Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Yuxin Ji April 12, 2022

Enriching and Evaluating Meaning Representations

By

Yuxin Ji

Jinho D. Choi
Adviser

Department of Quantitative Theory and Methods

Jinho D. Choi
Adviser

Marjorie Pak
Committee Member

Li Xiong
Committee Member

2022

Enriching and Evaluating Meaning Representations

By

Yuxin Ji

Jinho D. Choi
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Quantitative Theory and Methods

2022

Abstract

Enriching and Evaluating Meaning Representations
By Yuxin Ji

Meaning representations receive increasing attention in the field of computational linguistics in recent years. Works include developing frameworks to represent the meaning of a sentence and exploring schemes to extract document-level interpretations. Abstract Meaning Representation (AMR) is a semantic graph framework which fails to adequately represent a number of important semantic features, including number (singular and plural), 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 implemented on existing AMR corpora due to the labor costs associated with manual annotation. In addition to sentence-level, there are attempts to extend such representations to the document-level, one of which is on coreference resolution. In this paper, I develop an automated annotation tool which algorithmically enriches AMR graphs to better represent number, (in)definite articles, quantificational determiners, and intensional arguments. I compare the automatically produced annotations to gold-standard manual annotations and show that the automatic annotator achieves impressive results, even matching those of human annotators for certain tasks. [1] Through implementing the enriched structure to the large AMR 3.0 corpus and train models using the enriched graphs, I attested the feasibility of my proposals for enrichment. Additionally, I develop an annotation scheme for document-level coreference [2] and conduct a comparison study for the text type effects across news, fables, and a novel Reddit data. The experiment results indicate the need to develop schemes adjusted for each text type due to their distinct characteristics in language use and content.

[1] Code for the automatic annotation tool and evaluation metrics is available at `https://github.com/emorynlp/eAMR`

[2] The annotation guidelines for coreference is available at `https://github.com/Yuxin-Ji/Corefence-Annotation/blob/main/Coref-Guideline.md`

Enriching and Evaluating Meaning Representations

By

Yuxin Ji

Jinho D. Choi
Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Quantitative Theory and Methods

2022

Acknowledgments

My thanks to Jinho D. Choi, my adviser, who encouraged me to do the honors research in the first place and inspired me in numerous aspects of academic and professional lives. Special thanks to Gregor Williamson, who advised and collaborated with me in the stream of works on meaning representations. My other committee members, Marjorie Pak and Li Xiong were very supportive throughout this project, for which I am grateful. Additionally, I would like to thank my companion thesis students, Angela Cao and Yingying Chen, who worked on the larger meta-project of Reddit annotation comparison, with each focusing on different directions but providing useful feedback to each other, conducting annotations for each others' schemes, and supporting me mentally in the process of thesis writing. I also appreciate the help from Ph.D. students at the Emory NLP lab including Han He, who generously shared his knowledge in computer science and the great parsing tool ELIT, Sichang Tu, who provided support when I encountered technical problems, James Finch, who provided his insights in the early stage of theoretical formulations in this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

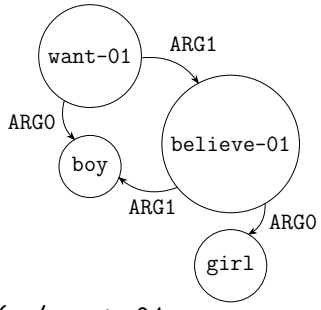
1.1 Introduction

Over the past decades, research in computational linguistics and natural language processing have achieved notable progress on representing language at the syntactic level such as the dependency and constituency tree structures. As the models for syntactic parsing started to reach high accuracy, the field began to shift its interest to meaning representation in recent years. Unlike syntactic-driven tasks that heavily rely on the fixed rules of language structure and grammar, meaning representation focuses on the meaning of language and requires more real world knowledge. Such representation brings one step forward in 'making sense' of natural language and allows the machine to understand human language more 'naturally'. Another potential advantage given by this is its cross-linguistic universality, since languages are often considered to be much similar in their content than surface forms [2].

However, because meaning representation encodes more information of natural language including a connection to the real world context, the task is inherently harder than syntactic representation. Despite existing endeavors in the field, neither the design nor the application of meaning representations are fully conquered. In this paper, I will enrich the meaning representation at two levels: the sentence level and the document level.

Presently, there are multiple existing works on the sentence-level meaning representation, of which graph-based representations gain the most interest due to their better expressiveness and representational adequacy[49]. Abstract Meaning Representation (AMR) is one of such semantic graph framework that represents natural language sentences in directed, acyclic graphs [11]. Nodes represent concepts, and labeled edges represent relations between concepts (1-b). AMRs are most commonly written in PENMAN format [57], 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. The abstractness and un-anchored nature of AMR distinguishes itself from other graphical meaning representations including Elementary Dependency Structures (EDS) [58] and Universal Conceptual Cognitive Annotation (UCCA) [1], which all present some degree of anchoring [59, 60]. This design of AMR is particularly useful in capturing the meaning of expressions that are not syntactically well-formed, which could be common in colloquial data like dialogues and online social media. Moreover, AMR has abundant resource in gold-standard manual annotation [46, 47, 48] and promising state-of-the-art parsers [23, 60, 72, 51, 14].

A major disadvantage, however, is that it can give rise to inter-annotator disagreement [13], as well as making the task of parsing harder [22, 52, 59, 60]. 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) [58] parser, has been shown to improve AMR parsing results [64].

Another feature of AMR is its underspecification with respect to grammatical and semantic features other than argument structure. The current AMR3.0 schema does not encode important semantic features such as tense and aspect, plurality, article, and scope. A consequence of this design choice is that AMR introduces ambiguity which is absent from its corresponding natural language 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 natural language understanding tasks beyond identifying argument structure and downstream applications such as natural language inference and generation.

Numerous proposals have been developed to improve the representational adequacy of AMR by enriching its graphical structure (see section 2.1). Despite various theoretical works on improving AMR, few of the suggested improvements have been adopted in other AMR-related research, meaning that there is a lack of annotated corpora for the proposed augmented structures. The gold standard AMR 3.0 corpus [48] of over 59,000 manually annotated sentences remains the major resource for parser training and evaluation. Considering the large 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 [25]. On the representational side, the challenge is to develop a design that is not only suitably expressive, but is also tractable for automatic annotation. Inspired by such observation, one goal of this study is to enrich meaning representation at the sentence-level by presenting a graph structure with augmented grammatical information about number, definiteness [66],

quantifiers [19, 50, a.o.] and intensional arguments [71], as well as introducing an automated rule-based annotator for augmenting AMR graphs with these information.

Additionally, this thesis aims to improve meaning representation beyond the sentence level. With the end goal of enabling machines to conduct language interactions naturally with human, a document-level understanding such as the referential or inferential knowledge is indispensable. Many current studies on meaning representation are refrained to the sentence-level, leaving the document-level applications to be explored. Among the many document-level applications, coreference, which is the common reference to the same entity or event, is one of most basic yet challenging tasks towards understanding the flow of a document or dialogue. For example, while it is easy for a human to keep track of a character in a novel, it is harder for the computer to understand the reference as the same entity may be represented with different surfaces forms and vice versa. Most of the previous works on coreference has focused on well-written text data such as news, with the informal text types less studied. The second goal of this paper is to enrich meaning representation at the document level by developing an annotation scheme for a novel Reddit dataset and compare its performance across three different text types: news, fable, and Reddit.

1.2 Thesis Statement

This thesis aims to entich meaning representation in two ways: at the sentence level and the document level. The research objectives and hypothesis are summarized as follows:

1. For the sentence-level enrichment, I proposes an augmentation of grammatical information to current AMR graph structure and introduce a rule-based method that add these information automatically. I hypothesize that with the enriched AMR graph structure, the parsing accuracy will be slightly lowered due to a more complex graph structure.[1]

[1]This study is conducted under the co-advise of Gregor Williamson and Jinho D. Choi and is submitted to the LREC 2022 conference as a first-authored paper.

2. For the document-level enrichment, I develops a coreference annotation schema for Reddit documents and presents a comparison study of annotation across different text types using the same annotation schema. I hypothesize that each text type will perform differently and have its unique features. Therefore, specific adjustments in annotation guidelines are needed for different text types at the moment.[2]

In chapter 2, I provide a summary of related works. In chapter 3, I present the efforts towards sentence-level enrichment, which is followed by the study of document-level coreference in chapter 4. The two studies are independent from each other in terms of experimental design and result analysis and are thus presented separately. However, both of the works are derived from the goal to enrich meaning representation towards better performance and broader application. Thus, an overall summary of the findings and contributions of the two studies will be presented in chapter 5.

[2]This study is conducted as part of a meta-project that involves two other undergraduate honors students who work on temporal and causal relations. The team is working on submitting a collaborated paper to the LAW Workshop in April.

Chapter 2

Related Works

2.1 Related Work

There has been a concerted effort towards improving the representational adequacy of AMR, as well as its recent offspring, Uniform Meaning Representation (UMR) [67] and Widely Interpretable Meaning Representations (WISeR) [33]. This strand of research endeavors to improve the expressive power of AMR either in terms of enriching its graphical structure [17, 31, 32, 62, 18, 20, 67] or by adding information during a subsequent translation step into a logical form (LF) in first-order logic or lambda-calculus [7, 19, 66, 50, 71]. Table 2.1 lists the phenomena addressed in these representative works.

Translation		**Richer Graph Structure**	
Artzi et al. [7]	(*coreference*)	Bonial et al. [17]	(*comparatives*)
Bos [19]	(*quantifier scope*)	Donatelli et al. [31]	(*tense and aspect*)
Stabler [66]	(*number, determiners*)	Donatelli et al. [32]	(*tense and aspect*)
Lai et al. [50]	(*quantifier scope*)	Pustejovsky et al. [62]	(*quantifier scope*)
Williamson et al. [71]	(*Intensionality*)	Bonial et al. [18]	(*speech acts*)
		Bos [20]	(*quantifier scope*)
		Van Gysel et al. [67]	(*quantifier scope*)

Table 2.1: Approaches to improving the representational adequacy of AMR

Each of these approaches has its 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 [48], is costly and time-consuming. Moreover, if the resulting graph structure cannot be mapped to a coherent model theoretical semantics, then the enriched graph will also be representationally inadequate. On the other hand, making use of a translation function with minimal revision to the graphical structure allows us to work with existing corpora. Ultimately, however, we would like to work with AMR graphs directly, avoiding the need for translation into symbolic logic. For these reasons, we take enriching of the graphical structure to be the long term goal, with the caveat that the graphs should be translatable into a model theoretic semantics with as few ad-hoc interpretation rules as possible.

2.1.1 Previous AMR Graph Enrichments

This research is inspired by WISeR [33], a project with the goal to create a frameless meaning representation suitable for dialogue interpretation, which I participated as a linguistic annotator. However, both AMR and WISeR under-represent several crucial grammatical and semantic features in their graph structures.

Despite its importance to natural language meaning, grammatical number has not received much attention in the AMR literature, with the notable exception of Stabler [66] who proposes various modifications to encode plurality, articles, and scope in AMR. In his AMRs, a grammatical number indicator is attached directly to the concept. Other information for noun concepts including articles and quantifiers are represented in the `:quant` role. The proposed changes are highlighted in red in (1-b).

(1) a. *A computer is on every desk.*

b.
```
(b / be-located-at-91
   :ARG0(c / computer.sg
      :quant a)
   :ARG1(d / desk.sg
      :quant every)
```

While the work of Stabler [66] is the starting point for the present work, it nonetheless has some weaknesses. Firstly, the augmentation of the singular marker for nouns is arguably redundant, since the absence of `.pl` can be interpreted as singular. Secondly, including articles in the `:quant` role is potentially undesirable since definites may be better thought of as referential rather than quantificational.

In addition to his enrichment of graphical structures, Stabler [66] proposes to generate quantifier scope from underspecified AMRs using the following algorithm: (i) attach quantificational determiners, represented as AMR constants, to a `:quant` role, (ii) translate ambiguous AMRs to unambiguous higher-order logic (HOL) representations using a deterministic finite state machine, (iii) produce possible interpretations of quantification scope by modifying the compositional order of the HOL functions, but not changing their contents, and (iv) add a definite constant `the` to help with coreference. This translation mechanism generates all logically possible readings from the various combinations of quantifiers in a sentence. This approach therefore generates unattested scope orderings for natural language sentences. Nevertheless, this approach can be employed provided it is combined with heuristics to constrain possible interpretations by filtering out undesirable scope orderings. Imposing soft constraints that define preference over possible readings would best be implemented by training a model on comprehensive scope-disambiguated corpora. 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 [39, 5, 65, 56], or are not yet publicly available [21].

Another approach to deal with quantifier scope is proposed by Bos [19] and Lai et al. [50] who develop a translation from AMR into continuation semantics [12]. A problem with this approach is that quantifiers may often take exceptionally wide scope, resulting in unattested interpretations of universal quantifiers in conjunctions, disjunctions, and across sentence boundaries [66].

Pustejovsky et al. [62] and Van Gysel et al. [67] encode scope through the introduction of

a special root node `(s / scope)` which indicates the scope order between quantifier phrases as numbered arguments of `scope`, as in (2).

(2)
```
(s / scope
   :pred(b / be-located-at-91
      :ARG0(c / computer)
      :ARG1(d / desk
         :quant (e / every))
   :ARG0 d
   :ARG1 c)
```

This allows for more control over scope orderings. But it is problematic for at least two reasons. First, it introduces inconsistency in the interpretation of the graph structure, since `scope` is introduced as an instance of a predicate, although it cannot naturally be understood as "an instance of scope". The same can be said for the representation of quantificational determiners themselves, which are probably better understood as AMR constants rather than concepts. Second, unlike most other semantic features represented by AMR, quantifier scope is not obviously associated with specific lexical items or particular construction in natural language. That is, there is no natural language span which corresponds to a scope node per se.

Other attempts to encode scope in AMR include Williamson et al. [71], a work done by our lab member which I participated in as a co-author, who represent the scope of intensional operators through the addition of a `:content` role in combination with a deterministic translation from AMRs to an intensional simply-typed lambda calculus (STLC). Our work addresses the non-veridical problem of intensional contexts in AMR. For example, a sentence like (3-a) would permit the inference of '*The boy is sick*' although it is not a guaranteed fact according to this sentence. We introduced a `:content` role that act as an intentional operator in the proposed semantic translation (3-c).

(3) a. *The boy believes that he is sick.*

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

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

In my enrichment scheme, described in section 3.1, I adopt and refine a number of these suggestions.

2.1.2 Automatic Enrichment Efforts for AMR

As mentioned in 1.1, few of the AMR-related enrichment proposals are adopted mainly due to the lack of annotated corpora with the augmented structures, without which the adequacy of these proposals could not be attested by the AMR parsers. Enriching the large pre-existing AMR 3.0 corpus manually would be laborious and costly, for which reason an automated method for augmentation is in demand. While some previous work has focused on classifying AMR labels for natural language sentences [25], there has been little attempt to systematically add these labels to the graph structure. Augmenting the AMR graphs requires additional steps in mapping the semantic features from sentence tokens to the abstract, unanchored, graphs.The methodology of this study is inspired by Chen et al. [25] who introduce a rule-based classifier for labeling grammatical aspect based on the UMR guidelines. The classifier uses part-of-speech (POS) tagging and lexical frames such VerbNet [43, 42]. It takes in a sentence and returns a list of events with labeled aspect. Unfortunately, this tool is tested only on a limited dataset with two annotated documents of relatively simple sentence structures and a limited list of VerbNet frames. It is thus unclear how it would perform with a larger dataset consisting of a wider range of grammatical constructions. Also, while this automatic annotation system handles the complicated aspect labeling procedure, it does not attempt to fit the extracted aspect information onto AMR/UMR graphs. Like Chen et al. [25], the present paper performs a rule-based identification of grammatical features. In addition, it performs the additional step of fitting the labels onto the corresponding AMR graph and could handle most of the complex situations in the AMR 3.0 corpus.

In this thesis, I focus on the representation of gammtical number (singular/plural),

(in)definite articles, quantifiers, and intensional arguments, all of which can provide important quantificational and referential cues for scope, coreference resolution, and natural language inference tasks.

2.2 Document-level Coreference

Coreference is a prevalent yet complicated phenomenon in natural language that requires an understanding from syntax and pragmatics as it is highly dependent on context. A concept may be represented in different forms and a same form may denote different concepts under different contexts. Consequently, a document-level or discourse-level interpretation is needed. For this reason, coreference resolution has long been a hard task for machines.

Past works on coreference annotation such OntoNotes often come together with annotated syntactic tree structures [40]. Recently, UMR [67] has attempted to incorporate document-level coreference into the graph structure of meaning representation. Adding the document-level information to meaning representation enables it to be more versatile in downstream dialogue management tasks. Unfortunately, UMR does not provide any annotated corpus that could be used for training. The present work is directly inspired by UMR's attempt to enrich the document-level representation and adopts and refines several of its concepts to the annotation scheme (see 4.1).

Another issue with coreference resolution is the lack of well-annotated corpora. Specifically, existing gold standard corpora largely focus on well-edited texts. The nowadays benchmark OntoNotes [40] and the prior efforts including MUC (1987-2001) [35] and ACE (1999-2005) [30] contain a variety of news and broadcast data in multiple languages. After the launch of OntoNotes, there are various attempts for coreference resolution in different genre and domains. LitBank [10] annotates coreference on English literature; PreCo [26] provides annotation on school examinations; Guha et al. [36] challenges the coreference in Quiz Bowl questions; Apostolova et al. [6] explores domain adaptation of coreference reso-

lution to biomedical reports. Texts from online forum discussion platforms like Reddit are less explored. Reddit data differ from the well-formulated data as it is messy and more colloquial. This characteristics makes it particularly interesting to study as its language is closer to the natural spoken language.

The enduring efforts in producing corpora of different genre and domains demonstrate the distinctive nature of these data in such way that each needs specific adaptations in annotation guidelines to achieve high performance. Parallel comparison across data sources thus received lesser attention due to the hardness of producing generic guidelines applicable for multiple data sources. Nevertheless, applying the same annotation scheme to different data could provide interesting insights in understanding the coreference task and data universals and guide future annotation designs. Existing work on this direction include GUM [73], a multi-layer corpus that annotates various data using the same guidelines and explores the effect of text types across news, interviews, travel guides, and how-to guides data.

The current study is part of a larger project at the Emory NLP Lab that aims to add to this stream of research by comparing the text type effects across news, fables, and the less studied Reddit data. In addition to my work investigating the coreference relationship, two other lab members work in parallel on document-level temporal and causal relations, in which I was involved as an annotator for their schemes. The three types of semantic relations (coreference, temporal, and causal) capture a good amount of the document-level information. While the development of annotation schemes and cross-text-type analyses for the three relations are done independently at the moment, we plan to investigate in the correlations across these relations upon completion of the thesis. For this paper, I will focus on the establishment of coreference annotation scheme for Reddit data and the comparison across different text types.

Chapter 3

Sentence-level Graph Structure

In this chapter, I present the study aiming to enrich the graph structure of AMRs at the sentence level. Following the related works discussed in section 2.1, I outline the basic cases for the proposed enrichments. In section **??**, I describe the structure of the automatic annotator, and discuss some of the more challenging constructions which arise due AMRs inherent abstraction from grammatical form. Next, in section 3.4 I report two annotation experiments and discuss the remaining difficulties for the automatic annotator. The first experiment reports Inter-Annotator Agreement (IAA) scores for gold-standard manual annotations, demonstrating the reliability of our enrichment scheme. The second experiment compares the output of our automatic annotator to manually produced annotations. Then, the parsing performance for the enriched graphs are presented and analyzed in section 3.5. Finally, in section 3.6, we discuss implications of the present approach on data production.

3.1 Enriched Graph Structure

In this section, I outline the enriched graph structure adopted in the present study. I describe simple cases for each feature, reserving discussion of the various mapping problems posed by AMR's abstraction from surface form to section 3.3.

3.1.1 Number

Plurality is an important grammatical feature in English that expresses important semantic information. In many cases, singular and plural adds important information because it is the only indicator of quantity. Even for noun phrases with a quantificational determiner, plurality is not entirely redundant. For example, the two cases in (1) can be differentiated only if plurality is marked.

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

Consequently, plurality should be represented in AMR to avoid the introduction of unwanted ambiguity. Stabler [66] represents both plural and singular nouns by appending a marker to the corresponding concept matching the noun's grammatical number, as in (2).

(2) 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 approach is not as efficient as it can be. Since, plurality is a binary attribute (in English), it is redundant to annotate both singular and plural explicitly. Instead, we can leave singular as the unmarked form and mark only the plural noun. This is similar to AMR's treatment of polarity, where the `:polarity` attribute is only ever marked for negative polarity (`:polarity -`) with the default interpretation being positive polarity. Thus, I adopted Stabler's representation of grammatical number [66] with this modification in my initial attempt. The initial version of my modified guideline in (3) adds information through minimal revision to the standard AMR, where a plural marker is appended only to plural noun concepts.

(3) **Enriched AMR: Number (Version 1)**

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.pl)))
```

However, the parsing experiment proofed the modification in (3) to be invalid and undesirable for most parsers. While the intention is to keep the plurality information as close to the concept as possible, such attachment of the plural marker as a string creates difficulty for the parser to learn concepts, particularly because it requires regular expression that conflicts with other matching tasks such as word senses. Therefore, I modified the scheme by encoding the plurality information in a new `:plural` role that would link to a concept only when it is plural, using the attribute `+`, as in (4).

(4) **Enriched AMR: Number (Version 2)**

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
         :plural +)))
```

3.1.2 Articles

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*'), (in)definite articles are typically referential. To avoid confounding the role of articles and quantificational determiners, we introduce a new `:def` role (for 'definite') with the attribute `+` and indefinite as `-`, as in (5). Note that unlike plurality, we represent both definite and indefinite articles to distinguish them from noun phrases without an article.

(5) **Enriched AMR: Articles**

a. *The boys give a girl some cookies.*

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

3.1.3 Quantifiers

The majority of work on quantifiers in AMR treats them as constants as opposed to concepts [19, 66, 50, 71]. As such, we replace quantificational concept arguments of a `:quant` role with a quantificational constant. It is also common 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 [19] and [50], we do not conflate universal quantifiers such as *every*, *all*, and *each*, as these may carry information about distributivity 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)
```

In this study we opt not to enrich the graph structure with a `scope` node, as in Pustejovsky et al. [62], for several reasons. Firstly, as mentioned above, the `scope` node requires an idiomatic interpretation. Secondly, use of `scope` node presupposes that any scope ambiguity of the corresponding natural language sentence can be resolved. However, this is not always possible for human annotators, let alone an automatic annotation tool. Finally, it is debatable whether we need to represent quantifier scope in AMRs at all, provided there is some independent mechanism of scope resolution (as in Minimal Recursion Semantics [28], Hole Semantics [16], or Glue Semantics [8]). Crucially, if AMRs are left underspecified for scope, no information is lost since the corresponding natural language sentence is also scopally ambiguous. Therefore, we assume that scope readings can be generated deterministically from AMR graphs in a manner similar to [66], and later filtered.

3.1.4 Intensionality

Finally, we note in our work [71] that AMR is unable to represent non-veridical environments [29]. To remedy this, we 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. This paper directly follows Williamson et al. [71]'s proposal to replace numbered arguments with the `:content` role where appropriate.

(8) **Enriched AMR: Intensionality**

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

Following the scheme just described, the sentence in (9-a) is represented as in (9-b).

(9) **Enriched AMR**

a. *A boy believes that the girls gave everyone some cookies.*

b.
```
(b / believe-01
   :ARG0 (b2 / boy
      :def -)
   :content (g / give-01
      :ARG1 (g2 / girl
         :plural +)
         :def +)
      :ARG1 (c / cookie
         :plural +)
         :quant some)
      :ARG2 (p / person
         :quant every)))
```

3.2 The Automatic Annotator

The automatic annotator 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 [34].[1] In sections 3.2.1-3.2.4, we describe the simpler cases of classification and mapping, before describing some of the numerous challenges in section 3.3.

3.2.1 Number

The automatic annotator searches for tokens identified by the Stanford CoreNLP parser[2] [55] as having the plural *noun* part-of-speech (POS) tag. The plural noun is then mapped to the corresponding alignment in the AMR graph. In the initial attempt (version 1 in (3)), the `.pl` marker is appended to concept argument of the instance assignment triple. For the revised scheme (version 2 in (4)), a new triple with the positive plural attribute is linked to the concept. Several abstract structures of AMR require special treatment, which are

[1] `https://github.com/goodmami/penman`

[2] `https://github.com/stanfordnlp/CoreNLP`

discussed in section 3.3.

3.2.2 Articles

Articles are identified through using a part-of-speech (POS) tag match and a string match. Tokens that are classified with a *DET* tag are then applied a string match on definite ('*the*') and indefinite ('*a/an*') articles. We then locate the span of head noun using the Stanford CoreNLP constituency parser [55] which was chosen, after experimenting with different constituency parsers including ELIT [38] and the Berkely Neural Parser [44], due to it's performance. Finally, a new AMR triple containing the article information is attached to the concept corresponding to the span of the head noun.

3.2.3 Quantifiers

The conversion for quantifiers utilize clues 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.

3.2.4 Intensionality

The annotator identifies intensionality through relevant lists of verbs and the constituency parsing structures. In most cases, appropriate uses of the `:content` role are identified using the MegaVeridicality dataset megaveridicality. Finite clauses are identified using MegaVeridicality version 1 [69], and non-finite clauses using version 2 [70]. We loop through the lemmatized tokens and search for ones that are in the MegaVeridicality dataset. We compared the NLTK [15] and LemmInflect[3] lemmatizer and found that LemmInflect performs better. An

[3] `https://github.com/bjascob/LemmInflect`

intensional context is identified by checking if the matched verb is followed by a sentential complement, which is signified by a corresponding verb phrase constituent containing SBAR or S labels.

We found that for verbs like '*say*', the sentence structure could not be correctly identified by the parser when it does not follow the normal sentence order (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. To deal with such cases, we instead look for sentences where the verb is not followed by a noun phrase and modify them as `:content` role exclusively.

3.3 Mapping Difficulties

This section lists some non-canonical cases of each phenomena which are handled by the automatic annotator, but which require additional mapping instructions. Discussion of cases which are not presently handled by the annotator are reserved to section 3.4.4.

3.3.1 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 [41] wherever possible. For instance, AMR uses a `person` concept as the argument of a predicate in order to represent agentive nouns (10) and patient nouns (11).

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

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

Other deverbal nouns are often represented through the use of an implicit `thing` argument.

(12) 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`.

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

The 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 `.pl` feature, `:def +/-` attribute, or `:quant` constant. However, the concept which most transparently corresponds to the surface string is not the root, for example *uncle* in (13-b). To solve this, we track back through the directed edges of the sub-graph to find the root node, which we then mark with the relevant feature or attribute.

3.3.2 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 either `:def` or `.pl` is redundant.

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

3.3.3 Intensional Transitive Verbs

In addition to attitude predicates present in the MegaVeridicality dataset, the automatic annotator 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 [63]. This can be seen in the following examples, which are semantically coherent despite the non-existence of unicorns in the actual world.

(15) *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, our automatic annotator 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 [61].

3.3.4 Other Intensional Operators

Besides attitude predicates and ITVs, we design our parser to handle modal auxiliaries, modal verbs, and intensional raising predicates. Consequently, our automatic annotator uniformly converts the following numbered arguments of certain modal concepts senses into a `:content` role. These are summarized in Table 3.1.

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`

Table 3.1: Numbered arguments of modal concepts which are uniformly converted to `:content`.

3.4 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 enrichment guidelines by doubly annotating 30 PENMAN graphs selected from the AMR 3.0 corpus. In the second experiment, we singly annotate an additional 96 graphs and compare our 126 manually annotated graphs to the output of our automatic annotation tool.

3.4.1 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 20 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 Inter-Annotator Agreement between the gold standard annotation by calculating F1 scores for the features of interest.

For the second experiment, we singly annotate the remaining 96 graphs, creating a dataset of 126 gold-standard human annotations. We use our automatic annotation tool to process the same 126 graphs and compare the output of our automatic annotator with our gold-standard annotations.

All annotations were carried out by the first and second authors using StreamSide [27], an open-source annotation tool for graph-based meaning representations.[4]

[4] `https://github.com/emorynlp/StreamSide`

3.4.2 Manual Annotation Results

In the first experiment, we doubly annotate 20 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 [24]. 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 our guidelines.

Table 3.2 presents the F1 scores and the data statistics for the 30 double annotations. The corpus contains around one grammatical number and article each per graph and one quantifier and intensional role per 2-3 graphs. The F1 scores for number, article, and quantifier range from 92.11 to 97.96, demonstrating the robustness of our annotation guidelines for human annotation. The IAA for intensionality is predictably lower, 72.73, due to the increased difficulty associated with correctly identifying intensional contexts. Moreover, 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 (Plural)	97.96	24	0.80
Articles	92.11	38	1.27
Quantifiers	96.55	12	0.40
Intensionality	72.73	11	0.37
Overall	92.05		

Table 3.2: Inter-Annotator Agreement (IAA) and count of enrichment types in the 30 doubly annotated AMR graphs.

3.4.3 Automatic Annotation Results

In the second experiment, we compare the output of our automatic annotator to 126 singly annotated gold-standard annotations. On average, each annotation in the gold-standard

dataset contains 1.36 plural numbers, 1.75 articles, 0.32 quantifier, and 0.71 intensional context, as shown in table 3.3. The frequent occurrence, particularly for intensional arguments, reflect the non-triviality of the enrichment tasks.

Task	**Count**	**Per-Annotation**
Number (Plural)	171	1.36
Articles	220	1.75
Quantifiers	40	0.32
Intensionality	90	0.71

Table 3.3: Absolute and per-annotation count of enrichment types in the 126 gold-standard annotations.

Table 3.4 presents the precision, recall, and F1 scores for the automatic annotator. For the 171 plural numbers identified in the gold annotations, the annotator failed to identify 18 of them. It also labeled 17 extra cases with plural that are not marked in the gold annotations, yielding an F1 score of 89.74. The sources of error originated mostly from incorrect alignment information and the parser's failure to identify the correct POS tags.

The F1 score for articles is 87.68, with high precision (95.19) and lower recall (81.28). The annotator failed to attach 41 out of the 220 gold (in)definite articles to the AMR graph and inserted 9 additional articles not identified by the gold. Potential causes for the false negative cases 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 section 3.3.

The performance for quantifiers is the best among all features and receives high score for both precision and recall. The annotator receives an F1 score of 72.53 for intensionality, which is only 0.20 lower than the score for manual annotation (72.73), which is a very promising result. Overall, the results demonstrate the high efficacy of the automatic annotator in enriching the AMR graphs on the targeted features, even to a level comparable to human annotators for the task of intensionality.

Task	FP	FN	Precision	Recall	F1
Number	17	18	90.00	89.47	89.74
Articles	9	41	95.19	81.28	87.68
Quantifiers	4	5	92.00	90.20	91.09
Intensionality	26	24	71.74	73.33	72.53
Overall	56	88	88.78	83.43	86.02

Table 3.4: The performance of automatic annotation tool on the 126 AMR graphs.

3.4.4 Analysis of Errors

In this section, we report on the errors observed for our automatic annotator. These limitations stem from a number of issues, among them are: imperfect annotation or alignment, limitations of the parsers, inadequacy of certain PropBank argument structures, and non-canonical or ungrammatical syntax.

Limitations of the POS tagger caused our automatic annotation tool 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 `(m / mathematics)`. Similarly, the constituency parser occasionally struggles with dialogue when it features interruption by a filler word, such as '*umm*' or '*err*', producing a disjoint constituency tree.

The abstract and un-anchored nature of AMR can sometimes present difficulties for the annotator 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.

Intensional arguments of event nominals are not presently handled by the annotator. While the event nominals are classified as nouns by the POS tagger, AMR represents them in predicative structures. For example, the event nominal *'attempt'* in the sentence *'his attempt to enter the store'* is represented as (16).

(16)
```
(a / attempt-01
   :ARG0 (h / he)
   :ARG1 (e / enter-01
      :ARG1 (s / store)))
```

Gerundive complements such as *'I remember swimming'* also pose a challenge since they are recognized as nouns by most constituency parsers, making it difficult to distinguish from non-intensional arguments like *'I remember Mary'*.

Finally, certain ITVs cannot be handled due to overloaded predicate senses in PropBank. For instance, the ITV *'look for'* has an intensional object position which is annotated using the `ARG1` of the predicate sense `look-01`. However, the same numbered argument is used to annotate the non-intensional *'look at'*, as seen in the description tag of its PropBank argument structure (17).

(17)
```
look.01
<role descr="thing looked at
 or for or on" f="gol" n="1">
```

3.5 Parsing Experiment

After evaluating the introduced annotator, the novel corpus is generated using the annotator. A parsing experiment is done to evaluate the effects of the enriched structures on parsing performance. In this section, I report the results when I train the AMR parser using the enriched structures.

3.5.1 Method

A parser from a NLP framework called ELIT [38] is adopted for the training. ELIT uses a graph sequence transduction decoder for the AMR graphs [37]. The data used for training and testing is the enriched AMR (we refer to it as AMR+) corpus generated from the AMR 3.0 corpus using the annotator introduced in section 3.2. Table 3.5 shows the data size and the count of each enriched feature. The model uses the same train-test split corpus

as the AMR 3.0 data. It is trained and evaluated using a single GPU for 30 epochs, and takes around 10 hours to complete. Each parser is trained for three times and the average performance is reported.

Set	Plural	Article	Content	Quantifier	Graphs
TRN	41,925	51,011	30,919	5,703	55627
DEV	2,222	2178	1,045	124	1,722
TST	2,004	2137	1,064	171	1,898
Total	46,151	55,326	33,028	5,998	59,247

Table 3.5: Number of semantic features in the enriched AMR corpus converted using the annotator.

3.5.2 Results

The parser is measured using Smatch [24], a standard measurement for AMR graphs. Table 3.6 shows the average and standard deviations on the original AMR 3.0 and enriched AMR corpora by running the ELIT parser [38]. The AMR+ model is trained with all four features added. In addition, a separate model is trained for each enrichment (article, plural, intensionality, and quantifier) to examine their performance separately. In addition to the overall Smatch score, table 3.6 also presents the fine-grained evaluation of the parsing performance.

	Smatch		**Fine-grained Evaluation**					
Dataset	**Labeled**	**Unlabeled**	**No WSD**	**Named Ent.**	**Negation**	**Concepts**	**Reentrencies**	**SRL**
AMR 3.0	82.5±0.2	**85.1**±0.2	82.9±0.1	**88.0**±0.3	71.7±0.7	89.7±0.2	70.5±0.1	78.6±0.2
AMR+ (All)	81.7±0.2	84.5±0.1	82.3±0.2	87.6±0.5	70.5±0.9	89.6±0.2	70.0±0.3	77.5±0.3
AMR+ (Article)	82.1±0.1	84.7±0.2	82.6±0.1	87.1±0.4	72.2±0.5	89.5±0.2	70.2±0.3	78.3±0.6
AMR+ (Plural)	**82.7**±0.2	**85.1**±0.2	**83.2**±0.2	87.3±0.3	71.7±0.6	89.7±0.1	**70.7**±0.2	**78.8**±0.3
AMR+ (Intensionality)	81.9±0.2	84.6±0.2	82.4±0.1	87.5±0.2	71.7±0.5	**89.8**±0.2	70.2±0.3	77.5±0.2
AMR+ (Quantifier)	82.1±0.2	84.7±0.2	82.7±0.2	87.3±0.3	**72.6**±0.9	89.3±0.2	70.3±0.1	78.4±0.4

Table 3.6: Parsing performance for AMR 3.0 and enriched AMR corpora.

Comparing the performance of AMR 3.0 and the AMR+, the overall AMR+ model with all four enriched features is less than 1% lower in Smatch score. The models for each feature display a variation in performance. Among all, the model for plurality performs the best. It outperforms or performs equally well as the AMR 3.0 model in the overall and many fine-grained evaluations like reentrencies and semantic role labeling (SRL). It shows that adding

plural information could improve the parsing results. For article, content, and quantifiers, the models are around 0.5% lower in overall Smatch score, which is not a significant drop. The slight decrease in parsing performance is understandable as the complexity of graph representation increases.

3.5.3 Analysis of Errors

There could be two major possible sources of error for the drop in parsing performance: 1) there are parsing errors in the enriched structures, and 2) the enriched structures cause errors in the unmodified structures. The first possibility shows that the change in performance is entirely due to the increased complexity. The second possibility would indicate that the augmented structures would influence the original structures as a side effect. While the first possibility is more desirable, we cannot rule out the possibility for the second. Thus, to find out the source of error, I compare the parsing performance on the unmodified structures using two approaches.

In the first approach, the enriched structures of all the predicted graphs are converted back to the original form. Table 3.7 shows the parser performance on the graphs after removing the enriched AMR structures. Plural increases the performance on original AMR structures by 0.3% and adds another 0.2% when plural information is attached. The results show that except for plurality, all other models perform slightly lower in parsing the original structures as well. The decrease in agreement, however, is less than 0.2%, which is relatively minor considering the potential fluctuation in the training process. Notably, a deficiency of this analysis is that additional errors are created when converting the enriched AMRs to the unmodified form. For example, while the annotator uses constituency parser to determine which numbered-argument role to convert to the enriched `:content` role, it is extremely hard to determine the numbered-argument role in the reverse direction. Thus, the present analysis suggests that while the enriched structures might affect the original structures, it would affect them only slightly.

Dataset	Smatch	Smatch (remove enrichment)
AMR3.0	82.5	
AMR+	**82.1**	82.0
AMR+ (Article)	**82.5**	82.4
AMR+ (Plural)	**83.0**	82.8
AMR+ (Intensionality)	82.2	**82.4**
AMR+ (Quantifier)	**82.5**	82.4

Table 3.7: Parsing performance after removing enriched structures.

Another approach is used for the analysis of source of error. Instead of converting all the predicted graphs back to the original AMR structures, this approach instead evaluate the predicted graphs that do not contain any enriched structures. In this analysis, the graphs predicted by the parser are separated into two sets and only those without the enriched structures are evaluated. This approach thus avoids errors in intermediate steps like removing the enriched structures in the first approach. The analysis results are presented in Table 3.8. The same set of graphs (n) is extracted from the predicted graphs of the AMR 3.0 parser as controls. The results show that the enriched AMR parsers outperform the AMR3.0 parser on predicting the original graph structure in all cases except for intensionality. In particular, the parser containing all the enrichment features improves the parsing of original structures by 1.2%.

Certainly, the set of graphs without enriched structures might have different characteristics than those with enriched structures, so we could not extend the conclusion to the entire dataset. However, it demonstrates the potential of the enriched AMR in improving the parsing performance.

Dataset	n	AMR3	AMR+
AMR+	396	84.2	**85.4**
AMR+ (Article)	803	83.4	**83.8**
AMR+ (Plural)	857	82.4	**82.6**
AMR+ (Intensionality)	804	**82.7**	82.0

Table 3.8: Parsing performance for graphs without the enriched structures. n is the number of such graphs in the test dataset with 1898 graphs.

The performance drops for the intensionality parser. A manual check of 50 randomly sam-

pled graphs shows that almost all disagreements in graph structures are due to either flipped or more compact argument structure. For example, instead of using the normal numbered-arguments, the intensionality parser tends to use the `:opX` option role. The might be due to the enriched `:content` role which shaded the numbered-argument structure. The similar error happens in the graphs with enriched structures as well. Surprisingly, the parser was able to predict the enriched structures correctly across all the parsers except for quantifier.

A closer inspection to the quantifier model shows that it does not predict the augmented structures at all. Suspecting that this error is due to the relatively low frequency of instances of quantifiers (table 3.5), we train an over-fitted model using only graphs containing quantifier information.

The Smatch score for this model is only 56.6, ruling out the possibility that the error is caused by small sample size. The error, then, originates from the parser. While we hypothesize that the parser fails to learn the converted quantifier structure because it needs specification in constant variables, there is no special treatment for the `:quant` role or constant in the algorithm. The reason that cause this error is under investigation and will be presented in future research.

3.6 Discussion

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

Even in the absence of further improvement, our automatic annotator could already be used to produce a large number of graphs, given its baseline performance. Along with

quality checks 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 [58, 4, 3]. Moreover, the parsing experiment shows that augmenting the graph with plural information will improve the parsing performance. Parsers trained using the enriched structures could also potentially improve the performance on the original AMR structures as well. While the enriched features of intensionality would slightly lower the parsing performance, they add important information desirable for downstream tasks like coreference and temporal relation identification.

Chapter 4

Document-level Coreference

In this chapter, I extend the enrichment of meaning representation to the document-level, with specific focus on coreference. I first present the annotation scheme with special accommodations for Reddit data in section 4.1.

4.1 Annotation Schemes

The current study consulted OntoNotes' guidelines for coreference annotation in identifying mentions and establishing the coreferential relationships, and made several adjustments to accommodate the unique properties discovered in forum discussions, namely Reddit.

Singletons: following OntoNotes, we do not allow singletons to be marked, except for two occasions, `doc-situation` and `post`, which will be introduced later in the section.

Entity: we largely follow OntoNotes for entity identification, where noun phrases and pronouns denoting entity concepts are labeled as `entity`.

Event: following OntoNotes, we annotate event concepts. Event mention identification has been a more challenging task than entity identification as it includes more diverse syntactic objects ranging from verbs to gerunds and noun phrases [53]. In hopes of differentiating entities and events conceptually, we label a mention with `event` as long as it refers to an event concept.

Generic mentions: we make a distinction between generic and specific mentions and follow PreCo's decision where generic mentions can directly coreference each other. While OntoNotes does not annotate coreference between generic mentions, we find such relation quite common in Reddit posts such as the generic *you*'s and decided to avoid losing this information.

Doc-situation and post: we add two mention types for characteristics commonly occurred in Reddit data. There are noticeable amount of cases where the entire situation described in the document is referred to, while it is hard to identify a single mention, or even an obvious set of mentions, that could well-summarize the situation described. Another domain-specific case is the use of a proximal demonstrative like *"this"* to refer to the Reddit post itself, which is uncommon in other texts.

Quantifiers and negated existentials: following the same rationale as Bamman et al. [10], we annotate all quantifiers including negations for consistency under situations like *"[No boy] took a picture of [himself]"*.

Appositive, attributive, and expletive uses: unlike OntoNotes who annotates appositive and attributive uses, we only focus on identity coreference for the current data.

Spans: unlike OntoNotes that mark the maximum span, we only annotate the syntactic heads of noun phrases. The noun phrases are

Subset: in addition to identity coreference, we mark subset-set relation of entities and events like "[the boy] got the lowest grade among [the students]" in a separate layer. Subset relations involving generics are not annotated. The subset-set relation would link the two identity coreference chains if exist, while at the same time allowing singletons to be involved.

4.2 Methods

Three text types including news, fables, and Reddit are annotated and compared. The data are collected from CNN daily mail [1], Aesops Fables [2], and Reddit posts accessed on 14th Feb 2022.

For each text type, 50 documents with length between 100 and 200 tokens ($100 < n < 200$) are selected and annotated, composing of 150 annotated documents for analysis. This document length is chosen to accommodate the relatively short fable data. The reddit data are collected from the college subreddits adn are filtered using the Profanity-Check Python library[3] to remove posts containing profanity.

Annotators are trained before entering the actual annotation. The training process involves (i) reading the annotation guideline, (ii) watching an tutorial video, (iii) completing training quizzes containing 70 quiz questions including corner cases, and (iv) three rounds of practice annotation of 5 to 10 documents each.

The 150 documents are doubly annotated by four trained annotators, including three undergraduate students and one postdoctoral researcher, all with professional training in linguistics. We use INCEpTION [45], a semantic annotation platform to conduct the training and annotation. Annotators rotate across text types in batches of 5 to avoid difference in Inter-Annotator Agreement score caused by familiarity with the guidelines and annotation tool.

[1] `https://huggingface.co/datasets/cnn_dailymail`
[2] Project Gutenberg: `https://www.gutenberg.org/cache/epub/21/pg21.txt`
[3] `https://github.com/vzhou842/profanity-check`

4.3 Results and Analysis

4.3.1 Inter-Annotator Agreement

The quality of annotation is measured through the standard evaluation metrics for coreference, including MUC [68], B^3 [9], and $CEAF_{\varphi 4}$ [54]. Table 4.1 presents the IAA scores acorss various text types.

Fables receive the highest avergae IAA of 82.45, followed by the Reddit data that obtain a score of 79.04. Surprisingly, the news data get the lowest score of only 66.29, though it is often the most commonly studied text type for coreference resolution. We noticed that this score for inter-annotator agreement is not desirable within the standard of coreference resolution across all text types. The discrepancy in annotations is partly due to difficulty in span identification, as we did not use any assisting layers with pre-identified spans. That being said, while we will work towards a better score in the future, the primary purpose of this study is to compare scores across text types.

Text type	**MUC**	**B^3**	**$CEAF_{\varphi 4}$**	**Avg.**
News	68.73	65.43	64.70	66.29
Fables	88.75	83.12	75.49	82.45
Reddit	85.29	80.21	71.61	79.04

Table 4.1: Comparison of coreference performance on different text types annotated using the same guidelines.

The unexpected score for news data is primarily due to the data quality. Currently, the CNN news data is used for it has an open copyright which enables us to make our annotated corpus publicly available. However, the data is relatively messy and contains titles and web links to other online news articles, which will affect the coherence of document, making it harder for the annotator to understand and identify the coreferential relations. Another source of error is from metonymy, which is common in this CNN news dataset. However, since metonymy is extremely rare in Reddit data, the guidelines do not provide specific instruction for it.

4.3.2 Challenges

The nature of the Reddit as a platform to share feelings and ask for help leads to relatively short post contents involving in describing situations (such as a problem to be solved) and referring back to the described situations. A Reddit post of around 100-200 tokens is often sufficient to describe one situation and the author will refer back to the entire situation to express feelings or ask for help and comments. Such reference is often vague and hard to find corresponding spans. For example, in 4.1, the first "this" refers to the whole situation causing the user stress instead of a specific event. To solve this issue, we add a new mention type called `doc-situation` for reference to vague document situations. These labels are often singletons. However, multiple occurrence of the same type can be linked by identity coreference just like standard cases of entity/event coreference.

I submitted my thesis for examination in Sep last year. The thesis was then sent to 2 external reviewers, one of whom thought my thesis was excellent. The second one ended up pulling out at the last minute so a replacement was found two months later. Also in mid-Dec a paper was published which literally destroys another paper that was published 2 years earlier, which I used as a reference and would ultimately lead to my most interesting findings. The replacement reviewer thought that the most significant bits of my analysis is garbage and recommended major revise. I've now been left to fully reanalyze half of the experimental results presented and then rewrite a quarter of the thesis, which they somehow expect me to re-submit within 8 weeks. This **doc-situation** has to be every PhD's worst nightmare. I know I'm just too stressed. Thank you all so much for reading this **post**.

Figure 4.1: Use of Document Situation and Post.

Another challenge specific to forum discussions is frequent reference to the post itself. While it is clear that there is a self-reference to document, there is no span in the text for annotators to establish coreference, shown in 4.1. Similar to `doc-situation`, we introduced a standalone mention type called `post` for mentions referring to the document itself.

The language used in Reddit posts are mostly informal and colloquial. Most of the existing gold-standard coreference data focus on well-edited text types such as newspapers and novels, which typically consist of professionally written and carefully edited text. Internet abbreviations such as "idk" for "I don't know" make it impossible to mark the first-person pronoun inside the abbreviation alone using token-based mention identification. Similarly,

missing punctuation, such as writing "I'll" as "Ill", can potentially lead to incorrect span identification.

Besides grammatical errors and abbreviations, another problem of Reddit data is the absence or unclear use of quotation marks. Direct quotation often involves a change in perspective and needs to be dealt with carefully for coreference. The misuse of quotation mark would cause problem for annotators to identify coreference correctly. Typically, it is often possible for annotators to resolve coreference based on the context. However, absence of the clear cue provided by quotation marks will make machine labelling tasks much harder.

4.3.3 Analysis

Table 4.2 shows the count of the mention types in the three text types. While all three types of data contain mostly entities, fables has the highest entity frequency and more mentions per document. This is explained by the text type's story-telling purpose that tends to frequently refer to the characters. News articles has more event labels than the other text types, which aligns with our intuition that news tend to refer to the news events happened in the past. Reddit post data contains a more dispersed mention types than the other two text types. In particular, Reddit presents more generic terms such as the generic *you* and *people* in general to describe certain situations.

Among the two labels deliberately introduced to accommodate to the Reddit data, `doc-situation` is only presented in Reddit texts. While news and fables might also describe a situation, they do not have much demand to refer back to that situation, or have certain mention that could represent the situation well due to the well-edited texts. Surprisingly,the `post` label is presented in both the Reddit and news data. In news, the mentions labeled as `post` are all of the form of a noun phrase: 'this report', which refers to the news document itself. In Reddit, the form of the `post` mentions are predominantly pronouns like 'this' or 'it', and occasionally noun phrases like 'this post'.

Table 4.3 shows the count and frequency of pronouns across text types. The coreference

	Reddit			News			Fables		
Category	n	**Frequency**	**Mean**	n	**Frequency**	**Mean**	n	**Frequency**	**Mean**
`entity`	828	88.2%	16.56	867	89.2%	17.34	1197	94.9%	23.94
`event`	62	6.6%	1.24	87	9.0%	1.74	36	2.9%	0.72
`generic`	32	3.4%	0.64	10	1.0%	0.20	25	2.0%	0.50
`doc-situation`	11	1.2%	0.22	0	0.00	0.0%	0	0.0%	0.00
`post`	5	0.5%	0.10	7	0.7%	0.14	1	0.1%	0.02
Total	938	100%	18.76	971	100%	19.42	1259	100%	25.18

Table 4.2: Count of mention types in different text types.

mentions are predominantly pronouns in Reddit (67.2%) and fables (58.8%), whereas this category is only a small fraction in news (14.2%). This sheer discrepancy indicates the different strategy and focus when annotating different text types. Moreover, almost half of the mentions (47.1%) identified in the Reddit data are first person pronouns and mentioned nearly 9 times per document on average.

	Reddit			News			Fables		
Category	n	**Frequency**	**Mean**	n	**Frequency**	**Mean**	n	**Frequency**	**Mean**
pronouns	631	67.2%	12.62	138	14.2%	2.76	741	58.8%	14.82
1st person pronouns	442	47.1%	8.84	9	0.9%	0.18	151	12.0%	3.02
Total	938	100%	18.76	971	100%	19.42	1259	100%	25.18

Table 4.3: Count of pronouns in different text types.

Table 4.4 shows the average number of coreference chains and their length. While news has the most number of chains per document (6.62), the average coreference chain length is around 3 mentions. The short chain length could justify why news has the largest distance of a mention to its nearest antecedent (31.0 tokens). Fables has the longest chains that reach an average of 5.71 mentions per chain and around 4 to 5 chains per document. Despite the short chain in news, each mention contains more token on average than Reddit and fables. Reddit has the shortest mentions, which is justified by the large proportion of first person pronouns that are only 1 token in length.

Given the strikingly large proportion of first person singular pronouns in Reddit data, we explore the coreference chain data for the first person singular pronouns as well. Identity chains where more than 75% of the mentions are first person singular pronouns are considered as first person pronoun chains and includes in this statistics. In Reddit, coreference chains

of first person singular pronouns is over two times longer than the average chain length and occurs around once per document. The median distance to antecedent for 1st person pronouns is smaller and indicates that the mentions are roughly evenly distributed across the document. In news and fables, there is only 1 out of the 50 annotated documents that contains a first person pronoun chain. This discovery coincides with our common knowledge that Reddit data are closer to the language used during natural communications, which is often conducted from a first person perspective.

	Reddit		**News**		**Fables**	
Task (Avg. #)	**Overall**	**1st PP**	**Overall**	**1st PP**	**Overall**	**1st PP**
chain per document	4.30	0.94	6.62	0.02	4.42	0.02
mention per chain	4.37	9.40	2.94	9.00	5.71	9.00
token per mention	1.25	1.00	1.85	1.00	1.43	1.00
distance to antecedent (median)	14.78	12.31	31.00	11.50	14.33	19.00

Table 4.4: Statistics on coreference annotation in different text types. Column '1st PP' shows statistics for first person pronouns.

Chapter 5

Conclusion

The current thesis aims to enrich meaning representations at both the sentence- and the document-level. At the sentence-level, this study starts from an existing meaning representation frame called AMR. Recent work on improving the representational adequacy of AMR has focused on enriching its graph structure. This paper presented an automatic annotation tool 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. The task is often non-trivial due to the abstract, or unanchored, 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 the automatic annotator, 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 in the form of F1 scores for specific labels, showing a high rate of agreement. Secondly, we compared the output of our automated annotator to gold-standard manual annotations. While the F1 scores of the automatic annotator were lower than the human annotators, they were nonetheless quite high, and in the case of intensional

arguments—the hardest classification task—the scores were as good as those of manually produced annotations, showing that the rule-based annotator was able to correctly classify the relevant features consistently.

Upon completion of the automated annotation tool, we convert the AMR 3.0 corpus using our tool and conducted a parsing experiment using the enriched AMR corpus. The result shows that the augmentation of additional grammatical and semantic features like intensionality and article would slightly lower the parsing performance, which aligns with the thesis hypothesis. Additionally, enriching the plural information will directly improve the parsing performance. Particularly, the parsing experiment demonstrates that while the proposal to attach the `.pl` marker to plural concepts is valid in improving the representational adequacy, it does not fit well to the current parser constructions. This finding provides insights to the various theoretical enrichment proposals in the importance of balancing between representational adequacy and implementation feasibility. I hope that the present study encourages further efforts to automatically augment existing AMR corpora, with the aim of producing large corpora of representationally adequate AMR that could be used for model training. [1]

At the document-level, while previous works focus mainly on extending coreference annotation to different text types and domains, not many compared the text type effects horizontally. This study presented a comparison of coreference annotation across different text types including news, fables, and Reddit. We conducted an annotation experiment where 50 documents are doubly annotated for each text types. We found that there is a difference in the IAA for different text types even though they are annotated using the same scheme. In addition, the three text types show a difference in language use and content. Several notable characteristics are that there are more event coreference relations in news data and more pronouns, especially first person pronouns, in Reddit data. The results verify my hypothesis that different text types will perform differently when using the same guideline. Therefore, text type-specific adjustments need to be made in order to reach higher agreement.

[1] All code for this paper is publicly available on the GitHub repository at `https://github.com/emorynlp/eAMR`.

The enriched graph structure enables a more adequate and informative meaning representation at the sentence-level. The comparative study of text type effects in coreference provides insights to developing more comprehensive and versatile guidelines in aiding document-level understanding of language. In future studies, the enrichment in graph structures could be incorporated into document-level interpretations, which would benefit the natural language understanding tasks as a whole.

Bibliography

[1] Omri Abend and Ari Rappoport. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, 2013.

[2] Omri Abend and Ari Rappoport. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, 2017.

[3] Lasha Abzianidze and Johan Bos. Thirty musts for meaning banking. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 15–27, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3302. URL `https://aclanthology.org/W19-3302`.

[4] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 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 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-2039`.

[5] Galen Andrew and Bill MacCartney. Statistical resolution of scope ambiguity in natural language. *Unpublished manuscript*, 2004.

[6] Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat, and Dina Demner-Fushman. Domain adaptation of coreference resolution for radiology reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 118–121, 2012.

[7] Yoav Artzi, Kenton Lee, and Luke Zettlemoyer. 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 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1198. URL `https://aclanthology.org/D15-1198`.

[8] Ash Asudeh and Richard Crouch. Glue semantics for hpsg. In *Proceedings of the 8th international HPSG conference, Stanford, CA. CSLI Publications*, 2002.

[9] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. Citeseer.

[10] David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.6`.

[11] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://aclanthology.org/W13-2322`.

[12] Chris Barker and Chung-chieh Shan. *Continuations and natural language*, volume 53. OUP Oxford, 2014.

[13] Emily M. Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK, April 2015. Association for Computational Linguistics. URL `https://aclanthology.org/W15-0128`.

[14] Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12564–12573, 2021.

[15] Steven Bird and Edward Loper. Nltk: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, 2004.

[16] Patrick Blackburn and Johannes Bos. *Representation and inference for natural language: A first course in computational semantics*. Center for the Study of Language and Information Amsterdam, 2005.

[17] Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 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 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1266`.

[18] Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France, May 2020. European

Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.86.

[19] Johan Bos. Squib: Expressive power of Abstract Meaning Representations. *Computational Linguistics*, 42(3):527–535, September 2016. doi: 10.1162/COLI_a_00257. URL https://aclanthology.org/J16-3006.

[20] Johan Bos. 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 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.dmr-1.2.

[21] Harry Bunt. 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 2020. European Language Resources Association. ISBN 979-10-95546-48-1. URL https://aclanthology.org/2020.isa-1.1.

[22] Jan Buys and Phil Blunsom. 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 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1112. URL https://aclanthology.org/P17-1112.

[23] Deng Cai and Wai Lam. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.119. URL https://aclanthology.org/2020.acl-main.119.

[24] Shu Cai and Kevin Knight. 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 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2131.

[25] Daniel Chen, Martha Palmer, and Meagan Vigus. 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 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.law-1.4.

[26] Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1016. URL https://aclanthology.org/D18-1016.

[27] Jinho D. Choi and Gregor Williamson. Streamside: A fully-customizable open-source toolkit for efficient annotation of meaning representations, 2021.

[28] Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332, 2005.

[29] Richard Crouch and Aikaterini-Lida Kalouli. Named graphs for semantic representation. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 113–118, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2013. URL https://aclanthology.org/S18-2013.

[30] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon, 2004.

[31] Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 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 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4912.

[32] Lucia Donatelli, Nathan Schneider, William Croft, and Michael Regan. Tense and aspect semantics for sentential AMR. *Proceedings of the Society for Computation in Linguistics*, 2(1):346–348, 2019.

[33] Lydia Feng, Gregor Williamson, Han He, and Jinho D Choi. Widely interpretable semantic representation: Frameless meaning representation for broader applicability. Under Review.

[34] Michael Wayne Goodman. 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 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.35. URL https://aclanthology.org/2020.acl-demos.35.

[35] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[36] Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1108–1118, 2015.

[37] Han He and Jinho D Choi. Levi graph amr parser using heterogeneous attention.

In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 50–57, 2021.

[38] Han He, Liyan Xu, and Jinho D. Choi. ELIT: Emory Language and Information Toolkit. *arXiv*, 2109.03903, 2021. URL `https://arxiv.org/abs/2109.03903`.

[39] Derrick Higgins and Jerrold M Sadock. A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96, 2003.

[40] Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, 2006.

[41] Paul R Kingsbury and Martha Palmer. From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer, 2002.

[42] Karin Kipper. Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. Master's thesis, University of Pennsylvania, 2005.

[43] Karin Kipper, Martha Palmer, and Owen Rambow. Extending PropBank with VerbNet Semantic Predicates. In *Proceedings of the AMTA Workshop on Applied Interlinguas*, 2002.

[44] Nikita Kitaev and Dan Klein. 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, 2018.

[45] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computa-

tional Linguistics, Juni 2018. URL http://tubiblio.ulb.tu-darmstadt.de/106270/. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

[46] Knight et al. Abstract meaning representation (amr) annotation release 1.0, 2014.

[47] Knight et al. Abstract meaning representation (amr) annotation release 2.0, 2017.

[48] Knight et al. Abstract meaning representation (amr) annotation release 3.0, 2020.

[49] Alexander Koller, Stephan Oepen, and Weiwei Sun. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-4002. URL https://aclanthology.org/P19-4002.

[50] Kenneth Lai, Lucia Donatelli, and James Pustejovsky. 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 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.dmr-1.1.

[51] Young-Suk Lee, Ramón Fernandez Astudillo, Tahira Naseem, Revanth Gangi Reddy, Radu Florian, and Salim Roukos. Pushing the limits of AMR parsing with self-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3208–3214, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.288. URL https://aclanthology.org/2020.findings-emnlp.288.

[52] Zi Lin and Nianwen Xue. Parsing meaning representations: Is easier always better? In *Proceedings of the First International Workshop on Designing Meaning Representations*,

pages 34–43, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3304. URL https://aclanthology.org/W19-3304.

[53] Jing Lu and Vincent Ng. Event coreference resolution: a survey of two decades of research. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5479–5486, 2018.

[54] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, 2005.

[55] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[56] Mehdi Manshadi, James Allen, and Mary Swift. 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 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-2025.

[57] Christian M. I. M. Matthiessen and John A Bateman. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter, 1991.

[58] Stephan Oepen and Jan Tore Lønning. Discriminant-based MRS banking. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/364_pdf.pdf.

[59] Stephan Oepen, Omri Abend, Jan Hajic, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, and Zdenka Uresova. 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 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-2001. URL https://aclanthology.org/K19-2001.

[60] Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 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 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-shared.1. URL https://aclanthology.org/2020.conll-shared.1.

[61] Palmer et al. Proposition bank (propbank) i, 2004.

[62] James Pustejovsky, Ken Lai, and Nianwen Xue. 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 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3303. URL https://aclanthology.org/W19-3303.

[63] Florian Schwarz. Intensional transitive verbs: I owe you a horse. *The Wiley Blackwell Companion to Semantics*, pages 1–33, 2020.

[64] Ziyi Shou and Fangzhen Lin. Incorporating eds graph for amr parsing. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 202–211, 2021.

[65] Prakash Srinivasan and Alexander Yates. 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, 2009.

[66] Edward Stabler. Reforming AMR. *International Conference on Formal Grammar*, pages 72–87, 2017.

[67] Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, pages 1–18, 2021.

[68] Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[69] Aaron Steven White and Kyle Rawlins. 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, 2018.

[70] Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. Lexicosyntactic inference in neural models. *arXiv preprint arXiv:1808.06232*, 2018.

[71] Gregor Williamson, Patrick Elliott, and Yuxin Ji. 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 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.law-1.17`.

[72] Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. Improving AMR parsing with sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2501–2511, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.196. URL `https://aclanthology.org/2020.emnlp-main.196`.

[73] Amir Zeldes. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.