QA-Adj: Adding Adjectives to QA-based Semantics

Leon Pesahov[1] **Ayal Klein**[1] **Ido Dagan**[1]
[1]Computer Science Department, Bar Ilan University
{leonpes92,ayal.s.klein}@gmail.com dagan@cs.biu.ac.il

Abstract

Identifying all predicate-argument relations in a sentence has been a fundamental research target in NLP. While traditionally these relations were modeled via formal schemata, the recent QA-SRL paradigm (and its extensions) present appealing advantages of capturing such relations through intuitive natural language question-answer (QA) pairs. In this paper, we extend the QA-based semantics framework to cover adjectival predicates, which carry important information in many downstream settings yet have been scarcely addressed in NLP research. Firstly, based on some prior literature and empirical assessment, we propose capturing four types of core adjectival arguments, through corresponding question types. Notably, our coverage goes beyond prior annotations of adjectival arguments, while also explicating valuable implicit arguments. Next, we develop an extensive data annotation methodology, involving controlled crowdsourcing and targeted expert review. Following, we create a high-quality dataset, consisting of 9K adjective mentions with 12K predicate-argument instances (QAs). Finally, we present and analyze baseline models based on text-to-text language modeling, indicating challenges for future research, particularly regarding the scarce argument types. Overall, we suggest that our contributions can provide the basis for research on contemporary modeling of adjectival information.

1 Introduction

A main challenge addressed by Natural Language Processing research is designing useful semantic representations, capturing and explicating important aspects of the meaning of a text. Numerous recent works illustrate how even in the era of strong end-to-end neural models, leveraging explicit semantic representations facilitates downstream processing of challenging tasks (Huang and Kurohashi, 2021; Mohamed and Oussalah, 2019; Zhu et al., 2021; Chen and Durrett, 2021).

Numerous semantic representations have been proposed and pursued (Abend and Rappoport, 2017). Traditionally, semantic representations rely on pre-defined schemata of linguistic classes, e.g. semantic roles or relations. Thus, mapping natural language onto its representations becomes a complex annotation task that requires significant linguistic expertise, causing challenges in data collection and utility in new domains and languages.

Recently, many researchers and practitioners seek to benefit from an explicit representation of text meaning while alleviating the reliance on hard-to-scale structured formalisms. For instance, Open Information Extraction (Banko et al., 2007) has gained popularity as a light-weight, NL-based alternative to Semantic Role Labeling (SRL) formalisms like PropBank (Palmer et al., 2005) or FrameNet (Baker et al., 1998). More recently, several works proposed using question-answer pairs (QAs) as an intermediate structure, e.g. in order to assess information alignment between texts for evaluating summarization quality (Eyal et al., 2019; Gavenavicius, 2020; Deutsch et al., 2021) and faithfulness (Honovich et al., 2021; Durmus et al., 2020). While these latter works utilize "general-purpose" question-answering datasets and models for generating the QAs, Klein et al. (2022a) put forward a systematically targeted QA-based semantic framework dubbed *QASem*. Pioneered by addressing verbal predicate-argument relations in QA-SRL (He et al., 2015), this framework integrates three systematic QA-driven representations, jointly covering semantic role labeling for verbs (He et al., 2015; FitzGerald et al., 2018; Roit et al., 2020), nominalizations (Klein et al., 2020) and informational discourse relations (Pyatkin et al., 2020).

However, current QA-based approaches lack principled coverage of adjectival information. In natural language text, adjectives carry vital information about the properties of entities, essential for many downstream NLP applications. For example,

Proceedings of the 4th International Workshop on Designing Meaning Representations, pages 74–88
June, 2023.

*Galbraith attacked the consensus for monetarist economics and argued that Keynesian economics were far **more relevant** for tackling the emerging crises.*

Question type	Question	Answer
Object	What was more relevant for something?	Keynesian economics
Comparison	Compared to what was something more relevant?	monetarist economics
Domain	What was something more relevant for?	tackling the emerging crises
Extent	To what degree was something more relevant?	far

Table 1: An example of QA-Adj question-answer pair.

in benchmarks of the widely-used sentiment analysis task (Pontiki et al., 2014), adjectives comprise 75% of the annotated "sentiment triggers".

In this work, we extend the QASem paradigm by capturing and explicating the fundamental aspects of adjectival information using natural-language question-answer pairs. Our representation, termed *QA-Adj*, consists of four adjective-related roles — object, comparison, domain, and extent. As we will see later, these roles provide a fairly complete representation of the core arguments of adjectives. Roles are annotated using question templates while arguments are captured as answers, as illustrated in Table 1. In addition to syntactical arguments, which are commonly available in prior semantic or syntactic representations, QA-Adj is designed to capture and explicate implicit arguments not immediately discernible from syntax, for example, stating (in Table 1) *Compared to what is something **more relevant**?*.

The main contributions of this paper are as follows: (1) We formulate a QA-based representation for capturing adjectival arguments, grouping them into four semantic categories; (2) we present a method for collecting low-cost, high-quality QA-Adj data through controlled crowdsourcing; (3) we create a QA-Adj dataset, comprising over 5K sentences and 12K QA pairs, assess its quality, and compare it to PropBank annotations for adjectival predicates (Bonial et al., 2014); (4) we finetune a baseline QA-Adj parser and evaluate its performance, providing a foundation for future model development.

Overall, our work provides an intuitive QA-based representation for explicitly capturing the semantics of adjectives, as well as a dataset and a parser for future research.

2 Background

2.1 Semantic Representations of Adjectives

Traditional logical approaches denote adjectives, as well as verbs, as predicates over entities — e.g., $\text{RED}(x) \wedge \text{BALL}(x)$ would represent *a red ball*, and $\text{AFRAID}(x, y)$ may denote that x *is afraid of* y. NLP semantic formalisms, however — such as PropBank (Palmer et al., 2005), Minimal Recursion Semantics (Copestake et al., 2005), semantic dependencies (Oepen et al., 2015) and more (Banarescu et al., 2013; Abend and Rappoport, 2013, inter alia) — commonly adopted the Neo-Davidsonian approach (Parsons, 1995). This approach decomposes predicative meaning into a set of binary relations between entities and events, labeled by semantic roles, e.g. $\text{FEAR}(e) \wedge \text{AGENT}(e, x) \wedge \text{THEME}(e, y)$.

While the SRL task has gained substantial attention, research thereof focuses primarily on the semantics of verbs or eventive nouns. Nevertheless, several computational resources include adjectives under their scope. In FrameNet (Baker et al., 1998) — a well-known SRL formalism — adjectives are listed in frames with their participants, or Frame Elements, in the same way verbs and nominals do. For example, the adjective *hungry* is listed under the BIOLOGICAL-URGE frame. Similarly, with the goal of complementing PropBank with information about new predicate types, Bonial et al. (2014) annotated adjectives in the PropBank corpus using both pre-existing and newly introduced framesets, along with corresponding semantic roles (see annotation examples in Table 2). In contrast, the formulation presented in this work targets four broad, generic semantic dimensions pertaining to any adjective, coupled with corresponding question templates, and does not require mapping adjectival predicates to a pre-defined inventory of frames. In Section 5.5, we further compare our approach to prior representations.

2.2 QA-Based Semantics

Semantic Role Labeling (Palmer et al., 2010) is typically perceived as answering argument role questions, such as *who*, *what*, *to whom*, *when*, or *where*, regarding a target predicate. For instance, PropBank's ARG0 for the predicate *say* answers the

Sentence	QA-Adj	PropBank
(1) There's so much punch packed into this combination that it's almost **scary**.	Object: What is scary? — There's so much punch packed into this combination + it Extent: To what degree is something scary? — almost	ARG0: it ARGM-EXT: almost
(2) Although these new rockets are probably more **expensive**, they will be able to go at a much greater range than it's shuttle cousins.	Object: What is expensive? — these new rockets + they Extent: To what degree is something expensive? — more expensive Comparison: Compared to what is something expensive? — it's shuttle cousins	ARG1: these new rockets ARGM-EXT: more
(3) The 69 year old Dr. Lopez was found **guilty**.	Object: Who is guilty? — The 69 year old + Dr. Lopez	ARG1: The 69 year old Dr. Lopez
(4) Wise decision to go through the private sector – NASA's budget may be kinda **tight** to fund a project like this.	Object: What might be tight? — NASA's budget Extent: To what degree might something be tight? — kinda Domain: What might something be tight to do? — to fund a project like this Comparison: Compared to what might something be tight? — the private sector's budget	ARG0: NASA's budget ARGM-EXT: kinda ARGM-PRP: to fund a project like this
(5) If anyone is interested in listening to this song, and in offering their opinion whether it be **positive** or negative, I'd appreciate it.	Object: What might be positive? — their opinion + it Domain: What might something be positive about? — this song	ARG1: it
(6) If you have any questions please feel **free** to call me (after Sat. the 26th, when I will return from a trip).	Object: Who is free to do something? — you Domain: What is someone free to do? — call me	ARG3: call me (after Sat. the 26th, when I will return from a trip).
(7) Should the Arctic Ocean become ice **free** in summer, it is likely that polar bears would be driven toward extinction.	Object: What might be free? — the Arctic Ocean	ARG1: the Arctic Ocean ARG2: ice
(8) That is what is about to happen with Judge Samuel Alito, in my opinion, because he has one tragic flaw – a very serious blind spot in his thinking – which makes him completely **unacceptable** for the position of Supreme Court Justice.	Object: Who is unacceptable for something? — Judge Samuel Alito + he Extent: To what degree is someone unacceptable? — completely Domain: What is someone unacceptable for? — the position of Supreme Court Justice	ARG1: him ARGM-EXT: completely ARG3: for the position of Supreme Court Justice
(9) She doesn't have the funds to continue without the grant, and without these treatments, her prognosis is **grim**.	Object: What is grim? — her prognosis without these treatments	ARG1: her prognosis ARG-MNR: without these treatments

Table 2: A sample of QA-Adj annotations, along with corresponding PropBank annotations for adjectives (Bonial et al., 2014, see §5.5 for the comparison). The **+** sign denotes multiple answers for the same question. While most QA-Adj QAs are similar to PropBank predicate-argument relations, many introduce additional information, including implicit or inferred relations (Ex. 2, 4, 5) and within-sentence coreference (Ex. 1, 5, 8). Annotation mistakes are rare, but include incorrect splitting of arguments (Ex. 3), incomplete QA-Adj answers (Ex. 6) and recall misses (Ex. 7).

question "Who **said** something?". QA-SRL (He et al., 2015) suggests that answering role questions is an intuitive means to solicit predicate-argument structures from non-expert annotators. In QA-SRL, annotators are presented with a sentence in which a target predicate has been marked, and are asked to generate questions and highlight the corresponding answers from the sentence. A question captures the semantic role, whereas answers to the question — which are spans from the sentence — denote the set of arguments associated with that role. The QA-based approach allows for a transparent representation, as the questions and answers can be understood by non-experts while providing an explicit account of the underlying meaning of the sentence. This laymen-intuitive definition of roles covers traditional cases of syntactically linked arguments, but also additional semantic arguments clearly *implied* by the sentence meaning (Roit et al., 2020).

QA-SRL has been demonstrated to be beneficial for various downstream tasks. It was shown to subsume open information extraction (OIE) (Stanovsky and Dagan, 2016b), making it possible to construct large supervised OIE dataset (Stanovsky et al., 2018) to serve as an intermediate structure for end applications. Additionally, QA-SRL and related QA-based semantic annotations (Michael et al., 2018) were shown to provide beneficial semantic signal through indirect supervision, resulting in improved performance on downstream tasks for modern pre-trained-LM encoders (He et al., 2020). Recently, QA-SRL was explicitly utilized as an intermediate representation for aligning predicate-argument relations across texts (Brook Weiss et al., 2021) and for detecting analogies through structure mapping (Sultan and Shahaf, 2022).

To address a broader semantic scope, the QA-SRL formalism, well suited for scalable crowdsourcing (FitzGerald et al., 2018), has been incrementally extended to account for discourse relations using semi-templated questions and answers (Pyatkin et al., 2020) as well as for deverbal nominalizations (Klein et al., 2020). These tasks, jointly denoted *QASem*, have been recently bundled by a unifying modeling framework and parsing tool (Klein et al., 2022a). In the QASem framework, each propositional predication relation — in the spirit of the aforementioned Neo-Davidsonian approach — is captured through a corresponding Question-Answer pair. In this work, we further

Question Type	PREFIX	WH	AUX	SBJ	DET	TRG	PP	OBJ	?
Object		Who	are		the	most suitable	for	something	?
Domain		What	is	someone		active	in		?
Comparison	Compared to	what	is	something		prominent			?
Extent	To what degree		is	something		popular			?

Table 3: Example questions illustrating our question templates.

extend the QASem paradigm to account for adjectives.

3 Task Formulation

In order to keep the task simple — both for annotation and for modeling — we consider all adjectives occurring in the sentence under the same formulation. We thus refrain from distinguishing different classes of adjectives (e.g. subsective, intersective or privative (Partee, 2007; Pavlick and Callison-Burch, 2016); superlative and comparative; etc.) or different syntactic realizations of adjectives — i.e. attributive vs. predicative (*the red ball* vs. *the ball is red*).

While free-formed questions have been proposed as a natural representation of semantic relations (Michael et al., 2018), prior works show that they yield inferior coverage relative to annotation schemes that systematically design restricted question templates, such as QA-SRL and QADiscourse (Pyatkin et al., 2020; Klein et al., 2020). Consequently, we adopt the template-based approach and design question templates corresponding to four core argument types of adjectival semantics that have practical value for downstream applications. The coverage of these templates was validated through the examination of prior linguistic works on adjectives (Huddleston and Pullum, 2002; Baker et al., 1998). See Table 3 for an illustration of each question template.

The most basic argument role for an adjective is the entity described by it, which corresponds to the predicated entity variable in logical representations and is captured by all other representation schemes as well. Our annotation scheme captures this argument role through the first question type (*What/Who is [ADJ]*), termed here **Object**.

In addition to Object, we adopt the three semantic dimensions of adjectives as identified by Ikeya (1995), namely — the *Thematic* dimension, the *Comparative* dimension, and the *Degree* dimension.

The *Thematic* dimension is mapped to the **Domain** question type in our scheme. Answers to this question type give a semantic specification to the adjective — For example, *good at dancing*, *lactose intolerant* and *former president*. To illustrate, two-place predicates (in first-order logic) would mostly fit their arguments into our Object and Domain roles. While often syntactically attached to the adjective, such answers can also occur as implicit arguments (Ex. 5 in Table 2).

The **Comparison** question type is aimed to capture the group or entity referenced by the adjective to which the object is being compared. These arguments are frequently implicit (e.g. Ex. 2, 4 in Table 2) and are therefore mostly neglected in prior formalisms that rely on syntax, such as PropBank.

Lastly, the **Extent** question type corresponds to the *Degree* dimension, that is, to what extent does the adjectival assertion holds. Such arguments can be realized by adverbs (e.g. *almost complete*, *very good*) or by more complicated constructions (*too political for my liking*, *competent enough for this job*, etc).

We note that our questions are designed at capturing semantic complements of the adjective meaning. In preliminary investigations, incorporating adverbial modifiers into the task scope was found to introduce annotation noise. Our role set thus omits adverbial modifiers, such as time and location (e.g. *By June, you'll be **capable** of programming by yourself*), leaving their investigation for future work.

In this paper, we focus on "core" adjectival arguments as laid down by Ikeya. See Appendix A.1 for a more elaborated discussion.

QA Format In the spirit of He et al. (2015), we define a small grammar over possible questions. The questions are constrained by a template with eight fields, $q \in$ **PREFIX** $\times$ **WH** $\times$ **AUX** $\times$ **SBJ** $\times$ **DET** $\times$ **TRG** $\times$ **PP** $\times$ **OBJ**, each associated with a set of possible options (see Table 3). Full descriptions for each field are provided in Table 8 in the Appendix.

Answers are selected from the words in the sentence but can be manually modified in order to make the answer appropriate and natural-sounding.

	Sentences	Adjectives	Total Roles		Object		Domain		Comparison		Extent	
			QAs	Answers	QAs	Answers	QAs	Answers	QAs	Answers	QAs	Answers
Train	3377	7266	8198	9080	6802	7654	613	627	412	426	371	373
Dev	668	750	951	1099	733	872	90	93	80	85	48	49
Test	1281	1659	2093	2398	1622	1914	176	178	189	199	106	107
Total	5326	9695	11242	12577	9157	10440	879	898	681	710	525	479

Table 4: Annotation statistics of the QA-Adj dataset.

We instruct the annotators to rewrite answers manually only when copying words from the sentence is insufficient for constructing a meaningful or grammatical answer, such as in Ex. 4 in Table 2 (*the private sector**'s** budget*). In addition, questions may have multiple answers, in order to better account for coordinations or co-referring entity mentions (Ex. 1, 5, 8 in Table 2).

We further guide our annotators to include restrictive modifiers (Stanovsky and Dagan, 2016a) in their answers, as these are considered an integral part of the noun phrase, e.g., the underlined modifier in *"She wore the **shiny** necklace that her mother gave her"*. Non-restrictive modifiers, which provide parenthetical information about the entity — e.g., *"The speaker thanked **former** president Obama, who just walked into the room"* — are not included in the answer span.

4 Dataset Construction

Preprocessing and annotation interface In this section, we describe the dataset creation process and in section 5 analyze its quality. We annotated over 5K sentences with 9K adjective mentions, across two domains: Wikinews and Wikipedia. We select sentences that are also covered by previous annotated QASem datasets (Roit et al., 2020; Pyatkin et al., 2020; Klein et al., 2020). In each sentence, we identify the target adjectives using SpaCy's POS-tagger. If an adjective is preceded by one of the words 'more', 'less', 'most', or 'least', then it is considered part of the target adjective. Table 4 shows the full data statistics.

We developed a Graphical User Interface (GUI) (See Appendix, Figure 1) deployed at Amazon's Mechanical Turk crowdsourcing platform. The worker, presented with a sentence with a marked adjective as a target, should generate question-answer pairs pertaining to this adjective. Questions are generated by filling templated slots using drop-down lists, whereas answers are selected by highlighting spans from the sentence, and manually corrected if needs be. The GUI also includes a short overview of the task and instructions, along with 5 annotation examples.

Annotator selection and training We adapted the controlled crowdsourcing process used by Roit et al. (2020) for QA-SRL. After establishing the task formulation and interface, the first two authors jointly annotated 60 instances as a seed gold set, for evaluating and guiding worker qualification. We then release a preliminary crowd-wide annotation round and contact workers who exhibit reasonable performance. They are asked to review our short guidelines, which highlight a few subtle aspects, and then annotate four qualification rounds, of 15-30 target adjectives each. Each round is followed by extensive feedback via email, pointing at errors and missed arguments, which are identified by automatic comparison to expert annotation. In total, this worker training process lasted approximately 8 weeks, and cost 240$, and is orders of magnitude shorter and simpler than training annotators for traditional semantic formalisms.

Annotation process During data collection, we observed that outcomes of a single crowd annotator tend to be of insufficient quality, especially with respect to capturing the rather infrequent roles of Domain, Comparison and Extent. To enhance the coverage of the evaluation set (dev & test), we aggregated QAs from two independent QA-generation workers and forwarded them to *consolidation*. In the consolidation task, a third worker reviewed and judged the aggregated generated annotations, producing a non-redundant consolidated set.

While aggregating annotations from multiple generators coped well with the coverage challenge, data precision was still mediocre for the non-Object roles, as opposed to the Object role, where precision was satisfactory (in Section 5.4 (Table 5) we report evaluated quality for each phase of the data collection process.). We hence employed an additional *expert verification* step pertaining to instances in which one of the non-Object arguments is provided. In this step, one of the first two authors of this paper reviewed the annotations, filtering or

	Generation (avg. of 2 workers)			Consolidation			Expert Verification (avg. of 2 experts)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Object	83.7	78.5	81.0	87.7	93.4	**90.4**	-	-	-
Domain	46.1	64.4	53.7	43.4	82.6	56.9	93.3	84.9	**88.9**
Comparison	61.4	44.7	52.6	64.1	75.4	69.2	91.7	77.1	**83.7**
Extent	49.5	67.7	57.1	67.5	80.6	73.4	86.6	80.6	**83.4**
Total	72.3	72.2	72.2	75.8	89.3	82.0	88.0	89.5	**88.8**

Table 5: Evaluating the different annotation stages against an expertly annotated reference set of 300 instances (See §5.2 for evaluation metrics). **Bold** numbers represent the final stage in the annotation process of the dev & test sets.

fixing answers to non-Object questions as required. Verification was applied on top of consolidated annotations for the dev and test sets (1010 out of 2409 adjective instances), and over single-generator annotations for the training set (2182 out of 7266 adjective instances).

Annotation cost Our annotators were paid 20¢ per instance in both the generation and consolidation steps, and a single expert verification assignment takes around 30 seconds. The resulting cost per instance in the development and test sets is 60¢ (2 generators + consolidator), along with around 30 seconds of expert review time. In the training set, the cost is 20¢ and 30 expert review seconds per instance. In total, creating the development and test sets costs 1445$ and 9 hours of expert verification, while the training set costs 1456$ and 19 expert hours, totalling 2891$ and 28 hours of expert review time for the entire dataset. This approach allowed us to efficiently collect a high-quality dataset for our QA-based representation of adjectival semantics.

5 Dataset Analysis and Quality

In this section, we report several analyses to quantify and establish the quality and coverage of the QA-Adj dataset. Additional Information about the joint distribution of different roles is reported in Appendix A.4.

5.1 Implicit Arguments

A key benefit of our laymen-intuitive annotation task is its aptitude to capture implicit arguments, that is, arguments that are harder to automatically read off of syntax (See Ex. 2, 4, 5 in Table 2 for illustrations). To quantify this aspect, we utilize a syntactic dependency parser[1] for measuring the proportion of implicit arguments on the evaluation sets. Following similar prior analyses (Klein et al., 2020), an argument is considered implicit if none of its words is connected to the predicate on an undirected dependency tree in a path of length ≤ 2.

We find that 17%, 30%, 49%, and 13% of the arguments are implicit for the Object, Domain, Comparison, and Extent roles, respectively. This demonstrates that many of our semantic arguments are hardly accessible from syntactic representations, especially for the Comparison and Domain roles. Focusing on the Object role, we further inspect that 91% of the instances have at least one explicit argument, which entails that most of the implicit arguments provide a (commonly more informative) coreferring mention of a syntactically-connected argument (e.g. Ex. 8 in Table 2).[2] In the remaining 9%, the object entity is connected through more complex linguistic constructions such as control or raising verbs (*The audience is asked to remain* ***silent***), adverbial clauses (*Argentina dropped three places to be ranked* ***sixth***) and coordinations (*Switzerland and Italy each moved down one, ranked eighth and* ***ninth*** *respectively*). In sum, relying on intuitive laymen annotations naturally yields many informative arguments that fall out of scope of more linguistically oriented representations.

5.2 Evaluation metrics

We use the same evaluation protocol both for dataset analysis (this section) and for model evaluation (Section 6). Given predicted QAs for all adjectives, we report precision and recall against the ground truth for each question type separately, as well as for the total set of predicted QAs. Following previous work on annotating semantic relations with QA pairs (Roit et al., 2020; Pyatkin et al., 2020; Klein et al., 2020), answers of the same question type are considered a match if the intersection over union (IOU) between the sets of tokens in each answer is greater than 0.3.

[1] We apply the same SpaCy model used for POS-tagging in preprocessing.

[2] In general, while multiple answers are rare for other role questions, 16% of the Object questions are answered by more than one answer, most commonly due to coreferring mentions.

5.3 Inter-Annotator Agreement

To assess the consistency of the annotated data, we measure the inter-annotator agreement on the dev & test sets, as well as expert-vs-expert agreement on data used as part of the validation and test sets for parser evaluation. The **Object** QA type macro-averaged F1 inter-annotator agreement is **76.0**, while for QA types **Comparison, Domain, Extent** it is **29.8, 37.0, 41.3**, respectively.

The main issue in disagreement arises from sentences that do not contain apparent adjectival arguments, especially in question types Comparison and Domain, where workers are inclined to ask questions either way, resulting in sometimes unnatural or overly implicit questions. To measure the expert-vs-expert agreement, we randomly sample 227 instances that underwent the consolidation process and contain at least one of the Comparison, Domain, or Extent roles. We perform the expert review step on them by each of the first 2 authors of this paper and compare the outputs. The expert-vs-expert F1, excluding the Object question type which was not reviewed in the expert-review step, reaches a reasonable **77.9** F1. Notably, the consolidation and expert review steps boost consistency significantly.

Agreement on restrictive modifiers We conjecture that a decent proportion of annotator disagreements arise from the difficulty to designate the proper argument span, which requires keeping within the span restrictive modifiers of the argument while omitting non-restrictive modifiers (Stanovsky and Dagan, 2016a). Therefore, we estimate the agreement between annotators on modifiers' restrictiveness by sampling from the final dataset 50 answers of the Object role that contain a restrictive modifier, as judged by the first author, and examining whether both annotators captured it. 26 modifiers were captured by both annotators, mostly simple prepositional phrases (e.g. *routes for complex molecules*), while 16 were captured only by one of the annotators. 8 were missed by both, but captured by the consolidator. Examining the missed modifiers, we find that many involve non-continuous span selection (which is feasible through the manual modification our interface enables on top of a copied sentence span). For example, in the sentence *"The alveolar letters had longer left stems, while retroflexes had* ***longer*** *right stems"*, the correct argument is *right stems of retroflexes*, while the annotators only captured *right stems*, omitting this implicit restrictive modifier which is nonetheless essential for demarcating the precise argument.

5.4 Dataset Assessment by Gold Reference Set

To ensure the quality of our annotation, we created a gold reference set consisting of 300 instances from the development set. The reference set should represent QA-Adj annotations of optimal quality. For this purpose, we take generated annotations along with their consolidation decisions (as described in §4) and manually correct them by each of the two first authors independently. We then reconciliate to resolve any disagreement.

We compare the annotations attained from the initial generation step, consolidation step, and single expert verification step against the reference set (Table 5). Results indicate that consolidation significantly boosts coverage, and confirm the high quality of our full annotation protocol (in bold).

5.5 Comparison with Other Formalisms

In this section, we compare QA-Adj to two common representations covering adjectival semantics — PropBank (Palmer et al., 2005) and Abstract Meaning Representation (AMR; Banarescu et al., 2013).

PropBank for adjectives One of the most widely used resources of English predicate-argument structure is PropBank, which has also incorporated adjectival predicates (Bonial et al., 2014). It is thus illuminating to examine the overlap and discrepancies between QA-Adj and PropBank. For this purpose, we collect QA-Adj annotations for 150 adjective instances from PropBank using the same annotation protocol as for the evaluation set (§4), yielding 296 answers (260 QAs) compared against 232 PropBank arguments. We employ our evaluation protocol (§5.2) to measure argument agreement between the two annotation schemes, and manually examine disagreements. Examples throughout this section are referring to Table 2.

Notably, the scope of adjectival arguments targeted by the two annotation schemes is somewhat divergent. Designed to explicate the syntactic-semantic interface, PropBank captures some syntactic markers (e.g. discourse, relative clause, negation, and modality) that cannot naturally answer role questions. It is worth mentioning that QA-Adj annotations incorporate information about negation and modality within the questions (see Ex. 5),

Sentence	Test set	Parser output
(1) Any deviation from this family model is considered a "**nontraditional** family".	Object: What is nontraditional? — a family + any deviation from this family model Comparison: Relative to what is something nontraditional? — this family model	Object: What is nontraditional? — family
(2) Regarding the lack of women members in the cabinet, Mr. Abbott said he was "**disappointed**".	Object: Who was disappointed? — he + Mr. Abbott Domain: What was someone disappointed about? — the lack of women members in the cabinet	Object: Who was disappointed? — Mr. Abbott Domain: What was someone disappointed about? — the lack of women members in the cabinet

Table 6: Comparison between QAs in the test set and the parser's output. Example 1 demonstrates an implicit argument that the parser missed, while in Example 2, the parser captured such an argument.

following the QA-SRL approach. In addition, PropBank includes many types of adverbial modifiers that are out of QA-Adj scope (§3). We thus exclude PropBank roles that pertain to syntactic markers or adverbials from our henceforth quantitative analysis (details in Appendix A.5) and focus on core argument roles.

QA-Adj covers **93.1**% of PropBank arguments, demonstrating that our task formulation and annotation substantially capture traditional predicate-argument relations. One source of disagreements are pronouns (Ex. 8), which PropBank captures in the form they appear in the sentence (e.g. *him*), while our flexible rewriting mechanism allows to capture them in the more natural subject form (i.e. *he*). Out of 16 PropBank arguments not covered by QA-Adj, only 6 reflect actual QA-Adj annotation misses (Ex. 7). Another source of disagreement (4 out of 16) is QA-Adj arguments that are split into multiple roles in PropBank's finer-grained annotation (Ex. 9).

On the other hand, PropBank arguments cover only **72.9**% of QA-Adj annotated answers. Out of 80 QA-Adj arguments that don't match PropBank annotations, 70 are correct but fall out of PropBank's scope. These include co-referring mentions (14; Ex. 1, 2, 5), implicit arguments (22; Ex. 2, 4, 5), and cases where PropBank arguments are split by our scheme into two distinct, co-referring answers (11; See Ex. 3).

While this analysis elucidates the relationship between QA-Adj and a more traditional semantic formalism, it also reaffirms the coverage of our QA-Adj annotations, demonstrating that non-experts can capture a major portion of the information found in PropBank. At the same time, relying on intuitive NL-based QAs introduces new types of implicit information that seem useful downstream, in addition to making the annotations cheaper, faster, and easier to replicate compared to expertly annotated formalisms.

Abstract Meaning Representation AMR is a comprehensive semantic representation designed to capture semantic aspects of complete sentences, including adjectival semantics, in an abstract, cross-language manner. It employs various mechanisms to account for adjectival semantics. For instance, the phrase "attractive spy" is represented with the corresponding verbal roleset, SPY :ARG0-OF ATTRACT-01, while for other adjectives, AMR defines specific framesets (e.g. SAD-02). The specification for semantic roles is predicate-specific, where, to cite AMR guidelines, "ARG0 often refers to the thing being described by the adjective, while ARG1 names the next most natural argument.".[3] These correspond to QA-Adj Object and Domain roles in most cases.

A later study (Bonial et al., 2018) expands the AMR lexicon with various constructions, including a HAVE-DEGREE-91 roleset, which handles degree adjectives and related constructions. Upon a close inspection, we find that QA-Adj Comparison and Degree roles capture most of the information within the HAVE-DEGREE-91 roles, though in a more coarse manner. See Appendix A.6 for an elaboration on the comparison to AMR.

6 Baseline Models

We devise an initial QA-Adj parser to serve as a baseline for future work on this task. We first apply the same preprocessing steps for identifying target adjectives as in our data collection procedure (§4). Then, following a prior QA-driven semantic parser (Klein et al., 2022a), we fine-tune the Text-to-Text Transformer model (T5; Raffel et al., 2020), which unifies multiple text modeling tasks, and achieves state-of-the-art results in various NLP benchmarks. We use Huggingface (Wolf et al., 2020) for fine-tuning the T5 model. A special token is marking the target adjective within the input sentence, while

[3] https://github.com/amrisi/amr-guidelines

Model Evaluation	Single Model Automatic			Role-specific Models Automatic			Role-specific Models Manual		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Object	82.2	75.1	78.5	78.2	75.4	76.8	86.3	84.4	85.3
Comparison	36.8	43.7	40.0	36.2	47.7	41.2	60.0	45.2	51.5
Domain	50.5	51.1	50.8	51.8	55.0	53.4	62.5	60.9	61.6
Extent	41.7	57.0	48.2	80.0	52.6	63.2	85.0	57.5	68.5
Total	72.5	69.9	71.2	71.5	70.6	71.0	73.7	62.5	67.6

Table 7: Baseline models evaluation. Automatic evaluation results are on the full test set, while manual evaluation is on a sample.

the output is formatted as *question: <**Q**> answer: <**A**>*. In case the semantic role is empty, the parser is to generate the special token [NO-QA].

In preliminary experiments, training a single model to generate all four QA pairs in one go has yielded poor results. We hypothesize this is due to the sparsity of the Domain, Comparison and Extent question types, which appear in 8%, 5%, and 5% of the training examples, respectively.

Therefore, to set up baseline results, we fine-tune an independent T5 model on each question type separately. The train set per question type consists of all instances which have the specific question type answered, along with random negative samples, i.e. empty QA instances. The ratio of negative samples is treated as a hyper-parameter of the model and is optimized on the development set.

Since holding a separate fine-tuned T5 model for every QA type is memory-consuming, we also fine-tune a single T5 model using the union of the training sets of each question type, using a different prefix for each QA type.

6.1 Results

Previous work on QA-based semantics has demonstrated that automatic argument-matching criteria can be too strict (Roit et al., 2020). Hence, to better estimate precision, we randomly select 40 generated QAs for each question type and assess their validity manually. Similarly, to estimate recall, we sample 40 annotated QAs of each question type and manually compare them to the parser's output.

Table 7 presents the automatic evaluation measures for a single parser trained on all roles, as well as automatic and manual evaluation of the role-specific models. Results indicate there is ample room for improvement, particularly on the more subtle roles of Comparison and Domain.

One factor contributing to the challenges in capturing these roles is the high prevalence of implicit arguments within them (Ex. 1 in Table 6), as demonstrated in our analysis (Section 5.1). As implicit arguments often rely on commonsense reasoning rather than syntactic structure, they may be more difficult for a model to identify. In future work, we aim to investigate methods for better capturing implicit arguments and explore the use of external knowledge sources to aid in this task.

7 Conclusion

In this work, we propose and realize a new approach to representing the semantics of adjectives using natural language question-answer pairs, focusing on four generic, core semantic dimensions. This intuitive representation enables high-quality yet scalable annotation through controlled crowdsourcing along with minimal expert verification. Our annotations explicate the fundamental aspects of an adjective's meaning in context, substantially overlapping with an expertly annotated SRL resource while adding previously uncovered implicit arguments.

We advocate utilizing QA-Adj downstream as an alternative for syntactical or semantic representations. As an example, recent works on aspect-based sentiment analysis use syntactic or semantic dependencies as scaffolds for enhancing domain transfer (Wang and Pan, 2019; Pereg et al., 2020; Klein et al., 2022b). Explicating relations between adjectives (sentiment/opinion terms) and their semantic objects (aspect terms) directly, QA-Adj is a worthwhile alternative to dependency representation.

Future works should explore methods for improving the baseline models presented in this work, such as prompt tuning (Lester et al., 2021) or multi-task learning with related QA-semantic tasks (Klein et al., 2022a). In addition, since the annotations are based on natural language and layman workers, it is appealing to transfer the scheme into various languages, possibly utilizing both machine translation and/or crowd annotations.

8 Limitations and Ethics

Unlike prior QASem annotation tasks, we empirically find that adding an expert verification step on a selective portion of the data — where more subtle roles are handled — is important for maintaining good precision. Indeed, despite putting efforts in making the task guidelines simple and intuitive, margins of the semantic space often introduce complexity that is hard to account for in a consistent manner without linguistic background. Although still significantly faster than full-fledged expert annotation, requiring an expert in the loop may pose a bottleneck to scaling annotations to large datasets and new domains and languages, which is a shortcoming of the current proposal.

Annotations were conducted on Amazon Mechanical Turk (MTurk) with an average pay of $12 per hour for all crowdsourcing data collection tasks. To maintain the anonymity of our workers, we do not collect personal information and do not keep any deanonymizing information such as MTurk IDs.

License The data collected in this work is licensed under the Creative Commons license.

Acknowledgements

This research was funded in part by grants from Intel Labs, the Planning and Budgeting Committee (PBC) of the Israeli Council for Higher Education under the National Data Science Competitive Program, and the Israel Science Foundation grant 2827/21.

References

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

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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

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*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019, Reykjavik, Iceland. European Language Resources Association (ELRA).

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9879–9894, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2021. Robust question answering through sub-part alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1251–1263, Online. Association for Computational Linguistics.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 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, Melbourne, Australia. Association for Computational Linguistics.

Mantas Gavenavicius. 2020. Evaluating and comparing textual summaries using question answering models and reading comprehension datasets. B.S. thesis, University of Twente.

Hangfeng He, Qiang Ning, and Dan Roth. 2020. QuASE: Question-answer driven sentence encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8743–8758, Online. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Akira Ikeya. 1995. Predicate-argument structure of English adjectives. In *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*, City University of Hong Kong, Hong Kong. City University of Hong Kong.

Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022a. Qasem parsing: Text-to-text modeling of qa-based semantics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7742—7756, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ayal Klein, Oren Pereg, Daniel Korat, Vasudev Lal, Moshe Wasserblat, and Ido Dagan. 2022b. Opinion-based relational pivoting for cross-domain aspect term extraction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 104–112, Dublin, Ireland. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Muhidin Mohamed and Mourad Oussalah. 2019. Srl-esa-textsum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing Management*, 56(4):1356–1372.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. 2015. SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado. Association for Computational Linguistics.

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

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*, 1st edition. Morgan and Claypool Publishers.

Terence Parsons. 1995. Thematic relations and arguments. *Linguistic Inquiry*, pages 635–662.

Barbara H Partee. 2007. Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive foundations of interpretation*, pages 145–161.

Ellie Pavlick and Chris Callison-Burch. 2016. So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

Oren Pereg, Daniel Korat, and Moshe Wasserblat. 2020. Syntactically aware cross-domain aspect and opinion terms extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1772–1777, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Gabriel Stanovsky and Ido Dagan. 2016a. Annotating and predicting non-restrictive noun phrase modifications. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1256–1265, Berlin, Germany. Association for Computational Linguistics.

Gabriel Stanovsky and Ido Dagan. 2016b. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2019. Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. *Computational Linguistics*, 45(4):705–736.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

A Appendices

A.1 Omitting Adverbial Modifiers

As mentioned in the paper body (§3), our role questions are designed toward semantic aspects that complement the meaning of the adjective. Consequently, in the spirit of the famous linguistic *argument//modifier* distinction (or *complement//adjunct*), we choose not to incorporate questions targeting generic ("adjunctive") adverbial information, such as temporal, causal, or locative modifiers of the copular phrase.

Field	Description	Values
PREFIX	Specific question type prefixes	Compared to, Relative to, To what degree
WH*	Question words	who, what
AUX*	Auxiliary verbs	is, was not, could be, ...
SBJ	Place-holder for subject position	someone, something
DET	Determiner	the
TRG*	The target adjective	tall, active, most accurate, ...
PP	Frequent prepositions	by, for, in, ...
OBJ	Placeholder for object position	someone, something

Table 8: The fields of question templates. WH, AUX and TRG are required; all other fields may be left empty.

This design choice arises from practical considerations. In preliminary investigations and crowdsourcing experiments, we have found the distinction between modifiers and the **Domain** role to be rather intricate, especially for non-linguist annotators. For example, locative or temporal descriptions that are commonly adverbial modifiers (*He was **hungry** this morning*) can in certain cases be semantic complements (***earlier** this morning*). Further, when supporting modifier questions like "When is something [ADJ]?" in the interface, non-expert annotators are often inclined to embrace loosely related and erroneous phrases as arguments. To illustrate, the instance *"If you have any questions please feel **free** to call me (after Saturday the 26th)"* (from Ex. 6 at Table 2) might be annotated with the inaccurate QA *"When should someone be free? — after Saturday the 26th"*.

A.2 Question Templates Description

Table 8 shows a full description of the 8 question fields comprising the four question templates (one per role), together with possible values that can fill each question field. Each field's exact set of optional values defines the role-dependent question template. In the table, we use three dots (...) to denote partial lists of values (the full lists would be released as supplementary material upon acceptance).

A.3 Annotation Task User Interface

Our Graphical User Interface (Figure 1) allows to create multiple answers per QA type (+ button). Upon answering the Object question, its answer is embedded in the other questions, making them more natural.

A.4 Arguments Joint Distribution

The sparsity of arguments corresponding to the question types Domain, Comparison, and Extent is a major challenge in our task and data (See Table 4). Indeed, as demonstrated in Section 6.1, this sparsity makes it difficult for a parser to accurately identify these roles. In our development set of 750 adjective instances, most (547) have only the Object question answered. There are 157 instances with two answered questions, 26 instances with three, and only three instances with all four questions answered.

A small minority of instances has no argument roles at all (17 out of 750 on dev). This is primarily due to POS-tagger erroneous adjective identification — for example, *Khufu's pyramid **complex** consists....* Annotators were instructed to leave empty such erroneous target adjectives, where our roles questions are not sound.

A.5 PropBank Roles Excluded from Comparison

Following our discussion in Appendix A.1, we need to account for the scope discrepancy between QA-Adj and PropBank prior to measuring their argument agreement. We thus exclude PropBank arguments capturing adverb, causation, temporal, location and relative clause roles, as well as markers of discourse, modality, and negation. The full list of PropBank's excluded roles, along with examples, can be seen in Table 9.

A.6 More details about AMR Comparison

Predicative vs. Attributive Adjectives AMR maintains a directionality distinction between predicative adjectives (*The marble is white*) and attributive adjectives (*The white marble*). Predicative adjectives would be the "root" of the sentence graph

Role Name	Argument	Sentence
ARGM-ADV	your career	The one department of life that may not quite be as hopeful as you'd like could be your career, where advancement may be slow and satisfaction **rare**.
R-ARGM-ADV	where	
ARGM-LOC	areas	There have been large numbers of population extinctions in Mexico and southern California in areas where the habitat is still **acceptable**.
R-ARGM-LOC	where	
ARGM-CAU	reasons	Also, from their own webpage, reasons why NASA is **important**, in a 5th-grade format.
R-ARGM-CAU	why	
R-ARG1	who	80 - Percentage of the Iraqi workforce who are **unemployed** a year after the war.
ARGM-MOD ARGM-NEG	may not	They may not be **familiar**, but they will be fascinating.
ARGM-TMP	when we love another person	We become most fully **human** when we love another person.
ARGM-DIS	Please	Please feel **free** to call me.

Table 9: Examples of PropBank roles omitted from comparison to QA-Adj.

(or clause subgraph), and the subject entity would be their :DOMAIN argument, e.g. WHITE :DOMAIN MARBLE. Attributive adjectives, on the other hand, are denoted as :MOD arguments of their target entity, e.g. MARBLE :MOD WHITE. This distinction is necessary for maintaining a fine-grained account of sentence meaning, as it captures the sentence focus, which may have pragmatical implications. In QA-Adj, and QASem in general, we take a more "informational" perspective on semantics (rooted in more traditional logical representations), thus wishing to abstract out surface realization details that do not modify the conveyed information.

Degree constructions Bonial et al. (2018) expands the AMR lexicon with various constructions. These include a HAVE-DEGREE-91 roleset, which handles constructions related to degree adjectives, such as comparatives, superlatives, or more idiosyncratic constructions, e.g. what they term 'Degree Consequence' (see Table for example annotations). The HAVE-DEGREE-91 roleset comprises the following semantic roles:

- ARG1: domain, entity characterized by attribute
- ARG2: attribute (e.g. tall)
- ARG3: degree itself (e.g. more/most, less/least)
- ARG4: compared-to
- ARG5: superlative: reference to superset
- ARG6: consequence, result of degree.

Compared to our scheme, ARG1 directly corresponds to the **Object** role, while ARG3 and ARG6 correspond to the **Extent** role. ARG4 and ARG5 align with the **Comparison** role. Examples illustrating this mapping are presented in Table 10. This comparison illustrates that the roles defined by our task are less fine-grained than those that can be found, at least in some contexts, in other semantic frameworks like AMR. Our choice of granularity is informed by our objective, aiming to facilitate streamlined non-expert annotation. Nevertheless, the comparison also demonstrates that our four roles adequately cover the most essential semantic roles of adjectival semantics.

Sentence	AMR	QA-Adj
(1) The watch is too **wide** for my wrist.	Arg1: watch Arg2: wide Arg3: too Arg6: my wrist	Object: What is wide? — The watch Extent: To what extent is something wide? too wide for my wrist
(2) The girl is **taller** than the boy.	Arg1: girl Arg2: tall Arg3: more Arg4: boy	Object: Who is taller? — The girl Comparison: Compared to whom is someone taller? — the boy

Table 10: Examples of AMR annotations for adjectives, using the specialized HAVE-DEGREE-91 roleset, along with corresponding QA-Adj annotations.

Figure 1: User interface for the Question-Answer Generation task.