Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 160–169
Marseille, 24 June 2022

Automatic Enrichment of Abstract Meaning Representations

Yuxin Ji†, Gregor Williamson‡, Jinho D. Choi‡
Emory University
Atlanta, GA 30322, USA
† Department of Quantitative Theory and Methods
‡ Department of Computer Science
{jessica.ji, gregor.jude.williamson, jinho.choi}@emory.edu

Abstract

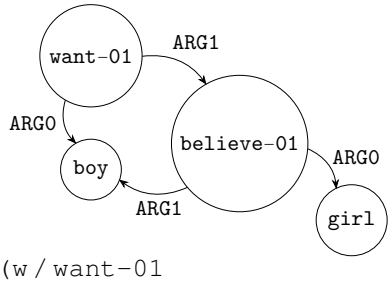
Abstract Meaning Representation (AMR) is a semantic graph framework which inadequately represent a number of important semantic features including number, (in)definiteness, quantifiers, and intensional contexts. Several proposals have been made to improve the representational adequacy of AMR by enriching its graph structure. However, these modifications are rarely added to existing AMR corpora due to the labor costs associated with manual annotation. In this paper, we develop an automated annotation tool which algorithmically enriches AMR graphs to better represent number, (in)definite articles, quantificational determiners, and intensional arguments. We compare our automatically produced annotations to gold-standard manual annotations and show that our automatic annotator achieves impressive results. All code for this paper, including our automatic annotation tool, is publicly available at https://github.com/emorynlp/EnrichedAMR/

Keywords: Abstract Meaning Representation (AMR), automatic annotation, automatic data enrichment

1. Introduction

Abstract Meaning Representation (AMR) is a semantic graph framework that represents natural language sentences in directed, acyclic graphs (Banarescu et al., 2013). Nodes represent concepts, and labeled edges represent relations between concepts (1-b). AMRs are most commonly written in PENMAN format (Matthiessen and Bateman, 1991), as shown in (1-c).

(1) a. *The boy wants the girl to believe him.*

b. 

c.
```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (b2 / believe-01
      :ARG0 (g / girl)
      :ARG1 b))
```

The primary function of AMR is to capture argument structure. Features of the graph need not be anchored to grammatical features of the natural language sentence. This has the advantage of allowing succinct representation of non-compositional aspects of meaning. A major disadvantage, however, is that it can give rise to inter-annotator disagreement (Bender et al., 2015), as well as making the task of parsing harder (Buys and Blunsom, 2017; Lin and Xue, 2019; Oepen et al., 2019; Oepen et al., 2020). Moreover, evidence show that more explicit grammatical information might improve AMR parsing performance. For example, bridging the gap between natural language and AMR, via preprocessing with an Elementary Dependency Structures (EDS) (Oepen and Lønning, 2006) parser, has been shown to improve AMR parsing results (Shou and Lin, 2021).

In addition to being abstract, AMR is under-specified with respect to a number of important semantic features. A consequence of this design choice is that AMR introduces ambiguity which is absent from the source sentence. For instance, the graph depicted in (1-b)/(1-c) is also the representation for (i) *'a boy wanted girls to have believed him'*, (ii) *'the boys will want a girl to believe them'*, etc. This radical under-specification can be problematic for NLU tasks beyond identifying argument structure.

In this paper, we report results from our Automatic (enriched) AMR Annotator, A^3. In section 2, we provide a background on existing approaches to improving the expressive capacity and representational adequacy of AMR. In section 3, we outline the proposed enrichments to be made by A^3. In section 4, we describe how the automatic annotator enriches existing graph structures, starting with the base cases before discussing more challenging constructions which arise as a result of AMR's abstraction from grammatical form. In section 5 we report two annotation experiments. In the first experiment, we calculate Inter-Annotator Agreement (IAA) scores for gold-standard manual annotations, demonstrating the reliability of the enrichment scheme. In the second, we compare the output of A^3 to manually produced annotations. Section 6 provides a comprehensive analysis of error types produced by A^3. Finally, in section 7, we discuss implications of the present approach on data production, before concluding in section 8.

2. Related Work

There has been a concerted effort towards improving the representational adequacy of AMR, as well as its recent

Translation		Richer Graph Structure	
Artzi et al. (2015)	(*coreference*)	Bonial et al. (2018)	(*comparatives*)
Bos (2016)	(*quantifier scope*)	Donatelli et al. (2018)	(*tense and aspect*)
Stabler (2017)	(*number, determiners*)	Donatelli et al. (2019)	(*tense and aspect*)
Lai et al. (2020)	(*quantifier scope*)	Pustejovsky et al. (2019)	(*quantifier scope*)
Williamson et al. (2021)	(*Intensionality*)	Bonial et al. (2020)	(*speech acts*)
		Bos (2020)	(*quantifier scope*)
		Van Gysel et al. (2021)	(*quantifier scope*)

Table 1: Approaches to improving the representational adequacy of AMR

offspring, Uniform Meaning Representation (UMR) (Van Gysel et al., 2021). This strand of research endeavors to improve the expressive power of AMR either in terms of enriching its graphical structure (Bonial et al., 2018; Donatelli et al., 2018; Donatelli et al., 2019; Pustejovsky et al., 2019; Bonial et al., 2020; Bos, 2020; Van Gysel et al., 2021) or by adding information during a subsequent translation step into a logical form (LF) in first-order logic or lambda-calculus (Artzi et al., 2015; Bos, 2016; Stabler, 2017; Lai et al., 2020; Williamson et al., 2021). Table 1 lists the phenomena addressed in these representative works.

Both of these approaches have their own merits. On the one hand, developing a richer graph structure allows us to directly represent meaning in the AMRs. However, revision of existing resources, such as the AMR 3.0 corpus (Knight et al., 2020), is costly and time-consuming. Moreover, unless the resulting graph structure can be mapped to a coherent model theoretical semantics, the enriched graph will not be any more representationally adequate than the original structure. On the other hand, making use of a translation function with minimal revision to the graphical structure allows us to work with existing corpora after translation into symbolic logical. However, we would ultimately like to work with AMR graphs directly, avoiding the need for translation into a logical language such as lambda calculus which can often be cumbersome for the purposes of computation. For these reasons, we take enriching the graph structure to be the ultimate goal, with the caveat that the graphs should have a model-theoretic semantic interpretation with as few ad-hoc interpretation rules as possible.

Despite various theoretical works on enriching AMR's graphical structure, there are no large-scale annotated corpora which implement these design features. The gold standard AMR 3.0 corpus (Knight et al., 2020) remains the major resource for parser training and evaluation. Considering the size of the AMR 3.0 corpus and the extensive cost for manual annotation, there is a clear need for efficient automatic annotation methods to augment the pre-existing data. The challenge, therefore, is to design graph structure which is not only suitably expressive but also tractable for automatic annotation.

While some previous work has focused on classifying AMR labels for natural language sentences (Chen et al., 2021), there has been no attempt to systematically add these labels to the graph structure. Enriching AMR graphs requires additional steps in mapping the semantic features from sentence tokens to the abstract (or unanchored) graphs. The methodology of this paper is inspired by Chen et al. (2021), who introduce a rule-based classifier for labeling aspect based on the UMR guidelines. The classifier uses part-of-speech (POS) tagging and lexical frames such VerbNet (Kipper et al., 2002; Kipper, 2005). It takes a sentence and returns a list of events labeled with aspectual information. Like Chen et al. (2021), we develop a rule-based classifier. However, our classifier performs the additional step of fitting the labels onto the corresponding AMR graph.

In this paper, we focus on the representation of grammatical number (singular/plural), (in)definite articles, quantifiers, and intensional arguments, all of which can provide important quantificational and referential cues for semantic scope, coreference resolution, and natural language inference tasks.

3. Enriched Graph Structure

In this section, we outline the enriched graph structure adopted in the present study. Here, we describe simple cases for each feature, reserving discussion of exceptional cases for section 4.5.

3.1. Representation of Number

In many cases, number marking adds important information because it is the only indicator of quantity. Even for noun phrases with a quantificational determiner, plurality is often informative. For example, the two cases in (2) can be differentiated only if plurality is marked.

(2) a. *Some boys painted the wall.*
b. *Some boy painted the wall.*

As such, plurality should ideally be represented in AMR to avoid the introduction of unwanted ambiguity. Stabler (2017) represents both plural and singular nouns by appending a marker to the corresponding concept matching the noun's grammatical number, as in (3).

(3) a. *The boy wants to go to the museums.*
b.
```
(w / want-01
    :ARG0 (b / boy.sg)
    :ARG1 (g / go-01
        :ARG0 b
        :ARG1 (m / museum.pl)))
```

However, this exact implementation is potentially problematic for a few reasons. Firstly, it is redundant to annotate both singular and plural explicitly. Instead, we can leave singular as the unmarked form, marking only plurals. Secondly, it is not uncommon for plural nouns to be represented by a predicate sense (e.g., *'the attempts'* ⇒ `attempt-01.pl`). However, most evaluation and processing scripts will be unable to process this notation since they rely on regex patterns to detect predicate senses.

We propose instead to add number as an additional attribute introduced by a `:number` role. We also do not abbreviate the marking, to better exploit the familiarity of AMR parsers built on pre-trained language models with natural language descriptions such as `plural` as opposed to the abbreviated `.sg` and `.pl`.

(4) **Enriched AMR: Number**

a. *The boy wants to go to the museums.*

b.
```
(w / want-01
    :ARG0 (b / boy)
    :ARG1 (g / go-01
        :ARG0 b
        :ARG1 (m / museum
            :number plural)))
```

This representation is also able to represent dual number marking, present in languages such as Slovene and Hebrew, with an additional attribute `dual`.

3.2. Representation of Definiteness

Definite and indefinite articles convey information which is useful for coreference resolution. While indefinite articles occasionally express quantity information (e.g., '*They could buy everyone a house*'), definite and indefinite articles are typically referential. To avoid confounding the role of articles and quantificational determiners, we introduce a new `:definite` role with the attribute + for definite and − for indefinite articles, as in (5).

(5) **Enriched AMR: Articles**

a. *The boy gave a girl some cookies.*

b.
```
(g / give-01
    :ARG0 (b / boy
        :definite +)
    :ARG1 (c / cookie
        :quant (s / some
        :number plural))
    :ARG2 (g / girl
        :definite -))
```

3.3. Representation of Quantifiers

The majority of work on quantifiers in AMR treats them as constants as opposed to concepts (Bos, 2016; Stabler, 2017; Lai et al., 2020; Williamson et al., 2021). As such, we aim to replace quantificational arguments of a `:quant` role with a quantificational constant. It is also common in existing corpora to see quantifiers annotated using the `:mod` role, in which case we replace it with `:quant` to maintain consistency, as in (6).

(6) **Enriched AMR: Quantifiers I**

a. *Every dog*

b.
```
(d / dog
    :mod (e / every))
```

c.
```
(d / dog
    :quant every)
```

Unlike Bos (2016) and (Lai et al., 2020), we do not conflate universal quantifiers such as *every*, *all*, and *each*, as these may vary in distributivity. Information which could be useful for downstream NLI tasks.

Next, AMR represents generalized quantifiers such as *someone*, *somebody*, *something*, *everyone*, *everybody*, *everything*, *no one*, *nobody*, and *nothing* as atomic concepts (7-b). However, this representation obscures the quantificational force of these noun phrases, so we decompose them as in (7-c).

(7) **Enriched AMR: Quantifiers II**

a. *Everyone*

b.
```
(e / everyone)
```

c.
```
(p / person
    :quant every)
```

We do not take a stance on whether or how to represent quantifier scope in the AMR graph structure. Unlike with the previous semantic features, if AMRs are left underspecified for scope, no information is lost since the corresponding natural language sentence is also scopally ambiguous. Provided there is some independent mechanism of scope taking, AMR can remain underspecified for scope as in Minimal Recursion Semantics Copestake et al. (2005), Hole Semantics Blackburn and Bos (2005), or Glue Semantics Asudeh and Crouch (2002), without loss of information. The scope of quantifier phrases could either be represented in an additional scope node layer (Pustejovsky et al., 2019; Van Gysel et al., 2021) or could be generated deterministically and filtered (Stabler, 2017). This could be done either manually or by training a parser on a large scope-disambiguated corpus. Unfortunately, the several existing scope-disambiguated corpora are either too small in size for robust machine learning and are not representative of complex scope interactions (Higgins and Sadock, 2003; Andrew and MacCartney, 2004; Srinivasan and Yates, 2009; Manshadi et al., 2011), or are not yet publicly available (Bunt, 2020). In anticipation of developments on this front, our changes to the representation of quantifier phrases remains flexible.

3.4. Representation of Intensionality

Finally, Crouch and Kalouli (2018) note that AMR is unable to represent non-veridical environments. For example, the following AMR will give rise to the inferences that there is a girl, and that she is sick.

(8) a. *The boy believes a girl is sick.*

b. `(b / believe-01`
`:ARG0 (b2 / boy)`
`:ARG1 (s / sick-05`
`:ARG1 (g / girl)))`

However, these inferences are not valid given the intensional nature of the attitude verb *'believe'*. To remedy this, Williamson et al. (2021) propose the addition of a `:content` role which is interpreted as an intensional operator responsible for representing the scope of modal predicates such as attitude verbs.

(9) $[\![\texttt{(x / P :content A)}]\!] =$
$\lambda w.\, \exists x.\, P(x) \wedge \text{content}(x)(\lambda w'.\, [\![\texttt{A}]\!](w')))$

We adopt Williamson et al. (2021)'s proposal to replace numbered arguments with the `:content` role where appropriate.

(10) **Enriched AMR: Intensionality**
a. *The boy believes a girl is sick.*
b. `(b / believe-01`
`:ARG0 (b2 / boy)`
`:content (s / sick-05`
`:ARG1 (g / girl)))`

Following the scheme just described, the sentence in (11-a) is represented as in (11-b).

(11) **Enriched AMR**
a. *A boy believes that the girls gave everyone some cookies.*
b. `(b / believe-01`
`:ARG0 (b2 / boy`
`:definite -)`
`:content (g / give-01`
`:ARG1 (g2 / girl`
`:definite +)`
`:ARG1 (c / cookie`
`:quant some`
`:number plural)`
`:ARG2 (p / person`
`:quant every)))`

4. The Automatic Annotator

Our automatic annotator, A^3 uses a combination of cues from the natural language sentence as well as its AMR in order to classify and map the target labels to the graph using the PENMAN parser (Goodman, 2020).[1] In sections 4.1-4.4, we describe the simpler cases of classification and mapping, before describing some of the numerous challenges in section 4.5.

4.1. Annotating Number

A^3 searches for tokens identified by the Stanford CoreNLP parser[2] (Manning et al., 2014) as having the plural *noun* part-of-speech (POS) tag. The plural noun is then mapped to the corresponding alignment in the AMR graph and the `plural` number attribute is appended to the triple. However, several abstract structures of AMR require special treatment. These are discussed in section 4.5.

4.2. Annotating Definiteness

Articles are identified through using a POS tag match. A string match for definite (*'the'*) and indefinite (*'a/an'*) articles is then used for tokens that are classified with a *DET* tag. A^3 then locates the span of head noun using the Stanford CoreNLP constituency parser (Manning et al., 2014) which was chosen due to it's performance, after experimenting with different constituency parsers including ELIT (He et al., 2021) and the Berkely Neural Parser (Kitaev and Klein, 2018). Finally, an appropriate `:definite` attribute is attached to the concept corresponding to the span of the head noun.

4.3. Annotating Quantifiers

The conversion for quantifiers utilize cues from the AMR graph alone and contains two steps. First, we identify quantifier concepts which are arguments of either a `:quant` or `:mod` role, before converting the quantificational concept to a constant. The second step decomposes generalized quantifiers by separating the concept and quantifier through a string match. The instance assignment for the original generalized quantifier is modified to the corresponding concept and the quantifier is attached to it as the attribute of the `:quant` role.

4.4. Annotating Intensionality

A^3 identifies intensionality through relevant lists of verbs and constituency structures. In most cases, appropriate uses of the `:content` role are identified using the MegaVeridicality dataset (White et al., 2018). Finite clauses are identified using MegaVeridicality version 1 (White and Rawlins, 2018), and non-finite clauses using version 2 (White et al., 2018). A^3 loops through the lemmatized tokens and searches for lemmas that are in the MegaVeridicality dataset. We compared the NLTK (Bird and Loper, 2004) and LemmInflect[3] lemmatizer and found that LemmInflect performs better. An intensional context is identified by checking if the matched verb is followed by a sentential complement, signified by a corresponding verb phrase constituent containing an SBAR or S label.

For speech verbs such as *'say'* or *'report'*, the sentence structure is not correctly identified by the parser when the complement clause has been fronted (e.g. '*The stock price doubled yesterday, as reported by the newspaper*'), which is not uncommon in the dataset, especially since AMR is sourced from news and broadcast data. To deal with these cases, A^3 instead looks for sentences where the verb is not followed by a noun phrase and annotates the object argument with a `:content` role.

[1] `https://github.com/goodmami/penman`
[2] `https://github.com/stanfordnlp/CoreNLP`
[3] `https://github.com/bjascob/LemmInflect`

4.5. Mapping Difficulties

Here, we list some non-canonical cases of each phenomena which are handled by A^3, but which require additional mapping instructions. We reserve discussion of cases which are not presently handled by our annotator to section 6.

4.5.1. Relational and Agentive/Patient Nouns

When enriching AMR with grammatical number and (in)definiteness, there are numerous non-trivial mapping problems posed by AMR's abstraction away from surface form. Most notably, AMR opts to express concepts using disambiguated predicate senses from PropBank (Kingsbury and Palmer, 2002) wherever possible. For instance, AMR uses a `person` concept to represent agentive nouns (12) and patient nouns (13).

(12) a. *Teacher*
b.
```
(p / person
   :ARG0-of (t / teach-01))
```

(13) a. *Employee*
b.
```
(p / person
   :ARG1-of (e / employ-01))
```

Other deverbal nouns may be represented through the use of an implicit `thing` argument.

(14) a. *An apology*
b.
```
(t / thing
   :ARG3-of (a / apologize-01))
```

Finally, AMR represents relational nouns using specialized concepts such as `have-rel-role-91` or `have-org-role-91`.

(15) a. *My uncles*
b.
```
(p / person
   :ARG0-of (h / have-rel-role-91
      :ARG1 (u / uncle)
      :ARG2 (i / i)))
```

These design choices create obvious problems for a naive mapping from grammatical features onto graph structure. In each case, we want to mark the root node of each of these (sub)-trees with a `plural` attribute, `:definite +/-` attribute, or `:quant` constant. However, the concept which most transparently corresponds to the surface string is not the root, for example *uncle* in (15-b). To solve this, A^3 tracks back through the directed edges of the sub-graph to find the root node, before marking it with the relevant attribute.

4.5.2. Name, Date, and Quantity Entities

We also observe exceptions for plural and definite markings for `name` and `date-entity` concepts, as well as `X-quantity` concepts. The `X-quantity` concept is typically introduced as a `:unit` and explicit quantity information is provided in the form of a real number. Similarly, for the case of `name` and `date-entity` concepts, the addition of a `:definite` or `:number` attribute is redundant.

(16) a. *Five dollars*
b.
```
(m / monetary-quantity
   :quant 5
   :unit (d / dollar))
```

4.5.3. Intensional Transitive Verbs

In addition to attitude predicates present in the MegaVeridicality dataset, A^3 is designed to map the numbered arguments of several Intensional Transitive Verbs (ITVs) to a `:content` role. ITVs are verbs that combine with a nominal direct object, but which do not permit an inference to the existence of the direct object in the world of evaluation (Schwarz, 2020). This can be seen in the following examples, which are semantically coherent despite the non-existence of unicorns in the actual world.

(17) *I* {*wanted/expected/desired/looked for*} *a unicorn*.

Since object arguments of ITVs are intensional regardless of whether their complement is a noun phrase or a sentential complement, A^3 converts the object argument of these predicates to a `:content` role. This mapping is defined for a non-exhaustive dictionary of the most common intensional transitive verbs (e.g. '*want*') and their intensional numbered argument as defined in their PropBank argument structure (Palmer et al., 2004).

4.5.4. Other Intensional Operators

Besides attitude predicates and ITVs, A^3 is designed to handle modal auxiliaries, modal verbs, and intensional raising predicates. Consequently, A^3 uniformly converts specific numbered arguments of modal predicate senses onto a `:content` role. These are summarized in Table 2.

Lexical item	Predicate Sense	Argument
need	need-01	:ARG1
can, might, could	possible-01	:ARG1
must (deontic)	obligate-01	:ARG2
must (epistemic)	infer-01	:ARG1
can	capable-01	:ARG2
seem	seem-01	:ARG1
allow	allow-01	:ARG1
permit	permit-01	:ARG1
should	recommend-01	:ARG1[4]

Table 2: Numbered arguments of modal concepts which are converted to `:content`.

5. Annotation Experiments

In this section, we report the methodology and results of two annotation experiments. In the first experiment, we measure Inter-Annotator Agreement (IAA) on the

[4] A^3 converts `:ARG1` of `recommend-01` to `:content` specifically when aligned with *'should'*, as this role may also used for non-intensional arguments of *'recommend'* e.g., *'I recommend this drink'*.

enrichment guidelines by doubly annotating 66 PENMAN graphs selected from the AMR 3.0 corpus. In the second experiment, we singly annotate an additional 60 graphs and compare the 126 manually annotated graphs to the output of our automatic annotation tool.

5.1. Method

To build our dataset, we first select up to 8 PENMAN graphs from each of the 12 datasets making up the (unsplit) AMR 3.0 corpus (excluding the guidelines). To ensure that the graphs contain relevant features, we restrict our dataset to graphs associated with a sentence of good-length (between 30 and 40 tokens), totalling 96 AMR graphs. We then select 30 additional graphs, from the same corpus (including the guidelines), which contain the relevant quantificational determiners or generalized quantifiers.

For the first experiment, we manually enrich 56 graphs from the good-length dataset and 10 graphs from the quantifier dataset for grammatical number, (in)definite articles, quantifiers, and the `:content` role. We compare IAA between the gold standard annotation by calculating F1 scores for the features of interest.

For the second experiment, we singly annotate the remaining 60 graphs and adjudicate among the doubly annotated graphs, creating a dataset of 126 gold-standard human annotations. We then processes the same 126 graphs using A^3 and we compare the output with our gold-standard annotations.

All annotations were carried out by the first and second authors using StreamSide[5] an open-source annotation tool for producing graph-based meaning representations (Choi and Williamson, 2021).

5.2. Manual Annotation Results

In the first experiment, two experienced annotators doubly annotate 56 graphs from our good-length dataset and a further 10 graphs from our quantifier dataset. The standard agreement metric for AMR graphs is the Smatch score of Cai and Knight (2013). However, this metric compares similarity between entire graphs. Calculating this score on our enriched graphs will give inflated scores due to the underlying similarity of the graphs used as the foundation for our annotations. Consequently, we present specific F1 scores calculated for each of the relevant features covered by the guidelines. Table 3 presents the F1 scores and the statistics for the 66 double annotations. This dataset contains around 1.7 grammatical number and article each per graph and one quantifier and intensional role per 2-3 graphs. The F1 scores range from 90.05 for (in)definite articles to 97.35 for the marking of plurals, demonstrating the robustness of our annotation guidelines for human annotation. The IAA for intensionality is surprisingly high (91.43) given the increased difficulty associated with correctly identifying intensional contexts. Unlike with number, articles, and quantifiers, there are a wide range of lexical items responsible for introducing a `:content` argument, as attitude predicates are a relatively open-class.

Task	F1	Count	Per-Annotation
Number	97.35	114	1.73
Articles	90.05	113	1.71
Quantifiers	95.45	20	0.30
Intensionality	91.43	53	0.80
All	93.52	300	4.55

Table 3: Inter-annotator agreement and count of enrichment types in the 66 doubly annotated AMR graphs.

5.3. Automatic Annotation Results

In the second experiment, we compare the output of the automatic annotator, A^3, to 126 singly annotated gold-standard annotations. Average count per annotation for each feature is provided in Table 4. The frequent occurrence of these semantic features highlight the need for representing them in meaning representations.

Task	Count	Per-Annotation
Number (Plural)	173	1.37
Articles	214	1.70
Quantifiers	41	0.33
Intensionality	102	0.81
All	530	4.21

Table 4: Count of enrichment types in the 126 gold-standard annotations.

Table 5 presents the precision, recall, and F1 scores for the automatic annotator.

Task	FP	FN	Precision	Recall	F1
Number	14	17	91.76	90.17	90.96
Articles	8	34	95.72	84.04	89.50
Quantifiers	2	2	96.00	96.00	96.00
Intensionality	15	24	84.54	77.36	80.79
All	39	77	92.26	85.79	88.91

Table 5: The performance of A^3 on 126 gold standard AMR graphs.

For the 173 plurals identified in the gold annotations, A^3 failed to identify 17 of them. It also labeled 14 extra cases with plural that are not marked in the gold annotations, yielding an F1 of 90.96. The sources of error originated mostly from incorrect alignment information and the parser's failure to identify the correct POS tags (see Table 6 in section 6). The F1 score for articles is 89.50, with high precision (95.72) and lower recall (84.04). A^3 failed to attach 34 out of the 214 (in)definite articles to the AMR graph and inserted 8 additional articles. Potential causes for the false negatives include failure to identify the correct head noun, incorrect alignment of the head noun, missing alignment of the head noun that disables attachment of articles, as well as incorrect article location due to mapping problems mentioned in

[5] https://github.com/emorynlp/StreamSide

section 4.5. The performance for quantifiers is the best among the features and scores highly for both precision and recall. Finally, A^3 achieves an F1 score of 80.79 for intensionality. While this score is lower than that of the other features, it is nonetheless quite high considering the degree of complexity of this classification task. Overall, the results demonstrate the efficacy of A^3 in enriching AMR graphs for the targeted features.

6. Analysis of Errors

In this section, we report on the errors made by A^3. These limitations stem from a number of issues. Among them are: imperfect annotation or alignment, limitations of the parsers, abstractness of the AMR graph, non-canonical or ungrammatical syntax, discrepancies between annotator judgements and the verb list, and inadequacies of certain PropBank argument structures. A percentage of error types made by A^3 is provided in Table 6, with specific examples provided in the text.

Limitations of the POS tagger caused A^3 to occasionally fail to label irregular plurals. For example, the tool correctly marks `person` for plural when aligned with *people*, but it fails to mark `phenomenon` for plural when aligned with *phenomena*. Moreover, *mathematics* is marked as plural by the automatic annotator even though it is associated with the concept `mathematics`. Lastly, the POS tagger fails to identify the head noun in *'the welfare rolls'* since *'rolls'* is treated as a verb instead of a plural noun.

The constituency parser struggles with dialogue when it features an interruption with a filler word, such as *'umm'* or *'err'*, producing a disjoint constituency tree. It may also struggle to correctly resolve syntactically ambiguous sentences. Lastly, there are a number of ungrammatical sentences in the dataset (e.g., *'For the time before everything is officially opened, opened, all , no cars can enter unless they have special permission'*) which lead to parsing errors.

The abstract and un-anchored nature of AMR can sometimes present difficulties for A^3 to map tokens to the corresponding concepts in the graph. For instance, *'according to'* is represented with the predicate `say-01` in AMR due to their similarity in meaning, though the token *'say'* does not appear anywhere in the sentence. Another example occurs with the noun phrase *'two men, deadly enemies to each other'* which is represented with two separate `man` concepts and thus should not be marked as plural in the graph.

Another source of error is discrepancies between the MegaVeridicality dataset and human annotator judgments about whether to mark an argument as intensional. For instance, the MegaVereridicality dataset contains some aspectual verbs which are not intensional such as *'continue'*.

Finally, certain ITVs have overloaded predicate senses in PropBank. For example, the ITV *'look for'* has an intensional object position which is annotated as `:ARG1` of `look-01`. However, the same numbered argument is used to annotate the non-intensional object argument of *'look at'*, as shown in the description tag of its PropBank argument structure (18).

(18)
```
look.01
  <role descr="thing looked at
   or for or on" f="gol" n="1">
```

7. Discussion

While the agreement scores of A^3 are impressive, there is nonetheless a gap in quality between the annotation tool's output and our manual annotations. Nevertheless, we expect this gap to inevitably shrink with the development of better parsers, and several of the remaining problems can be solved through the production of hand-written mapping dictionaries, similar to the ones we created for modal auxiliaries and common ITVs but at a larger scale.

Given its baseline performance A^3 can already be used to enrich a large number of graphs, which can then be quality checked by trained human annotators. This semi-automated approach affords a means of producing gold-standard meaning representations at a rate which far surpasses creating manual annotations from scratch (Oepen and Lønning, 2006; Abzianidze et al., 2017; Abzianidze and Bos, 2019).

8. Conclusion

Recent work on improving the representational adequacy of AMR has focused on enriching its graph structure. In this paper, we presented an automatic AMR annotation tool, A^3, designed to enrich AMR graphs to better represent a number of important semantic features including number, (in)indefiniteness, quantificational determiners, and intensional arguments. This task involves correctly identifying an appropriate label, before mapping it onto an existing AMR graph. This task is often non-trivial due to the abstract, or un-anchored, nature of AMR graphs. Our tool thus utilizes a number of cues provided by several state of the art parsers.

To demonstrate the effectiveness of the enrichment scheme as well as that of A^3, we presented two annotation experiments. The first involves manually producing doubly annotated graphs which are enriched for the semantic features mentioned above. IAA was calculated for specific labels, showing a high rate of agreement. Secondly, we compared the output of A^3 to gold-standard manual annotations. The F1 scores of the automatic annotator are close to that of human annotators except when identifying intensional arguments which is by far the hardest classification task. It is our hope that the present paper encourages further efforts to automatically augment existing AMR corpora, with the aim of producing large corpora of representationally adequate Abstract Meaning Representations. All code for this paper is publicly available on our repository at `https://github.com/emorynlp/EnrichedAMR/`.

Source of Error	Plural	Article	Quantifier	Intensionality
Incorrect or missing alignment	32.26%	19.05%		2.56%
POS tagger fails to identify correct tag	48.39%	16.67%		
Constituency parser error		50.00%		33.33%
Ambiguous/ungrammatical syntax			25.00%	
Abstractness of AMR graph	19.35%	14.28%	75.00%	30.77%
Verb list discrepancies				23.08%
Overloaded predicate sense for ITV				10.26%
Total	100%	100%	100%	100%

Table 6: Percentages of error types made by A^3

9. Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Alexa AI.

10. Bibliographical References

Abzianidze, L. and Bos, J. (2019). Thirty musts for meaning banking. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 15–27, Florence, Italy, August. Association for Computational Linguistics.

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain, April. Association for Computational Linguistics.

Andrew, G. and MacCartney, B. (2004). Statistical resolution of scope ambiguity in natural language. *Unpublished manuscript*.

Artzi, Y., Lee, K., and Zettlemoyer, L. (2015). Broad-coverage CCG semantic parsing with AMR. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1699–1710, Lisbon, Portugal, September. Association for Computational Linguistics.

Asudeh, A. and Crouch, R. (2002). Glue semantics for hpsg. In *Proceedings of the 8th international HPSG conference, Stanford, CA. CSLI Publications*.

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, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bender, E. M., Flickinger, D., Oepen, S., Packard, W., and Copestake, A. (2015). Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK, April. Association for Computational Linguistics.

Bird, S. and Loper, E. (2004). Nltk: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217.

Blackburn, P. and Bos, J. (2005). *Representation and inference for natural language: A first course in computational semantics*. Center for the Study of Language and Information Amsterdam.

Bonial, C., Badarau, B., Griffitt, K., Hermjakob, U., Knight, K., O'Gorman, T., Palmer, M., and Schneider, N. (2018). Abstract Meaning Representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Bonial, C., Donatelli, L., Abrams, M., Lukin, S. M., Tratz, S., Marge, M., Artstein, R., Traum, D., and Voss, C. (2020). Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France, May. European Language Resources Association.

Bos, J. (2016). Squib: Expressive power of Abstract Meaning Representations. *Computational Linguistics*, 42(3):527–535, September.

Bos, J. (2020). Separating argument structure from logical structure in AMR. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 13–20, Barcelona Spain (online), December. Association for Computational Linguistics.

Bunt, H. (2020). Annotation of quantification: The current state of ISO 24617-12. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 1–12, Marseille, May. European Language Resources Association.

Buys, J. and Blunsom, P. (2017). Robust incremental neural semantic graph parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1215–1226, Vancouver, Canada, July. Association for Computational Linguistics.

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, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chen, D., Palmer, M., and Vigus, M. (2021). AutoAspect: Automatic annotation of tense and aspect for uniform meaning representations. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 36–45, Punta Cana, Dominican Republic, November 11. Association for Computational Linguistics.

Choi, J. D. and Williamson, G. (2021). Streamside: A fully-customizable open-source toolkit for efficient annotation of meaning representations.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Crouch, R. and Kalouli, A.-L. (2018). Named graphs for semantic representation. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 113–118, New Orleans, Louisiana. Association for Computational Linguistics.

Donatelli, L., Regan, M., Croft, W., and Schneider, N. (2018). Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Donatelli, L., Schneider, N., Croft, W., and Regan, M. (2019). Tense and aspect semantics for sentential AMR. *Proceedings of the Society for Computation in Linguistics*, 2(1):346–348.

Goodman, M. W. (2020). Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online, July. Association for Computational Linguistics.

He, H., Xu, L., and Choi, J. D. (2021). ELIT: Emory Language and Information Toolkit. *arXiv*, 2109.03903.

Higgins, D. and Sadock, J. M. (2003). A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96.

Kingsbury, P. R. and Palmer, M. (2002). From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer.

Kipper, K., Palmer, M., and Rambow, O. (2002). Extending PropBank with VerbNet Semantic Predicates. In *Proceedings of the AMTA Workshop on Applied Interlinguas*.

Kipper, K. (2005). Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. Master's thesis, University of Pennsylvania.

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686.

Lai, K., Donatelli, L., and Pustejovsky, J. (2020). A continuation semantics for Abstract Meaning Representation. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 1–12, Barcelona Spain (online), December. Association for Computational Linguistics.

Lin, Z. and Xue, N. (2019). Parsing meaning representations: Is easier always better? In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 34–43, Florence, Italy, August. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Manshadi, M., Allen, J., and Swift, M. (2011). A corpus of scope-disambiguated English text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 141–146, Portland, Oregon, USA, June. Association for Computational Linguistics.

Matthiessen, C. M. I. M. and Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter.

Oepen, S. and Lønning, J. T. (2006). Discriminant-based MRS banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Oepen, S., Abend, O., Hajic, J., Hershcovich, D., Kuhlmann, M., O'Gorman, T., Xue, N., Chun, J., Straka, M., and Uresova, Z. (2019). MRP 2019: Cross-framework meaning representation parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1–27, Hong Kong, November. Association for Computational Linguistics.

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, Online, November. Association for Computational Linguistics.

Pustejovsky, J., Lai, K., and Xue, N. (2019). Modeling quantification and scope in Abstract Meaning Representations. In *Proceedings of the First International*

Workshop on Designing Meaning Representations, pages 28–33, Florence, Italy, August. Association for Computational Linguistics.

Schwarz, F. (2020). Intensional transitive verbs: I owe you a horse. *The Wiley Blackwell Companion to Semantics*, pages 1–33.

Shou, Z. and Lin, F. (2021). Incorporating eds graph for amr parsing. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 202–211.

Srinivasan, P. and Yates, A. (2009). Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1465–1474.

Stabler, E. (2017). Reforming AMR. *International Conference on Formal Grammar*, pages 72–87.

Van Gysel, J. E., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C.-R., et al. (2021). Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, pages 1–18.

White, A. S. and Rawlins, K. (2018). The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, pages 221–234.

White, A. S., Rudinger, R., Rawlins, K., and Van Durme, B. (2018). Lexicosyntactic inference in neural models. *arXiv preprint arXiv:1808.06232*.

Williamson, G., Elliott, P., and Ji, Y. (2021). Intensionalizing Abstract Meaning Representations: Non-veridicality and scope. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 160–169, Punta Cana, Dominican Republic, November 11. Association for Computational Linguistics.

11. Language Resource References

Knight et al. (2020). *Abstract Meaning Representation (AMR) Annotation Release 3.0*. distributed via Linguistic Data Consortium, ISLRN 676-697-177-821-8.

Palmer et al. (2004). *Proposition Bank (PropBank) I*. distributed via Linguistic Data Consortium, ISLRN 874-058-423-080-1.

White et al. (2018). *The MegaVeridicality Dataset*. available at: `http://megaattitude.io/projects/mega-veridicality/`.