Proceedings of the LREC 2020 Workshop "Citizen Linguistics in Language Resource Development", pages 35–39
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

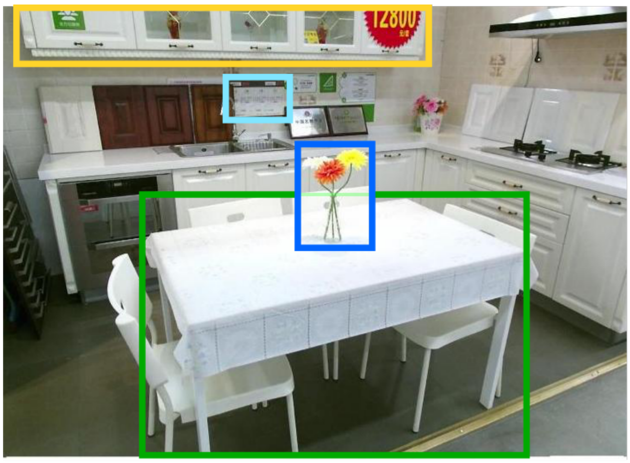
Leveraging Non-Specialists for Accurate and Time Efficient AMR Annotation

Mary Martin, Cecilia Mauceri, Martha Palmer and Christoffer Heckman
University of Colorado Boulder
Department of Computer Science, 430 UCB
Boulder, Colorado 80309-0430
`<first>.<last>@colorado.edu`

Abstract

Abstract Meaning Representations (AMRs), a syntax-free representation of phrase semantics (Banarescu et al., 2013), are useful for capturing the meaning of a phrase and reflecting the relationship between concepts that are referred to. However, annotating AMRs is time consuming and expensive. The existing annotation process requires expertly trained workers who have knowledge of an extensive set of guidelines for parsing phrases. In this paper, we propose a cost-saving two-step process for the creation of a corpus of AMR-phrase pairs for spatial referring expressions. The first step uses non-specialists to perform simple annotations that can be leveraged in the second step to accelerate the annotation performed by the experts. We hypothesize that our process will decrease the cost per annotation and improve consistency across annotators. Few corpora of spatial referring expressions exist and the resulting language resource will be valuable for referring expression comprehension and generation modeling.

Keywords: Abstract Meaning Representation, crowd-annotation, spatial referring expressions



The **calendar** is hanging below the **cupboards**.

```
(h / hang-01
  :arg0 () #agent, entity causing thing to be suspended
  :arg1 (c1 / calendar) #thing suspended
  :arg2 () #suspended from
  :location (b / below-01
    :op1 (c2 / cupboards))
```

The **flowers** in the middle of the **table**

```
(s / sit-01
  :arg1 (f / flowers) #thing sitting
  :arg2 (m / middle-01 #location or position
    :part-of (t / table))
```

Figure 1: Two referring expressions with their AMR parses. The color-coded bounding boxes and entity mentions indicate correspondences between the image and text.

1. Introduction

The relationship between the linguistic and visual representations of the same information is non-trivial. Not only is “a picture worth a thousand words”, but there are also many possible ways to describe the same configuration of objects, i.e. the cupboard is *above* the sink or the sink is *below* the cupboard. Different syntax may also be used to communicate the same meaning. We need a linguistic representation where two expressions with the same underlying meaning have the same representation in order to build a correspondence between the text and image that can be used for visual question answering and referring expression comprehension and generation. AMRs (Abstract Meaning Representations) are one such representation.

Abstract Meaning Representations are a novel, natural language representation which is defined purely by the phrase’s semantics. The novelty of this data structure lies in its ability to provide a single abstraction that can represent a number of different phrases. AMRs accomplish this through the use of relations and concepts that form a logical tree structure, as opposed to syntactic representations such as those produced through dependency and constituency parsing.

Using the AMR structure, we seek to annotate the object relationships from a corpora of spatial referring expressions. This representation effectively harnesses the spatial information in a given natural language sentence that is formulated based on a human’s perception of the scene. AMR representations of spatial referring expressions will allow future research to explore how visual features relate to spatial relationships. Unfortunately, AMRs are expensive to annotate. There is no automated tool that has been deemed consistent enough to effectively create AMR parses of natural language sentences as there are with dependency and constituency parses. AMRs require annotators to derive the exact meaning of certain entities or ”concepts” through context. This aspect, along with in-depth guidelines for structuring the trees, requires annotators to undergo extensive training.

Luckily, there are parts of the AMR annotation process that don’t require expert knowledge. For example, it does not require training for humans to identify object relationships in phrases. Volunteers also do not require training in order to derive meaning from phrases and respond to queries such as ”who is doing what to whom?” (Banarescu et al., 2013). We propose to divide the AMR annotation pipeline into two parts; the first part using crowd-workers and volunteer annotators, and the second, AMR experts. The intention of the tasks presented for non-specialist annotators is to create the closest possible result to an AMR without the need for domain specific knowledge. This approximate AMR can then be used as a starting point for expert annotation, limiting the role of experts to the more challenging annotation decisions. We hypothesize that this two step annotation will improve consistency and efficiency of annotation.

2. Related Work

2.1. Crowdsourcing Annotations

Crowdsourcing annotations is a common method for sourcing data for linguistics experiments and tasks. Techniques such as those used to annotate Question Answer (QA) Meaning Representations distribute the annotation process over multiple annotators in order to gain sufficient coverage when producing QA pairs (Michael et al., 2018). Methods for Semantic Role Labeling (SRL) in CROWD-IN-THE-LOOP improve upon previous practices for SRL by enabling annotators to produce gold-labeled training data without the need for expert involvement (Wang et al., 2017). We will take a similar approach to crowdsourcing in order to optimize the quality of data gathered by non-specialist volunteers, though we cannot eliminate the need for expert involvement. As opposed to splitting tasks for phrase coverage, we choose to split based on whether an annotation step requires expert knowledge.

2.2. Related Datasets

A few existing visual referring expressions datasets provide entity and relationship annotation. Flickr30k Entities includes annotations which link entity mentions and bounding boxes (Plummer et al., 2017). SentencesNYUv2 similarly aligns entity mentions and bounding boxes, and additionally provides adjective and preposition parsing (Kong et al., 2014). Visual Genome’s region and scene graphs are most similar to AMRs (Krishna et al., 2017). Like AMRs, scene graphs are a formal representation of objects, relationships, and attributes. Like AMRs, they organize these elements in a graph structure and are syntax independent. In contrast to scene graphs, AMRs provide greater differentiation between roles than scene graphs do. To our knowledge, there is no dataset which pairs images and AMRs.

3. Proposed Method

Our goal is annotation, similar to that shown in Figure 1, consisting of referring expressions parsed into AMRs and linked to object bounding boxes. We source our referring expressions and bounding boxes from the SUN-Spot dataset (Mauceri et al., 2019). The challenge is to parse these referring expressions and link the entities to bounding boxes at low cost.

To complete this task, we propose an AMR annotation pipeline with three steps: (1) automated text preprocessing, (2) annotation by non-specialists, and (3) annotation by experts. With each step, the difficulty of the annotation tasks increase. We hypothesize that by ordering tasks in order of increasing difficulty, we can minimize the cognitive load of the annotators at each step, thus speeding annotation, decreasing overall cost, and improving consistency across annotators. The following sections detail each part of the pipeline.

3.1. Text Preprocessing

In order to structure the data for efficient annotation, we have implemented an automated text preprocessing function. This simple preprocessing step isolates certain parts of speech to assist with recognition of objects and spatial relationships. Automated preprocessing is done using the Stanford CoreNLP Toolkit (Manning et al., 2014) and Stanford Part-of-Speech (POS) Tagger (Toutanova et al., 2003). We intend to adopt some of the preprocessing techniques applied to phrases when generating the SentenceNYUv2 dataset (Kong et al., 2014). These techniques include using Stanford’s coreference system to predict clusters of coreference mentions in order to identify pronouns. This can assist with identifying pronouns as they relate to objects in scenes (Clark and Manning, 2016).

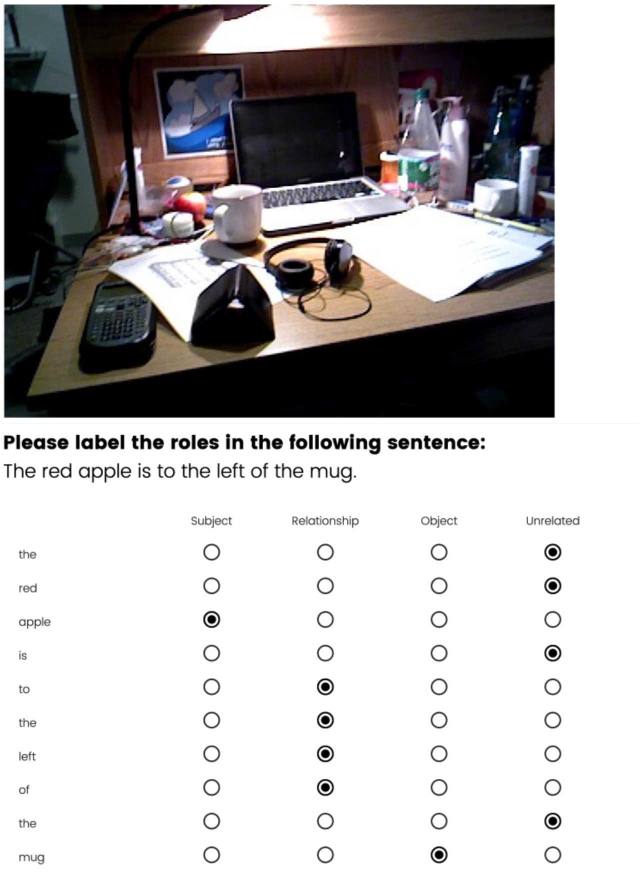
The text preprocessing step also removes words from the phrase that are not relevant to the creation of an AMR. Such parts of speech include articles and conjunctions. In order for the phrase to be represented using a syntax-free graph, words in the sentence must pass through a lemmatizer. The lemmatizer reduces words to their root. This standardizes verb representation.

The preprocessing function also seeks to automate portions of the AMR annotation task which can produce inconsistent parses when manually performed by volunteers and workers. With the goal of consistency in mind, it is important to recognize where human error may occur in any process. We mitigate this by taking advantage of automated NLP tools that are accurate and easy to implement. The output of this function indicates important POS that highlight roles of words as they relate (or do not relate) to spatial relationships.

3.2. Annotation by Non-specialist Annotators

The next phase of annotation is performed by non-specialist annotators, such as crowd-workers and citizen scientist volunteers. Their job is twofold; the non-specialist annotators perform an initial pass identifying argument roles, and they label correspondences between object mentions in text and the location in the image.

In the final AMR annotation, words will be assigned to argument roles. However, argument roles are not familiar to most non-linguists. In order to provide a simplified annotation tool to non-specialist annotators, we chose a succinct set of familiar word classes that are analogous to argument roles. These classes include “subject”, “relationship”, “object” and “unrelated”. Annotators are asked to classify all words in the processed phrase into one of these classes using a simple multiple choice interface. The proposed interface takes a similar form to that shown when decomposing QA-SRL questions into slot-based representations (FitzGerald et al., 2018). A mockup of our proposed inter-



Please label the roles in the following sentence:
The red apple is to the left of the mug.

	Subject	Relationship	Object	Unrelated
the	○	○	○	●
red	○	○	○	●
apple	●	○	○	○
is	○	○	○	●
to	○	●	○	○
the	○	●	○	○
left	○	●	○	○
of	○	●	○	○
the	○	○	○	●
mug	○	○	●	○

Figure 2: Example of annotation interface for approximate role labeling used by non-specialist annotators.

face is shown in Figure 2. We chose the word class role ”relationship” in place of ”preposition” in order to give annotators the choice to group chunks of words as a ”relationship”. During this annotation task, the annotators are provided with the full phrase, processed phrase and the original image for reference.

In the next annotation task, annotators label correspondences between the text and image. Our goal in annotating this dataset is to relate spatial relationships in images and referring expressions. Therefore, we wish to annotate any object mentions in the referring expression with links to the corresponding bounding box in the image. Highlighting the “subject” and “object” annotations from the previous step, we ask annotators to click on the corresponding object in the image. A similar task was successfully used to validate the referring expressions during the SUN-Spot dataset collection (Mauceri et al., 2019).

3.3. Annotation by Experts

The creation of AMRs from raw, unprocessed phrases is a time-consuming task because of the extensive set of guidelines that exist to create consistency between parses. To assist with this, experts will receive AMR proposals generated from the previous annotation steps instead of raw text. We hypothesize that approving, rejecting, and editing proposed AMRs is faster and easier than full annotation. The challenge is how to create appropriate proposals from the rough grained approximate roles provided by the non-specialist annotators.

Approximate Roles

```
Subject: apple
Relationship: to the left
Object: mug
```

Mapped to

```
(b / be-01
  arg1: (a / apple)
  arg2: (m / mug)
  location: (t / to the left)))
```

Corrected

```
(b / be-01
  arg1: (a / apple)
  arg2: (l / left
    op1: (m / mug)))
```

Figure 3: The approximate role labels are mapped to the AMR structure for review by experts. In this example, the subject and object roles are mapped to arg1 and arg2 and the relationship role is mapped to location. However, in the correct AMR, the relationship should be arg2. The expert must approve or reject the mapped AMRs. Rejected mapped AMRs are then hand-corrected.

In this generation process, the structure of spatial referring expressions comes to our assistance. Spatial referring expressions have two typical forms; either they contain a copula with a be-verb, or they use a position verb like “hang” or “sit”. In both cases, the arg1 tends to be the subject of the referring expression, and the arg2 is either the location preposition or the object of the sentence. Using simple rules like these, we can establish a rule-based mapping for a large portion of our dataset. The expert annotator’s role is to correct this mapping as shown in figure 3.

The data that the experts are presented with includes the full phrase, the processed phrase, the approximate argument role of each word, and the links between entities in the sentence and corresponding image. This data is meant to capture a simplified form of the relationship between the objects in the text and image domains. Through eliminating extraneous words and predetermining the roles of entities, we seek to introduce consistency and efficiency to this step in the pipeline. Consistency among a large number of examples is key in introducing a dataset that may act as ground truth when determining AMR parses of a variety of phrases.

An important aspect of this method is ensuring that the annotation pipeline provides improvements in consistency and efficiency as proposed. To assess the effectiveness of the process in these respects, we intend to compare the expense of annotating data from the perspective of the expert annotator. This involves evaluating the change in the time that it takes to complete one AMR, as well as qualitatively evaluating the change in the difficulty of the task based on feedback from the annotators. Ideally, an experiment such as this should yield results that indicate a significant decrease in annotation time, improvements in data quality, and a smoother process.

4. Future Work

4.1. Using Language Resources for Efficient Text Pre-processing

When designing tasks for annotation by non-specialists, phrase pre-processing has the potential to affect an annotator's interpretation of the phrase. For example, in a given word role classification task, identifying prepositions with multiple words may prove to be a challenge. Annotators must determine the words that define the spatial relationship between multiple objects. This presents a problem because interpretations of words that define relationships between objects may be inconsistent among annotators. A solution for this potential problem would be to present annotators with complete preposition phrases for role classification. In practice, this may involve chunking, for example ”next to” instead of ”next” and ”to”, in order to definitively demonstrate that the role of these words is a ”relationship”. Additionally, we intend to incorporate suggestions from expert annotators to develop ways to format the annotated phrases that will convert most directly to an AMR. In conjunction to taking an iterative approach for improving the data quality with expert feedback, we seek to improve the pipeline by automating much of the process if possible.

4.2. Using paired AMRs and RGB-D Images for Multi-modal Deep Learning

The graph structure of Abstract Meaning Representations makes them a suitable data structure for use with graph transformer networks, a variation of Graph Neural Networks (Scarselli et al., 2009). Graph Transformer Networks allow for the representation of heterogenous graph structures for machine learning tasks with graph structured input data (Yun et al., 2019). In this case, ”heterogenous” refers to graphs with multiple edge types. The SUN-Spot dataset contains color images with an additional depth channel or RGB-D images. Through pairing AMRs and images where objects act as nodes on a graph and edges represent their spatial relationships, we hope to learn the relationship between the spatial relationships in phrases and depth images. Incorporating depth allows us to derive the locations of objects relative to others in the scene.

4.3. Automated AMR Parsing

Though the goal of annotating a referring expressions dataset is to capture spatial relationships in language, creating a large corpus of AMR-phrase pairs lends itself to other tasks. With an accumulation of phrases and corresponding ground truth AMR trees, this data would be well suited for a machine learning problem involving the automation of phrase parsing. A similar method has been used to automate Question Answer driven Semantic Role Labeling with successful results through a combination of phrase preprocessing and machine learning (FitzGerald et al., 2018).

5. Conclusion

We proposed an annotation pipeline with the goal of increasing efficiency in an expensive and time consuming process. By adopting and iteratively improving this method, our intention is to create a corpus that enables research involving solving problems in domains where AMRs have not previously been applied. In future work, we intend to demonstrate the benefits of linking this type of text abstraction to corresponding scenes. With this data, we will use deep neural networks to learn the connection between spatial relationships in natural language sentences using the RGB-D scenes that they are gathered from. Tangentially, we hope to move closer to a process for fully automated AMR parsing.

Acknowledgments

This work is partially supported by the National Science Foundation award number 1849357.

6. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

FitzGerald, N., Michael, J., He, L., and Zettlemoyer, L. (2018). Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060.

Kong, C., Lin, D., Bansal, M., Urtasun, R., and Fidler, S. (2014). What are you talking about? text-to-image coreference. In *CVPR*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

Michael, J., Stanovsky, G., He, L., Dagan, I., and Zettlemoyer, L. (2018). Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, Jan.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 173–180, USA. Association for Computational Linguistics.

Wang, C., Akbik, A., Chiticariu, L., Li, Y., Xia, F., and Xu, A. (2017). CROWD-IN-THE-LOOP: A hybrid approach for annotating semantic roles. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1913–1922, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. (2019). Graph transformer networks. In *Advances in Neural Information Processing Systems*, pages 11960–11970.

7. Language Resource References

Mauceri, C., Palmer, M., and Christoffer, H. (2019). SUN-Spot: An RGB-D Dataset With Spatial Referring Expressions. In *International Conference on Computer Vision Workshop on Closing the Loop Between Vision and Language*.