Czech has demonstrated (Xue et al., 2014).

There are always two sides to every coin and affording annotators unconstrained freedom to make up new concepts can lead to inconsistent and unusable annotation without carefully designed guidelines that specify when a new concept can be inferred and when a discontinuous pattern can be mapped to a single concept/relation. Although annotating meaning representation independently of syntactic structures serves to speed up annotation, in automatic meaning representation parsing, morpho-syntactic structures often serve as important clues that can be used to derive the semantic representation. Some minimal correspondence between the two representations needs to be established in order to make use of the syntactic structure when developing meaning representation parsers. When conducting automatic AMR parsing, it is customary to explicitly provide the correspondence between word tokens in the sentence to the concepts and relations in its AMR, that is, the alignment between the input sentence and its AMR. Since this alignment is not provided in the English AMR Bank (Banarescu et al., 2015), AMR parsing researchers have to develop a word-to-concept aligner as the first step in AMR parsing. This can be done via either a supervised or unsupervised approach. For example, Flanigan et al. (2014) develops a rule-based aligner by independently annotating the alignment between word tokens and AMR concepts for a small corpus that can be used to extract alignment rules. The alignment F-score of this aligner is about 90%. Pourdamghani et al. (2014) develops an EM-based aligner that yields similar performance without any manual alignment. While these aligners may seem to be very accurate, a 10% error rate in alignment imposes a serious limitation on the overall AMR parsing accuracy as errors in alignment will propagate to subsequent steps.

In this article, we present the CAMR Corpus, a growing Chinese AMR corpus[2] that currently has 10,149 sentences annotated with meaning representations. We adopt the AMR approach of representing the meaning of a sentence as a rooted, directed acyclic graph, and we also adopt the AMR philosophy of annotating the meaning representation independently of syntactic structures, even

[2]https://catalog.ldc.upenn.edu/LDC2019T07

though the data we annotated are drawn from the Chinese TreeBank that already has syntactic annotation (Xue et al., 2005). In the meantime, we have also made a number of adaptations. First, rather than letting users of the corpus perform their own word-to-concept alignments, we incorporated this as an integral part of the annotation. We show in Section 3 that incorporating this alignment for a language like Chinese is straightforward and has a number of advantages. Second, while in English AMR discourse relations such as temporal and causal relations are annotated in a variety of ways, we use a dedicated set of *abstract concepts* to annotate discourse relations. This "modular" approach makes it easier for users to examine and use different aspects of the CAMR Corpus. Third, we added a few labels to the English AMR label set to account for a few Chinese-specific linguistic phenomena. In general, however, the label set used in the English AMR Bank works surprisingly well in our CAMR annotation and readily applies to Chinese data. This bodes well for this annotation framework to be applied to additional languages.

The rest of the article is organized as follows. In Section 2 we present an overview of the CAMR annotation framework that integrates word-to-concept and word-to-relation alignments. We start with a presentation of the AMR annotation specification and then outline our extensions. In Section 3 we describe how we perform alignment between the concepts/relations in AMR and word tokens in sentences. We illustrate how to handle a few well-known Chinese-specific constructions in CAMR in Section 4. In Section 5, we present results on our CAMR annotation experiments, as well as a quantitative analysis of the proportions of non-tree graphs. We describe related work in Section 6 and conclude our article with a summary of our contribution in Section 7.

2 Overview of the CAMR annotation framework

CAMR inherits the core principles of the AMR annotation in that it represents the meaning of a sentence as a single-rooted, directed, acyclic graph. The nodes of the graph are concepts and the edges represent the relations between concepts. In this section, we first provide some background and discuss how word senses and semantic roles for verbal and nominal predicates are defined in PropBank (Palmer et al., 2005) and the Chinese PropBank (Xue and Palmer,

2009), and then describe the composition of concepts and relations in an AMR (or CAMR) graph, which makes heavy use of PropBank and Chinese PropBank senses and semantic roles.

2.1 Background: Propbank and Chinese Propbank

Because AMR makes heavy use of the predicate senses and semantic roles defined in PropBank and likewise, CAMR uses the predicate senses and semantic roles in the Chinese PropBank, we will first briefly describe how the senses and semantic roles are defined in the two PropBanks so that the reader can more easily understand how AMR and CAMR concepts and relations are defined.

PropBank makes the distinction between *core* arguments and *adjunctive* arguments of a predicate. A core argument is one that is conceptually essential to (one sense of) a predicate, while an adjunctive argument is one that provides additional information that is not necessarily essential or unique to that predicate. For example, in the sentence "The girl wants to study in New York", there are two predicates: "wants" and "study". "wants" has two core arguments, "the girl" and "to study in New York", and "study" has one core argument, "the girl". "In New York" is a location that is non-essential to "study" and like time, it is not unique to "study" and can potentially be applied to many different types of arguments. In a given sentence, not all the core arguments of a predicate have to actually occur. PropBank defines a set of semantic roles for each core argument of a predicate sense and uses them to label arguments that are actually realized in a sentence. These roles range from 0 to 5, and are prefixed by *Arg*.

Table 1 gives the senses as well as the semantic roles for each sense of the English verbal predicate *want* and Chinese predicate 想. The predicate *want* has only one sense in PropBank, and it has 5 semantic roles, *Arg0* "wanter", *Arg1* "thing wanted", *Arg2* "beneficiary", *Arg3* "in_exchange_for", and and *Arg4* "from". The Chinese predicate 想 has three senses. Each sense has two semantic roles. Depending on the sense, each set of roles are interpreted differently even though they have the same role labels. For example, *Arg1* of 想-01 refers to thoughts of *Arg0* while *Arg1* of 想-02 refers to thing that *Arg0* misses. In this sense, the interpretation of the semantic roles are specific to each sense of the predicate.

The senses and the semantic roles of the core arguments are

defined for each verbal or nominal predicate in a language and they collectively constitute a valency lexicon for the language called "frame files", as each predicate has its own file. When annotating AMR or CAMR, these senses and semantic roles are consulted. This is illustrated in Figure 3, which has the AMR annotation for "The girl wants to study in New York" and its Chinese translation "女孩 想 在 纽约 上学". Node labels "want-01" and "想-02" are word senses defined in the PropBank and Chinese PropBank frame files, while edge labels *Arg0* and *Arg1* are semantic roles defined for those senses.

2.2 AMR and CAMR Concepts

Now that we have explained how senses and the semantic roles for verbal and nominal predicates are defined, we are ready to present AMR and CAMR concepts. For the sake of clarity in exposition, we find it useful to distinguish *lexical concepts* from *abstract concepts*. Lexical concepts are grounded to word tokens in a sentence, while abstract concepts are not necessarily linked to a specific lexical item. An abstract concept may be inferred from the context, or it may be an abstract characterization of one of more lexical items (e.g., *person*, *city*). This is a meaningful distinction because while the former is specific to each language, the latter is to a large extent language-independent, as evidenced by the fact that the set of abstract concepts defined in AMR readily apply to Chinese.

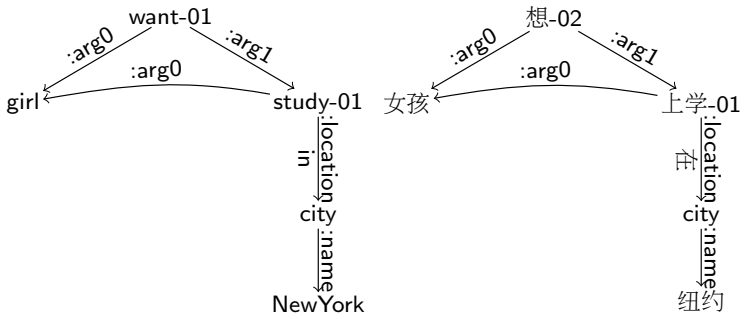


FIGURE 1 A CAMR graph and its corresponding graph

AMR uses two types of lexical concepts: i) sense-disambiguated lexical items, and ii) lemmatized words. For AMR, the sense-

TABLE 1 Semantic roles for core arguments of *want* in PropBank and 想 in the Chinese PropBank

want	想		
want-01 arg0: wanter arg1: thing wanted arg2: beneficiary arg3: in-exchange-for arg4: from	想-01 (think) arg0: people described arg1: thoughts of arg0	想-02 (want) arg0: people described arg1: thing arg0 wants	想-03 (miss) arg0: people described arg1: entity arg0 misses

disambiguated lexical items are typically verbal and nominal predicates drawn from the PropBank while for Chinese the sense-disambiguated lexical concepts are verbal and nominal predicates drawn from the Chinese PropBank. Verbal predicates in Chinese also include adjectives, which are considered to be "stative" verbs.

The sense information has not been defined for all words in the two languages. When sense definitions for a word is not available, its lemmaitzed form is used as concept. For example, in English AMR, the concepts for non-predicative nouns and adjectives are typically their lemmas as their senses have not been defined. There is no principled reason why adjectives cannot be sense-disambiguated as well, and it is simply a matter of availability. As senses are defined for these words, they can certainly be used in the AMR annotation.

The lexical concepts in the AMR graph of Figure 3 include "want-01" and "study-01", while the corresponding CAMR concepts are "想-02" and "上学-01". Concepts that are not sense-disambiguated include "girl" in AMR and "女孩" in CAMR. Notice that these lexical concepts are language-specific, and there is no attempt to establish any connection between the lexical concepts for language to those of another. The practical consequence for this is that each language can be annotated with AMR on its own without considering the vocabulary used for another language.

In contrast, abstract concepts are to a large extent language-independent. In CAMR annotation, we adopted all the abstract concepts while proposing a few new abstract concepts that we believe are needed to account for the linguistic facts of Chinese. The AMR abstract concepts mainly include i) entity types ii) quantity, iii) polarity, modality, and mode values. For example, in Figure 3, "city" is an abstract concept that represents the type of the named entity "New York". It should be noted that only named entities (in the form of proper nouns) project abstract concepts and there is an implicit hierarchy in the types of named entities that are used as abstract concepts in AMR. A more specific entity always has precedence over a more general named entity. For example, *city* is preferred over *location* in (3) because the former is a more specific named entity than the latter. *Location* is only used when the none of the more specific categories for location is appropriate.

Perhaps paradoxically, AMR concepts can also be used to represent real-world semantic relations. For example, one abstract

AMR concept is called "have-org-role-91", and it represents a real-world relation between an office-holder, the organization, title of the office held, and the responsibility of the office [3]. Similar concepts include "be-located-at-91", and the full list of such concepts are provided in Table 2.

One of the more significant differences between AMR and CAMR is how temporal and discourse relations are annotated. Since for the moment AMR is a sentence-level meaning representation, here we only discuss intra-sentential discourse relations to the exclusion of inter-sentential relations. In AMR, discourse relations are represented with a combination of abstract concepts (e.g., *and*, *or*, *contrast.01*) and relations (*:cause*, *:condition*, *:concession*, *:purpose*). This dichotomy reflects the syntactic realization of the two types of relations in English. Discourse relations represented as concepts are typically realized syntactically as coordination constructions while discourse relations represented as relations are typically syntactic subordination constructions. One drawback of this approach is that it makes it harder for users of the annotated AMR data to examine all instances of discourse relations.

In CAMR, we represent all discourse relations as concepts and we adopt the 10 discourse relations defined in the Chinese Discourse TreeBank (CDTB) (Zhou and Xue, 2015). These 10 discourse relations include *and*, *or*, which are also used in AMR, but they also include *causation*, *condition*, *contrast*, *expansion*, *purpose*, *temporal*, *progression*, *concession*. Some of these discourse relations, e.g., *causation*, *condition*, *purpose*, and *concession* are treated as relations in AMR, while others are not part of the AMR vocabulary (*expansion*, *progression*, and *temporal*). In particular *temporal* represents the temporal precedence of a sequence of discourse segments while *progression* means one argument represents a progression from the other, in extent, intensity, scale, etc. As CDTB discourse relations are formal predicates that take two or more discourse segments as their arguments, the argument labels are meaningful as well. (1) is an example of temporal relation. The arguments are arranged in chronological order, with *Arg1* temporally preceding *Arg2*, and *Arg2* temporally preceding Arg3. (2) is an example of *condition* relation.

[3] https://www.isi.edu/ ulf/amr/ontonotes-4.0-frames/have-org-role-v.html

TABLE 2 List of abstract concepts used in CAMR

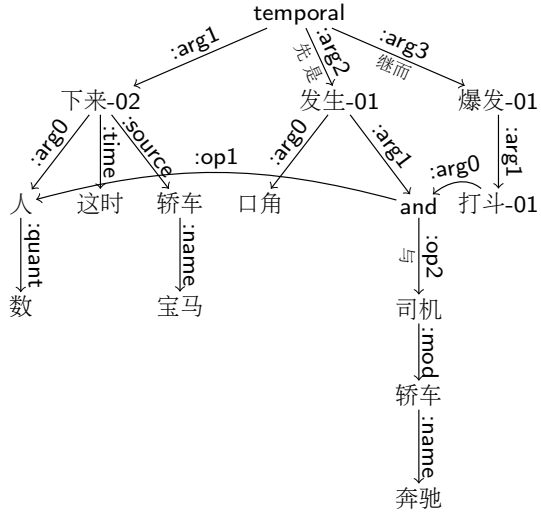
Type	Abstract concepts	Num
Named Entity (108)	thing	1
	person, family, animal, language, nationality, ethnic-group, regional-group, religious-group	8
	organization, company, government-organization, military, criminal-organization, political-party, school, university, research-institute, team, league	11
	location, city, city-district, county, local-region, state, province, country, country-region, world-region, continent, ocean, sea, lake, river, gulf, bay, strait, canal, peninsula, mountain, volcano, valley, canyon, island, desert, forest, moon, planet, star, constellation	29
	facility, airport, station, port, tunnel, bridge, road, railway-line, canal, building, theater, museum, palace, hotel, worship-place, market, sports-facility, park, zoo, amusement-park	20
	event, incident, natural-disaster, earthquake, war, conference, game, festival	8
	product, vehicle, ship, aircraft, aircraft-type, spaceship, car-make, work-of-art, picture, music, show, broadcast-program	12
	publication, book, newspaper, magazine, journal	5
	naturalobject	1
	molecular-physical-entity, small-molecule, protein, protein-segment, amino-acid, macro-molecular-complex, enzyme, rna, pathway, gene, dna-sequence, cell, cell-line, organism, disease	15
	law, treaty, award, food-dish, dynasty	5

Type	Abstract concepts	Num
∗discourse	and, or, ∗causation, ∗condition, ∗contrast, ∗temporal, ∗concession, ∗progression, ∗purpose, ∗expansion, multi-sentence	11
subjectivity	-(polarity), +(polite), possible	3
mode	interrogative, expressive, imperative	3
unknown	amr-unknown	1
quantity	monetary-quantity, distance-quantity, area-quantity, volume-quantity, temporal-quantity, frequency-quantity, speed-quantity, acceleration-quantity, mass-quantity, force-quantity, pressure-quantity, energy-quantity, power-quantity, voltage-quantity, charge-quantity, potential-quantity, resistance-quantity, inductance-quantity, magnetic-field-quantity, magnetic-flux-quantity, radiation-quantity, concentration-quantity, temperature-quantity, score-quantity, fuel-consumption-quantity, seismic-quantity	26
91 concept	have-concession, have-condition, be-destined-for, have-frequency, have-instrument, be-located-at, have-manner, have-mod, have-name, have-part, have-polarity, have-purpose, have-quant, be-from, have-subevent, include, be-temporally-at, rate-entity	18
	Total	179

∗ marks new concepts added to CAMR

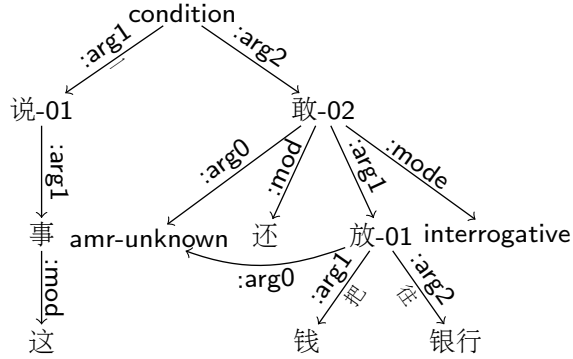
(1) 这时[1] ，[2] 宝马[3] 轿车[4] 上[5] 下来[6] 数[7] 人[8] ，[9]
this time , BMW sedan up come down several people ,
"At this time, several people came down from the BMW sedan,"
与[10] 奔驰[11] 轿车[12] 司机[13] 先[14] 是[15] 发生[16] 口角[17] ，[18]
with Benz sedan driver first is happen quarrel ,
"first quarreled with the driver of the Benz sedan, "
继而[19] 爆发[20] 打斗[21] 。[22]
then erupt fighting
"then a fighting broke out between them."

```
x24/temporal
    :arg1 x6/下来-02
        :arg0 x8/人
            :quant x7/数
        :time x1/这时
        :source x4/轿车
            :name x1/宝马
    :arg2(x14_x15/先是) x16/发生-01
        :arg0 x17/口角
        :arg1 x39/and
            :op2(x10/与) x13/司机
                :mod x12/轿车
                    :name x11/奔驰
            :op1 x8
    :arg3(x19/继而) x20/爆发-01
        :arg1 x21/打斗-01
            :arg0 x39
```



(2) 这[1] 事[2] 一[3] 说[4] ，[5]
this thing once said ,
"Once the thing was said, "
谁[6] 还[7] 敢[8] 把[9] 钱[10] 往[11] 银行[12] 放[13] 呀[14] ？[15]
who also dare BA money to bank put ah ?
"who would dare to put the money in the bank?"

```
x19/condition
    :arg1(x3/一) x4/说-01
        :arg1 x2/事
            :mod x1/这
    :arg2 x8/敢-02
        :arg0 x6/amr-unknown
        :mod x7/还
        :arg1 x13/放-01
            :arg0 x6
            :arg1(x9/把) x10/钱
            :arg2(x11/往) x12/银行
        :mode x14_x15/interrogative
```



2.3 Relations

Like AMR, the CAMR relations include semantic roles as well as nominal relations. In computational linguistics, semantic roles come in different favors, and a survey of these different approaches can be found in Bai and Xue (2016). The three representative approaches include the Lyrics/VerbNet types of semantic roles which are defined independently of the types of predicates, the FrameNet styles of semantic roles which are defined with respect to specific *frames*,

and the PropBank-style of semantic roles which are defined with respect to specific predicates. Propbank uses predicate-specific numbered roles for the core arguments of each predicate, verbal and nominal, and uses more general roles for adjunctive arguments, which are not specific to a predicate. AMR adopts this PropBank approach for labeling the semantic roles for the core arguments, but substantially expands the set of semantic roles for adjunctive arguments. It also adds semantic relations that are typically not considered to be semantic roles. In CAMR, we also adopt the PropBank approach to represent semantic roles for core arguments, and use 6 semantic role labels for core arguments (*Arg0-Arg5*) as they are defined in the Chinese Proposition Bank, and 44 labels for adjunctive arguments and other semantic relations largely taken from the AMR label set.

TABLE 3 The full set of semantic relations used in CAMR

:accompanier, :age, ∗:aspect, :beneficiary, :cause , :compared-to, :consist-of, :cost , ∗:cunit, :degree, :destination, :direction, :domain, :duration, :example, :extent, :frequency, :instrument, :li, :location, :manner, :medium, :mod, :mode, :name, :ord, :part, :path, ∗:perspective, :polarity, :polite, :poss, :purpose, :quant, :range, :source, :subevent, :subset, :superset, ∗:tense, :time, :topic, :unit, :value

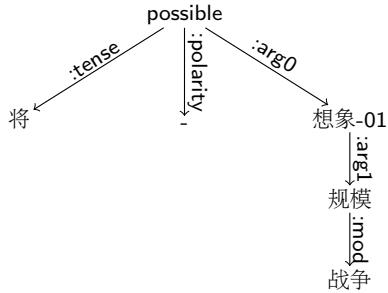
∗ marked the new relations added to CAMR

A full set of semantic relations are listed in Table 3, with relations added in CAMR prefixed by ∗. *:cunit* is introduced to represent Chinese classifiers that are discussed in more detail in Section 4 when we discuss Chinese-specific constructions. An example of *:cunit* can be found in (11). We also introduced *:tense* and *:aspect* in CAMR annotation as we believe these two categories are important to make the AMR representation more expressive and more faithful to the meaning expressed by the sentences. While tense and aspect are realized in English as morphological inflections, specifically as suffixes on verbs, in Chinese they are realized as stand-alone lexical items or particles. For example, 将 (will) is a lexical item that indicates tense, while 着 (Progressive), 了 (Complete) and 过 (Complete) are aspect markers. We should note that tense and aspect are only annotated when an overt lexical marker exists. Unlike English

where each finite verb is morphologically inflected for tense, in Chinese only a small proportion of verbs are associated with an overt lexical tense or aspect marker, so in practice, only a small proportion of verbal predicates are annotated with tense and aspect.

(3) 战争[1] 规模[2] 将[3] 无法[4] 想象[5]
war scale will unable imagine
"The scale of the war will be unimaginable."

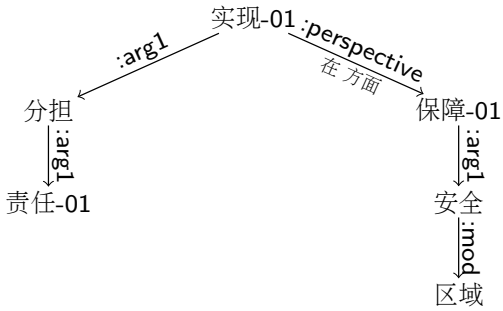
```
x8/possible
    :polarity x4/-
    :tense x3/将
    :arg0 x5/想象-01
        :arg1 x2/规模
            :mod x1/战争
```



Another non-core relation we added is *perspective*. It is not the core argument of a verb, but it indicates the perspective of the statement. This is illustrated in (4).

(4) 在[1] 区域[2] 安全[3] 保障[4] 方面[5] 实现[6] 责任[7] 分担[8]
at area security ensure aspect achieve responsibility share
"Achieve responsibility sharing in ensuring regional security"

```
x6/实现-01
    :arg1 x8/分担-01
        :arg1 x7/责任
    :perspective(x1_x5/在……方面) x4/保障-01
        :arg1 x3/安全
            :mod x2/区域
```

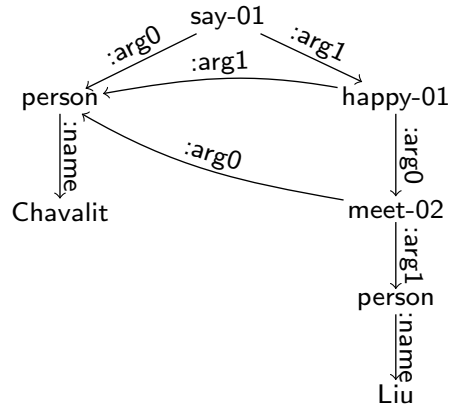


3 Sentence-to-CAMR alignment

As we briefly mentioned in the introduction section, one hallmark of AMR annotation is the decoupling of the strict correspondence between the word tokens in a sentence and the concepts and relations in AMR. However, for automatic AMR parsing, the process of taking a sentence as an input and producing an AMR representation for it as an output, alignment between word tokens in a sentence and concepts/relations in AMR is essential to the effective modeling of the derivation process of how a sentence is transformed into its AMR. It is worth noting, however, that alignment does not reverse the effect of decoupling the strict correspondence between word tokens in a sentence and concepts and relations in an AMR graph. Alignment is performed only if it is possible — in some cases a word token may not map to any concept or relation in the AMR graph, while in other cases a concept or relation may not map to any word token. In (5), for example, the word "that" does not map to any concept or relation, so it cannot be aligned. Similarly, the concept *person* is an abstract concept that cannot be aligned. However, in cases where a word token can be aligned to a concept or relation, it should be aligned to aid the automatic parsing process.

(5) Chavalit said that he was happy to meet Liu.

```
s/say-01
  :arg0 p/person
      :name c/Chavalit
  :arg1 h/happy-01
      :arg0 m/meet-02
          :arg0 p
          :arg1 p2/person
              :name l/Liu
      :arg1 p
```



The word-to-concept/relation alignment is not integrated into the English AMR annotation process, mainly out of concern that it will slow down AMR annotation too much and it is too complex to provide support for this when developing an annotation tool. For example, it is non-trivial to automatically generate the

concept from an English word due to the fact that English words are often morphologically inflected. There was also a hope that the alignment can be learned automatically in an unsupervised manner with EM-based algorithms, just like word alignment between different languages can be learned without the need for manual annotation. Although this expectation has been partially born out in the work of Pourdamghani et al. (2014), we argue that an error rate of around 10% is too much of a deficit in the AMR parsing process to achieve an AMR parser that is as accurate as possible. In order for AMR parsing accuracy to approach that of syntactic parsing where there is an inherent alignment between the word tokens in a sentence and the leaf nodes of a syntactic parse, starting with accurate word-to-concept/relation alignment is crucial. With this in mind, we have decided to incorporate alignment into the CAMR annotation process. Chinese has an advantage in this regard as it has very limited morphological inflection and generating lemmatized concepts is relatively straightforward. It is also worth noting that unlike word alignment in parallel text for training Machine Translation systems, where the volume of parallel text is too large to realistically perform manual alignment on, we do not expect to the amount of AMR annotation will ever reach that scale and manual alignment is feasible.

In the rest of the section, we will present our alignment approach and then discuss some of the details in word-to-concept and word-to-relation alignment.

3.1 Alignment approach

Our general approach is to integrate alignment into the Chinese AMR annotation process, starting with the development of an annotation tool that allows annotators to input the index of a word token instead of the concept or relation itself. The annotation tool presents a text for annotation one sentence at a time. As the annotator inputs the index of a word token, the annotation tool will automatically retrieve the word token based on its index and generate the concept for it. It also generates an ID for the concept using the index of the word token, thus establishing the alignment between the AMR concepts. When generating the concept, the tool will have to perform automatic lemmatization, which fortunately is very straightforward for Chinese where there is little inflectional

morphology. In many cases, the lemma is the concept, in which case the annotator does not have to do anything further. In other cases, the lemma needs to be sense-disambiguated when the senses for the lemma are defined. This is the case with verbal or nominal predicates, the senses of which are defined in the Chinese Propbank frame files. In this case, the tool allows the annotator to revise the concept by adding the sense ID to the lemma. The lemma also needs to be revised when a word does have morphological inflections in a limited number of cases or when the word is misspelled.

We illustrate this process with the example in (6). 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 within the same CAMR. For example, the IDs of 喜欢-01 and 唱-01 are "x2" and "x5" respectively, indicating that they are aligned to the 2nd and 5th word of the sentence. For abstract concepts that do not correspond to any word token, they are assigned IDs that have a value greater than the total number of word tokens in the sentence. For example, in (6) *person* is an abstract concept that is essentially an entity type for the word token that has the ID"x4", 邓丽君, and is not aligned to any word token in the sentence, so we assign it the ID "x8", an ID that is greater than the maximum length of the sentence. The functional word 的 (DE), which does not correspond to any concept in the AMR graph, is aligned to the relation *:arg1-of*. Table 4 shows what the annotator enters as input in the annotation interface in order to generate the CAMR graph for the Chinese sense[4]. This example also serves to show that while the CAMR of a sentence diverges from the word tokens due to the existence of abstract concepts or word tokens that do not map to a concept, it is still useful to provide alignment annotation when it is plausible, for purposes of training automatic CAMR parsers.

(6) 他[1] 喜欢[2] 听[3] 邓丽君[4] 唱[5] 的[6] 歌[7]
He like listen Lijun Deng sing DE song
"He likes to listen to the songs sung by Lijun Deng"

[4]The annotation interface allows the user to choose the sense ID ("喜欢-01") of a predicate ("喜欢") from a list of possible senses of the predicate.

x2/喜欢-01
    :arg0 x1/他
    :arg1 x3/听-01
        :arg1 x7/歌
            :arg1-of(x6/的) x5/唱-01
                :arg0 x8/person
                    :name x4/邓丽君

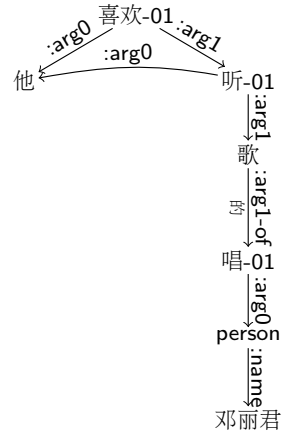


TABLE 4 Annotator's inputs in CAMR Annotation Toolkit

Annotator's Inputs	Generated AMR Graph
root :top x2	x2/喜欢-01
x2 :arg0 x1	:arg0 x1/他
x2 :arg1 x3	:arg1 x2/听-01
x3 :arg1 x7	:arg1 x7/歌
x7 :arg1-of(x6) x5	:arg1-of(x6/的) x5/唱-01
x5 :arg0 person	:arg0 x8/person
x8 :name x4	:name x4/邓丽君

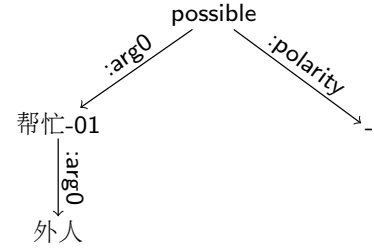
This approach outlined here is an extension of the alignment approach described in Li et al. (2016), where only concepts are aligned but relations are not. In addition to its benefits to automatic AMR parsing, our new alignment scheme also has other benefits. (i) Using the concept IDs accelerates the manual annotation by about 10∼20%. It reduces the time needed to input the word form and to shift the input methods between English and Chinese. (ii) The annotation tool also keeps track of which words in the sentence have been "covered" at any point during AMR annotation 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 annotators to miss some words. (iii) With the alignment, it is easy to determine which words are omitted, which concepts are inferred, and whether a word is aligned to a concept or relation.

3.2 Word-to-Concept alignment

Since AMR abstracts away from surface forms of a sentence, there are 5 basic types of abstraction: insert, delete, replace, merge and split (see Table 5). Some word tokens are considered to be devoid of meaning and are not represented in the AMR. Words that are not represented in AMR include determiners such as "a","an", "the", and infinitive marker "to". On the other hand, there are also abstract concepts in AMR that are not grounded to any specific lexical item and are inferred from the context. In some cases, one word token is analyzed into multiple AMR concepts. For example, the English word "protector" is represented in a similar way to "person who protect" in AMR. In other cases, multiple word tokens in a sentence may represent a single AMR concept. These word tokens do not even have to be contiguous. For example, the discontinuous Chinese words 帮 ··· 忙 are merged to one single concept 帮忙. So other than straightforward one-to-one mappings between word tokens and AMR concepts, there are also complex alignment patterns such as one-to-zero, zero-to-one, one-to-many and many-to-one alignments. In many ways, this is not too different from word alignment between two languages. As we mentioned briefly above, having this alignment is important to AMR parsing. Word-to-concept alignment is essential to this process, not unlike the role of word alignment to statistical machine translation.

In addition to one-to-one, one-to-zero, and zero-to-one alignments, there are also one-to-many and many-to-one alignments between word tokens in a sentence and concepts in its AMR. The following is the AMR for Example (7) where one AMR concept is aligned to two word tokens that are also discontinuous. This is a case of split verbs that we will discuss in Section 4. The word tokens are "帮 ··· 忙" and the AMR concept is simply 帮忙. Its ID is a concatenation of the indices of the two word tokens "x2_x5" .

(7) 外人[1] 帮[2] 不[3] 上[4] 忙[5]
outsiders help no up business
"Outsiders cannot help"




```
x4/possible
    :arg0 x2_x5/帮忙-01
        :arg0 x1/外人
    :polarity x3/-
```

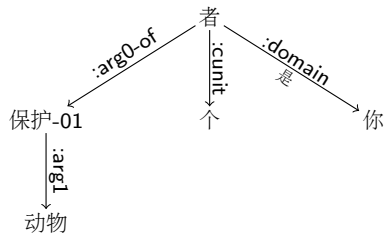
(8) is an example where one word is aligned to multiple con-

TABLE 5 Abstraction Types

Abstraction Types	English Example	Chinese Example
Insert	the young →young (person)	唱歌的→唱歌的(人)
Delete	the boy →boy	解释道→解释
Replace	like→resemble-01	宛如→像-01
Merge	as ···as···→compared-to	帮···忙→帮忙
Split	protector→person :arg0-of protect-01	保护者→者:arg0-of 保护-01

cepts. This usually happens when the word has a complicated internal structure and each morpheme corresponds to an AMR concept. Chinese has very little derivational or inflectional morphology, but compounding is a highly productive morphological process.

(8) 你[1] 是[2] 个[3] 动物[4] 保护-者[5]
you are CL animal protector
"You are an animal protector"



x5_3/者
:arg0-of x5_1_2/保护-01
:arg1 x4/动物
:cunit x3/个
:domain(x2/是) x1/你

In (8), the compound word 保护者 (protector) has 3 characters and corresponds to two AMR concepts: 保护 (protect) and 者 (person). In this case, we represent the alignment with the character offsets within the compound word. Notice that character offsets, unlike word indices, are not prefixed with "x". This is how we differentiate word indices from character offsets. For example, the concept ID for 保护 is "x5_1_2", meaning that it is aligned with the first two characters of the fifth word. Similarly, the ID for the concept 者 is "x5_3", meaning that it is aligned with the third character of the fifth word.

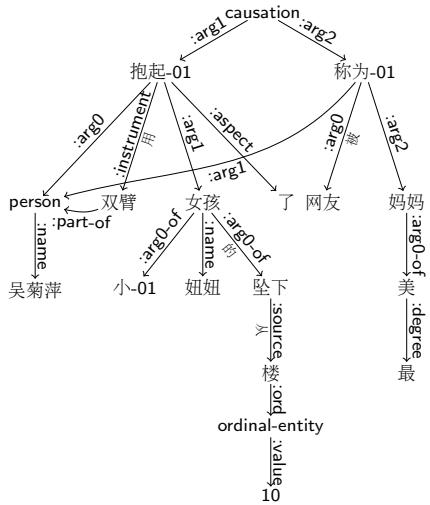
3.3 Word-to-Relation alignment

In addition to word-to-concept alignment, we also align words to relations. Relations are typically signaled by function words. For example, in the English sentence "he walks in the room", "in" indicates the *:location* where he walks. Similarly, in (9), the Chinese case marker 用 (*with*) is aligned to *:instrument*, and 被 (*by*) is aligned to *:arg0*. We argue that it is necessary to annotate functional words because they are manifestations of the semantic relations between two words. In other words, these words are the relation markers.

(9) 吴菊萍[1] 用[2] 双臂[3] 抱起[4] 了[5]
Juping Wu with arms pick up ASP
从[6] 十[7] 楼[8] 坠下[9] 的[10] 小[11] 女孩[12] 妞妞[13] ，[14]
from tenth floor fall DE little girl Niuniu ,
"Juping Wu picked up the little girl Niuniu who fell from the tenth floor with her arms,"

被[15] 网友[16] 称为[17] “[18] 最[19] 美[20] 妈妈[21] ”[22] 。[23]
by netizens call “ most beautiful mother ” .
“and was called ‘the most beautiful mother’ by netizens.”

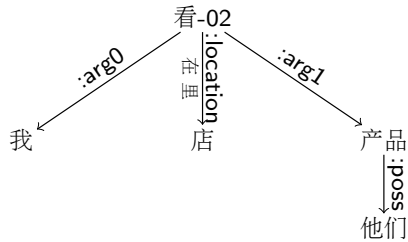
```
x31/causation
    :arg1 x4/抱起-01
        :arg0 x34/person
            :name x1/吴菊萍
        :aspect x5/了
        :arg1 x12/女孩
            :arg0-of x11/小-01
            :name x13/妞妞
            :arg0-of(x10/的) x9/坠下
                :source(x6/从) x8/楼
                    :ord x43/ordinal-entity
                        :value x7/10
        :instrument(x2/用) x3/双臂
            :part-of x34
    :arg2 x17/称为-01
        :arg0(x15/被) x16/网友
        :arg1 x34
        :arg2 x21/妈妈
            :arg0-of x20/美-01
                :degree x19/最
```



Many-to-one mappings also happen in word-to-relation alignment and like concept alignment we represent many-to-one alignments by concatenating word indices when two or more function words in conjunction express the same semantic relation. For example, in (10), "在 ··· 里" means "in", which is aligned to the relation *:location*.

(10) 我[1] 在[2] 店[3] 里[4] 看[5] 他们[6] 产品[7]
I at store in look their products
"I look at their products in their store"

x5/看-02
:arg0 x1/我
:location(x2_x4/在……里) x3/店
:arg1 x7/产品
:poss x6/他们



4 Chinese specific constructions

Even though we use the same annotation convention and mostly the same vocabulary as used in the English AMR, we still need to specify how to annotate Chinese-specific constructions that are not in English so that these constructions are consistently annotated. Due to the limitation of space, we only describe six such constructions: number and classifier construction, serial verb construction, headless relative construction, verb complement (VC) construction, split verb construction, and reduplication. We will also discuss how to represent discourse relations in Chinese AMR, an area where there are significant adaptations.

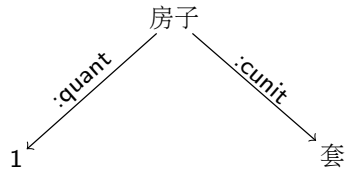
4.1 Number and classifier construction

When a number modifies a Chinese noun or verb, it is always followed by a classifier. A classifier can be a measure word like 公斤,

which has an equivalent word in English, "kilogram" . However, there is also another type of classifier which does not have an English equivalent. It serves as a cognitive measure of things and its meaning is hard to represent. The word 套 in (11) is such an example. It is also very idiosyncratic in the type of nouns it can modify. For example 套 can be used to modify house and furniture, but not other things such as apples or cars. They are generally referred to as "individual classifiers" in Chinese linguistics. As AMR is concerned with the abstract meaning, we keep the measure words in the AMR representation and annotate the individual classifiers as *:cunit* relations in a CAMR graph. Notice that the numbers are also normalized to Arabic numerals.

(11) 一[1] 套[2] 房子[3]
a CL house
"A house"

x3/房子
:quant x1/1
:cunit x2/套



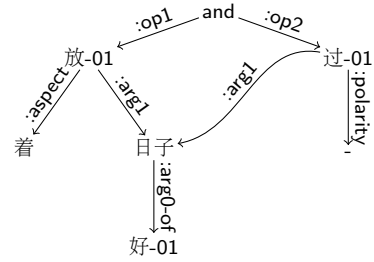
4.2 Serial-Verb construction

Serial-verb constructions are very common in Chinese. It is characterized by having several verbs in a sequence, but it is sometimes very hard to determine the grammatical relations between them. For example, in some cases one verb modifies another while in other cases the two are semantically equally important as in a coordinate structure. We choose to avoid making this hard decision for now for the sake of consistent annotation and consider these verbs to be in a coordination structure and create a non-lexical "and" concept to connect them. It is worth noting that Chinese linguistics researchers differ as to what counts as a serial verb construction, which is really a descriptive term that does not have a generally agreed-upon scope of linguistic phenomena that it applies to. Serial verb constructions, when defined broadly, can include cases where any two or more verb phrases occurring in a sequence. This broader interpretation of serial verb construction will include examples in (1), which we interpret as a temporal relation or (2), which we interpret as a discourse relation of condition and consequence. What we consider to be a serial verb construction is narrower in scope, and only include cases like (12), where the relation between the

serial verbs is hard to define.

(12) 放[1] 着[2] 好[3] 日子[4] 不[5] 过[6]
leave ASP good life not live
"Do not want to settle with living a good life"

x8/and
:op1 x1/放-01
:aspect x2/着
:arg1 x4/日子
:arg0-of x3/好-01
:op2 x6/过-01
:arg1 x4
:polarity x5/-

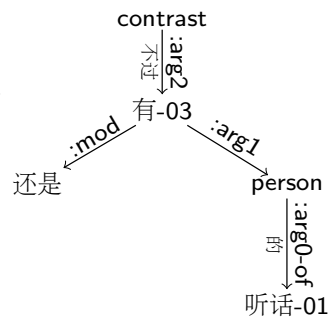


4.3 Headless relative construction

Headless relative constructions are relative constructions without an explicit noun head. Syntactically it is realized as a relative clause followed by 的 (DE), a function word that serves multiple purposes, one of which is to serve as the marker of a relative clause. The dropped noun head of the relative clause could play any roles with regard to the verb in the relative clause: agent, patient, instrument, location, etc. When doing CAMR annotation, we use an abstract concept to represent the dropped noun head. In (13), for example, the abstract noun head is a "person", and it is Arg0 of the verb 听话 (obedient).

(13) 不过[1] 还是[2] 有[3] 听话[4] 的[5]
but still have obedient DE
"But there are still obedient people"

x8/contrast
:arg2(x1/不过) x3/有-03
:mod x2/还是
:arg1 x9/person
:arg0-of(x5/的) x4/听话-01



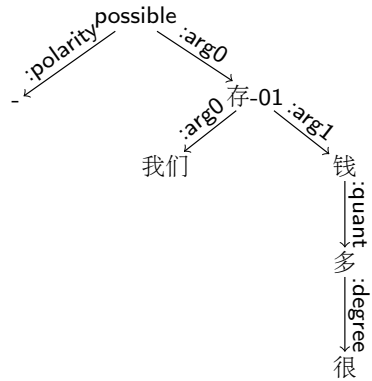
4.4 Verb-Complement construction

A Verb-Complement (VC) construction is composed of a verb followed by another verb that indicates possibility, result, etc. The function word 得 (DE) can optionally come between those two

words. In AMR annotation, we make the meaning of the construction explicit using abstract concepts or relations. In (14), for example, the VC construction has a modal meaning, represented by "possible" , although there isn't one word that specifically means possible. This meaning comes from the VC construction. In (15), there is a causal relationship between the two verbs 挑动 (provoke) and 失控 (out of control), represented as a *:cause* relation between the two verbs.

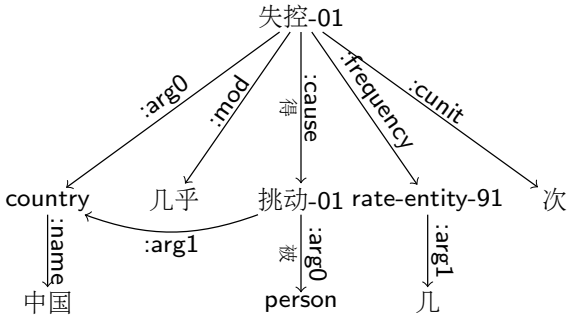
(14) 我们[1] 存[2] 不[3] 了[4] 很[5] 多[6] 钱[7]
we save not possible very much money
"We can not save a lot of money"

```
x4/possible
    :polarity x3/-
    :arg0 x2/存-01
        :arg0 x1/我们
        :arg1 x7/钱
            :quant x6/多
                :degree x5/很
```



(15) 中国[1] 几[2] 次[3] 被[4] 挑动[5] 得[6] 几乎[7] 失控[8]
China several times by provoke DE almost out of control
"China has been provoked almost out of control several times"

```
x8/失控-01
    :arg0 x10/country
        :name x1/中国
    :mod x7/几乎
    :cause(x6/得) x5/挑动-01
        :arg1 x10
        :arg0(x4/被) x11/person
    :frequency x12/rate-entity-91
        :arg1 x2/几
    :cunit x3/次
```

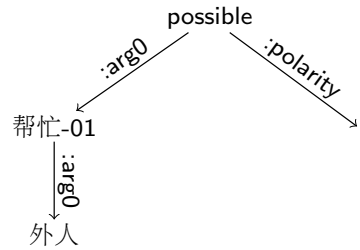


4.5 Split verb construction

A "split verb" is a verb whose two parts can be separated by other words. 帮忙 (help) is a typical example. When it is separated, it takes the form of a verb (帮) followed by an object (忙), separated by some modifiers. Its syntactic representation is quite a paradox: on the one hand, the semantics of the two parts are not separable, and it simply means "help"in its totality. On the other hand, it takes the form of a verb–object construction, and needs to be represented that way. AMR solves this paradox by just representing the entire construction as one concept, 帮忙, regardless of whether it is split or not.

(16) 外人[1] 帮[2] 不[3] 上[4] 忙[5]
outsiders help no up business
"Outsiders cannot help"

x4/possible
:arg0 x2_x5/帮忙-01
:arg0 x1/外人
:polarity x3/-



4.6 Reduplications

There are two types of reduplications in Chinese. In the first type of reduplications (17a-17b), the reduplicated form has roughly the same meaning as the root form. The reduplication has either an aspectual meaning that the root form does not have (17a), or has its meaning intensified (17b). For the moment, we do not represent such subtle aspectual meanings or intensification. In the second type, however, the reduplicated form clearly adds meaning to its root form (17c, 18). We annotate their actual meaning by adding an abstract concept. The root form is in brackets in the following

examples:

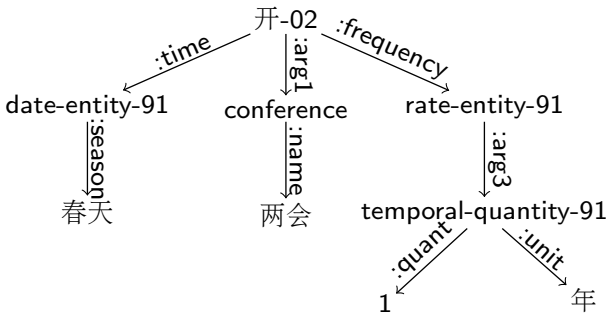
(17) a. 说 说 说
say say say

b. 干 干 净 净 干净
dry dry clean clean clean

c. 年 年 every 年
year year every year

(18) 年年[1] “[2] 两[3] 会[4] ”[5] 春天[6] 开[7]
year year “ two sessions ” spring hold
“Every year, the two sessions are held in spring”

x7/开-02
:time x10/date-entity-91
:season x6/春天
:arg1 x13/conference
:name x3_x4/两会
:frequency x15/rate-entity-91
:arg3 x16/temporal-quantity-91
:quant x17/1
:unit x1/年



5 Corpus statistics

The CAMR corpus[5] currently includes 10,149 sentences from the CTB8.0 with a total word count of 227,661 and character count of 347,750. The average word count per sentence is 22.43, and the average concept count per sentence is 19.24. Among the concepts of a

[5] https://catalog.ldc.upenn.edu/LDC2019T07

sentence, 16.35 of them are *concrete* concepts that are aligned to a specific lexical item and 2.88 of them are *abstract* concepts that do not necessary correspond to a lexical item. Rather, they are inferred from the context or are categorizations of a named entity. An average sentence has 22.50 relations, and the fact that there are more relations (which are labeled arcs that connect concepts which are labeled nodes in the AMR graph) than concepts suggest that there are re-entrancies. On average, there are 1.99 re-entrancies per sentence in the CAMR corpus. These basic statistics are summarized in Table 6. The CAMR corpus also includes 1,562 sentences from the Chinese version of *the Little Prince*, which has shorter sentences and simpler AMRs.

The predicate-argument structure annotation in the CAMR corpus is based on the frame files for the Chinese Proposition Bank (CPB) 3.0. The frame files define the senses (called *framesets*) of each verbal or nominal predicate in Chinese, as well as the set of arguments for each predicate sentence. The frame files include 24,510 Chinese predicates and 26,650 framesets. Two linguistics under-graduate students were trained to perform the annotation. To evaluate annotation consistency, each annotator completed the annotation for all of the 1,562 sentences from the Chinese translation of *the Little Prince* and 500 sentences from CTB, and the inter-annotator agreement (IAA) is 0.83, as calculated by Smatch toolkit (Cai and Knight, 2013). The rest of the sentences are single-annotated.

5.1 Non-tree Graphs

One distinctive characteristic of AMR annotation is that it allows re-entrancy, which means that the mathematic objects used to represent AMR can be non-tree graphs. This has profound implications for the class of algorithms that can be used to parse AMRs. In this subsection we take a deeper look at the proportion of AMR graphs that are non-tree graphs, and compare the proportions of non-tree graphs in English AMR corpus and CAMR corpus. In the CAMR corpus, about 53% of the sentences only have simple tree structures and do not have re-entrancies. The remaining 47% of the sentences are non-tree graphs that have at least one instance of re-entrancy. Table 7 presents a comparison of re-entrancies between the CAMR corpus and the English AMR corpus. As can be seen from the ta-

TABLE 6 Basic statistics of the CAMR corpus

sentences	words	concepts	relations	re-entrancies	non-tree graphs	abstract concepts	sentences with abstract concepts
10,149	227,661	195,282	228,410	9,449	4,741 (46.71%)	26,269	9,039 (88.95%)
characters	words per sent	concepts per sent	relations per sent	re-entrancies per sent	concrete concepts per sent	abstract concepts per sent	words aligned to relations
347,750	22.43	19.24	22.50	1.99	16.35	2.88	29,533 (12.97%)

ble, 49% of the English sentences have non-tree graphs while for Chinese that ratio is 45%.

table 7 The re-entrance arcs in AMR corpus

AMR Corpus	# of sentences	# of sentences with re-entrancies	ratio
eng_bolt	1,062	722	0.68
eng_dfa	1,703	898	0.53
eng_mt09sdl	204	137	0.67
eng_proxy	6,603	2,954	0.45
eng_xinhua	741	423	0.57
eng_Little prince	1,562	663	0.43
eng_total	11,875	5,797	0.49
chs_Little Prince	1,562	576	0.36
chs_CTB	10,149	4,741	0.47
chs_total	11,711	5,317	0.45

The re-entrancy arcs are mainly caused by *argument sharing*, meaning multiple predicates sharing one argument. In a tree structure, a concept can only be dominated by one other concept, while in AMR, a concept can be dominated by two or more predicates, in which case the AMR will be a non-tree graph. The number of predicates sharing one argument ranges from 1 to 12, meaning that a concept could be dominated by as many as 12 other concepts in the CAMR corpus. Perhaps unsurprisingly, the probability of an AMR being a non-tree graph is highly correlated with the length of the sentence. Figure 2 illustrates how the ratio of an AMR being a non-tree graph grows as the number of words in the sentence increases. The longer the sentence is, the more likely it will be a non-tree graph, with only a few exceptions.

5.2 Inverse Relations

AMR is formally a single-rooted, directional, and acyclic graph, and the property of being single-rooted is made possible to a large extent by the use of inverse relations. For each relation (e.g., *arg0*), there is also a corresponding inverse relation (e.g., *arg0-of*) that allows the dominance relation between two concepts to be switched. This is illustrated in Example 19, where the concept *person* is an argument of the predicate 听话-01. Typically the predicate con-

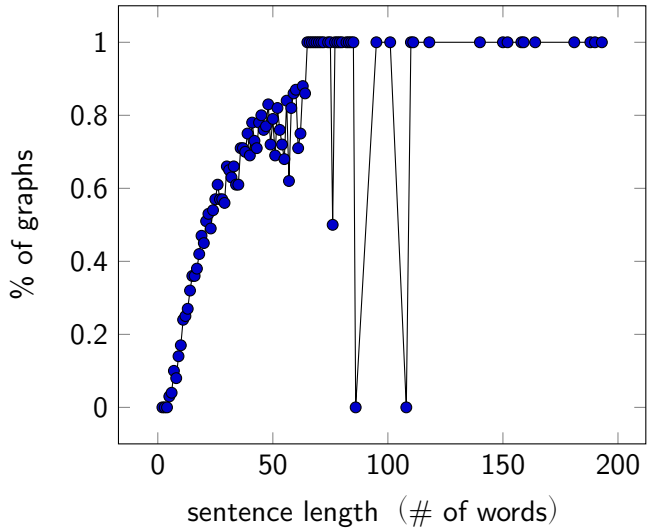


FIGURE 2 The ratio of sentences being non-tree graphs

cept dominates the argument but in an inverse relation the dominance relation is reversed. As should be clear from Example 19, the AMR for the sentence would no longer be single-rooted if the inverse relation is not used. In other words, AMR trades off the use of larger set of relations for a simpler structure (single-rooted vs. multi-rooted). It should be noted for AMR, there is no semantic difference in how a relation and its inverse counterpart are interpreted. An inverse relation is only used when it is necessary to maintain the single-rootedness of the graph. Trained annotators can recognize syntactic constructions (e.g., the relative construction) where inverse relations are typically needed so their recognition is not an obstacle for the annotator.

(19) 不过[1] 还是[2] 有[3] 听话[4] 的[5]
but still have obedient DE
"But there are still obedient people"

```
x8/contrast
    :arg2(x1/不过) x3/有-03
        :mod x2/还是
        :arg1 x9/person
            :arg0-of(x5/的) x4/听话-01
```

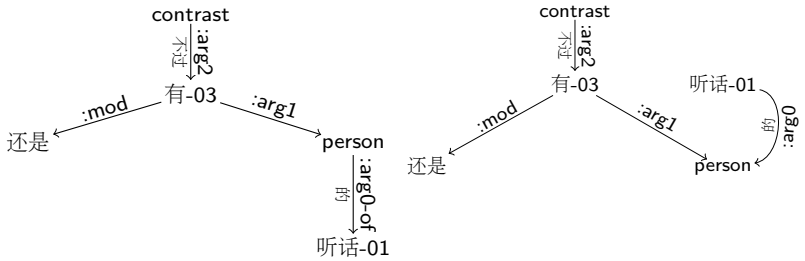
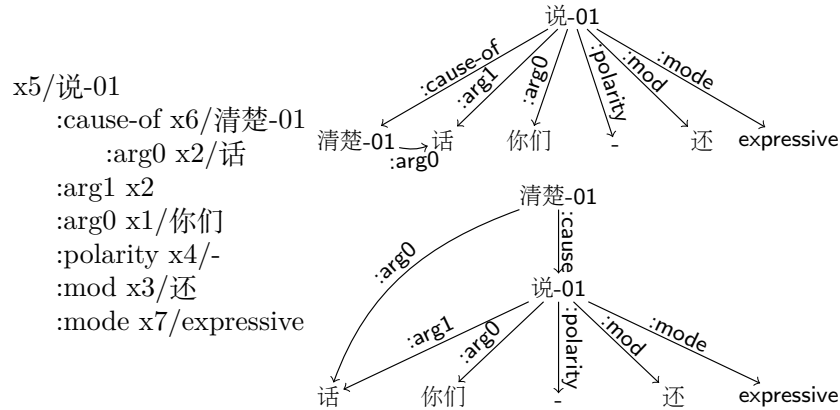


FIGURE 3 A CAMR graph and its corresponding graph

In addition to being crucial to ensuring that AMRs are single-rooted graphs, inverse relations can also be interpreted as reflecting the focus of the speaker/writer, although this interpretation cannot consistently applied. This is illustrated in Example (20). Depending on which concept is more prominent, either 说-01 or 清楚-01 can be the focus and be the dominating concept, leading to different AMR graphs, although the two graphs are semantically equivalent.

(20) 你们[1] 话[2] 还[3] 没[4] 说[5] 清楚[6] 呢[7]
you words still never say clearly EMPHASIS
"You have not made it clear yet"



In the CAMR corpus, 27 types of inverse relations are attested with 11,338 instances. The *:arg0-of* and *:arg1-of* relations are most common, accounting for almost 90% of the instances in the corpus. The distribution of the inverse relations are presented in Table 8.

Table 9 presents a comparison of the use of inverse relations between the CAMR corpus and the AMR corpus. The AMR statistics

TABLE 8 Top 10 inverse relations

Rank	Relations	Count	Portion
1	:arg0-of	8,136	71.76%
2	:arg1-of	2,057	18.14%
3	:cause-of	404	3.56%
4	:part-of	172	1.52%
5	:arg2-of	165	1.46%
6	:time-of	80	0.71%
7	:instrument-of	79	0.70%
8	:location-of	73	0.64%
9	:cost-of	48	0.42%
10	:purpose-of	25	0.22%

are from Kuhlmann and Oepen (2016), based on the AMR version LDC2014T12. Table 9 shows the proportion of non-tree graphs between the AMR corpus and the CAMR corpus are very close. When the inverse relations are "nominalized", the number of multi-rooted graphs jumps from zero to 57.59% for the CAMR corpus and to 77.5% for the AMR Corpus, indicating that the use of inverse relations is crucial to maintaining the single-rootedness property of AMR. The proportion of non-tree graphs also jumps from 46.71% to 74.72% for the CAMR corpus, and from 47.52% to 81.4% for the AMR corpus.

TABLE 9 The Comparison of inverse relations between Chinese and English

Comparison	Chinese	English
Sentences	10,149	10,309
Non-tree Graphs	46.71%	47.52%
Graphs without inverse relations	74.72%	81.4%
Multi-rooted sentences without inverse relations	57.59%	77.5%

6 Related work

The work we report in this article is obviously most closely related to the English AMR project, which itself is built on over a decade of research on semantic annotation that focused on different meaning components, the most notable of which is the predicate-argument structure annotation of the PropBank (Palmer et al., 2005) and Chinese Propbank (Xue and Palmer, 2009). The AMR annotation

also builds on the entity and relation annotation of the Automatic Content Extraction (ACE) (Doddington et al., 2004), as well as the annotation of discourse relations in the Penn Discourse TreeBank (Prasad et al., 2008) and the Chinese Discourse Treebank (Zhou and Xue, 2015).

The work presented here is also related to other flavors of whole-sentence meaning representation such as the Minimum Recursion Semantics (MRS) (Copestake et al., 2005) and the Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), both of which have been used in building annotated semantic resources. For example, MRS has been used in HPSG-based frameworks in generating semantically annotated resource such as the Lingo Redwoods Initiative (Oepen et al., 2004), while DRT has been adopted in building semantically annotated resources such as the Groningen Meaning Bank (Bos et al., 2017).

There are several efforts constructing the Chinese semantic dependency resources. Li et al. (2004) reported parsing experiments on a one million word Chinese corpus annotated with semantic dependencies, but their dependency structure is tree-based rather than graph-based. Chen and Ji (2011) described a three thousand sentence corpus annotated with semantic graphs. Corpora annotated with semantic graphs also include those reported in Ding et al. (2014) and Zheng et al. (2014). These semantic resources vary in the types of semantic relations they use, but they all differ from the work we report here in that they define semantic relations between word tokens instead of abstract concepts.

Our work is also related to efforts in building AMR resources for languages other than English. These include efforts in Spanish (Migueles-Abraira, 2017) and Czech (Xue et al., 2014), but these efforts are still preliminary and do not amount to an annotated corpus of significant size.

7 Conclusion and future work

In this article, we presented our effort in developing the Chinese AMR (CAMR) corpus, which consists of 10,149 sentences selected from the Chinese Treebank. Our general approach was to adopt the AMR strategy of annotating the meaning representation of each sentence independently of other layers of linguistic analysis for the sake of scalability, while developing detailed specifications as to how

to annotate each linguistic construction to ensure consistent annotation. On one hand, we have found that the AMR specifications, consisting of the graph structure, the abstract concepts and relations, readily applies to the CAMR annotation almost in its entirety. On the other hand, we also extended the AMR specifications by devising a consistent way to annotate discourse relations as well as tense and aspect. Another departure from the AMR approach is that we integrate word-to-AMR concept and relation alignment to the CAMR annotation process. The inter-annotation agreement shows that our approach is effective. A quantitative analysis of the CAMR corpus shows that 46.71% of the AMRs are non-tree graphs. In addition, the AMRs of 88.95% of the sentences have abstract concepts inferred from the context of the sentence but do not correspond to a particular word or phrase in a sentence, and the average number of such inferred concepts per sentence is 2.88. We believe this corpus will prove to be crucial resource in advancing the state of the art in Chinese semantic parsing and in Chinese AMR parsing in particular. In the future, we plan to annotate additional data of other genres as part of this on-going project. We will also develop automatic Chinese AMR parsers.

Acknowledgments

This work is the staged achievement of the projects supported by National Social Science Foundation of China (18BYY127) and National Science Foundation of China (61772278).

References

Bai, Xiaopeng and Nianwen Xue. 2016. Generalizing the semantic roles in the chinese proposition bank. *Language Resources and Evaluation* 50(3):643–666.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. Sofia, Bulgaria.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2015. *Abstract Meaning Representation (AMR) 1.2.2 Specification*. https://github. com/amrisi/amr-guidelines/blob/master/amr.md.

Böhmová, A., J. Hajič, E. Hajičová, and B Hladká. 2003. The prague dependency treebank. In A. Abeillé, ed., *Treebanks. Text, Speech and Language Technology*, vol. 20, pages 103–127. Dordrecht: Springer.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In N. Ide and J. Pustejovsky, eds., *Handbook of Linguistic Annotation*, vol. 2, pages 463–496. Springer.

Cai, Shu and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752. Sofia, Bulgaria: Association for Computational Linguistics.

Chen, Bo and Donghong Ji. 2011. Chinese semantic parsing based on dependency graph and feature structure. In *International Conference on Electronic and Mechanical Engineering and Information Technology*, vol. 4, pages 1731–1734.

Copestake, Ann, Flickinger Dan, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation* 3(2-3):281–332.

Ding, Yu, Yanqiu Shao, Wanxiang Che, and Ting Liu. 2014. Dependency graph based chinese semantic parsing. *Lecture Notes in Computer Science* 8801:58–69.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of LREC*, vol. 2, pages 837–840.

Flanigan, Jeffrey, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the* 52^{nd} *Annual Meeting of the Association for Computational Linguistics*, pages 1426–1436.

Kamp, Hans and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, vol. 42. Springer Science & Business Media.

Kuhlmann, Marco and Stephan Oepen. 2016. Squibs: Towards a catalogue of linguistic graph banks. *Computational Linguistics* 42(4):819–827.

Li, Bin, Yuan Wen, Q. U. Weiguang, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with chinese amrs. In *Linguistic Annotation Workshop Held in Conjunction with ACL*, pages 7–15.

Li, Mingqin, Juanzi Li, Zuoying Wang, and Lu Dajin. 2004. A statistical model for parsing semantic dependency relations in a chinese sentence. *Chinese Journal of Computers* 27(12):1679–1687.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In A. Meyers, ed., *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31. Boston, Massachusetts, USA.

Migueles-Abraira, Noelia. 2017. *A Study Towards Spanish Abstract Meaning Representation*. Master's thesis, University of the Basque Country, Basque Autonomous Community.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christopher D Manning. 2004. Lingo redwoods. *Research on Language and Computation* 2(4):575–596.

Oepen, Stephan, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–105.

Pourdamghani, Nima, Yang Gao, Ulf Hermjakob, and Kevin Knight. 2014. Aligning english strings with abstract meaning representation graphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 425–429. Doha, Qatar: Association for Computational Linguistics.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*, pages 2961–2968.

Pustejovsky, James, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, vol. 2003, pages 647–656. Lancaster, UK.

Saurí, Roser and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation* 43(3):227–268.

Xue, Nianwen, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of english amrs to chinese and czech. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Xue, Nianwen and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Natural Language Engineering* 15(1):143–172.

Xue, Nianwen, Fei Xia, Fudong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11(2):207–238.

Zheng, Lijuan, Yanqiu Shao, and Erhong Yang. 2014. Analysis of the non-projective phenomenon in chinese semantic dependency graph. *Journal of Chinese Information Processing* 28(6):41–47.

Zhou, Yuping and Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources & Evaluation* 49(2):1–35.