Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5938–5945
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Align-smatch: A Novel Evaluation Method for Chinese Abstract Meaning Representation Parsing based on Alignment of Concept and Relation

Liming Xiao[1], Bin Li[1], Zhixing Xu[1], Kairui Huo[1], Minxuan Feng[1], Junsheng Zhou[2] and Weiguang Qu[2]
[1]School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu, China
[2]School of Computer, Electronics and Information, Nanjing Normal University, Nanjing, Jiangsu, China
lmxiao1@foxmail.com, libin.njnu@gmail.com, xzx0828@live.com,
kairui.huo.nj@gmail.com, fennel_2006@163.com, {zhoujs, wgqu}@njnu.edu.cn

Abstract

Abstract Meaning Representation is a sentence-level meaning representation, which abstracts the meaning of sentences into a rooted acyclic directed graph. With the continuous expansion of Chinese AMR corpus, more and more scholars have developed parsing systems to automatically parse sentences into Chinese AMR. However, the current parsers can't deal with concept alignment and relation alignment, let alone the evaluation methods for AMR parsing. Therefore, to make up for the vacancy of Chinese AMR parsing evaluation methods, based on AMR evaluation metric smatch, we have improved the algorithm of generating triples so that to make it compatible with concept alignment and relation alignment. Finally, we obtain a new integrity metric Align-smatch for parsing evaluation. A comparative research then was conducted on 20 manually annotated AMR and gold AMR, with the result that Align-smatch works well in alignments and more robust in evaluating arcs. We also put forward some fine-grained metric for evaluating concept alignment, relation alignment and implicit concepts, in order to further measure parsers' performance in subtasks.

Keywords: Abstract Meaning Representation, Align-smatch, Concept Alignment, Relation Alignment

1. Introduction

With the growing maturity of morphological analysis and syntactic analysis techniques, natural language processing in general has advanced to the level of semantic analysis. As the crux part, sentence-level meaning has occupied the core position of semantic analysis research (Sun et al., 2014). To address the lack of whole-sentence semantic representation and the domain-dependent problem of sentence semantic annotation, Banarescu et al. proposed a domain-independent whole-sentence semantic representation method in 2013, which is Abstract Meaning Representation (AMR) that can abstract the meaning of a sentence with a single-rooted, acyclic and directed graph (Banarescu et al., 2013). AMR not only depicts the phenomenon of argument sharing formed by a noun governed by multiple predicates but also allows to supplement the latent semantic to represent the sentence meaning completely. The great capability of semantic representation renders AMR the widespread attention upon its introduction, and countless articles on various aspects such as AMR automatic parsing and AMR transformation applications have emerged as well.

Li et al. (2016) introduced AMR into Chinese and made corresponding adjustments according to the linguistic characteristics of Chinese such as adding labels representing semantic relations like "aspect" and "quantifier", specifying the treatment of unique structures like clutch words and most importantly, retaining function words as nodes or labeling them on directed arcs. And in particular, they also proposed the alignment of concept, relation and words in the original sentence when it comes to the representation method (Li et al., 2019). For example, Figure 1 compares the two versions of the sentence "他想去北京看演出。(*He wants to see the show in Beijing.*)" in which the Chinese AMR allows relation alignment and concept alignment while English cannot. Finally, a set of annotation specifications for Chinese AMR was designed, and a Chinese AMR corpus with the size of about 20,000 sentences was constructed (Li et al., 2016) (Li et al., 2017b) (Li et al., 2017a) (Wen et al., 2018) (Dai et al., 2020). With the expansion of the corpus size, more and more scholars have been involved in the automatic parsing of Chinese AMR (Wang et al., 2018) (Gu, 2018) (Wu et al., 2019) (Damonte and Cohen, 2017) (Blloshmi et al., 2020). The parsing system they have developed is able to predict and output the corresponding Chinese AMR structure of a given sentence. The parsing accuracy reached 0.81 of F1 in the parsing evaluation task of cross-semantic representation methods released by the International Conference on Natural Language Learning (CoNLL) in 2020, which is the best result so far (Oepen et al., 2020) (Samuel and Straka, 2020).

These results, however, still could not reflect the real

He wants to see the show in Beijing.	他[1]想[2]在[3]北京[4]看[5]演出[6]。[7]
Unaligned	**Aligned**
(x0/want-02 :arg0 (x1/he) :arg1 (x2/see-01 :arg0 x1 :arg1 (x3/show) :location (c/city :name (x4/name :op1 Beijing))))	(x2/想-02 :arg0() (x1/他) :arg1() (x5/看-01 :arg0() x1 :arg1() (x6/演出) :location(x3/在) (x10/city :name() (x4/name :op1 北京))))

Figure 1: The alignment of relation and concept in Chinese AMR

level of Chinese AMR parsing. Because all tests and evaluations including CoNLL are based on English AMR, which apparently are not suitable for Chinese AMR. The integrity metrics such as smatch (Cai and Knight, 2013) and SemBleu (Song and Gildea, 2019) are not compatible with the adjustments made by Chinese AMR, for example, the evaluation corpus all removed the alignment information, which means the alignment information of concept and relation in Chinese AMR, especially the function words labeled on directed arcs, were not parsed and evaluated. Besides, some AMR fine-grained metrics (Damonte et al., 2016) (Cai and Lam, 2019) do not involve alignment information as well. The evaluation tool MTool (Oepen et al., 2020) by far is the only one that can evaluate the concept alignment information of Chinese AMR and yet cannot measure the relation alignment information still.

In order to fill the gap in alignment information of Chinese AMR parsing evaluation and provide new standards and directions for the future development of Chinese AMR parsing work, we introduce the Alignsmatch metric based on the smatch metric. To take evaluation needs of specific tasks into account, we also propose a total of three fine-grained metrics including concept alignment metric, relation alignment metric and implicit concept metric to serve Chinese AMR parsing evaluation even better.

2. Alignment of Concept and Relation

AMR is to abstract the words in a sentence and the connections between words into "concepts" and "relations", which are reflected in the AMR graph as "nodes" and "edges" respectively. To be specific, words are abstracted as concept nodes and relations between words as directed arcs with semantic role labels. Thus, AMR can be formally defined as a triple (*T, N, A*), where $T \subseteq N$, representing the root node, is the center of the sentence and generally the main predicate *N*, represents the node, which contains information about concepts, attributes, etc., and is represented as a triple: (*instance* (*node name, concept*)). $A \subseteq N \times N$ denotes the directed arc consisting of the source node of the arc N_R, the target node of the arc N_T and the semantic role label *R*, hence representing as the triple (*R* (N_R, N_T)).

This abstract semantic representation way enables AMR to add, delete and modify concept nodes, and to complement the implicit concept annotation. For example, in Figure 2, AMR abstracts the implicit concept "*country*" as a node when annotating the named entity "中国*(China)*". On the other hand, this also means that AMR has difficulty in providing a mapping relationship between concepts and words for it has no concept alignment. Given that AMR was originally designed based on English and words in English have morphological changes while concepts do not, the initial letter of each word was normally used as the number of the concept node in AMR annotation processing. It leads to the inability of computers to directly trace concepts back to their source and to restore the order of sentences from AMR, and brings great difficulties to AMR parsing (Li et al., 2017b).

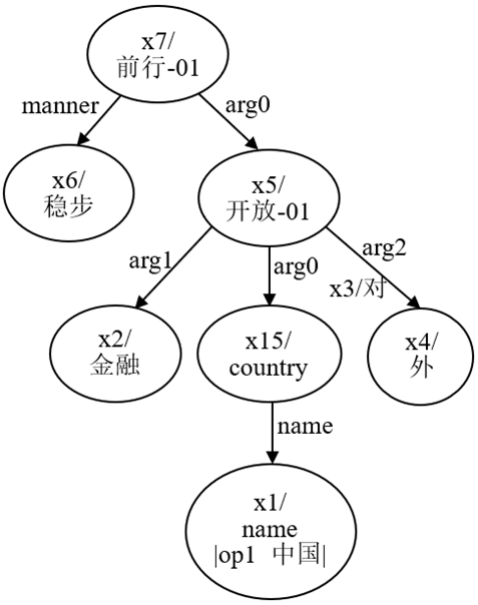

Figure 2: Chinese AMR of "China's finance keeps opening up to the foreign and moving forward"

To solve this problem, Li et al. (2016) proposed an efficient framework incorporating concept-to-word alignment to achieve concept alignment for Chinese AMR. By assigning a number to each word in the original sentence after word separation according to the principle of linear ordering, each concept node is also assigned a corresponding number. The numbering takes the form of "*x*" + a number, and if the number is not greater than the total number of words in the sentence, it represents the number of the word aligned with that node. As shown in Figure 3, the bold part is the aligned number, which ensures that most of the concept nodes are aligned with the words in the original sentence. Concepts that are not aligned are implicit concepts that are complemented, such as the concept "*country*" in the "*x15*" node, which is given a number greater than the total number of words in the sentence.

In addition to concept alignment, AMR also omits function words such as prepositions and articles that are less meaningful, assuming that they do not contribute very much to the semantic. However, function words are quite useful for connecting contexts. Function words in Chinese, as a matter of fact, contain rich semantic information. Hence, Chinese AMR chooses to retain function words for annotation. Function words indicating aspect meaning and mood meaning of the sentence are generally treated as concept nodes while function words indicating the relation between content words are regarded as mappings of semantic relations and labeled on the directed arcs together with semantic

中国	金融	对	外	开放	稳步	前行
x1	x2	x3	x4	x5	x6	x7

Concept-to-word Unaligned	Concept-to-word Aligned
(x0 / 前行-01 :manner() (x1 / 稳步) :arg0() (x2 / 开放-01 :arg1() (x3 / 金融) :arg2() (x4 / 外) :arg0() (c / country :name() (x5 / name :op1 中国))))	(**x7** / 前行-01 :manner() (**x6** / 稳步) :arg0() (**x5** / 开放-01 :arg1() (**x2** / 金融) :arg2() (**x4** / 外) :arg0() (x15 / country :name() (**x1** / name :op1 中国))))

Figure 3: Concept alignment annotation

Relation-to-word Unaligned	Relation-to-word Aligned
(x7 / 前行-01 :manner() (x6 / 稳步) :arg0() (x5 / 开放-01 :arg1() (x2 / 金融) :arg2() (x4 / 外) :arg0() (x15 / country :name() (x1 / name :op1 中国))))	(x7 / 前行-01 :manner() (x6 / 稳步) :arg0() (x5 / 开放-01 :arg1() (x2 / 金融) :**arg2(x3/对)** (x4 / 外) :arg0() (x15 / country :name() (x1 / name :op1 中国))))

Figure 4: Relation alignment annotation

role labels (Dai et al., 2020). Function words on the directed arcs are also numbered, which would achieve relation alignment by completing the alignment of semantic relations with the words in sentences. As shown in Figure 4, the bold part is that the function word "对(*to*)" representing the relation between the content word "开放(*open up*)" and "外(*foreign*)" in the original sentence is aligned with the relation *arg2* using number *x3* in AMR.

3. Align-smatch

To promote the further development of Chinese AMR parsing, we believe that we should first provide corresponding evaluation metrics to assess the quality of a Chinese AMR parsing system considering that no parser by far can handle the alignment of concept and relation in Chinese AMR. To this end, we propose the Align-smatch metrics based on smatch to evaluate the accuracy of Chinese AMR parsers in general. In terms of fine-grained metrics, we also propose the concept alignment metric, the relation alignment metric and the implicit concept metric to evaluate the performance of parsers on alignment of subtasks.

3.1. Related Work

The integrity metric usually returns a value between 0 and 1 to measure how well two AMR graphs match, among which the smatch metric is by far the most widely used. For two AMR graphs to be matched, smatch first renames the nodes of AMR graphs and

	Triples	Quantity
Nodes	instance (a0, 前行-01), instance (a1, 稳步-01), instance (a2, 开放-01), instance (a3, 金融), instance (a4, country), instance (a5, 外), instance (a6, name)	7
Directed arcs	manner (a0, a1), arg0 (a0, a2), arg0 (a2, a4), arg1 (a2, a3) arg2 (a2, a5), name (a4, a6)	6
Property	TOP (a0, "top"), op1 (a6, 中国)	2

Table 1: Smatch triples

transforms each AMR graph into a set of triples, then performs a greedy search using the Hill-climbing algorithm to obtain the maximum number of triples matching the two sets, and finally returns the *Precision*, *Recall* and *F1*.

Each triple set generated by smatch generally contains three triple categories: for a node N_1, there is the triple (*instance* (N_1, *C*)) representing the concept of the node, and there is the triple (*P* (N_1, V)) representing the property and value of the node. In particular, when *P* = *"TOP"*, node N_1 is the vertex. The directed arc between node N_1 and node N_2 is represented by the triple (*R* (N_1, N_2)), where *R* represents the semantic role, node N_1 is the source node and node N_2 is the target node.

Taking the Chinese AMR in Figure 2 as an example, Table 1 lists the set generated by smatch with a total of fifteen triples. The nodes are renamed by smatch and the concept alignment information is missing. The function word "对(*to*)", which is labeled on the directed arc and the relation alignment information is also not reflected. Apparently, smatch metric is not suitable for Chinese AMR.

Besides, there are two shortcomings of smatch itself also. First, when comparing triples of directed arcs, smatch only considers whether the semantic role labels are the same but does not examine whether the concepts of nodes are consistent. This can easily lead to an awkward situation where two AMRs with completely different semantics yet get high scores (Song and Gildea, 2019). Second, smatch adds *TOP* property triples for the root node of each AMR graph but does not consider whether the concepts of the two root nodes are the same when comparing them, which makes it possible for two AMRs with different root node concepts to have *TOP* property triples to match. Taking Table 1 and Figure 6 for example, these two sentences with completely different semantics can reach about 40% of *F1* in smatch, including *arg0* and *arg1* both matching

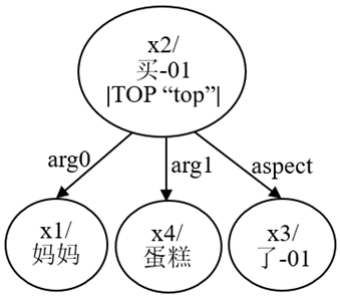

Figure 5: Chinese AMR of "Mum has bought cakes"

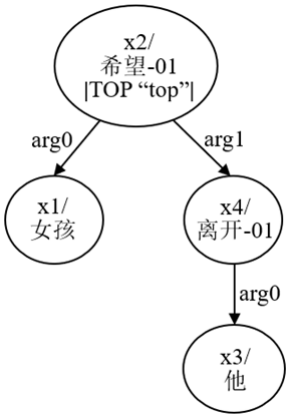

Figure 6: Chinese AMR of "The girl hopes he can leave"

with *TOP* property, which doesn't make sense.

3.2. Align-smatch Metrics

To address the above issues, we first made two fixes to smatch. One is that when matching directed arc triples, we stipulate that the source nodes and target nodes must be the same under the premise that the semantic roles are the same, otherwise they cannot be matched. Another one is to move the triple representing the root node from the property to the category of directed arcs, written as *TOP* (*a0*, *a0*), which is considered as a directed arc pointing to itself as shown in Figure 7. This makes it possible to examine whether the concepts of root nodes are consistent according to the first fix when matching root nodes, thus avoiding the above problem of different root nodes but matching with each other. The smatch values are reduced to 0.13 after the first fix and become 0 after the second fix, which is more inline with our intuition. In the following chapters, we refer to the modified version as "*FIX*".

We then incorporate a triple representing concept alignment and relation alignment and a quadruple representing relation alignment in smatch. If the sentence contains *L* words and the index $I \leq L$ of node N_1, there is a property triple anchor (N_1, *I*) with *P = "anchor"* and *V = I* that represents the concept alignment information of node N_1. For relation alignment, we already have a triple containing semantic role labels and only need to represent the words corresponding to semantic relations again. We consider the word W_A (Word in Arc) and its index value *I* on a directed arc as a whole, represented by a quadruple ((W_A, *I*), N_1, N_2), where N_1 are the source nodes and N_2 are the target nodes of the directed arc. Taking the directed arc from node *x5* to node *x4* in Figure 2 as an example, it carries the semantic role label *arg2* and a preposition "对(*to*)", so it is represented as *arg2* (*x5*, *x4*) and (对, *3*) (*x5*, *x4*). These two tuples indicate that both *arg2* and "对(*to*)" are on this directed arc, representing the relation alignment information successfully.

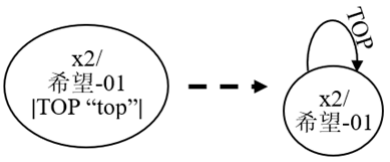

Figure 7: The directed arc of the root node, "希望(*hope*)"

	Triples	Quantity
Nodes	instance (a0, 前行-01), instance (a1, 稳步), instance (a2, 开放-01), instance (a3, 金融), instance (a4, country), instance (a5, 外), instance (a6, name)	7
Arcs	**top (a0, a0)**, manner (a0, a1), arg0 (a0, a2), arg0 (a2, a4), arg1 (a2, a3) arg2 (a2, a5), **(对, 3) (a2, a5)**, name (a4, a6)	8
Property	**anchor (a0, 7),** **anchor (a1, 6),** **anchor (a2, 5),** **anchor (a3, 2),** **anchor (a5, 4),** **anchor (a6, 1),** op1 (a6, 中国)	7

Table 2: Alignment-smatch tuples

Table 2 shows the tuples formed by Align-smatch. Compared to Table 1, Table 2 represents the root node in the directed arc category instead of in the node property category; adds triples representing concept alignment in the node property category; and adds quadruples representing relation alignment in the directed arc category. We now believe that Align-smatch tuple is a more complete representation of Chinese AMR.

To generate the triples representing the alignment information in Table 2, we improved the algorithm for generating triples by smatch so that it can recognize

Algorithm 1 Chinese AMR → tuples

Input: Chinese AMR $w = w_0 \ldots w_n$
Output: a set of tuples $T_w = \{t_1 \ldots t_m\}$

```
1: T_w ← {}
2: for i ← w_0 to w_n do
3:     if i = ) or ( or : or / then
4:         state ← 0 to 3
5:         t_m ← Action(state, s)
6:         T_w.Push(t_m)
7:     else
8:         s ← i
9: end for
10: return T_w
```

Figure 8: The pseudo-code of Algorithm 1

the alignment information. The new algorithm uses the Shift-Reduce method to process the character W_n of Chinese AMR sequentially, using a stack *Node_Stack (NS)* to store nodes, a stack *Arc_Stack (AS)* to store words on directed arcs, a stack *Relation_Stack (RS)* to store semantic roles, and a buffer (*Buffer*) to relay semantic roles. Finally, the corresponding state codes (*State*) are generated by four Chinese AMR symbols ")", "(", ":" and "/", so that the corresponding operations (*Action*) can be performed on the string s, the stack and the buffer, and the Chinese AMR inputted is transformed into a tuple set T_W. Figure 8 is the pseudo-code of the tuple set generation algorithm.

After transforming Chinese AMR corpus (Gold AMR) G_g and the parser's output (Parsed AMR) G_p into two triple sets T_g and T_p by Algorithm 1, we follow Hill-Climbing algorithm in smatch to obtain the best triple match numbers for the two sets.

The Hill-Climbing algorithm first initializes the matching of nodes in T_g and T_p triples to get the number of triples and the set of node mappings for the first match. For example, if T_g has a triple (*instance* (*a0*, 前行(*move forward*)-*01*)) and T_p has a triple (*instance* (*b0*, 前行(*move forward*)-*01*)), which has the same concept, they form the mapping pair (*a0, b0*). Then, the initial mapping set is searched for a better match by two operations "*Swap*" and "*Move*".

"*Swap*" occurs between two mapping pairs, e.g. (*a0, b0*), (*a2, b2*), and after swapping each other's mapping objects, two new pairs of (*a0, b2*), (*a2, b0*) are formed. "*Move*" occurs between a mapping pair and an unmatched node, e.g., if there is an (*a0, b0*) mapping pair and an unmatched node *b5*, then an attempt is made to construct the mapping pair (*a0, b5*).

Eventually, the optimal number of matches of two triple sets is searched by multiple initialization to avoid falling into local optimum. The default number of initialization (Cai and Knight, 2013) is 5.

Based on the matching results, the *Precision*, *Recall*, and *F1* will be returned to evaluate the accuracy of the parser. Thus, the alignment information of concept and relation is incorporated into smatch metric and can be evaluated for Chinese AMR, which we name Align-smatch.

3.3. Three Fine-grained Metrics

The integrity metrics alone are not sufficient to evaluate the performance of the parsing system for it is poorly readable, coarse-grained and cannot reflect the performance of the parsing system on single dimension sub-tasks such as concept recognition and relation recognition, nor can they reflect the current problems of the parsing system, which is not conducive to further improvement of the parsing system. Therefore, AMR parsing evaluation needs fine-grained metrics to evaluate the parsing system from a view of multi-dimension so that it can meet the requirement of specific tasks.

Damonte et al. proposed their fine-grained metrics from nine dimension (Damonte et al., 2016): *Unlabeled*, *No WSD*, *NP-only*, *Reentrancy*, *Concepts*, *Named Ent.*, *Wikification*, *Negations* and *Semantic Role Labeling*. Based on these and combined with the characteristics of Chinese AMR, we now propose alignment metrics of concept, relation and implicit concept.

Concept alignment metrics focus on measuring the matching degree of concept alignment information of two Chinese AMRs, G_g and G_p, and can examine the performance of the parser in the concept alignment subtasks. The node *N* extracted from Chinese AMR, the index value *I* of the concept mapping to the word, and the concept *C* altogether are obtained and converted into a triple (*C* (*N, I*)) to calculate the concept alignment metrics. When the concepts are the same, if the index values of the two triples are also the same, the number of correct triples is added one. In the end, P_{CA}, R_{CA} and *F1* are returned.

$$P_{CA} = \frac{num(True\_CA\_Triples(G_P))}{num(All\_CA\_Triples(G_P))}$$

$$R_{CA} = \frac{num(True\_CA\_Triples(G_P))}{num(All\_CA\_Triples(G_g))}$$

$$F1_{CA} = \frac{2 \times (P_{CA} \times R_{CA})}{(P_{CA} + R_{CA})}$$

Concept alignment metrics can also be used for English AMR parsing evaluation. In recent years, with the rise of deep learning, a great number of studies have made parsers automatically acquire alignment information by the aid of the encoder-decoder model based on attention mechanism, thus discarding external aligners that tend to bring error propagation (Cai and Lam, 2019) (Barzdins and Gosko, 2016) (Konstas et al., 2017) (Zhang et al., 2019). Although the parsing performance are improved without using aligners, the alignment accuracy of parsers remains unknown. Therefore, a concept alignment metric is in demand to evaluate the performance of the parser in alignment subtasks. Relation alignment metrics examine the parser's ability to correspond words to directed arcs. The word W_A on a directed arc, the index value *I* of the word, the source node N_1 and target node N_2 of the directed arc are

Metrics \ Groups	A-G	B-G
$F1_{CA}$	0.86	0.88
$F1_{RA}$	0.57	0.51
$F1_{I}$	0.75	0.85

Table 3: Results of fine-grained metrics

Metrics \ Groups	A-B	A-G	B-G
Smatch	0.74	0.78	0.84
Concept-smatch	0.77	0.81	0.87
Concept-smatch (**FIX**)	0.76	0.79	0.87
Align-smatch	0.73	0.78	0.83
Align-smatch (**FIX**)	0.71	0.76	0.83

Table 4: Results of integrity metrics

transformed into a quadruple ((W_A, *I*), N_1, N_2), then we calculate the matching of two relation alignment quadruple. When the word on arcs is the same with the index, if the nodes of target and source are the same, the number of the correct quadruple is added one. As always, P_{RA}, R_{RA} and *F1* are returned.

$$P_{RA} = \frac{num(True\_RA\_Quadruple(G_P))}{num(All\_RA\_Quadruple(G_P))}$$

$$R_{RA} = \frac{num(True\_RA\_Quadruple(G_P))}{num(All\_RA\_Quadruple(G_g))}$$

$$F1_{RA} = \frac{2 \times (P_{RA} \times R_{RA})}{(P_{RA} + R_{RA})}$$

In Chinese AMR, there are also nodes that do not have concept alignment. These nodes are implicit concepts that do not appear in the sentence, e.g., "上海(*Shanghai*)" implies the concept of "*city*". Implicit concepts are closely related to the semantic integrity of sentences and the named entity recognition, so it is inevitable to examine the parser's ability to generate implicit concepts. Therefore, we propose an implicit concept metric, which is obtained by counting the number of nodes without concept alignment in two AMR graphs and then returns P_I, R_I and *F1*.

$$P_I = \frac{num(True\_Unaligned\_Nodes(G_P))}{num(All\_Unaligned\_Nodes(G_P))}$$

$$R_I = \frac{num(True\_Unaligned\_Nodes(G_P))}{num(All\_Unaligned\_Nodes(G_g))}$$

$$F1_I = \frac{2 \times (P_I \times R_I)}{(P_I \times R_I)}$$

4. Comparison Test

We randomly selected 20 Chinese AMR sentences with relation alignment from the CAMR 1.0 corpus as the standard corpus (*G*). Since there is no Chinese AMR parser with alignment information, we selected two annotators to re-annotate these 20 sentences to obtain the control corpus *A* and *B*.

We combined the control corpus and the standard corpus to obtain three groups of experimental subjects: *A-B*, *A-G*, and *B-G*. We used smatch, concept-smatch with concept alignment, and Align-smatch with alignment of concept and relation to evaluate a total of three metrics, respectively, in which concept-smatch and Align-smatch were evaluated once more in modified version (*FIX*). The result is shown in Table 3.

Under the same index, we can see *B-G* >*A-G* >*A-B*, indicating that annotation *B* is more standard. Scores of *A-B* are all smaller than the consistency criterion of 0.83 required by AMR (Banarescu et al., 2013), indicating that annotation *A* and B are less consistent.

Under the same set of subjects, concept-smatch is greater than smatch before and after modification. This is mainly attributed to the Chinese AMR annotation platform's support for concept alignment annotation. The annotator only needs to enter the word number to achieve concept alignment, and there are also word highlighting warnings to prevent missing words (Li et al., 2017b), which improves the annotation accuracy of concept alignment information and thus the evaluation score.

Align-smatch is generally lower than smatch and concept-smatch before and after correction, and the addition of relation alignment pulls down the evaluation score. Each Chinese AMR adds about two relation alignment quadruples on average, but the score decreases by 3%-4% compared to concept-smatch, which indicates that annotation *A* and *B* have errors in labeling words on directed arcs, mainly including:(1) The function word framework. For example, the word "所发出的信息(*the message which is sent*)" on the directed arc is a framework "所...的(*that thing*)", but the word "所(*that*)" is omitted.(2) The source or target node of the directed arc. For example, the conjunction "虽(*although*)" in "只可惜这些人虽有一颗心(*Pity that although these people have a heart*)" should be marked between the relation node "*contrast*" and the concept node "有(*have*)", but it is incorrectly marked on the directed arc between "可惜(*pity*)" and "有(*have*)".(3) Adverbs. Chinese AMR distinguishes most adverbs from function words by labeling them as concept nodes. For some adverbs with weak semantics, the annotator may mark them incorrectly. For example, "一下(*roughtly*)" in "昨晚计划了一下(*I roughly made a plan last night*)" is an adverb of frequency, which is used after the verb and yet actually less meaningful. Chinese AMR treats it as a node concept after relation "*frequency*", while annotation *B* treats it as a word aligned with relation "*frequency*" thus it cannot be matched.

The annotation *B* also has omission problem in relation alignment information. The average number of arcs la-

beled with relation alignment information is only 1.6, which is 0.6 less than the standard corpus. Thus, the Align-smatch score of *B*-*G* decreases more than that of *A*-*G* compared to concept-smatch.

The scores of the modified version are not higher than those of the regular version, for it is stricter but more reasonable for the matching of directed arcs.

It's noteworthy that the evaluation scores of *B*-*G* do not fluctuate before and after the modification, while *A*-*G* is adjusted downward by about 2%. This may be because the directed arcs in *B* fit the standard corpus more closely while the directed arcs in *A* may have errors in the source or target nodes.

Table 4 shows the performance of *A*-*G* and *B*-*G* under these three fine-grained metrics. Both have high scores on the concept alignment metric, which is consistent with the trend reflected in the integrity metric. Scores also show that even with the help of the annotation platform, errors still occur while annotating concept alignment information. In particular, it is easy to mislabel homonyms and words with long-distance dependencies in the sentence.

Scores of *A* and *B* on the relation alignment metric are much lower than those of the concept alignment metric, denoting that relation alignment may be more difficult to label. Scores of *A* on the relational alignment metric are about 6 points higher than those of *B*, suggesting that *A* is better at labeling relation alignment information, which echoes the results of the control tests on integrity metrics.

B scored about 10 points higher than *A* on the implicit concept metric. Apparently, *B* is better at filling out the implicit concepts of the sentences. *A*, on the other hand, suffered from mislabeling. Among the specific error categories, *A* has about 70% of the complex sentence relation concepts mislabeled, and *B* has about 67% of the named entity concepts mislabeled.

We believe that the replenishment of fine-grained metrics will allow us to evaluate every Chinese AMR parser in a more comprehensive and diverse way in the future, helping to reveal the pros and cons of each parser and thus promoting the development of Chinese AMR parsing.

5. Conclusion

There is no AMR parsing evaluation metric for Chinese AMR yet that can be compatible with the alignment information of concepts and relations, which hinders the further development of Chinese AMR for the evaluation results cannot truly reflect the level of Chinese AMR parsing. Consequently, this paper proposes an integrity metric for Chinese AMR parsing, Align-smatch, based on smatch metric and merged with triples describing alignment information of concept and relation, and performs two control tests between the manually annotated corpus and the standard corpus. The results demonstrate that the consistency of the two manually annotated corpora is lower, the corpus *B* is more standard, and the modified evaluation metrics are more reasonable.

This paper also proposes a total of three fine-grained metrics including concept alignment metrics, relation alignment metrics and implicit concept metrics to present the real performance of Chinese AMR parser in the subtasks of concept alignment, relation alignment, and implicit concept generation. The three fine-grained metrics reflect that *A* is better at labeling relation alignment information and *B* is better at labeling implicit concepts.

Our next step is to apply these four evaluation methods for Chinese AMR parser currently under development that contains alignment information to fully evaluate the performance of automatic analysis of Chinese AMR. Research on Universal Abstract Meaning Representation (UMR) (Žabokrtskỳ et al., 2020) is also in full swing. We would like to make Align-smatch compatible with AMR of other languages to promote the construction of cross-language AMR parsing and evaluation tools.

6. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Barzdins, G. and Gosko, D. (2016). Riga at semeval-2016 task 8: Impact of smatch extensions and character-level neural translation on amr parsing accuracy. *arXiv preprint arXiv:1604.01278*.

Blloshmi, R., Tripodi, R., and Navigli, R. (2020). Enabling cross-lingual amr parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500.

Cai, S. and Knight, K. (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.

Cai, D. and Lam, W. (2019). Core semantic first: A top-down approach for amr parsing. *arXiv preprint arXiv:1909.04303*.

Dai, Y., Dai, R., Feng, M., Li, B., and Qu, W. (2020). Representation and analysis of abstract meaning representation of chinese function words based on relation alignment. *Journal of Chinese Information Processing*, 34(4):21–29.

Damonte, M. and Cohen, S. B. (2017). Cross-lingual abstract meaning representation parsing. *arXiv preprint arXiv:1704.04539*.

Damonte, M., Cohen, S. B., and Satta, G. (2016). An incremental parser for abstract meaning representation. *arXiv preprint arXiv:1608.06111*.

Gu, M. (2018). Research on chinese amr parsing using transition-based neural network. Master's thesis, Nanjing Normal University.

Konstas, I., Iyer, S., Yatskar, M., Choi, Y., and Zettlemoyer, L. (2017). Neural amr: Sequence-to-sequence models for parsing and generation. *arXiv preprint arXiv:1704.08381*.

Li, B., Wen, Y., Qu, W., Bu, L., and Xue, N. (2016). Annotating the little prince with chinese amrs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15.

Li, B., Wen, Y., Bu, L., Qu, W., and Xue, N. (2017a). A comparative analysis of the amr graphs between english and chinese corpus of the little prince. *Journal of Chinese Information Processing*, 31(1):50–57.

Li, B., Wen, Y., Song, L., Bu, L., Qu, W., and Xue, N. (2017b). Construction of chinese abstract meaning representation corpus with concept-to-word alignment. *Journal of Chinese Information Processing*, 31(6):93–102.

Li, B., Wen, Y., Song, L., Qu, W., and Xue, N. (2019). Building a Chinese AMR bank with concept and relation alignments. In *Linguistic Issues in Language Technology, Volume 18, 2019 - Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing*. CSLI Publications, July.

Oepen, S., Abend, O., Abzianidze, L., Bos, J., Hajic, J., Hershcovich, D., Li, B., O'Gorman, T., Xue, N., and Zeman, D. (2020). Mrp 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22.

Samuel, D. and Straka, M. (2020). Ufal at mrp 2020: Permutation-invariant semantic parsing in perin. *arXiv preprint arXiv:2011.00758*.

Song, L. and Gildea, D. (2019). Sembleu: A robust metric for amr parsing evaluation. *arXiv preprint arXiv:1905.10726*.

Sun, M., Liu, T., et al. (2014). Frontiers of language computing. *Journal of Chinese Information Processing*, 28(1):1–8.

Wang, C., Li, B., and Xue, N. (2018). Transition-based chinese amr parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 247–252.

Wen, Y., Song, L., Wu, T., Li, B., Zhou, J., and Qu, W. (2018). Research on non-projective structure based on the chinese abstract meaning representation corpus. *Journal of Chinese Information Processing*, 32(12):31–40.

Wu, T., Gu, M., Zhou, J., Qu, W., Li, B., and Gu, Y. (2019). Chinese amr parsing using transition-based neural network. *Journal of Chinese Information Processing*, 33(4):1–11.

Žabokrtskỳ, Z., Zeman, D., and Ševčíková, M. (2020). Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3):605–665.

Zhang, S., Ma, X., Duh, K., and Van Durme, B. (2019). Amr parsing as sequence-to-graph transduction. *arXiv preprint arXiv:1905.08704*.