ACCoRD: A Multi-Document Approach to Generating Diverse Descriptions of Scientific Concepts

Sonia K. Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl
Sophie Johnson, Jonathan Borchardt, Daniel S. Weld, Tom Hope, Doug Downey
Allen Institute for Artificial Intelligence
{soniam,kylel,daniel,chandrab,baileyk,
sophiej,jonathanb,danw,tomh,dougd}@allenai.org

Abstract

Systems that can automatically define unfamiliar terms hold the promise of improving the accessibility of scientific texts, especially for readers who may lack prerequisite background knowledge. However, current systems assume a single "best" description per concept, which fails to account for the many potentially useful ways a concept can be described. We present ACCoRD, an end-to-end system tackling the novel task of generating *sets* of descriptions of scientific concepts. Our system takes advantage of the myriad ways a concept is mentioned across the scientific literature to produce distinct, diverse descriptions of target scientific concepts in terms of different reference concepts. To support research on the task, we release an expert-annotated resource, the ACCoRD corpus, which includes 1,275 labeled contexts and 1,787 hand-authored concept descriptions. We conduct a user study demonstrating that (1) users prefer descriptions produced by our end-to-end system, and (2) users prefer multiple descriptions to a single "best" description.

1 Introduction

Readers of scientific papers often encounter unfamiliar concepts, which impedes their understanding (Portenoy et al., 2022). This is because papers assume *a priori* knowledge, and often lack definitions for the scientific terms that they use. While readers may turn to external encyclopedic resources like Wikipedia, these contain descriptions for only a small fraction of scientific concepts (King et al., 2020), which has motivated the development of systems that automatically extract or generate descriptions for scientific concepts. Unfortunately, current systems only surface a single "best" result for all users, which is often extracted from a single input document (Espinosa-Anke and Schockaert, 2018; Vanetik et al., 2020; Veyseh et al., 2019; Kang et al., 2020). The one-best description may
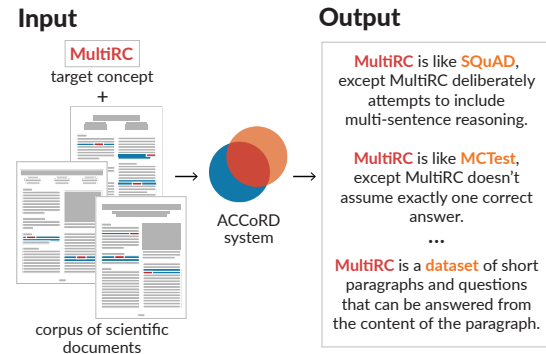

Figure 1: *ACCoRD's approach to the Description Set Generation task.* Given a corpus of papers and a target concept to be described (red, e.g. MultiRC), our system produces a diverse *set* of descriptions. These are generated using mentions of the target concept in terms of myriad other reference concepts (orange) from extracted contexts (blue), resulting in a diverse set of descriptions.

not be accessible for all readers, given varying background knowledge.

Scientific concepts can be described in multiple distinct ways, and our approach is based on our hypothesis that a *set of descriptions* is more useful for users than a single description. Humans learn new concepts by understanding how they relate to other, known concepts (Rumelhart and Ortony, 1977; Spiro, 1980; NRC, 2018), and providing multiple descriptions allows us to highlight multiple such relationships, contributing to a more complete understanding. Furthermore, providing multiple descriptions increases the number of potentially helpful connections between a new concept and concepts within the user's specific background knowledge (see Figure 1), increasing accessibility. This relational approach to human concept learning has been formalized through the lens of Analogical Transfer Theory (Gentner, 1983; Kurtz et al., 2001; Gentner and Goldin-Meadow, 2003) and has long been employed as a tool in scientific discourse and education (Treagust et al., 1992; Heywood, 2002).
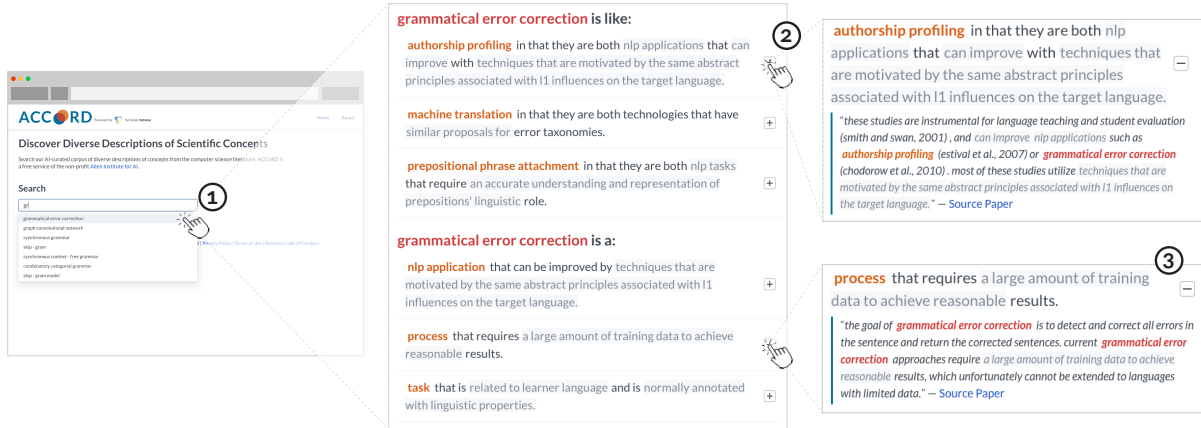
Figure 2: *Demo screenshots.* (1) Users search for a target scientific concept from a pre-defined list and are shown cards for top reference concepts used to describe the target concept in terms of a particular relation. (2) Users click to expand the cards to see the extracted snippet (context) that produced the generated concept description, in addition to a link to the source paper. (3) Spans of text that are shared between the extracted context and generated description are highlighted to facilitate easy comparison.

Our work expands upon the notion of a description in the context of description generation systems to include analogy-like descriptions that are currently not captured by either scientific definition (Kang et al., 2020) or relation extraction (Wadden et al., 2019) systems.

In this work, we present Automatic Comparison of Concepts with Relational Descriptions (ACCoRD) – an end-to-end system that tackles the novel task of producing a set of distinct descriptions for a given target concept.[1] Given text from scientific papers, our system first **extracts** all sentences from across the corpus that contain a description of the concept in terms of any other concept. Then, conditioned on the extractions, ACCoRD **generates** succinct, self-contained descriptions of the concepts' relationship using GPT-3 (Brown et al., 2020) in the few-shot setting. The system finally **selects** a smaller, yet diverse subset of descriptions that captures the richness of a concept's usages by including multiple relation types and reference concepts.

Our contributions are:

1. We introduce Description Set Generation (DSG), the novel task of generating multiple distinct descriptions of a single target concept. In support of this task, we release the ACCoRD corpus, an expert-annotated resource of 1,275 labeled contexts and 1,787 hand-authored concept descriptions.

2. We present ACCoRD, an end-to-end system for DSG that outputs a diverse set of descriptions for concepts in the computer science domain.

3. We conduct a user study demonstrating that users prefer multiple descriptions over a single "best" description, and that they prefer our system's generated concept descriptions over those of an extractive baseline.

2 Description Set Generation

2.1 Task definition

We introduce the task of Description Set Generation (DSG) as follows: Given a large corpus of N scientific documents, a target concept to be described and a desired output size $|S|$, the goal of DSG is to output a set S of succinct, self-contained, and distinct descriptions of the target concept. A schematic of this task is provided in Figure 1. Unlike prior work, which defines the task in terms of a single output description per scientific concept (Jin et al., 2013; Kang et al., 2020), DSG proposes outputting a set of descriptions. One can view DSG as a generalization of the format used in prior work (i.e., single-description outputs are sets with $|S| = 1$).

2.2 Approach

DSG is an open-ended task, and many possible description sets could form valid output for a given concept. To facilitate the generation of descriptions that are useful and factual, in this work, we focus

---

[1]System demo, code, and data set available at github.com/allenai/ACCoRD

on descriptions that meet three criteria: (i) They are derived from an extracted snippet of a scientific document, referred to as the *context*, which contains the target concept. In our experiments, the contexts are limited to 1-2 contiguous sentences. (ii) They must mention another concept, referred to as the *reference* concept, which is mentioned in the extracted context and is related to the target concept by one of the four relations in {is-a, is-like, part-of, used-for}. This relation must also be reflected in the extracted context. (iii) The description must contain an *elaboration*, or a span of text that further explains the specified relation between the target and reference concepts. For example, a description cannot only say that "SQuAD is like TriviaQA" but it must also specify that they "are both reading comprehension data sets." These elaborations must be supported by the associated extracted context.

The description criteria described above enabled us to build a system that produced many descriptions preferred by users, as we show in our experiments. However, the DSG task is more general than our specific formulation, and experimenting with richer description formats in DSG is an important item of future work.

3 Data set

To support work on DSG, we compile and release the ACCoRD corpus. The data set consists of 1,275 labeled contexts and 1,787 hand-authored concept descriptions, and induces diversity among these concept descriptions in two key ways. First, our data set allows for concept descriptions beyond the typical is-a relation. Second, a single target concept is allowed to be described in terms of any number of other concepts in the source text.

3.1 Data set construction

To construct the ACCoRD corpus, we consider the abstract, introduction, and related works sections of 698 CS papers from S2ORC (Lo et al., 2020). We automatically identify candidate contexts of 1-2 contiguous sentences with at least one significant CS concept by performing simple string matching against a high-precision set of CS concepts from ForeCite (King et al., 2020). ForeCite assigns a score to candidate concept string using the citation graph, based on the intuition that when papers containing a string all tend to cite the same paper, that string is more likely to be a bona-fide scientific con-

Extracted context	Hand-authored descriptions
word embedding is a word representation that captures semantic and syntactic similarities between words. it has been widely utilized for a variety of tasks, such as sentence classification [42], relation classification [41], and sentiment analysis [38], since the introduction of word2vec software. (Shi et al., 2019)	[sentence classification, relation classification] is a task that word embedding has been utilized for since the introduction of word2vec software.
	sentence classification is like [relation classification, sentiment analysis] in that they are both tasks that word embedding has been used for since the introduction of word2vec software.
	relation classification is like [sentence classification, sentiment analysis] in that they are both tasks that word embedding has been used for since the introduction of word2vec software.
	word representation has been used for [sentence classification, relation classification, sentiment analysis] since the introduction of word2vec software.

Table 1: *Sample entry from the ACCoRD corpus.* The ACCoRD annotation procedure uniquely allows each positively-labeled context to yield multiple concept descriptions for target ForeCite concept(s) (red) present in an extracted context. Diversity among these concept descriptions is induced through multiple relation types (yellow) and distinct reference concepts (green), each with an elaboration that specifies the relationship between the target and reference concepts (blue).

cept. We filter to concepts with a ForeCite score ≥ 1.0.

Annotators were instructed to assign a binary label to each candidate context indicating whether the context sentence(s) contained a description of the target ForeCite concept in terms of any other concept in the context. This component of our annotation task was found to have high inter-annotator agreement with a Cohen's κ of 0.658.

For each extracted context that was assigned a positive label, annotators were instructed to author as many descriptions of the target ForeCite concept that follow criteria 1-3 above.[2] These criteria uniquely allow each positively-labeled context

---

[2] In a small fraction of cases the hand-labeled descriptions deviate from the criteria. First, in $< 4.6\%$ of examples, we allowed annotators to specify a reference concept not explicitly mentioned in the extracted context. These were limited to obvious cases; e.g. "neural network" is a reference concept for target concept "recurrent neural network." Second, $< 0.4\%$ of examples do not contain an elaboration. The majority of these cases are of the used-for relation, where the reference concept and elaboration are a single entity (e.g. "gav is used for *query processing in stable environments*.")
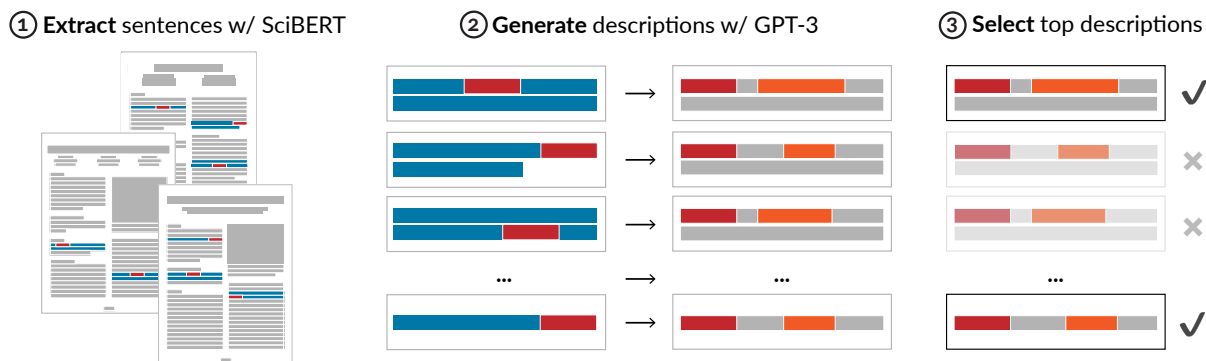
Figure 3: *ACCoRD system implemetation.* Our system (1) **extracts** context sentences (blue) from scientific documents that describe a target scientific concept (red) in terms of another using SciBERT (Beltagy et al., 2019) finetuned on the ACCoRD corpus, (2) **generates** succinct, self-contained, and distinct descriptions of the target's relationship to each reference concept (orange) from the extracted contexts using GPT-3 (Brown et al., 2020) in the few-shot setting, and (3) **selects** a final description set involving multiple relation types and reference concepts.

to yield multiple concept descriptions if a target concept was described in terms of multiple other concepts in the source text, if a single concept pair can be described using multiple relations, or if the extraction contained multiple target concepts (see Table 1).

4 System overview

The ACCoRD system has 3 pipeline stages: (1) **extract** sentences that describe one scientific concept in terms of another, (2) **generate** succinct, self-contained descriptions of the concepts' relationship, and (3) **select** the resulting descriptions to produce a final set of top descriptions for each concept (see Figure 3).

Extraction We build a two-stage SciBERT-based (Beltagy et al., 2019) extractor to identify sentences that describe a target concept in terms of another concept. In the first stage, a binary classification model trained on the binary labels from the ACCoRD corpus extractions identifies reasonable candidates containing a description of the target concept in terms of another concept. In the second stage, a multilabel classification model trained on the relation types in ACCoRD predicts a relation type for a given extracted context, such that we can assign it to the appropriate prompt for the few-shot generation model described below. The inputs to both models have the target scientific concepts demarcated following Wu and He (2019). We select optimal hyperparameters using cross-validation on the ACCoRD training set. This two-stage approach was found to outperform using a single multi-label classifier in preliminary experiments.

Generation The positive-predicted outputs of our extraction model are used as input to our generation model, which produces succinct, self-contained summaries of the concept relationship described in the extracted text. We use GPT-3's `davicini-instruct-beta` model (Brown et al., 2020) in the few-shot setting. We provide the model with a prompt that includes the instruction "Describe the provided concept in terms of another concept in the text" along with five hand-picked (extraction, ground truth concept description) example pairs from the ACCoRD corpus. We hand-select example pairs for each relation type, and for each query extraction, provide the examples of the relation type predicted by the multilabel classifier. This model maps each extraction to a single concept description.

Each generated description is then post-processed to heuristically identify the reference concept, using noun chunking and regular expressions based on our description templates. We then apply additional heuristics to the descriptions as a first-pass filter for low-quality GPT-3 generations, e.g. removing descriptions that have any mention of unresolved references like "our work," descriptions that erroneously contain a reference concept that is an author's name, and descriptions with more than one occurrence of the target concept (to prevent descriptions of the target concept in terms of itself).

Selection Having obtained a candidate set of descriptions for each concept, we attempt to identify a smaller, easily-consumable set that is most likely to be informative and globally descriptive of the

target concept using a selection process. We first filter our descriptions to those that involve a reference concept from ForeCite (King et al., 2020). For each (target, relation) pair, we then choose the most frequent k reference concepts among the descriptions. This ranking method reflects the intuition that the reference concepts that are more helpful for understanding a target concept will also be used relatively more frequently to describe that concept. We then select a top description for each (target, reference, relation) triple by selecting the one with the highest prediction score from our multilabel extraction model. Building sets of descriptions that contain distinct (target, reference, relation) triples allows us to maximize the number of potentially helpful connections that can be made between the target concept and concepts known to the user (see analysis in Section 4.1).

4.1 ACCoRD generated descriptions are diverse

By identifying concept descriptions across the scientific literature, our system is able to capture a diversity of descriptions for a given target concept (see Table 2). We measure this diversity for a set of 150 popular NLP concepts using two metrics: the number of candidate descriptions prior to the selection stage of our system and the number of unique reference concepts contained in those descriptions.[3]

For descriptions involving the `compare` and `is-a` relations, we find an average of 153 and 373 candidate descriptions per target concept, respectively (see Appendix, Figure 7). These candidate descriptions contain an average of 15 unique reference concepts per target concept for `is-a` descriptions and 11 for `compare` descriptions (see Figure 4). Thus, our system captures a wealth of information from the scientific literature that is not retained by a "single best" approach.

5 User study

The experiments in the previous section show that our system produces meaningful diversity in generated descriptions. In order to answer key questions about the utility of our methods, we perform a user study with the full end-to-end system.

---

[3]We report these statistics per relation type exhibited in the description. For brevity, we restrict this to the two most commonly observed relation types.
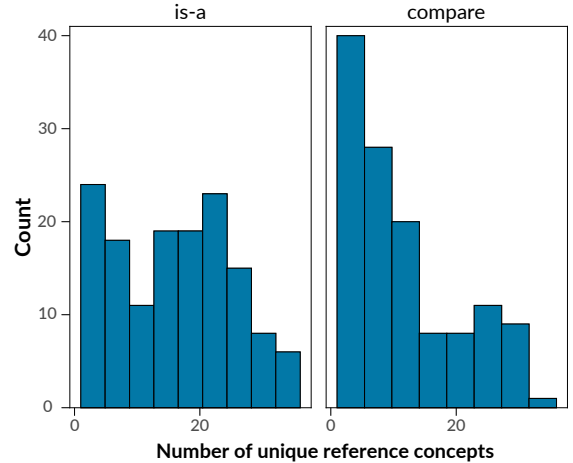

Figure 4: *Distribution over number of unique reference concepts per target among 150 popular NLP concepts.* For each target concept, ACCoRD produces candidate descriptions involving a variety of reference concepts and relations.

- **RQ1:** Which method of producing concept descriptions do users most prefer?

- **RQ2:** Does there exist a single "best" description per user?

5.1 Study description

In our study, participants were asked to imagine they were reading a section in a paper and came across a scientific concept they wanted to learn more about. Participants were asked their preferences for different *sets* of descriptions, as well as their preferences for the individual descriptions.

Participants We recruited 22 participants with native or bilingual English proficiency whose areas of proficiency within computer science included natural language processing (NLP). Participants were recruited through Upwork. All participants had at least a bachelor's degree in computer science. 8 participants additionally had a Master's degree in computer science and 3 had obtained a PhD in computer science. 14 participants indicated having up to three years of experience in NLP, five had 4-6 years of experience, and three others indicated having more than 7 years of experience.

Design Our study consisted of two parts: the first was designed to understand users' preferences for *sets* of descriptions, and the second aimed to understand their preferences for individual descriptions within the sets. For both parts, participants were asked to imagine they were reading a section in a paper and came across a scientific concept that

Extraction	Generation
some such models, including variational autoencoders (vaes) and generative adversarial networks (gans) [goodfellow et al., 2014, kingma and welling, 2013, rezende et al., 2014], learn an explicit low-dimensional manifold that approximates a natural signal class. (Asim et al., 2020)	variational autoencoder is like generative adversarial network in that they are both models that learn an explicit low-dimensional manifold that approximates a natural signal class.
we then compare the performance of rbm with other standard representation learning algorithms, including principal or independent component analysis (pca, ica), autoencoders (ae), variational autoencoders (vae), and their sparse variants. we show that rbms, due to the stochastic mapping between data configurations and representations, better capture the underlying interactions in the system and are significantly more robust with respect to sample size than deterministic methods such as pca or ica. (Tubiana et al., 2019)	variational autoencoder is like autoencoders in that they are both methods for representation learning and are more robust with respect to sample size than deterministic methods such as pca or ica.
such controllable text generation models help in a wide range of application, e.g., dialogues systems [17] . backed by the recent success of deep generative models, such as variational autoencoders (vaes) [10] , generative adversarial nets (gans) [5] , and autoregressive models [12] , existing models have made progress towards controllable text generation [15, 8, 9, 18]. (Xu et al., 2019)	variational autoencoder is like generative adversarial net in that they are both deep generative models that have made progress towards controllable text generation.
generative models in combination with neural networks, such as variational autoencoders (vae), have gained tremendous popularity in learning complex distribution of training data by embedding them into a low-dimensional latent space. traditional vaes usually incorporates simple priors, e.g., a single gaussian, for regularizing latent variables. (Zhao et al., 2019b)	variational autoencoder is a generative model that is used in combination with neural networks to learn complex distribution of training data by embedding them into a low-dimensional latent space.
recently, deep generative models such as variational autoencoders (vaes) (rezende et al., 2014) have become increasingly popular for modelling real-valued data, such as images. (Zhao et al., 2019a)	variational autoencoder is a deep generative model that is used for modelling real-valued data, such as images.
latent variable models such as variational autoencoders (kingma and welling, 2013) tend to better capture the global feature representation in data, but do not offer an exact density estimate as they maximize a lower bound of it. implicit generative models such as gans have recently become popular for their ability to synthesize realistic data (karras et al., 2018; engel et al., 2019). (Das et al., 2019)	variational autoencoder is a latent variable model that does not offer an exact density estimate.

Table 2: *Sample ACCoRD system output for target concept "variational autoencoder."* Extracted contexts are synthesized into generations that describe the target concept (red) in terms of a reference concept (green) using a specified relation type (yellow).

they wanted to learn more or be reminded about. A hypothetical interface was proposed that could provide multiple descriptions of these concepts while reading, in a pop-up box format.

We selected a set of 20 popular NLP concepts with a ForeCite score greater than 1.0. For each concept, we obtained three sets of 6 descriptions each designed to test the various components of our approach:

- **generate-stratify** the output of our complete system: generated descriptions that were selected according to our ranking and filtering methods. This set was comprised of the top 3 descriptions for each of the relation classes `compare` and `is-a`.

- **extract-stratify** the raw extractions corresponding to the generations in *generate-stratify*.

- **generate-naive** the output of the generation component of our system, but without the final stratified selection step. Instead, the top 6 descriptions for this set were selected based on the prediction scores output by the extractive step of our system.

In the first part of the study, for each concept tested, participants were were first asked to indicate their level of expertise with the concept on a 5-point scale ranging from 1 = "I do not know this concept" to 5 = "I know the concept and could explain it to someone else." They were then asked to read the three sets of descriptions for the given concept and select the description set that they deemed most
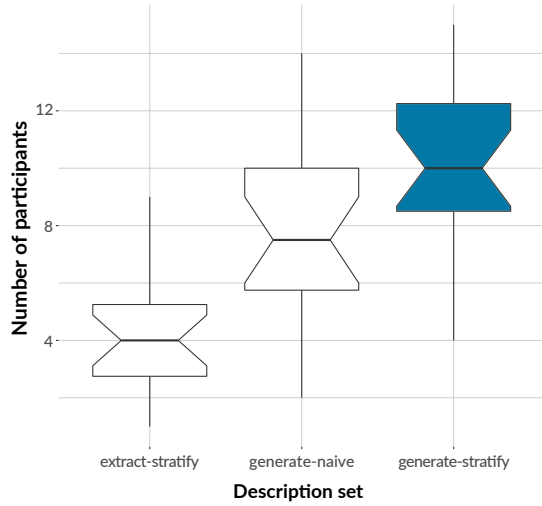
Figure 5: *User preferences for description sets.* Participants strongly preferred descriptions sets that contained generated descriptions (*generate-stratify*, *generate-naive*) over the set that contained extracted text snippets (*extract-stratify*). Participants' preference for the set produced by ACCoRD's more sophisticated description selection component was less pronounced, but still resulted in a higher minimum preference count than *generate-naive*.

helpful for the proposed task. At the end of the section, participants were asked to describe how they evaluated/determined their preference for the sets of descriptions in a free-response style question. In particular, we asked them to articulate which features of the description sets were important in determining a preference, and how they evaluated the combination of the descriptions.

In part two of the study, we were interested in understanding users' preferences at the individual description level. Participants were shown each of the 6 descriptions from our complete system's output (*generate-stratify*) and asked to indicate their level of preference for the description in one of the following three ways: "I would want to see this description of the concept," "No preference/opinion," "I would not want to see this description of the concept." At the end of the section, participants were asked to respond to two free-response style questions explaining (1) why they preferred certain descriptions over others and the qualities that made them better or worse and (2) how their criteria may have differed when rating *sets* of descriptions compared to *individual* descriptions.

5.2 Results

RQ1: **Users prefer our system's generated description over baselines** The three description

sets we tested were aimed at understanding users' preferences for the individual components of our system. In particular: (1) whether people preferred the final, summarized concept description over the raw extractions, and (2) whether our stratified selection method of filtering the concept descriptions was preferred.

As shown in Figure 5, aggregated over the responses for all 20 concepts in our study, participants strongly preferred both versions of the generated description sets, which received a median score of 7.5 for *generate-naive* (95% CI = [5.9, 9.1]) and 10.0 for *generate-stratify* ([8.4, 11.6]) compared to the raw extractions from *extract-stratify* at 4.0 ([2.9, 5.1]). These results also show a preference for the description set obtained using ACCoRD's stratified selection method *generate-stratify* over *generate-naive*.

RQ2: **There is no single "best" description per user** ACCoRD's approach is based on the hypothesis that users prefer a set of descriptions to a single "best" description per concept. Our findings support this idea: When presented with multiple individual descriptions from our end-to-end system, *generate-stratify*, participants on average preferred around 3 descriptions for a given target concept ($\mu = 3.41$, 95% CI = [3.03, 3.79]).

5.3 Qualitative analysis

An analysis of the free-text responses from study participants generally confirmed the results of our quantitative findings, while shedding more light on users' considerations in evaluating concept descriptions.

Users prefer concise descriptions Participants most consistently articulated some preference for shorter, more concise, and more direct descriptions of the target concept ($n = 11$). This provides strong support for the generative component of our system, though, a number of users ($n = 5$) noted that the generations were not always accurate (see Section 7). Additionally, though participants appreciated the simplicity and conciseness of the generated descriptions, many ($n = 6$) noted referencing the extracted text for additional context, confirming a design choice of our system to display each generation with its source text (see Figure 2).

Many users prefer analogical descriptions Our work expands the notion of a description in the context of description generation systems, to

include analogy-like descriptions that are currently not captured by either scientific definition (Kang et al., 2020) or relation extraction (Wadden et al., 2019) systems. A number of participants ($n = 9$) noted that descriptions that drew connections between other concepts in this fashion were helpful, in particular because they could ease learning and memorization of the concept (P18), reflected their own process when trying to synthesize new information (P19), and helped make sense of the many similar model architectures (P14).

6 Related Work

Cognitive theories From children to experts, readers of all levels have been found to actively employ structured background knowledge in the process of comprehending a text (Spiro, 1980; Bazerman, 1985). This background knowledge, often referred to as "schemata," "scripts," or "frames" in cognitive theories of knowledge representation, can be thought of as the data structure for storing concepts in human memory (Rumelhart and Ortony, 1977). These schemata contain the network of interrelations that hold for a given concept. Cognitive theories of learning have asserted that effective ways of describing a new concept to someone take advantage of such schemata, by grounding new descriptions within the network of concepts they are already familiar with (NRC, 2018). Systems that operate within the paradigm of providing a single "best" result for all users, as many definition generation systems do, limit the accessibility of technical knowledge to diverse audiences (Teevan et al., 2010). These considerations motivate our system and novel task definition, which extends the conventional description generation setting to include multiple target descriptions for a single concept.

Data sets While definition and relation extraction data sets address individual components of DSG, they each lack vital components that prevent them from being used as training data for the diverse, multi-relation view of concept descriptions that our task requires. Relevant to the scientific, computer science domain that we study is the W00 (Jin et al., 2013) data set, a corpus of 2,512 sentences from 234 workshop papers from the 2000 ACL Conference. While this data set includes definitions that generalize beyond common lexico-syntactic patterns, they are ultimately restricted to definitions involving a typical `is-a`

relationship between target and reference concept because of their traditional notion of a definition. Data sets for evaluating scholarly relation extraction systems, like SciERC (Luan et al., 2018), do connect concept pairs using a variety of relation types. However, they do not include differentia between the concept pairs that elaborate on the concepts' relationships, making them unsuitable as data for the concept description setting we investigate.

Methods While previous work has investigated extracting and generating definitions of scientific concepts (Espinosa-Anke and Schockaert, 2018; Vanetik et al., 2020; Veyseh et al., 2019; Kang et al., 2020), they focus on producing a *single* canonical description for each concept. These methods also approach definition generation from a single-document perspective, which doesn't account for the multitude of ways a concept might be described outside of the context of that paper. In contrast, we aim to preserve multiple distinct descriptions that take advantage of the corpus-wide mentions of a given concept. In addition, we look to the related works sections of papers to identify descriptions that involve comparative relationships between concepts beyond the typical `is-a` relationship.

7 Future work

7.1 Improving generation quality

Our results show that a generation component that produces succinct, direct descriptions of a target concept is helpful for a user-friendly system for DSG. However, our qualitative feedback suggests that this is also an important area for future work, as poor generation quality was often cited as the reason users preferred the set with only extracted descriptions.

An analysis of a sample ($n = 100$) of ACCoRD system outputs revealed particular areas for improvement. 48% of the sample contained at least type of error. The majority of the error came from the generation component of our system and included generations that were an inaccurate synthesis of extracted context; generations that were technically correct, but unhelpful because they lacked important details; incoherent generations; instances where the second sentence of extraction was appended verbatim, often resulting in one of the other error types; and generations where irrelevant details were appended (see Table 4 for proportions

and an example of each error type). In a minority of cases, system errors were due to issues at the extractive stage. These error types included poor delimitation of the target ForeCite concept within candidate extractions that resulted in an inaccurate generation; low quality extracted text that made it difficult to produce a high quality generation; and difficulties in delimiting sentences in scientific text using available methods (see Table 5). While these errors are of varying degrees of severity, further work on methods that address these issues at scale will be important to expanding the proof-of-concept system we propose here.

7.2 Controllable generation

Beyond resolving errors in generation, future work might investigate methods for controllable generation that are better tailored to user needs. For example, in our user study free-text responses, participants suggested that users may require different kinds of descriptions based on the type of concept being described. In particular, two participants (P14, P21) noted that they preferred set *generate-naive*, which contained more "canonical" descriptions, for simple, standalone data set concepts that could be explained straightforwardly. On the other hand, for more complex method and system-based concepts, like RoBERTa, GPT, and LSTM, users expressed preference for descriptions that made comparisons to other concepts (as produced by our complete system).

7.3 Potential for personalization

While we showed in Section 5.2 that participants often had multiple preferred descriptions per concept, a question remains—Are these preferences similar or different *across* users? To investigate, we compute the Fleiss' κ score measuring agreement in participant preference votes across the six available descriptions for each concept, and find this to be low on average across concepts ($\mu = 0.06$, 95% CI = [0.04, 0.09]). Likewise, only a minority of users ($\mu = 0.34$, 95% CI = [0.31, 0.36]) listed the top-voted description among their preferred ones, on average. The high variation in preferences across participants suggests potential for personalization in the DSG task. In this section, we investigate avenues that future researchers might explore in order to operationalize personalization within description generation systems.

We consider how level of expertise in a concept might affect a user's preferences over descrip-

tions. In Figure 6, we plot the Fleiss' κ scores for user preferences over descriptions against the average level of self-reported expertise of the concepts. While we do observe low agreement in preferences overall, interestingly the lowest agreement scores are found for concepts for which participants mostly self-rated as having low expertise. Fitting a linear model to the data, we find the estimated slope coefficient is significantly greater than zero ($b = 0.03628, p < 0.05$).[4] In terms of qualitative evidence, we found in interviews that user study participants might even disagree on the manner in which expertise should affect ideal description output. For example, P9 suggested that analogical descriptions would be most useful for concepts with low familiarity. In contrast, P14 stated that they would be more helpful for concepts with some level of expertise.
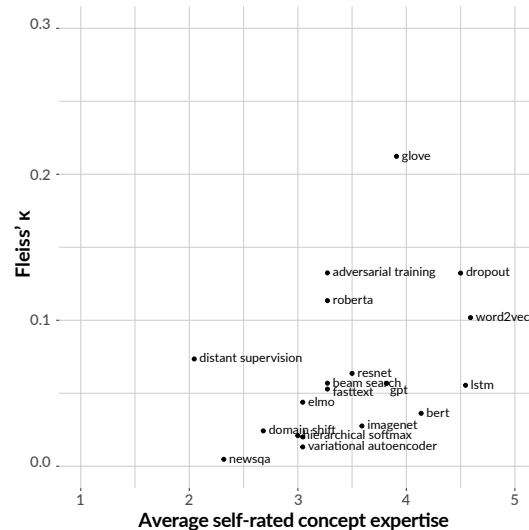

Figure 6: *User agreement (Fleiss' κ) in description preferences for each concept versus average concept expertise level.* We find low agreement in preferences overall, with the lowest agreement scores for concepts for which participants also indicated low expertise.

8 Conclusion

We have presented ACCoRD, an end-to-end system for the novel task of Description Set Generation (DSG): generating multiple distinct descriptions of a single target concept. In experiments, our methods were preferred over baseline approaches

---

[4]We further investigated this by, for each concept, segmenting participants into high and low expertise groups (i.e. above or below the global median expertise of 4.0) and computing average agreement within those groups. The difference in mean agreement between high ($\mu = 0.08$) and low ($\mu = 0.05$) expertise groups was not statistically significant.

and produce a diversity of generated concept descriptions. We also release the ACCoRD corpus to facilitate development of future systems for DSG. We hope that such systems will help increase the accessibility of scientific literature for people with diverse background knowledge.

Acknowledgements

References

Mohammed N. A. Ali, Guanzheng Tan, and Aamir Hussain. 2019. Boosting arabic named-entity recognition with multi-attention layer. *IEEE Access*, 7:46575–46582.

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. 2019. Chasing ghosts: Instruction following as bayesian state tracking. In *NeurIPS*.

Muhammad Asim, Ali Ahmed, and Paul Hand. 2020. Invertible generative models for inverse problems: mitigating representation error and dataset bias. *ArXiv*, abs/1905.11672.

Charles Bazerman. 1985. Physicists reading physics. *Written Communication*, 2:23 – 3.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yin-Wen Chang, Alexander M. Rush, John DeNero, and Michael Collins. 2014. A constrained viterbi relaxation for bidirectional word alignment. In *ACL*.

Hari Prasanna Das, P. Abbeel, and Costas J. Spanos. 2019. Dimensionality reduction flows. *ArXiv*, abs/1908.01686.

Luis Espinosa-Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. 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 378–385, New Orleans, Louisiana. Association for Computational Linguistics.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cogn. Sci.*, 7:155–170.

Dedre Gentner and Susan Goldin-Meadow. 2003. *Language in mind: Advances in the study of language and thought*. MIT press.

David Heywood. 2002. The place of analogies in science education. *Cambridge Journal of Education*, 32:233 – 247.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL*.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the ACL Anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA. Association for Computational Linguistics.

Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel Weld, and Marti A. Hearst. 2020. Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 196–206, Online. Association for Computational Linguistics.

Daniel King, Doug Downey, and Daniel S. Weld. 2020. High-Precision Extraction of Emerging Concepts from Sientific Literature. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, Virtual Event, China. ACM.

Kenneth J. Kurtz, Chun-Hui Miao, and Dedre Gentner. 2001. Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, 10(4):417–446.

Chu-Cheng Lin and Jason Eisner. 2018. Neural particle smoothing for sampling from conditional sequence models. In *NAACL*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Boyu Lu, Frank Shyu, Yun-Nung (Vivian) Chen, Hung yi Lee, and Lin-Shan Lee. 2017. Order-preserving abstractive summarization for spoken content based

on connectionist temporal classification. *ArXiv*, abs/1709.05475.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Sainik Kumar Mahata, Avishek Garain, Adityar Rayala, Dipankar Das, and Sivaji Bandyopadhyay. 2019. Jumt at wmt2019 news translation task: A hybrid approach to machine translation for lithuanian to english. *ArXiv*, abs/1908.01349.

NRC. 2018. How people learn: Brain, mind, experience, and school: Expanded edition. In *The National Academies Press*, Washington, DC.

Jason Portenoy, Marissa Radensky, Jevin D. West, Eric Horvitz, Daniel S. Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In *CHI*.

Alexandre Reiffers-Masson and Vincent Labatut. 2017. Opinion-based centrality in multiplex networks: A convex optimization approach. *Network Science*, 5:213 – 234.

David Rumelhart and Andrew Ortony. 1977. *Schooling and the acquisition of knowledge*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Weijia Shi, Muhao Chen, Pei Zhou, and Kai-Wei Chang. 2019. Retrofitting contextualized word embeddings with paraphrases. In *EMNLP*.

Rand J. Spiro. 1980. Schema theory and reading comprehension: New directions.

Jun Suzuki, Kevin Duh, and Masaaki Nagata. 2011. Distributed minimum error rate training of smt using particle swarm optimization. In *IJCNLP*.

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2010. Potential for personalization. *ACM Trans. Comput.-Hum. Interact.*, 17(1).

David F. Treagust, Reinders Duit, Paul Joslin, and Ivo E. Lindauer. 1992. Science teachers' use of analogies: observations from classroom practice. *International Journal of Science Education*, 14:413–422.

Jérôme Tubiana, Simona Cocco, and Rémi Monasson. 2019. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural Computation*, 31:1671–1717.

Natalia Vanetik, Marina Litvak, Sergey Shevchuk, and Lior Reznik. 2020. Automated discovery of mathematical definitions in text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2086–2094, Marseille, France. European Language Resources Association.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2019. A joint model for definition extraction with syntactic connection and semantic consistency. *CoRR*, abs/1911.01678.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pretrained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2361–2364, New York, NY, USA. Association for Computing Machinery.

Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. Unsupervised controllable text generation with global variation discovery and disentanglement. *ArXiv*, abs/1905.11975.

He Zhao, Piyush Rai, Lan Du, Wray L. Buntine, and Mingyuan Zhou. 2019a. Deep generative models for sparse, high-dimensional, and overdispersed discrete data.

Qingyu Zhao, Nicolas Honnorat, Ehsan Adeli, and Kilian M. Pohl. 2019b. Truncated gaussian-mixture variational autoencoder. *arXiv: Learning*.

Huangjie Zheng, Lingxi Xie, Tianwei Ni, Ya Zhang, Yanfeng Wang, Qi Tian, Elliot K. Fishman, and Alan Loddon Yuille. 2018. Phase collaborative network for multi-phase medical imaging segmentation. *ArXiv*, abs/1811.11814.
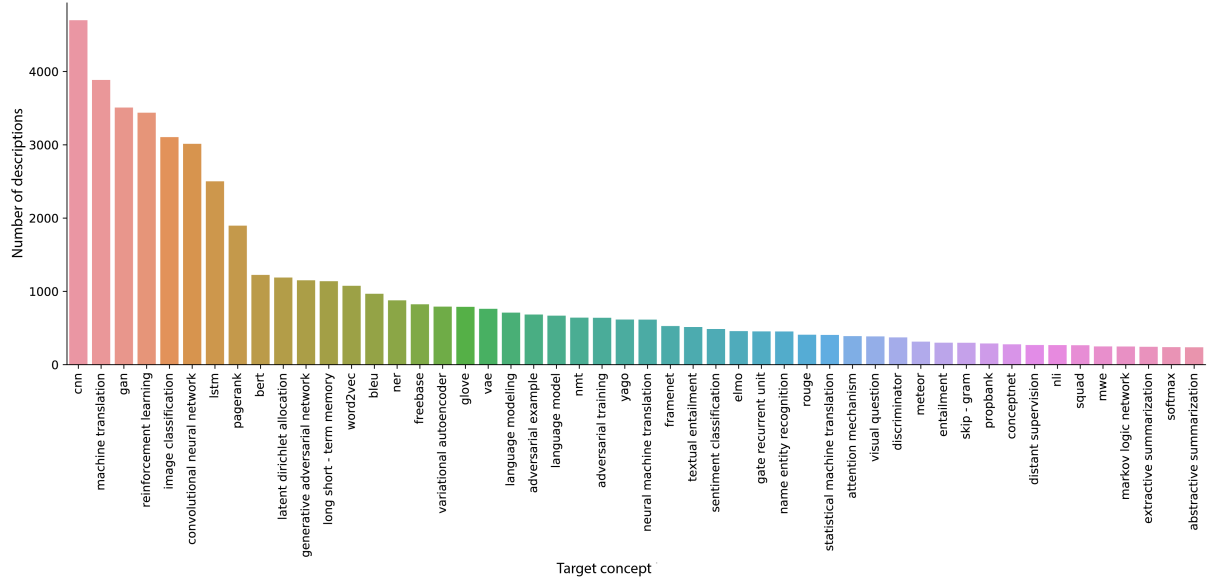
Appendix

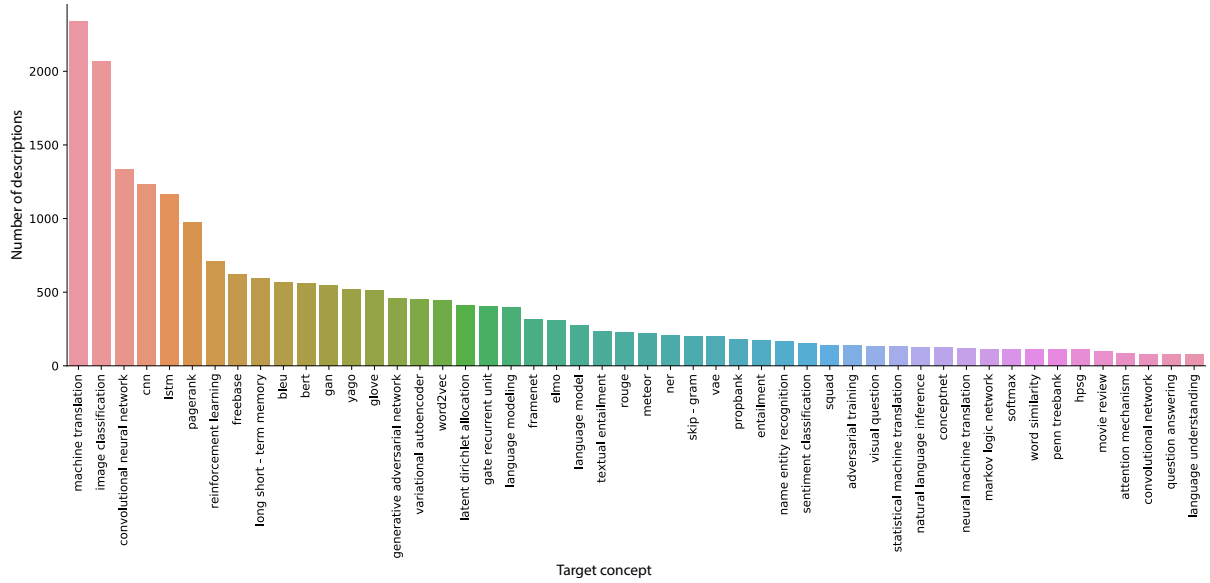**ACCoRD addresses a meaningfully novel task**

To verify that the ACCoRD corpus addresses a novel task that is not well-captured by existing resources, we compare our system's results on ACCoRD to those of existing state-of-the-art scientific definition and relation extraction systems. For our definition extraction baseline, we test HEDDEx (Kang et al., 2020) trained on W00 (Jin et al., 2013), a similarly-sized corpus of definition sentences from workshop papers from the 2000 ACL Conference. Since HEDDEx was originally only intended to produce a single canonical definition of scientific terms and symbols at the sentence-level, we also evaluate its performance on the subset of ACCoRD that was marked as containing an "is-a" relationship between the reference and target concept, to more faithfully evaluate its potential. For our relation extraction baseline, we test DyGIE++ (Wadden et al., 2019) trained on SciERC (Luan et al., 2018), a scientific relation extraction data set. Table 3 shows these results for the union of the 1- and 2-sentence source text settings in ACCoRD, as our qualitative conclusions remained unchanged across these settings. Our model trained on ACCoRD outperforms models that target related tasks, even when they beat a baseline that always assigns positive labels, suggesting that our data set addresses an importantly different task.

Model	Train set	F1
HEDDEx	W00	0.329
HEDDEx$_{is\text{-}a}$	W00	0.449
DyGIE++	SciERC	0.532
SciBERT	**ACCoRD**	**0.624**
Positive baseline		0.484

Table 3: Results for our extractive model and relevant baselines on the ACCoRD test set ($n = 674$). Our model trained on ACCoRD outperforms models that target related tasks, even when they beat a baseline that always assigns positive labels, suggesting that our data set addresses an importantly different task.

(a) `compare` relation

(b) `is-a` relation

Figure 7: Number of candidate descriptions for 50 target NLP concepts, for each relation type present in our system demo, prior to the selection stage of ACCoRD. Our system identifies, extracts, and generates approximately twice as many candidate descriptions of the `compare` relation than of the `is-a` relation.

Error type	Percentage	Sample extraction	Sample generation
Generation is an inaccurate synthesis of extraction	24%	in the modern context, mt systems can be categorized into statistical machine translation (smt) and neural machine translation (nmt). smt has had its share in making mt very popular among the masses. (Mahata et al., 2019) in the nlp community, beam search and importance sampling are common. unfortunately, beam search considers only the approximate-top-k taggings from an exponential set (wiseman and rush, 2016) , and importance sampling requires the construction of a good proposal distribution (dyer et al., 2016) . (Lin and Eisner, 2018)	neural machine translation is like statistical machine translation in that they are both types of mt systems and have had their share in making mt very popular among the masses. beam search is like importance sampling in that they are both methods that consider only the approximate-top-k taggings from an exponential set and require the construction of a good proposal distribution.
Generation is technically correct, but unhelpful because it lacks important details	15%	however, these approaches lack geometric priors and contain no mechanism for reasoning about the likelihood of alternative trajectories -a crucial skill for the task, e.g., 'would this look more like the goal if i was on the other side of the room?'. due to this limitation, many previous works have resorted to performing inefficient first-person search through the environment using search algorithms such as beam search [4, 6] . (Anderson et al., 2019)	beam search is a search algorithm that is used for first-person search through the environment.
Generation is incoherent	8%	in recent years, generative models like vae and gan [20, 11] and their extensions have become popular and have been applied to various scenarios due to their impressive performance. since the arterial and venous phases form different data distribution, to build a relation between them refers to the domain adaptation using generative models, where pix2pix [19] , cyclegan [50] and unit [28] etc. are typical models in this field. (Zheng et al., 2018) the trained model can be fine-tuned for downstream nlp tasks such as question answering and language inference without substantial modification. bert outperforms previous state-of-the-art models in the eleven nlp tasks in the glue benchmark (wang et al., 2018) by a significant margin. (Jawahar et al., 2019)	vae is like gan in that they are both generative models that have been applied to various scenarios due to their impressive performance and since the arterial and venous phases form different data distribution, to build a relation between them refers to the domain adaptation using generative models. question answering is like language inference in that they are both nlp tasks that bert outperforms previous state-of-the-art models in the eleven nlp tasks in the glue benchmark (wang et al., 2018) by a significant margin.
Second sentence of extraction is appended verbatim, resulting in a difficult to read, incoherent, or inaccurate generation	6%	recent statistical machine translation (smt) systems employ a linear combination of several model components, such as translation models, language models, and reordering models. translation is performed by selecting the most-likely translation, which is the candidate translation with the highest score based on the different model components. (Suzuki et al., 2011)	language model is like translation model in that they are both model components that are used in recent statistical machine translation (smt) systems and translation is performed by selecting the most-likely translation, which is the candidate translation with the highest score based on the different model components.

Table 4: *Error analysis of ACCoRD system output: generation stage.* Common generation error types for a sample ($n = 100$) of ACCoRD system outputs include generations that are an inaccurate synthesis of source text; generations that are technically correct but unhelpful because they lack important details; incoherent generations; and instances where the second sentence of extraction is appended verbatim, often resulting in a difficult to read, incoherent, or inaccurate generation. A final error type not shown was, appending irrelevant information, which occurred in 5% of the sample. Multiple error types can be present in the same generation.

Error type	Percentage	Sample extraction	Sample generation
Poor target concept delimitation resulting in inaccurate generation	6%	in order to ensure accurate word alignments, most systems employ a post-hoc symmetrization step to combine directional word aligners, such as ibm model 4 (brown et al., 1993) or hidden markov model (hmm) based aligners (vogel et al., 1996) . several authors have proposed bidirectional models that incorporate this step directly, but decoding under many bidirectional models is np-hard and finding exact solutions has proven difficult. (Chang et al., 2014)	ibm model is like hidden markov model in that they are both directional word aligners and have been used to create post-hoc symmetrization steps.
Extracted text is missing details or contains unresolved references that make it difficult to produce a high quality generation	6%	most of them generalize widespread existing unilayer measures such as degree [29, 17, 1] , betweenness [28, 34, 11] , closeness [29, 35] , eigenvector [17, 33, 1] , pagerank [31, 20, 13] or hits [26] . these generalizations rely on the adaptation of unilayer concepts to the multilayer case. (Reiffers-Masson and Labatut, 2017)	pagerank is like degree, betweenness, closeness, eigenvector and hits in that they are all unilayer measures that generalize to the multilayer case.
		[22] . their system achieved good results on various labeling tasks, including ner and pos, without depending on any manual feature engineering. (Ali et al., 2019)	ner is like pos in that they are both labeling tasks that the system achieved good results on without depending on any manual feature engineering.
Poor sentencization of scientific text	4%	various approaches have been proposed for sequence-tosequence problems, such as speech recognition [15, 16] , neural machine translation [17, 18] , and dialogue modeling [19, 20] . among various deep-learning-based methods, cho et al. (Lu et al., 2017)	neural machine translation is like speech recognition in that they are both approaches for sequence-tosequence problems and among various deep-learning-based methods, cho et al.

Table 5: *Error analysis of ACCoRD system output: extraction stage.* Common extraction error types for a sample ($n = 100$) of ACCoRD system outputs include poor delimitation of the target ForeCite concept within candidate extractions, resulting in an inaccurate generation; low quality extracted text that results in low quality generation; and poor sentence delimitation of scientific text. Multiple error types can be present in the same generation.