Abstract Meaning Representations as Linked Data

Gully A. Burns(✉), Ulf Hermjakob, and José Luis Ambite

Information Sciences Institute, University of Southern California,
Marina del Rey, CA 90292, USA
burns@isi.edu

Abstract. The complex relationship between natural language and formal semantic representations can be investigated by the development of large, semantically-annotated corpora. The "Abstract Meaning Representation" (AMR) formulation describes the semantics of a whole sentence as a rooted, labeled graph, where nodes represent concepts/entities (such as PropBank frames and named entities) and edges represent relations between concepts (such as verb roles). AMRs have been used to annotate corpora of classic books, newstext and biomedical literature. Research on semantic parsers that generate AMRs from text is progressing rapidly. In this paper, we describe an AMR corpus as Linked Data (AMR-LD) and the techniques used to generate it (including an open-source implementation). We discuss the benefits of AMR-LD, including convenient analysis using SPARQL queries and ontology inferences enabled by embedding into the web of Linked Data, as well as the impact of semantic web representations directly derived from natural language.

Keywords: Linked linguistic data · Abstract Meaning Representation · AMR · Sembank · Biological pathways

1 Introduction

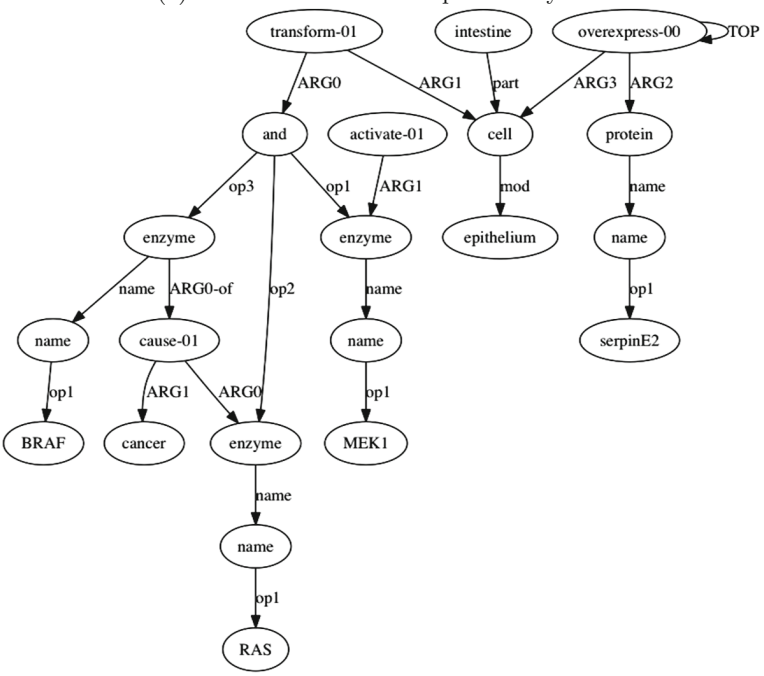
The Abstract Meaning Representation (AMR) formulation follows the success of using high-quality parallel corpora for machine translation and using Penn Treebank for statistical parsing. Such a "sembank of simple, whole sentence semantic structures" [1] seeks to accelerate natural language understanding research. AMRs abstract away from syntactic variation (so that different ways of saying the same thing map to the same AMR). Consider the text *"SerpinE2 is overexpressed in intestinal epithelial cells transformed by activated MEK1 and oncogenic RAS and BRAF"* [2]). Figure 1 shows this sentence's AMR, and Fig. 2 shows its translation into RDF (AMR-LD). Using AMR-LD with an AMR parser permits semantic representations to be generated directly from text. AMR semantic parsing is an active area of research (e.g., [3–6]).

A significant corpus of AMRs have been generated in the Natural Language Processing (NLP) community. Initially, AMRs were developed as annotations of

© Springer International Publishing AG 2016
P. Groth et al. (Eds.): ISWC 2016, Part II, LNCS 9982, pp. 12–20, 2016.
DOI: 10.1007/978-3-319-46547-0_2

```
# ::id a_pmid_2094_2929.39 ::amr-annotator SDL-AMR-09 ::preferred
# ::tok SerpinE2 is overexpressed in intestinal epithelial cells
#       transformed by activated MEK1 and oncogenic RAS and BRAF
(o / overexpress-01
     :ARG2 (p / protein
          :name (n2 / name :op1 "serpinE2"))
     :ARG3 (c / cell
          :mod (e2 / epithelium)
          :part-of (i / intestine)
          :ARG1-of (t / transform-01
               :ARG0 (a / and
                    :op1 (e3 / enzyme
                         :name (n3 / name :op1 "MEK1")
                         :ARG1-of (a2 / activate-01))
                    :op2 (e4 / enzyme
                         :name (n4 / name :op1 "RAS")
                         :ARG0-of (c2 / cause-01
                              :ARG1 (c3 / cancer)))
                    :op3 (e / enzyme
                         :name (n5 / name :op1 "BRAF")
                         :ARG0-of c2)))))).
```

(a) AMR in its native s-expression syntax



(b) ARMICA graphical rendering of the AMR

Fig. 1. Abstract meaning representation (AMR): biological pathway example

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ac: <http://amr.isi.edu/rdf/core-amr#> .
@prefix pb:
       <http://verbs.colorado.edu/propbank/framesets-english-aliases#> .
@prefix ae: <http://amr.isi.edu/entity-types#> .
@prefix at: <http://amr.isi.edu/rdf/amr-terms#> .
@prefix up: <http://www.uniprot.org/uniprot/> .
@prefix pfam: <http://pfam.xfam.org/family/> .
@prefix pm: <http://www.ncbi.nlm.nih.gov/pubmed/>
@base : <http://amr.isi.edu/amr_data/> .
@prefix : <a_pmid_2094_2929.39#> .

<a_pmid_2094_2929.39> rdf:type ac:AMR .
<a_pmid_2094_2929.39> ac:has-tokens "SerpinE2 is overexpressed in ..." .
<a_pmid_2094_2929.39> ac:has-id "pmid_1177_7939.53" .
<a_pmid_2094_2929.39> ac:has-annotator "SDL-AMR-09" .
<a_pmid_2094_2929.39> ac:is-preferred "true"^^xsd:boolean .
<a_pmid_2094_2929.39> ac:has-file "alignment-release-bio.txt" .
<a_pmid_2094_2929.39> ac:in-document pm:20942929 .
<a_pmid_2094_2929.39> amr:has-root :o .

:o rdf:type pb:overexpress-01 ;
   pb:overexpress-01.ARG2 :p ;
   pb:overexpress-01.ARG3 :c .
:p rdf:type ae:protein ;
   rdfs:label "serpinE2" ;
   ac:xref up:P07093 ;
   ac:xref up:GDN_HUMAN .
:c rdf:type ae:cell ;
   ac:mod :e2 ;
   ac:part-of :i ;
   pb:ARG1-of :t .
:e2 rdf:type at:epithelium .
:i rdf:type at:intestine .
:t rdf:type pb:transform-01 ;
   pb:transform-01.ARG0 :a .
:a rdf:type ac:and ;
   ac:op1 :e3 ;
   ac:op2 :e4 ;
   ac:op3 :e .
:e3 rdf:type ae:enzyme ;
    rdfs:label "MEK1" ;
    ac:xref up:MP2K1_HUMAN ;
    ac:xref up:Q02750 ;
    pb:activate-01.ARG1-of :a2 .
:a2 rdfs:type pb:activate-01 .
:e4 rdf:type ae:enzyme ;
    rdfs:label "RAS" ;
    ac:xref pfam:PF00071 ;
    pb:cause-01.ARG0-of :c2 .
:c2 rdf:type pb:cause-01 ;
    pb:cause-01.ARG1 :c3 .
:c3 rdf:type at:cancer .
:e rdf:type ae:enzyme ;
   rdfs:label "BRAF" ;
   pb:cause-01.ARG0-of c2 .
```

Fig. 2. AMR-LD example in Turtle syntax

the short novel "The Little Prince". Subsequent releases of core AMR data were derived from newswire and discussion forum sources. In August 2015, the AMR group released 19,572 AMRs through the Linguistic Data Consortium (LDC) licensing process (LDC2015E86: DEFT Phase 2 AMR Annotation R1). Detailed annotation guidelines and active curation tools are available from [7].

In this paper we focus on an *open* AMR corpus produced by the R2L2K2 project at USC/ISI within DARPA's Big Mechanism program [8], which is machine reading the literature on biological pathways in cancer research. In March 2016, the R2L2K2 team (with contributions from USC/ISI, University of Colorado, and SDL) released version 0.8 of the BioAMR corpus consisting of 6,452 sentences. We present this corpus as AMR-LD and describe our approach to convert AMRs into linked data, in effect, linking a rapidly expanding area in natural language research to the semantic web.

2 Linking Abstract Meaning Representation

Converting AMR resources to linked data proceeds in two steps. First, we identify and/or define ontologies that capture the semantics of the original AMR concepts and relations (including the corresponding namespaces). Second, we use entity linkage techniques [9] to map to well-known web resources.

2.1 Representing and Linking AMR Concepts and Relations

The semantics of AMRs is primarily defined by the types of nodes in the graph, which are PropBank frames [10] and AMR entity types. Each PropBank frame defines the applicable roles and their specific meaning. For example, in Fig. 1 the type of the node labeled `t` is the PropBank frame `transform-01` (defined as to "change, cause a change in state"), which has four roles: `ARG0`, the causer of transformation (Agent); `ARG1`, the thing changing (Patient); `ARG2`, the end state (Result); and `ARG3`, the start state (Material). In our example, the agent of the transformation `ARG0` is a set of enzymes (BRAF, MEK1, RAS).

We use a simple meta-model consisting of (1) a main concept class `AMR-Concept` with subclasses (`AMR-PropBank-Frame`, `AMR-Entity-Type`, and `AMR-Term`) and (2) a main relation `AMR-Role` with a subrelation (`AMR-PropBank-Role`). We also define namespaces to make these distinctions explicit, as follows:

AMR-Core (ac) includes the core constructs of the AMR specification: the `AMR` concept, metadata associated with AMRs (e.g., `has-annotator`, `in-document`), and AMR-specific roles (e.g., `mod`, `part-of`). We also define a `xref` role to link to external entities (cf. Sect. 2.2).

AMR-PropBank-Frame (pb) includes PropBank frames (e.g. `transform-01`, `activate-01`), the roles used by PropBank Frames (e.g., `ARG0, ARG1, ...`), and their inverses (e.g., `ARG0-of, ARG1-of, ...`) with corresponding inverse role assertions (e.g., `pb:ARG0-of owl:inverseOf pb:ARG0`).

AMR-Entity-Type (ae) includes all named entity types corresponding to common concepts in a domain, (`person`, `organization`, and `location` in general news text, or `enzyme` or `cell` in biomedical text).

AMR-Term (at). AMR parsing tools and human curators are free to create additional entity types, even if those types are not predefined in the AMR-Entity-Type namespace. For example, concepts, such as `cancer` and `intestine` in Fig. 2, are not registered as AMR entity types.

In our translation of AMRs to AMR-LD we closely follow the AMR design. One representational structure we deliberately altered was the naming convention used in the core AMR formulation. This involves the use of a `:name` role to create an instance of a `name` concept containing one ore more `:opN` roles that contain the string tokens of the name, for example, *e.g.*, `:name (n3 / name :op1"MEK1")`. In AMR-LD, we replace this with an standard `rdfs:label` property.

A feature of the PropBank roles `:ARG0`, `:ARG1`, `:ARG2`, `:ARG3` is that their precise semantics may change from frame to frame. Generally, `:ARG0` is the Agent, and `:ARG1` is the Patient. However, this is not always the case. The semantics of `:ARG2`, `:ARG3`, etc., is even more variable. In our presentation in the paper we will show only properties using the `:ARGN` roles. However the tool we describe in Sect. 2.3 can also generate frame-specific roles, like `transform-01.ARG0`. The rationale is to attach precise semantics to roles of different frames, as needed. For example, stating that `transform-01.ARG0` is a subproperty of `vnrole:26.6.1--Agent` role (while not all `ARG0`'s may be agents).

2.2 Representing and Linking AMR Entities

A crucial feature of the AMR-LD representation is that it explicitly links to well-known entities in the Semantic Web using the `xref` property. For example, in Figs. 1 and 2 the AMR node `p` labeled "serpinE2" corresponds to the UniProtKB protein `GDN_HUMAN` and its synonymous identifier `P07093`. Similarly the entity `e4` labeled "RAS" corresponds to entity `PF00071` in the protein family ontology [11]. We could have used `owl:sameAs` to indicate linkages. However, given the strong semantics of `owl:sameAs` and the difficulty of accurately performing entity linkage, we decided to use a more relaxed property like `ac:xref`. The `ac:in-document` property also provides links into the literature. For example, the AMR `<a_pmid_2094_2929.39>` comes from the PubMed article `pm:20942929`. These linkages embed AMR data into the Semantic Web and can significantly enhance the value of AMR corpora by leveraging existing ontologies, as well as provide an entry point into linguistic resources for semantic web applications.

We developed an entity linkage algorithm for common bioentities based on their labels. First we collected protein and chemical names from existing databases, specifically the UniProt knowledge base [12], proteins appearing in pathways in Pathway Commons [13], and chemicals from NCBI's PubChem. Then, we mapped entities appearing in BioAMRs to these resources. For short (protein) names, like "BRAF", we use a combination of string similarity metrics, such as edit distance, and Jaccard similarity over n-grams. For efficiency we include a

blocking algorithm based on prefixes of the protein names. Our implementation used the FRIL [14] record linkage system. For long (protein, chemical) names, such as "Cbl E3 ubiquitin ligase", we used traditional information retrieval techniques, such as TF-IDF cosine similarity.

2.3 AMR-LD Open-Source Conversion Software

We developed a Python library for translating the original AMR representation to RDF. The library is hosted on GitHub [15]. The tool provides extensions to connect to different record linkage algorithms. In our development of the bio AMR-LD corpus, we used the L2K2R2 project bioentity mapping web service [16]. We applied this system to the Bio-AMR v0.8 data [7] to generate the publicly available AMR-LD resource at [17]. The conversion proceeds as follows:

1. Generate URLs for AMR elements, qualified by appropriate namespaces.
2. Add RDFS classes to represent AMR Concepts, Entities, Frames and Roles.
3. Convert entity names to standard `rdfs:label` elements.
4. Define elements from the AMR base language, AMR named entity vocabulary, and PropBank frame repository.
5. Link to well-known semantic web entities using `xref` properties.

3 Querying and Reasoning with AMR-LD

An advantage of using linked data standards is that it facilitates data analysis by using existing query and reasoning engines. For example, consider the SPARQL query in Fig. 3 to identify sentences in papers that contain activated entities and their types. Note that we take advantage of path-following queries of SPARQL 1.1 using over the ac:Role predicate which is a super-property of the AMR properties. Similar queries can point to sentences containing specific proteins or chemicals. Leveraging the ontology of AMR types either directly, or through linkage to external ontologies, enables more sophisticated queries. For example, if our ontology includes an axiom that states that `ae:enzyme rdfs:subClassOf ae:protein`, then a query (similar to the one in Fig. 3(a) for sentences) can retrieve proteins of any kind. Finally, although our focus has been on Linked Data, having AMR data represented as an RDF graph can facilitate using Graph databases (e.g., neo4j) that implement more complex graph algorithms, such as shortest paths, or centrality measures.

4 Related Work

Cimiano et al. 2014 [18] describe how semantic web methods can be used in natural language interpretation systems as an implementation of discourse relation theory (DRT) [19]. This "*meaning representation*" captures the semantics of simple sentences and phrases. This work is not driven by the goal of building a sembank and is driven by data generated from syntactic parse trees. In

```
select ?pmid ?sentence ?activated ?entityType
where {?amr rdf:type ac:AMR .          ?amr :has-pmid ?pmid .
       ?amr :has-tokens ?sentence .    ?amr :root ?aroot .
       ?aroot ac:Role* ?actFrame .     ?actFrame rdf:type pb:activate-01 .
       ?actFrame pb:ARG1 ?actEntity .  ?actEntity rdfs:label ?activated .
       ?actEntity rdf:type ?entityType }
```

pmid	sentence	activated	entityType
14656721	As was the case for p38 , ERK1 @/@ 2 was both rapidly and persistently activated in neutrophils exposed to H2O2 (Fig . 2A) .	ERK12	ae:enzyme
15156153	Expression of either Hes1 or Hes5 significantly activated the STAT reporter gene construct in both E13 neuroepithelial cells and MNS @-@ 70 cells ,	STAT	ae:gene

Fig. 3. SPARQL query over AMR-LD to identify sentences in papers that contain activated entities and their types, and some results

other work, the FRED system provides a live system for Semantic Web Machine Reading with a live interface [20]. FRED uses Boxer [21] to generate a RDF representation derived from DRT. Additional components within the FRED architecture include semantic representations for sentiment, citation, and type definitions. Another important framework for semantic representation is Hobb's logical form methodology driven by abductive reasoning [22], which also makes use of the Boxer system.

5 Discussion

The development of AMR as a semantic representation of complete English sentences is blossoming. Current curation efforts provide training data for automatic AMR parsing systems and metrics for evaluating the quality of AMRs ('Smatch' scores [23], which measure overlap between AMR graphs, seen as triples, normalized to 100). Current performance of AMR parsing is $\tilde{6}7.1$ smatch, with human inter-annotator in the 79–83 range [24]. Researchers are using AMRs to build knowledge bases of biological pathways [25]. Additional linkage approaches for AMRs over news text are available [26]. Our dual goal in providing AMR as Linked Data is to empower NLP researchers with a representation naturally suited for inference and embedding in the web knowledge graph, and to provide Semantic Web researchers connections to vast text/linguistic resources.

In future work we plan to investigate improved algorithms for linking AMR data to both entities and concepts from external ontologies, leveraging our work on semantic similarity [27], and on mapping ontologies of Linked Data [28].

Acknowledgments. This work was supported by grant W911NF-14-1-0364.

References

1. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation for sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 178–186, Sofia, Bulgaria, Assoc. Computational Linguistics (2013)
2. Bergeron, S., et al.: The serine protease inhibitor serpinE2 is a novel target of ERK signaling involved in human colorectal tumorigenesis. Mol. Cancer **9**, 271 (2010)
3. ISI software page. http://www.isi.edu/natural-language/software/
4. Vanderwende, L., et al.: An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In: NAACL Demonstrations, pp. 26–30. ACL (2015). http://www.aclweb.org/anthology/N15-3006
5. Flanigan, J., et al.: Generation from abstract meaning representation using tree transducers. In: NAACL: Human Language Technologies, pp. 731–739. ACL (2016). http://www.aclweb.org/anthology/N16-1087
6. Rao, S., Vyas, Y., Daume, H., Resnick, P.: Parser for abstract meaning representation using learning to search. In: Proceedings of SemEval 2016 (2016)
7. AMR project website. http://amr.isi.edu/
8. Cohen, P.R.: DARPA's big mechanism program. Phys. Biol. **12**(4), 045008 (2015)
9. Naumann, F., Herschel, M.: An Introduction to Duplicate Detection. Morgan and Claypool Publishers, New York (2010)
10. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (2005)
11. Pfam: home page. http://pfam.xfam.org
12. UniProt home. http://www.uniprot.org/
13. Pathway commons homepage. http://www.pathwaycommons.org/
14. Jurczyk, P., Lu, J.J., Xiong, L., Cragan, J.D., Correa, A.: FRIL: a tool for comparative record linkage. AMIA Ann. Symp. Proc. **2008**, 440–444 (2008)
15. AMR-linked data github repository. https://github.com/BMKEG/amr-ld/
16. L2K2R2 bioentity mapping web service. http://dna.isi.edu:7080/grounding/
17. Burns, G., Ambite, J.L., Hermjakob, U., The AMR Development Team: Biomedical abstract meaning representation as linked data (v0.8.1). figshare (2016). https://dx.doi.org/10.6084/m9.figshare.3206062.v1
18. Cimiano, P., Unger, C., McCrae, J.P.: Ontology-Based Interpretation of Natural Language. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, New York (2014)
19. Kamp, H.: A theory of truth and semantic representation. In: Groenendijk, J., Janssen, T., Stokhof, M. (eds.) Formal Methods in the Study of Language. Mathematical Centre, Amsterdam (1981)
20. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: Teije, A., et al. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 114–129. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33876-2_12. http://wit.istc.cnr.it/stlab-tools/fred/
21. Bos, J.: Wide-coverage semantic analysis with boxer. In: Bos, J., Delmonte, R. (eds.) Semantics in Text Processing (STEP), pp. 277–286. College Publications, London (2008)
22. Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P.A.: Interpretation as abduction. Artif. Intell. **63**(1–2), 69–142 (1993)

23. Cai, S., Knight, K.: Smatch: an evaluation metric for semantic feature structures. In: Proceedings 51st Annual Meeting of the Association for Computational Linguistics, vol. 2, Short Papers, Sofia, Bulgaria, pp. 748–752 (2013)
24. Pust, M., Hermjakob, U., Knight, K., Marcu, D., May, J.: Parsing English into abstract meaning representation using syntax-based machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 1143–1154. Association for Computational Linguistics, September 2015
25. Garg, S., Galstyan, A., Hermjakob, U., Marcu, D.: Extracting biomolecular interactions using semantic parsing of biomedical text. In: Proceedings of AAAI (2016)
26. Pan, X., Cassidy, T., Hermjakob, U., Ji, H., Knight, K.: Unsupervised entity linking with abstract meaning representation. In: Proceedings of North American Chapter Association for Computational Linguistics, Denver, Colorado, pp. 1130–1139 (2015)
27. Ashish, N., Dewan, P., Ambite, J.-L., Toga, A.W.: GEM: the GAAIN entity mapper. In: Ashish, N., Ambite, J.-L. (eds.) DILS 2015. LNCS, vol. 9162, pp. 13–27. Springer, Heidelberg (2015). doi:10.1007/978-3-319-21843-4_2
28. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012. LNCS, vol. 7649, pp. 427–443. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35176-1_27

http://www.springer.com/978-3-319-46546-3

The Semantic Web – ISWC 2016
15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II
Groth, P.; Simperl, E.; Gray, A.J.G.; Sabou, M.; Krötzsch, M.; Lecue, F.; Flöck, F.; Gil, Y. (Eds.)
2016, XXVIII, 456 p. 107 illus., Softcover
ISBN: 978-3-319-46546-3