AMR Beyond the Sentence: the Multi-sentence AMR corpus

Tim O'Gorman[1], Michael Regan[3], Kira Griffitt[2], Ulf Hermjakob[4], Kevin Knight[4], Martha Palmer[1]
[1]University of Colorado Boulder
{ogormant,mpalmer}@colorado.edu
[2]Linguistic Data Consortium
kiragrif@ldc.upenn.edu
[3]University of New Mexico
reganman@unm.edu
[4]Information Sciences Institute, University of Southern California
{ulf,knight}@isi.edu

Abstract

There are few corpora that endeavor to represent the semantic content of entire documents. We present a corpus that accomplishes one way of capturing document level semantics, by annotating coreference, implicit roles and bridging relations on top of gold Abstract Meaning Representations of sentence-level semantics. We present the methodology of developing this corpus, alongside analysis of its quality and a plausible baseline for comparison. It is hoped that this Multi-Sentence AMR corpus (MS-AMR) illustrates a feasible approach to developing rich representations of document meaning, useful for tasks such as information extraction and question answering.

1 Introduction

Although Abstract Meaning Representation (AMR) (Banarescu et al., 2013) shows promise for a range of tasks such as summarization (Liu et al., 2015; Viet et al., 2017) and information extraction (Garg et al., 2016), it is restricted to capturing the semantics of individual sentences. For many purposes, when examining the semantics of a document, one also needs access to cross-sentence information such as coreference.

We suggest that the AMR approach to semantic representation has useful characteristics for an extension to discourse-level representations. AMR represents sentence meaning in a simple, readable semantic graph, such that annotators may directly mark coreference relations upon the AMR graph itself. AMR also annotates implicit roles in some within-sentence contexts, and the PropBank predicate annotations provide a resource for extending those implicit roles annotations to a document level.

The Multi-Sentence Abstract Meaning Representation (MS-AMR) corpus is a corpus annotated on top of existing gold AMRs, extending them with this additional information. By linking those AMRs together, it presents an integrated representation of the meaning of an entire document or discourse, as the addition of the coreference, implicit role reference and bridging relations across each AMR helps to build a larger representation of the entire propositional content of the document. Because these MS-AMR representations are annotated directly onto the variables within an AMR semantic representation, it is also a different task from traditional coreference, event coreference or implicit role coreference tasks, and results in a fundamentally different kind of data. We present a baseline system and inter-annotator agreement scores, which we hope will illuminate the nature and quality of the dataset, and outline methods for how to score MS-AMR system outputs.

2 Annotation Methodology

2.1 Background: Within-sentence Abstract Meaning Representation

AMRs are directed acyclic graph representations of sentence meaning (Banarescu et al., 2013), designed to capture the important meaning elements of a sentence while abstracting away from syntactic details

Proceedings of the 27th International Conference on Computational Linguistics, pages 3693–3702
Santa Fe, New Mexico, USA, August 20-26, 2018.

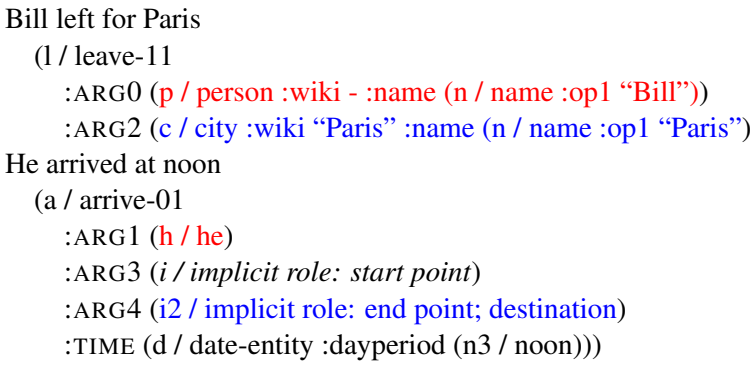

Figure 1: Example of MS-AMR annotation; annotators link coreferent variables (such as marking a relation between between *p* and *h* (in red)) and implicit roles, here linking the destination (arg4) in the second sentence to the previous variable *c* (in blue)

of how that meaning was expressed. AMRs capture basic representations of semantic roles (using the numbered arguments of PropBank (Palmer et al., 2005)), as well as within-sentence coreference, named entity and entity linking information. Thus, a simple sentence such as "Bill left for Paris" can be represented as in the first AMR in Figure 1, with "leave-11" denoting the physical departure sense of "leave", numbered arguments such as "arg2" denoting semantic roles that are unique to that sense (e.g. for leave-11, arg2 is the destination), and "Bill" and "Paris" both having named entity labels (person and city) as well as links to Wikipedia when possible. One can therefore think of an AMR as a graph of the meaning of a sentence, with "variables" such as *l*, *p* and *c* being nodes in that graph, and "/" denoting an "instance-of" relation, showing that *p* is a thing of type *person* and that *c* is a thing of type *city*.

2.2 Annotating Multi-sentence Coreference Over AMRs

Annotating multi-sentence AMR has fundamentally different characteristics from coreference annotation over surface forms. Annotators start with gold AMR annotations (using the upcoming AMR public release), and add coreference relations (and related annotations such as bridging and implicit roles) as a layer on top of those gold AMRs. Therefore, rather than annotating spans, annotators label clusters of variables that refer to the same entity or event over the AMRs constituting a document. Figure 1 illustrates a basic example of coreference in MS-AMR, showing both coreference and an example of implicit role coreference.

This annotation is structured so that alongside the annotation of explicit coreference information, annotators can also capture implicit role information. Because within-sentence AMR has all predicates annotated with PropBank senses, we have access to the lexicon with a list of all numbered arguments we might expect for that predicate. We can therefore produce a list of the numbered arguments that are not explicitly filled in the text and show these unfilled roles to annotators. These unfilled roles are temporarily added during annotation – as is illustrated with the arg3 and arg4 of the predicate "arrive-01" in Figure 1 – and can be added to coreference clusters just like any other variable. This is similar to prior works in implicit role annotation (Gerber and Chai, 2010) in that we are using semantic role inventories to prompt annotators with possible implicit roles, while adding the innovation of fitting this within a coreference task. However, while previous annotations prompted annotators with an implicit role and asked them to look through prior text for its referents, this annotation fits implicit roles into the task of coreference labeling. An example of the actual act of annotation with the Anafora toolkit(Chen and Styler, 2013) for these additional implicit role options can be seen in Figure 2.

This annotation also labels some examples of "bridging" coreference relations (Clark, 1977; Poesio et al., 1997). We annotate two more common bridging relations, part/whole relations (as in example 1) and set/member relations (as in example 2), with a focus upon those between named entities and common nouns.

1. *I think this shows that pretty much every President can do any design thing they want with both* **the**

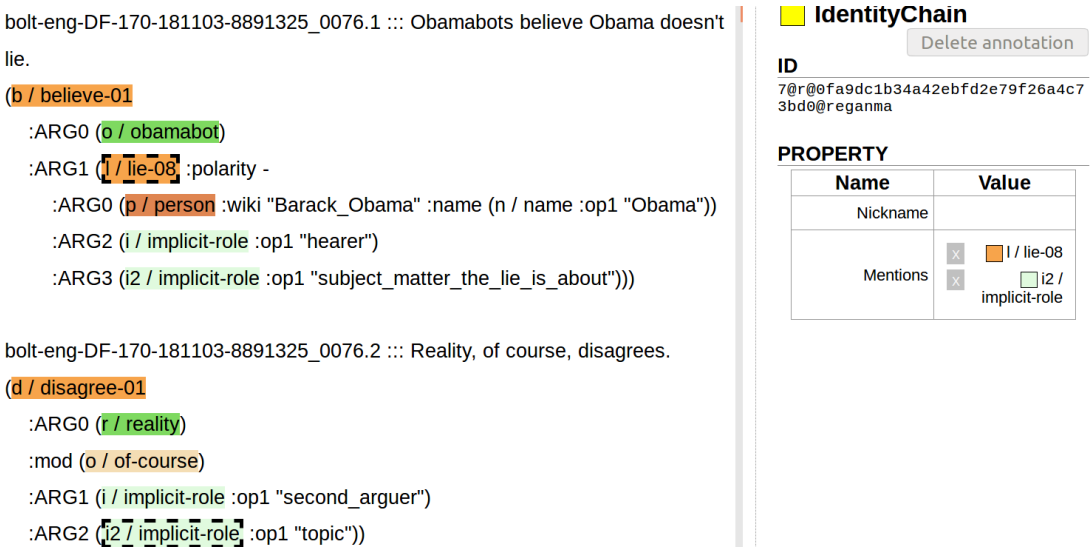

Figure 2: Annotation interface, illustrating implicit role links. Annotators click on boxes within the AMR (left) to add them to coreference chains (full chains shown on the right), as with the link between the implicit topic ("i2") and the earlier "l / lie" mention.

Residence *and* **the Office Wings** (both given part/whole relation to "the White House")

2. *I also liked* **"Deception point"**. *So I have read* **all four of his books** *and enjoyed them.* (set/member relation)

Those examples illustrate the emphasis upon capturing part/whole and set/member relations that require contextual understanding; annotators were instructed not to link part/whole relations that are only knowable through world knowledge, specifically those between named, wikified entities (such as knowing that Damascus is part of Syria).

This annotation also captures more event coreference phenomena than what is captured in OntoNotes-style coreference annotations (Pradhan et al., 2011). While those prior annotations focused upon nominal coreference, capturing verbal mentions only occasionally (when they were coreferent with a nominal mention), multi-sentence AMR annotators were instructed to link together coreferent variables regardless of their part of speech. Furthermore, because of the AMR normalization of surface-form variation, complex details regarding how to represent an event (such as the span to use for light verbs) is already normalized into single PropBank rolesets during AMR annotation. Annotation guidelines are publicly available at `https://github.com/timjogorman/Multisentence-AMR-guidelines/`.

2.3 Annotation Toolkit and Pipeline

MS-AMRs are annotated using the Anafora toolkit (Chen and Styler, 2013), a web-based system designed for coreference and temporal annotation. An example of the actual annotation interface is shown in Figure 2. As MS-AMR representations provide an annotation layer on top of gold AMRs – rather than as a change to the format of the AMRs themselves – the output of this annotation does not modify the AMRs. THis results in a set of stand-off annotations linking to each variable within a given AMR. We assert this is a useful characteristic of any modifications of the AMR corpus, as it allows the AMRs to remain compatible with existing parsing or generation systems.

Some coreference information was already provided during within-sentence AMR annotation, as AMR annotations are annotated with links between named entities and the relevant Wikipedia ID. This gold "wikification" information was provided as pre-annotations, and allowed annotators to focus upon more difficult coreference links involving pronouns, common nouns, verbs and implicit roles.

Quality control tested for consistency and coverage of the data. First and second person pronouns were checked against speaker metadata to confirm that annotators were properly keeping track of discourse

	Train	Test	Double Annotation
Files	284	9	43
AMRs	7826	201	588
Tokens	122000	3700	8200
Coreference Chains	3810	87	381
Implicit Roles	2386	67	371
Bridging Relations	1792	54	160

Table 1: Basic statistics about size of the MS-AMR corpus

participants, and certain highly anaphoric elements (such as "he" and "she") flagged whenever not annotated as anaphoric. As the AMR corpus was also being corrected and revised during this annotation to improve predicate coverage and treatment of comparatives (Bonial et al., 2018), annotations were stored with their concept labels and double-checked for changes in the underlying AMR.

Final data (included in an upcoming AMR public release) includes a description of each AMR document (as defined by LDC segmentations of multi-thread discourse into tractable discussions), defined as an ordered list of AMRs identified by their IDs. Each coreference cluster identifies explicit mentions by the ID within the normal AMR release and the variable within that AMR; implicit roles are identified by the identity of the predicate they are an argument of, with a label for the numbered argument that is implicit.

2.4 Corpus Profile

Multi-Sentence AMR was annotated over previously annotated gold AMRs of colloquial written English (from multi-post discussion forums and web blogs), filtered for discussions with fewer than 100 sentences. The corpus is composed of 293 annotated documents in total, with an average of 27.4 AMRs (429 words) per document, covering roughly 10% of the total AMR corpus. Counts for different types of mentions are listed in Table 1.

A small test split is also defined, annotated over documents from the test split of the AMR corpus. This is a small test set for systems to report preliminary results; it is hoped that this can result in shared tasks that might test against larger sets of unseen data. Annotation of this corpus is complete, and will be released in the next public release of AMR data through the Linguistic Data Consortium.

2.5 Related Work

Densely annotated datasets in which semantic data and coreference have been represented in the multiple layers of the OntoNotes corpus (Pradhan et al., 2011) and the Prague Czech-English Dependency Treebank (PCEDT) (Čmejrek et al., 2004). This MS-AMR data differs primarily in that the data is more directly joined into a single set of connected graphs, rather than many different layers of annotation. Annotations such as ACE and ERE also capture roles and entity annotations alongside coreference, but only cover a small portion of the total semantics of a document, filtering only the elements relevant to a task-specific ontology (Song et al., 2015; Bies et al., 2016).

Li Song and Xue (2018) annotates dropped pronouns over Chinese AMR annotators, but only deals with implicit roles in specific constructions. Annotations independent of AMR have provided implicit role annotations with PropBank or NomBank(Palmer et al., 2005; Meyers et al., 2004) semantic roles, but both resources were limited to a small inventory of 5-10 predicate types, rather than all implicit arguments in a text(Gerber and Chai, 2012; Moor et al., 2013).

Bridging information has also been annotated in a range of recent corpora (Nedoluzhko et al., 2009; Poesio et al., 2008; O'Gorman et al., 2016; Roesiger, 2018) and resulted in systems (Poesio et al., 2004) capable of predicting such relations. Some bridging relations were also captured by within-sentence AMR annotations, encoded by the "include-91" relation.

While MS-AMR focuses upon capturing the *propositional* content of a document, it does not capture other dimensions of discourse annotation, such as rhetorical structure. Corpora such as RST (Carlson

et al., 2003), SDRT (Baldridge et al., 2007), GraphBank (Wolf et al., 2004) or PDTB (Miltsakaki et al.,) therefore capture dimensions of meaning that differ from the propositional content captured in this corpus.

3 Measuring MS-AMR Corpus Quality

For a subset of the training data, we took each document (i.e. the sequence of AMRs that reflect a document) and double-annotated it, to measure inter-annotator agreement and check for persistent errors. These additional annotations are listed in Table 1 under the "Double Annotation" column, and provided in the release.

3.1 Coreference Chain Quality

When dealing with two MS-AMR annotations done over the same set of gold AMRs – as with inter-annotator agreement data here – we can treat a given variable in each AMR as a possible "mention", just as one might treat a span of text in a document. With that assumption, we may therefore measure MS-AMR IAA coreference scores using standard means of measuring coreference quality. One current approach – used because it is the standard for scoring coreference systems – is to use an average of the BCUB (Bagga and Baldwin, 1998), MUC (Vilain et al., 1995) and CEAF-E (Luo, 2005) metrics, referred to as the "CoNLL-2012 F1 score" (calculated using the reference implementation of Pradhan et al (2014a)).

Under those assumptions, the annotations get a CoNLL-2012 F1 of 69.86, using the reference implementation for these scores. For comparison, more traditional span-based coreference annotations (O'Gorman et al., 2016) found inter-annotator agreement of 65.5 for event F1, and 70.4 for entity F1 (CoNLL F1 score). This is therefore in the rough range of where one might expect human performance to be. However, this does not have actual comparability to other annotation schemes, as this data uses both event and entity coreference, and does not encompass within-sentence coreference (which is already provided in gold AMR annotations).

3.2 Agreement when underlying AMRs are not identical

Because MS-AMR was annotated on top of gold AMRs, most inter-annotator agreement pairs were annotated over an identical set of within-sentence AMRs for the document. This has the advantage of separating multi-sentence AMR disagreements from other AMR disagreements, but it left open the question of whether that agreement would change drastically if the underlying AMRs were different. Starting with documents where two separate annotators provided different AMRs for each sentence, we separately annotated two multi-sentence AMR annotations for that same document, each annotated on top of a different set of AMRs.

The challenges in evaluating this kind of agreement presage the challenges of measuring the quality of system-predicted MS-AMR. In traditional annotations of coreference over surface forms, one can assume that two mentions are identical when they refer to the same span of text. However, with two separate AMRs, one cannot directly infer whether two mentions reference the same span, but must determine a mapping between the variables of each AMR. The SMATCH (Cai and Knight, 2013) algorithm, used for evaluating AMRs, calculates such a mapping. Using SMATCH to align those AMRs and coreference scores from the reference implementation of the scoring metrics (Pradhan et al., 2014b), we find the CoNLL-2012 F1 to be 66.72. Such a score is limited in that it was measured over a very small exploration set (40 AMRs), but it provides some preliminary suggestion that the identical underlying AMRs used in the IAA numbers above are not dramatically inflating the inter-annotator agreement.

3.3 Implicit Role Agreement

Implicit roles have been annotated in prior corpora, but those annotations have been limited in size or coverage (Ruppenhofer et al., 2010; Gerber and Chai, 2010). The most comparable prior annotation is that of Gerber and Chai (2010), which looked at 10 nominal predicates in WSJ data, presenting annotators with numbered arguments without explicit mentions and instructions to link these arguments to

Relation Type	Resource	Metric	Score
Coreference	This work	CoNLL-F1	69.9
	RED-Entities (O'Gorman et al. 2016)	CoNLL-F1	70.4
	RED-Events (O'Gorman et al. 2016)	CoNLL-F1	65.5
Implicit Role	This work	Cohen's κ	0.59
	Gerber and Chai (2010)	Cohen's κ	0.64

Table 2: Agreement, alongside agreement in similar corpora (not perfectly comparable)

previously mentioned terms. That annotation resulted in a Cohen's kappa (Cohen, 1960) of κ=0.64. The current annotations result in a lower kappa of κ=0.59, which is a testament to the difficulty of the task.

As found in Gerber and Chai (2010), the bulk of this type of error is in the task of discerning whether a given referent is implicit at all. When both annotators agreed that a given implicit role was present, κ=0.85.

The implicit roles that annotators disagreed on were often those whose referents could be construed either as a specific referent in context or as a generic reference, most commonly with the non-focused element in a communication or mental state verb – such as the person being interested in "that's interesting", or the cause of "I laughed out loud". A similar issue involved the recipient or listener role for verbs like "say-01" or "ask-01", which can sometimes be inferable from context but are low in prominence.

4 Measuring MS-AMR similarity — scoring system performance

4.1 A baseline implementation

We present a simple baseline that hints at a lower bound for the task. We use the publicly available version of the Brandeis transition-based AMR parser (Wang et al., 2016) combined with an off-the-shelf coreference system (Clark and Manning, 2016) using an AMR-to-surface-form aligner (Flanigan et al., 2014) to convert the surface coreference to links between AMR nodes.

4.2 Simple evaluation of system prediction

As mentioned in section 3.2, one hurdle in evaluating system predictions comes from the absence of clearly alignable "mentions" for use in coreference metrics. We can use the SMATCH metric on each individual sentence within an AMR document to resolve this, as calculating a SMATCH score involves determining the highest scoring alignment between variables within a system prediction and a gold AMR. We can then score against that mapping.

The simple baseline outlined above was evaluated according to this metric, and we found a 27.79 CoNLL-F1 average (evaluating against the inter-annotator agreement data, which we use as a development set). While some amount of this may reflect simple differences such as the shift in domain to discussion forum data, it also underlines the inherent challenge of the task.

4.3 Evaluation with Document-level SMATCH

One hope with MS-AMR data is to encourage people to produce systems that go from strings to a representation of the meaning of a document. We therefore may want to have metrics that do not simply score coreference, but which score the entire quality of a resultant knowledge graph. One method to accomplish this goal is to use these multi-sentence AMR annotations to combine all the AMRs of a document into a single large "document graph", and to score entire document graphs using the SMATCH metric.

We follow prior work in abstractive summarization (Liu et al., 2015), which originally outlined simple methodologies for combining AMRs into a larger AMR for a document. This fundamentally involves a combination of two things – concatenating all sentences together under a new root node, and merging variables that are coreferent into a single variable.

Earlier work on this (Liu et al., 2015) merged only identical named entities and date-entity subgraphs, and therefore did not run into the complexities in terms of merging documents with similar

Bill left for Paris
(l / leave-11
:ARG0 (p / person :wiki - :name (n / name :op1 “Bill”))
:ARG2 (c / city :wiki “Paris” :name (n / name :op1 “Paris”)
He arrived at noon
(a / arrive-01
:ARG1 (h / he)
:ARG4 (i2 / implicit role: end point; destination)
:TIME (d / date-entity :dayperiod (n3 / noon)))

Merged form:
(m / multi-sentence)
:snt1 (l / leave-11
:ARG0 (p / person :wiki - :name (n / name :op1 “Bill”))
:instance-of “he”
:ARG2 (c / city :wiki “Paris” :name (n / name :op1 “Paris”)
:snt2 (a / arrive-01
ARG1 p
ARG4 c
:TIME (d / date-entity :dayperiod (n3 / noon)))

Figure 3: Example of merging multiple documents into a single AMR.

but non-identical information. When two triples are identical – such as twice labeling “Hillary Clinton as “:instance-of person”, or adding multiple “:wiki” links to “Hillary_Clinton” – we merge those redundant bits of information. However, if a variable has different concepts or relations (e.g. “person” in one AMR and “woman” in another), the additional concepts are added as additional “instance-of” relations to the resultant merged entity. Figure 3 illustrates how the merging of AMRs is actually enacted into an output AMR.

We then score these two document graphs using the SMATCH metric, originally proposed to score single-sentence AMRs. SMATCH scores each relation within an AMR graph, as measured as a triple containing the variable, the relation, and the variable or constant that is being linked to. These triples can reflect an actual relation, such as saying that there is a TIME relation between variable *a* and variable *d*, or can express the “instance-of” relations captured by the “/” in AMR, such as *a* being an instance-of “arrive-01”. The SMATCH metric calculates an F1 measuring how many of the triples within the gold AMR correspond to a triple in the system AMR. The issue is that AMR variable names are somewhat arbitrary — “a” and “d” in one AMR do not necessarily correspond to “a” and “d” in another AMR — and therefore one needs to calculate how to map the variables within another AMR onto the variables within the gold AMR; Cai and Knight (2013) introduced a hill-climbing method allowing one to calculate the alignment that gives the highest F1 over these triples. Importantly for multi-sentence AMR, we can utilize the same metric over a much larger AMR graph, although this becomes computationally expensive with larger document sizes.

We therefore score these document AMRs by treating them as single AMRs to be scored using SMATCH. As such, this is an evaluation of both the quality of within-sentence AMRs and coreference information. Table 3 illustrates the performance of the baseline system, compared with inter-annotator agreement scores. Following the suggestion to approach coreference evaluation based on how it works on edge cases (such as leaving all mentions as singletons) (Recasens and Hovy, 2011), we also report scores for the document-level graphs when no cross-sentence coreference information is used. The “AMR=identical” condition illustrates most instances of double annotation in our data, where both versions were annotated over the same AMRs; these receive very high SMATCH scores, due to the identical underlying AMRs. The “AMR=human” condition illustrates a small set where two different annotations of each AMR were used; this therefore represents human performance at this task. One can see that there

System type	AMR	Coreference	Double-annotation data	Test
baseline	system	none	53.0	43.2
baseline	system	system	53.6	44.0
IAA	human	none	58.0	
IAA	human	human	68.9	
IAA	identical	none	78.5	80.6
IAA	identical	system	80.1	82.9
IAA	identical	human	87.3	

Table 3: Agreement and baseline performance using SMATCH over document graphs. The AMR=human condition is evaluated on a much smaller set of AMRs where we have two annotated AMRs for each sentence.

is still a quite meaningful gap between this performance and the baseline system performance.

Notably, the simple baseline systems do not have dramatic increases over simply scoring the document SMATCH of a sequence of AMRs with no coreference. Basic analysis of this shows that some of this error is a consequence of the discussion genre: a baseline coreference model that did not track speakers in the AMRs makes important errors in coreference chains regarding "I" and "you". However, one can also see from Table 3 that the MS-AMR score is still very dependent upon the quality of the within-sentence AMRs, as one might expect from a measure of the general quality of a document representation.

5 Conclusions

We present here a new corpus of coreference, partial coreference and implicit roles on top of the Abstract Meaning Representation corpus. While this is fundamentally an extension of AMR, we frame this in relation to prior coreference work and propose a methodology for evaluating multi-sentence AMR system predictions against these gold annotations. Such an annotation does not fully capture all of the information one might hope to represent about the meaning of a document, as these representations leave unmarked a great deal of information about temporal and aspectual structure, discourse structure or information structure. However, this corpus illustrates a methodology of annotating new layers of meaning on top of AMR representations, which might be extended to such other representations.

Acknowledgements

This work is supported by Defense Advanced Research Projects Agency (DARPA) under DEFT-FA-8750-13-2-0045 . We would like to thank annotators at the University of Colorado – Boulder and the Linguistic Data Consortium for their diligent annotations and insightful feedback on the guidelines.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

Jason Baldridge, Nicholas Asher, and Julie Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft*, 26(2):213–239.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of Linguistic Annotation Workshop*.

Ann Bies, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, and Tim O'Gorman. 2016. A comparison of event representations in deft. In *Proceedings of the Fourth Workshop on Events*, pages 27–36.

Claire Bonial, Bianca Badaru, Kira Griffitt, Ulf Hermjakob, and Kevin Knight. 2018. Abstract meaning representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shu Cai and Kevin Knight. 2013. Smatch: an Evaluation Metric for Semantic Feature Structures. In *ACL (2)*, pages 748–752.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Wei-Te Chen and Will Styler. 2013. Anafora: a web-based general purpose annotation tool. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2013, page 14. NIH Public Access.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Herbert H Clark. 1977. *Bridging. Thinking: readings in cognitive science*. Cambridge: Cambridge University Press.

Martin Čmejrek, Jan Hajič, and Vladislav Kuboň. 2004. Prague czech-english dependency treebank: Syntactically annotated resources for machine translation. In *In Proceedings of EAMT 10th Annual Conference*. Citeseer.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1426–1436.

Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In *AAAI*, pages 2718–2726.

Matthew Gerber and Joyce Y. Chai. 2010. Beyond NomBank: A study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592.

Matthew Gerber and Joyce Y. Chai. 2012. Semantic role labeling of implicit arguments for nominal predicates. *Computational Linguistics*, 38(4):755–798.

Sijia Ge Bin Li Junsheng Zhou Weiguang Qu Li Song, Yuan Wen and Nianwen Xue. 2018. An easier and efficient framework to annotate semantic roles: Evidence from the chinese amr corpus. In Kiyoaki Shirai, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), may.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. Annotating noun argument structure for NomBank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 803–806.

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. The penn discourse treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004*.

Tatjana Moor, Michael Roth, and Annette Frank. 2013. Predicate-specific annotations for implicit role binding: Corpus annotation, data analysis and evaluation experiments. In *Proceedings of the 10th International Conference on Computational Semantics*.

Anna Nedoluzhko, Jiří Mírovskỳ, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 108–111. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

M. Palmer, P. Kingsbury, and D. Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 1–6. Association for Computational Linguistics.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014a. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the Association for Computational Linguistics*.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014b. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 30. NIH Public Access.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Ina Roesiger. 2018. Bashi: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), may.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Lai Dac Viet, Nguyen Le Minh, and Ken Satoh. 2017. Convamr: Abstract meaning representation parsing for legal document. *arXiv preprint arXiv:1711.06141*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Chuan Wang, Sameer Pradhan, Xiaoman Pan, Heng Ji, and Nianwen Xue. 2016. Camr at semeval-2016 task 8: An extended transition-based amr parser. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1173–1178.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2004. Discourse graphbank. *Linguistic Data Consortium, Philadelphia*.