SemEval-2017 Task 9: Abstract Meaning Representation Parsing and Generation

Jonathan May and Jay Priyadarshi
Information Sciences Institute
Computer Science Department
University of Southern California
jonmay@isi.edu, jpriyada@usc.edu

Abstract

In this report we summarize the results of the 2017 AMR SemEval shared task. The task consisted of two separate yet related subtasks. In the parsing subtask, participants were asked to produce Abstract Meaning Representation (AMR) (Banarescu et al., 2013) graphs for a set of English sentences in the biomedical domain. In the generation subtask, participants were asked to generate English sentences given AMR graphs in the news/forum domain. A total of five sites participated in the parsing subtask, and four participated in the generation subtask. Along with a description of the task and the participants' systems, we show various score ablations and some sample outputs.

1 Introduction

Abstract Meaning Representation (AMR) is a compact, readable, whole-sentence semantic annotation (Banarescu et al., 2013). It includes entity identification and typing, PropBank semantic roles (Kingsbury and Palmer, 2002), individual entities playing multiple roles, as well as treatments of modality, negation, etc. AMR abstracts in numerous ways, e.g., by assigning the same conceptual structure to *fear* (v), *fear* (n), and *afraid* (adj). Figure 1 gives an example.

In 2016 an AMR parsing shared task was held at SemEval (May, 2016). Task participants demonstrated several new directions in AMR parsing technology and also validated the strong performance of existing parsers. We sought, in 2017, to focus AMR parsing performance on the biomedical domain, for which a not insignificant but still relatively small training corpus had been produced. While sentences from this domain are quite

```
(f / fear-01
  :polarity "-"
  :ARG0 ( s / soldier )
  :ARG1 ( d / die-01
          :ARG1 s ))
```

The soldier was not afraid of dying.
The soldier was not afraid to die.
The soldier did not fear death.

Figure 1: An Abstract Meaning Representation (AMR) with several English renderings. Example borrowed from Pust et al. (2015).

formal compared to some of those evaluated in last year's task, they are also very complex, and have many terms unique to the domain. An example is shown in Figure 2. We continue to use Smatch (Cai and Knight, 2013) as a metric for AMR parsing, but we perform additional ablative analysis using the approach proposed by Damonte et al. (2016).

Along with parsing into AMR, it is important to encourage improvements in automatic *generation* of natural language (NL) text from AMR. Humans favor communication in NL. An AI that is able to parse text into AMR at a quality level indistinguishable from humans may be said to understand NL, but without the ability to render its own semantic representations into NL no human will ever be able to appreciate this.

The advent of several systems that generate English text from AMR input (Flanigan et al., 2016b; Pourdamghani et al., 2016) inspired us to conduct a generation-based shared task from AMRs in the news/discussion forum domain. For the generation subtask, we solicited human judgments of sentence quality. We followed the precedent established by the Workshop in Machine Translation (Bojar et al., 2016) and used the Appraise solicitation system (Federmann, 2012), lightly mod-

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 536–545,
Vancouver, Canada, August 3 - 4, 2017.

```
Interestingly, serpinE2 mRNA and protein
were also markedly enhanced in human CRC
cells exhibiting mutation in
<i>KRAS </i>and <i>BRAF</i>.

(e / enhance-01 :li 2
 :ARG1 (a3 / and
  :op1 (n6 / nucleic-acid
    :name (n / name :op1 "mRNA")
   :ARG0-of (e2 / encode-01
    :ARG1 p))
  :op2 (p / protein
    :name (n2 / name :op1 "serpinE2")))
 :manner (m / marked)
 :mod (a2 / also)
 :location (c / cell
  :ARG0-of (e3 / exhibit-01
   :ARG1 (m2 / mutate-01
    :ARG1 (a4 / and
     :op1 (g / gene
       :name (n4 / name :op1 "KRAS"))
     :op2 (g2 / gene
       :name (n5 / name :op1 "BRAF")))))
  :mod (h / human)
  :mod (d / disease
    :name (n3 / name :op1 "CRC")))
 :manner (i / interesting))
```

Figure 2: One of the simpler biomedical domain sentences and its AMR. Note the italics markers in the original sentence are preserved, as they are semantically important to the sentence's understanding.

ified, to gather human rankings, then TrueSkill (Sakaguchi et al., 2014) to elicit an overall system ranking.

Since the same training data and tools are available to both subtasks (though, in the case of the generation subtask, the utility of the Bio-AMR corpus is unclear), we will describe all the resources for both subtasks in Sections 2 and 3 but then will handle descriptions and ablations for the parsing and generation subtasks separately, in, respectively, Sections 4 and 5. Readers interested in only one of these subtasks should not feel compelled to read the other section. We will reconvene in Section 6 to conclude and discuss hardware, as we continue the tradition established last year in the awarding of trophies to the declared winners of each subtask.

2 Data

LDC released a new corpus of AMRs (LDC2016E25), created as part of the DARPA DEFT program, in March of 2016. The new corpus, which was annotated by teams at SDL, LDC, and the University of Colorado, and supervised by Ulf Hermjakob at USC/ISI, is an extension of previous releases (LDC2015E86, LDC2014E41 and LDC2014T12). It contains 39,260 sentences (subsuming, in turn, the 19,572 AMRs from LDC2015E86, the 18,779 AMRs from LDC2014E41, and the 13,051 AMRs from LDC2014T12), partitioned into training, development, and test splits, from a variety of news and discussion forum sources. Participants in the generation task *only* were provided with AMRs for an additional 1,293 sentences for evaluation; the original sentences were also provided, as needed, to human evaluators during the human evaluation phase of the generation subtask (see Section 5.2). These sentences and their corresponding AMRs were sequestered and never released as data before the evaluation phase.

We also made available the Bio-AMR corpus version 0.8, which consists of 6,452 AMR annotations of sentences from cancer-related PubMed articles, covering 3 full papers[1] as well as the result sections of 46 additional PubMed papers. The corpus also includes about 1000 sentences each from the BEL BioCreative training corpus and the Chicago Corpus. The Bio-AMR corpus was partitioned into training, development, and test splits. An additional 500 sentences and their AMRs were sequestered until the evaluation phase, at which point the sentences were provided to parsing task participants *only*. Table 1 summarizes the available data, including the split sizes.

3 Other Resources

We made the following resources available to participants:

- The tokenizer (from Ulf Hermjakob) used to produce the tokenized sentences in the training corpus.[2]
- The AMR specification, used by annotators in producing the AMRs.[3]
- The JAMR (Flanigan et al., 2014)[4] and CAMR (Wang et al., 2015a)[5] parsers, as strong parser baselines.

[1] PMIDs 24651010, 11777939, and 15630473

[2] http://alt.qcri.org/semeval2016/task8/data/uploads/tokenizer.tar.gz

[3] https://github.com/kevincrawfordknight/amr-guidelines/blob/master/amr.md

[4] https://github.com/jflanigan/jamr

[5] https://github.com/c-amr/camr

Corpus	Domain	Train	Dev	Test	Eval
LDC2016E25	News/Forum	36,521	1,368	1,371	N/A
Bio-AMR v0.8	Biomedical	5,452	500	500	N/A
Parsing evaluation set	Biomedical	N/A	N/A	N/A	500
Generation evaluation set (LDC2016R33)	News/Form	N/A	N/A	N/A	1,293

Table 1: A summary of data used in this task; split sizes indicate the number of AMRs per sub-corpus.

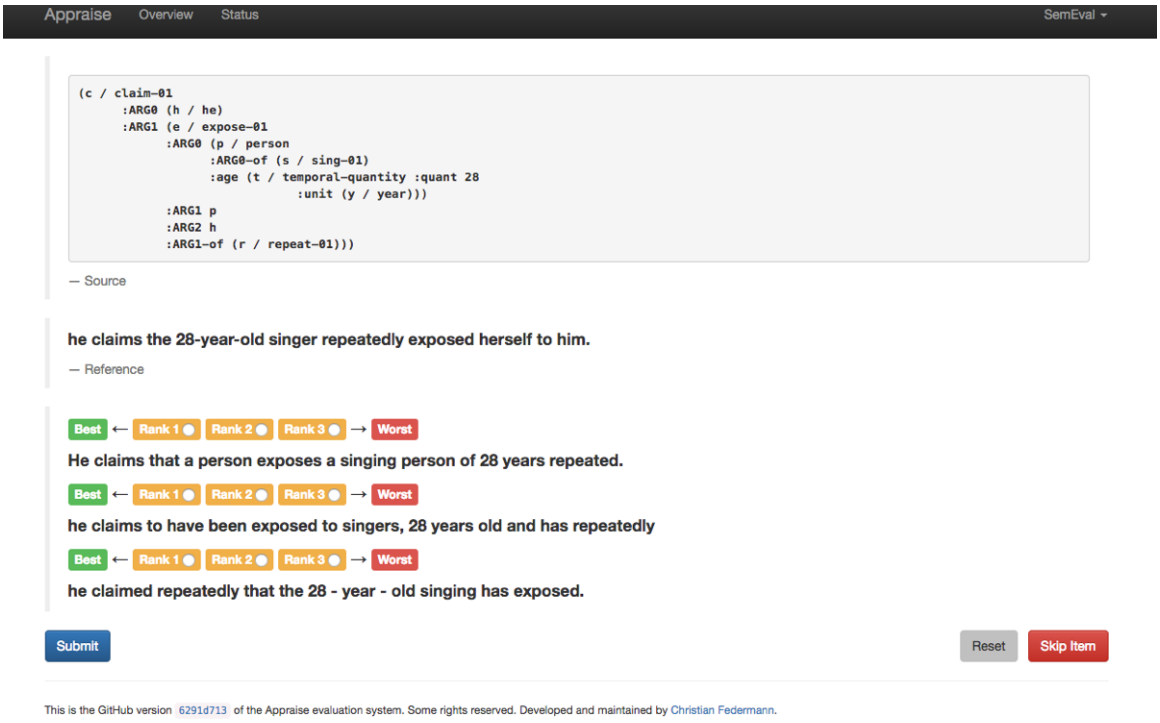

Figure 3: The Appraise interface, adapted for AMR generation evaluation.

- The JAMR (Flanigan et al., 2016b) generation system, as a strong generation baseline.
- An unsupervised AMR-to-English aligner (Pourdamghani et al., 2014).[6]
- The same Smatch (Cai and Knight, 2013) scoring script used in the evaluation.[7]
- A Python AMR manipulation library, from Nathan Schneider.[8]

4 Parsing Sub-Task

In the parsing sub-task, participants were given 500 previously sequestered Bio-AMRs and were asked to produce AMR graphs. Main results and ablative results are shown in Table 2.

4.1 Systems

Five teams participated in the task, a noticeable decline from last year's task, which saw eleven full participants. One team submitted two systems for a total of six distinct systems. Two teams were repeats from last year: CMU and RIGOTRIO (previously RIGA). Below are brief descriptions of each of the various systems, based on summaries provided by the system authors. Readers are encouraged to consult individual system description papers or relevant conference paper descriptions for more details.

4.1.1 The Meaning Factory (van Noord and Bos, 2017)

This team submitted two parsers. TMF-1 is a character-level sequence-to-sequence deep learn-

[6] http://isi.edu/~damghani/papers/Aligner.zip
[7] https://github.com/snowblink14/smatch
[8] https://github.com/nschneid/amr-hackathon

	Smatch	Unlab.	No WSD	NER	Wiki
TMF-1	0.46	0.5	0.46	0.51	0.46
TMF-2	0.58	0.63	0.58	0.58	0.4
UIT-DANGNT-CLNLP	**0.61**	**0.65**	**0.61**	**0.66**	0.35
Oxford	0.59	0.63	0.59	**0.66**	0.18
CMU	0.44	0.47	0.44	0.48	**0.59**
RIGOTRIO	0.54	0.59	0.54	0.46	0

(a) Main parsing results and four ablations

	Smatch	Neg.	Concepts	Reent.	SRL
TMF-1	0.46	0	0.63	0.29	0.43
TMF-2	0.58	0.24	0.76	0.35	0.54
UIT-DANGNT-CLNLP	**0.61**	0.24	**0.78**	0.37	0.56
Oxford	0.59	0.27	0.74	**0.43**	**0.57**
CMU	0.44	**0.33**	0.65	0.27	0.41
RIGOTRIO	0.54	0.31	0.71	0.34	0.51

(b) Main parsing results and four other ablations

Table 2: Main parsing results: For Smatch, a mean of ten runs with ten restarts per run is shown; standard deviation was about 0.0003 per system. For the remaining ablations, a single run was used.

ing model[9] similar to that of Barzdins and Gosko (2016), but with a number of pre- and post-processing changes to improve results. TMF-2 is an ensemble of CAMR (Wang et al., 2015b) models trained on different data sets and the seq-to-seq model to find the best CAMR parse.

4.1.2 UIT-DANGNT-CLNLP (Nguyen and Nguyen, 2017)

This team implemented two wrapper layers for CAMR (Wang et al., 2015a). The first layer standardizes and adds additional information to input sentences to eliminate the weakness of the dependency parser observed when parsing scientific quotations, figures, formulas, etc. The second layer wraps the output data of CAMR. It is based on a prebuilt list of (biology term–AMR structure) pairs to fix the output data of CAMR. This makes CAMR deal with unknown scientific concepts better.

4.1.3 Oxford (Buys and Blunsom, 2017)

This is a neural encoder-decoder AMR parser modeling the alignment between graph nodes and sentence tokens explicitly with a pointer mechanism. Candidate lemmas are predicted as a preprocessing step so that the lemmas of lexical node labels are factored out of the graph linearization.

[9] https://www.tensorflow.org/tutorials/seq2seq/

4.1.4 CMU

This was the same JAMR parsing system used in last year's evaluation (Flanigan et al., 2016a). The participants declined to submit a new system description paper.

4.1.5 RIGOTRIO (Gruzitis et al., 2017)

This team extended their CAMR-based AMR parser from last year's shared task (Barzdins and Gosko, 2016) with a gazetteer for recognizing as named entities the biomedical compounds frequently mentioned in the biomedical texts. The gazetteer was populated from the provided biomedical AMR training data.

4.2 Quantitative Ablation

We made use of the analysis scripts produced by Damonte et al. (2016) to conduct a more fine-grained ablation of scores. As noted in that work, Smatch provides full-sentence analysis but some aspects of an AMR are more difficult to parse correctly than others. The ablation study considers only (or excludes) an aspect of the AMR and then calculates Smatch (or F1, when no heuristic matching is needed) with that limitation in place. Ablation scores are shown in Table 2. The ablations are:[10]

[10] see Damonte et al. (2016) for more details

- Unlabeled: All argument labels (e.g. `ARG0`, `location`) are replaced with a single common label
- No WSD: Propbank frames indicating different senses (such as `die-01` vs `die-02`) are conflated
- NER: Only named entities are scored; that is, in both reference and hypothesis AMR, only nodes with an incoming arc labeled `name` are considered.
- Wiki: Only wikifications are scored; this is achieved in a manner similar to NER but with the incoming arc labeled `wiki`.
- Negation: Only concepts with an outgoing `polarity` arc are considered. In practice this arc is only used to indicate negation.
- Concepts: Only concepts, not relations, are scored.
- Reentrancies: Only concepts with two or more incoming relations are scored. Reentrancies occur when a concept has several mentions in a sentence, or where an 'inverted' relation (one that ends in `-of`) occurs, implying inverse dependency. In practice the latter is much more often the cause of a re-entrancy.
- Semantic Role Labeling (SRL): only relations corresponding to roles in PropBank, i.e. those named `ARG0` and the like, are scored.

The ablation results show that superior performance in Smatch correlates with superior performance in the Unlabeled, No-WSD, NER, and Concepts performance. Additionally, Figure 4, which plots each ablation score against Smatch and induces a linear regression, shows that six of the eight ablation sub-metrics are well correlated with Smatch; only wikification and negation are not. Wikification is generally handled as a separate process on top of overall AMR parsing; this may explain that discrepancy. We have no great explanation for negation's weak correlation but note that it is generally considered a difficult task in semantics.

4.3 Discussion

It is interesting to note that the top-scoring system was, as in last year's shared task, based on

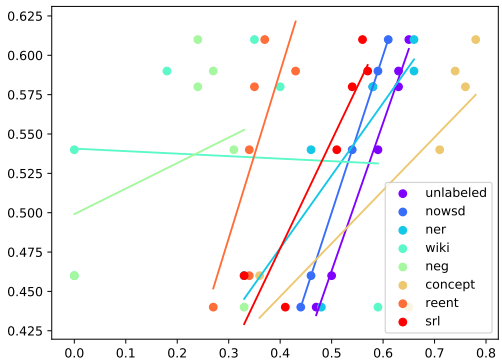

Figure 4: Relationship between each of the eight quantitative ablation studies from Damonte et al. (2016) and Smatch; six of the eight metrics are well-correlated with Smatch.

CAMR (Wang et al., 2015b). It is also interesting to note that, in the Oxford team's submission, once again, a pure neural system is nearly as good as the CAMR system, despite having rather little data to train on. The Oxford system appears to be quite different from last year's neural submission (Foland and Martin, 2016) but nevertheless is a strong competitor. Finally, the top-scoring system, that of UIT-DANGNT-CLNLP, got a 0.61 Smatch, while last year's top scoring systems (Barzdins and Gosko, 2016; Wang et al., 2016) scored a 0.62, practically the same score. This, despite the fact that the evaluation corpora were quite different. One might expect the biomedical corpus to be easier to parse than the news/forum corpus, since its sentences are rather formal, and do not use slang or incorrect syntax. On the other hand, the sentences in the biomedical corpus are on average longer than those in the news/forum corpus (on average 25 words in bio vs. 14.5 in news/forum) and the biomedical corpus contains many unknown words, corresponding to domain terminology not in general use (1-count words are 9% of tokens in bio training, vs. 7.2% in news/forum). The news/forum corpus has, in its forum content, colloquialisms and writing variants that are very difficult to automatically analyze. Perhaps the relatively 'easy' and 'hard' parts of each corpus canceled each other out, yielding corpora that were about the same level of difficulty to parse. Nevertheless, it is somewhat concerning that AMR parsing quality appears to have stalled, as parsing performance remains in the low 0.60 range.

5 Generation Sub-Task

As AMR provides full-sentence semantics, it may be a suitable formalism for semantics-to-text generation. This subtask explored the suitability of that hypothesis. Given that AMRs do not capture non-semantic surface phenomena nor some essential properties of realized text such as tense, we incorporated human judgments into our evaluation, since automatic metrics against a single reference were practically guaranteed to be inadequate.

5.1 Systems

Four teams participated in the task. We also included a submission from Pourdamghani et al. (2016) run by the organizer, though *a priori* declared that submission to be non-competitive due to a conflict of interest. Below we provide short summaries of each team's approach.

5.1.1 CMU

This was the JAMR generation system described in (Flanigan et al., 2016b). The participants declined to submit a system description paper.

5.1.2 Sheffield (Lampouras and Vlachos, 2017)

This team's method is based on inverting previous work on transition-based parsers, and casts NLG from AMR as a sequence of actions (e.g., insert/remove/rename edges and nodes) that progressively transform the AMR graph into a syntactic parse tree. It achieves this by employing a sequence of four classifiers, each focusing on a subset of the transition actions, and finally realizing the syntactic parse tree into the final sentence.

5.1.3 RIGOTRIO (Gruzitis et al., 2017)

For generation, this team's approach was to write transformation rules for converting AMR into Grammatical Framework (Ranta, 2004) abstract syntax from which semantically correct English text can be rendered automatically. In reality the approach worked for 10% of AMRs. For the submission the remaining 90% AMRs were converted to text using the JAMR (Flanigan et al., 2014) tool.

5.1.4 FORGe (Simon Mille and Wanner, 2017)

UPF-TALN's generation pipeline comprises a series of rule-based graph-transducers, for the syntacticization of the input graphs (converted to CoNLL format) and the resolution of morphological agreements, and an off-the-shelf statistical linearization component.

5.1.5 ISI

This was an internal, non-trophy-eligible submission based on the work of Pourdamghani et al. (2016). It views generation as phrase based machine translation and learns a linearization of AMR such that the result can be used in an off-the-shelf Moses (Koehn et al., 2007) PBMT implementation.

5.2 Manual Evaluation

We used Appraise (Federmann, 2012), an open-source system for manual evaluation of machine translation, to conduct a human evaluation of generation quality. The system asks human judges to rank randomly selected systems' translations of sentences from the test corpus. This in turn yields pairwise preference information that can be used to effect an overall system ranking.

For the purposes of this task we needed to adapt the Appraise system to admit nested representations of AMRs, and to be compatible with our IT infrastructure. A screen shot is shown in Figure 3.

5.3 Scoring

We provided BLEU as a potentially helpful automatic metric but consider several metrics induced over pairwise comparisons induced by manual evaluation to be the "true" evaluation metric for the purposes of trophy-awarding:

- Win+tie percentage: This is simply the percentage "wins" (better pairwise comparisons) plus "ties" (equal comparisons) of the total number of its pairwise comparisons. This metric was largely used to induce rankings from human judgments through WMT 2011.
- Win percentage: This is a "harsher" version of Win+tie; the percentage is $\frac{\text{wins}}{\text{wins}+\text{ties}+\text{losses}}$. Essentially, ties are judged as losses. This was used in WMT 2011 and 2012.
- TrueSkill (Sakaguchi et al., 2014). This is an adaptation of a metric developed for player rankings in ongoing competitions such as on Microsoft Xbox Live. The metric maintains estimates of player (i.e., generation system)

	Win	Win+Tie	Trueskill	BLEU
RIGOTRIO	**54.91**	**81.49**	**1.07**	18.82
CMU	50.36	72.48	0.85	**19.01**
FORGe	43.64	57.43	0.45	4.74
ISI	26.05	38.39	-1.19	10.92
Sheffield	8.38	21.16	-2.20	3.32

Table 3: Main generation results: The three manually-derived metrics agree on the systems' relative rankings.

	Win	Win+Tie	Trueskill
RIGOTRIO	**53.00**	**79.98**	**1.03**
CMU	50.02	71.91	0.819
FORGe	44.49	58.57	0.458
ISI	26.40	38.60	-1.172
Sheffield	9.46	22.84	-2.132

Table 4: Human judgments of generation results after self-judgments are removed: The results are fundamentally the same

ability as Gaussian distributions and rewards events (i.e., pairwise rankings of outputs) that are unexpected, such as a poorly ranked player outperforming a highly-ranked player, more than expected events.

We note that the three metrics derived from human pairwise rankings agree with the relative ordering of the submitted systems' abilities on the evaluation data, while the BLEU metric does not. It is not terribly surprising the BLEU does not correlate with human judgment; it was designed for a very different task.

Since the participants in this task were also judges in the human evaluation, we were somewhat concerned that implicit bias might lead to a skewing of the results, even though system identification was not available during evaluation. We thus removed all judgments that involved self-scoring and recalculated results. The results, shown in Table 4, show little difference from the main results.

5.4 Qualitative Analysis

The generation task was quite challenging, as generation from AMR is still a nascent field. Table 5 shows an example of a single AMR and the content generated by each system for it, along with the number of wins, ties, and losses per system by the human evaluations (note: not all segments were scored for all systems, and not all systems received the same number of comparisons). Some systematic errors, such as incorporating label text into the generation, could lead to improvements, as could a stronger language model; generated output is often disfluent.

6 Conclusion

Both biomedical AMR parsing and generation from AMRs appear to be challenging tasks; perhaps too challenging, as the number of participants in either subtask was significantly lower than the participation rate from a year ago. However, we observed that AMR parsing quality on the seemingly more difficult biomedical domain was no worse than that observed on the news/forum domain. In fact, the same fundamental technology that dominated in last year's evaluation once again reigned supreme. A concern that Smatch was too coarse a metric to evaluate AMRs was not borne out, as scores in an ablation study tracked well with the overall Smatch score. We are pleased to award the parsing trophy to the UIT-DANGNT-CLNLP team, which added domain-specific modification to the strong CAMR (Wang et al., 2015b) parsing platform.

On the generation side, it seems that there is still a long way to go to reach fluency. We note that BLEU, which is often used as a generation metric, is woefully inadequate compared to human evaluation. We hope the analysis presented here will lead to better generation systems in the future. It was clear from the human evaluations, however, that the RIGOTRIO team prevailed and will receive the generation trophy.

Acknowledgments

We were overjoyed to be offered volunteer human judgments by Nathan Schneider and his class at Georgetown: Austin Blodgett, Emma Manning, Harry Eldridge, Joe Garman, Lucia Donatelli, Sean MacAvaney, Max Kim, Nicholas Chapman, Mohammad Ali Yekataie, and Yushi Zhao. Thanks to Marco Damonte and Shay Cohen for reaching out regarding Smatch ablations and providing scoring code. The human evaluations would not have been possible without the use of the Appraise system, which was shared by Christian Federmann and Matt Post. Many thanks to the AMR creation team: Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. Thanks also to

```
(a / and
  :op1 (r / remain-01
    :ARG1 (c / country :wiki "Bosnia_and_Herzegovina"
      :name (n / name :op1 "Bosnia"))
    :ARG3 (d / divide-02
      :ARG1 c
      :topic (e / ethnic)))
  :op2 (v / violence
    :time (m / match-03
      :mod (f2 / football)
      :ARG1-of (m2 / major-02))
    :location (h / here)
    :frequency (o / occasional))
  :time (f / follow-01
    :ARG2 (w / war-01
      :time (d2 / date-interval
        :op1 (d3 / date-entity :year 1992)
        :op2 (d4 / date-entity :year 1995)))))
```

Source	Text	W	T	L
Reference	following the 1992-1995 war bosnia remains ethnically divided and violence during major football matches occasionally occurs here.			
RIGOTRIO	following the 1992 1995 war, bosnia has remained an ethnic divide, and the major football matches occasionally violence in here.	1	2	1
CMU	following the 1992 1995 war , bosnia remains divided in ethnic and the occasional football match in major violence in here	2	0	0
FORGe	Bosnia and Herzegovina remains under about ethnic the Bosnia and Herzegovina divide and here at a majored match a violence.	0	0	4
ISI	following war between 1992 and 1995 , the country :wiki bosnia_and_herzegovina bosnia remain divided on ethnic and violence in football match by major here from time to time	3	0	1
Sheffield	Remain Bosnia ethnic divid following war 920000 950000 major match footbal occasional here violency	0	2	0

Table 5: Examples of an AMR from the Generation subtask and each system's generation for it (W = # wins, T = # ties, L = # losses in human evaluation).

the SemEval organizers: Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif Mohammad, Daniel Cer, and David Jurgens. We also gratefully acknowledge the participating teams' efforts. This work was sponsored by DARPA DEFT (FA8750-13-2-0045).

References

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 the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, pages 178–186. http://www.aclweb.org/anthology/W13-2322.

Guntis Barzdins and Didzis Gosko. 2016. RIGA at SemEval-2016 task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Jan Buys and Phil Blunsom. 2017. Oxford at SemEval-

2017 task 9: Neural AMR parsing with pointer-augmented attention. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 748–752. http://www.aclweb.org/anthology/P13-2131.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2016. An incremental parser for abstract meaning representation. *CoRR* abs/1608.06111. http://arxiv.org/abs/1608.06111.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics* 98:25–35.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016a. CMU at SemEval-2016 task 8: Graph-based AMR parsing with infinite ramp loss. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016b. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 731–739. http://www.aclweb.org/anthology/N16-1087.

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)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1426–1436. http://www.aclweb.org/anthology/P14-1134.

William Foland and James H. Martin. 2016. CU-NLP at SemEval-2016 task 8: AMR parsing using LSTM-based recurrent neural networks. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics, San Diego, California.

Normunds Gruzitis, Didzis Gosko, and Guntis Barzdins. 2017. RIGOTRIO at SemEval-2017 task 9: Combining machine learning and grammar engineering for AMR parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Language Resources and Evaluation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/P07-2045.

Gerasimos Lampouras and Andreas Vlachos. 2017. Sheffield at SemEval-2017 task 9: Transition-based language generation from AMR. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Jonathan May. 2016. Semeval-2016 task 8: Meaning representation parsing. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1063–1073. http://www.aclweb.org/anthology/S16-1166.

Khoa Nguyen and Dang Nguyen. 2017. UIT-DANGNT-CLNLP at SemEval-2017 task 9: Building scientific concept fixing patterns for improving CAMR. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Nima Pourdamghani, Yang Gao, Ulf Hermjakob, and Kevin Knight. 2014. Aligning English strings with Abstract Meaning Representation graphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 425–429. http://www.aclweb.org/anthology/D14-1048.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating English from Abstract Meaning Representations. In *Proceedings of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, Edinburgh, UK, pages 21–25.

Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. 2015. Parsing English into Abstract Meaning Representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1143–1154. http://aclweb.org/anthology/D15-1136.

Aarne Ranta. 2004. Grammatical framework. *Journal of Functional Programming* 14(2):145189. https://doi.org/10.1017/S0956796803004738.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 1–11. http://www.aclweb.org/anthology/W14-3301.

Roberto Carlini Alicia Burga Simon Mille and Leo Wanner. 2017. FORGe at SemEval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Rik van Noord and Johan Bos. 2017. The meaning factory at SemEval-2017 task 9: Producing AMRs with neural semantic parsing. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, Vancouver, Canada.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015a. Boosting transition-based AMR parsing with refined actions and auxiliary analyzers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 857–862. http://www.aclweb.org/anthology/P15-2141.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015b. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 366–375. http://www.aclweb.org/anthology/N15-1040.

Chuan Wang, Nianwen Xue, Sameer Pradhan, Xiaoman Pan, and Heng Ji. 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)*. Association for Computational Linguistics, San Diego, California.