MiST: a Large-Scale Annotated Resource and Neural Models for Functions of Modal Verbs in English Scientific Text

Sophie Henning[1,2] **Nicole Macher**[1] **Stefan Grünewald**[1,3] **Annemarie Friedrich**[1]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Center for Information and Language Processing, LMU Munich, Germany
[3]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
sophieelisabeth.henning@de.bosch.com macher.nicole@gmail.com
stefan.gruenewald|annemarie.friedrich@de.bosch.com

Abstract

Modal verbs (e.g., *can*, *should* or *must*) occur highly frequently in scientific articles. Decoding their function is not straightforward: they are often used for hedging, but they may also denote abilities and restrictions. Understanding their meaning is important for various NLP tasks such as writing assistance or accurate information extraction from scientific text.

To foster research on the usage of modals in this genre, we introduce the MiST (**M**odals **I**n **S**cientific **T**ext) dataset, which contains 3737 modal instances in five scientific domains annotated for their semantic, pragmatic, or rhetorical function. We systematically evaluate a set of competitive neural architectures on MiST. Transfer experiments reveal that leveraging non-scientific data is of limited benefit for modeling the distinctions in MiST. Our corpus analysis provides evidence that scientific communities differ in their usage of modal verbs, yet, classifiers trained on scientific data generalize to some extent to unseen scientific domains.

1 Introduction

Each year, an estimate of 1.5 million scientific articles are published (Knoth et al., 2020); hence, the construction of knowledge graphs (KGs) from scholarly texts for aggregating and navigating research findings is an active research area (Chandrasekaran et al., 2020; Knoth et al., 2020; Nastase et al., 2019; Demner-Fushman et al., 2019, 2020). Professional academic writing makes ample use of *hedges*, linguistic devices indicating uncertainty, because scientific propositions are usually considered as valid only until they are overwritten by newer findings (Hyland, 1998). Distinguishing valid solutions to problems from unverified and/or potential solutions is a crucial step in information extraction (IE) from scientific text (Heffernan and Teufel, 2018) as KGs should at least mark untested hypotheses as such (see Figure 1). Yet, with the notable exception of BioScope (Szarvas et al., 2008), prior work in this area is limited.

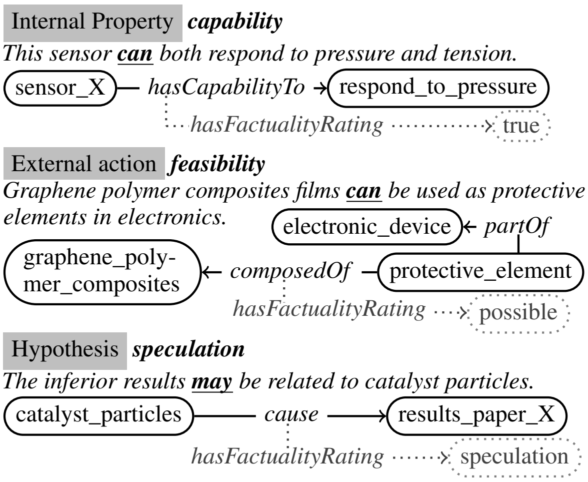

Internal Property *capability*
This sensor ***can*** *both respond to pressure and tension.*
sensor_X
hasCapabilityTo
respond_to_pressure
hasFactualityRating
true
External action *feasibility*
Graphene polymer composites films ***can*** *be used as protective elements in electronics.*
electronic_device
partOf
graphene_polymer_composites
composedOf
protective_element
hasFactualityRating
possible
Hypothesis *speculation*
The inferior results ***may*** *be related to catalyst particles.*
catalyst_particles
cause
results_paper_X
hasFactualityRating
speculation

Figure 1: **Modal verbs** perform various **functions in scientific text** affecting KG representations.

In this paper, we focus on **modal verbs**, a frequently used device for signaling hedging in academic discourse (Hanania and Akhtar, 1985; Getkham, 2011). Other functions of modals include indicating abilities or restrictions. Their meaning depends on the sociopragmatic context (Yamazaki, 2001), i.e., here on the conventions of the community of a particular academic field. Successful academic writing requires correct community-specific use. As shown in Figure 1, understanding the different notions has relevance to KG population (e.g., Luan et al., 2018; Friedrich et al., 2020). Computational modeling of the functions of modal verbs also has applications in language learning and writing assistance software (Römer, 2004).

Prior work in computational linguistics targeting modal verbs (Ruppenhofer and Rehbein, 2012; Rubinstein et al., 2013; Pyatkin et al., 2021; Marasović et al., 2016) has primarily worked with data from the news domain. The annotation schemes of these datasets largely follow distinctions estab-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1305 - 1324
December 7-11, 2022 ©2022 Association for Computational Linguistics

lished in the linguistic literature (see, e.g., Kratzer, 1981; Palmer, 2001; Von Fintel, 2006; Portner, 2009), differentiating between the following coarse-grained modal senses: (a) *epistemic* expresses judgments about the factual status of a proposition, (b) *deontic* relates to permission, obligation, and requirements, and (c) *dynamic* refers to internal abilities or conditions. Our work differs from all of these approaches (a) in that we are the first to address the domain of scientific writing, and (b) in that we do not primarily study modal *senses*, but instead focus on the *pragmatic function* of modal verbs, i.e., our aim is to capture an author's *reason* for using a particular modal verb in a context.

With this paper, we release MiST (**M**odals **In** **S**cientific **T**ext), a manually annotated **corpus** for investigating the usage of modal verbs in scientific text. Our **multi-label** annotation scheme for *modal functions* covers semantic, pragmatic, and rhetorical reasons for an author's use of a modal, with a focus on sub-distinctions that are crucial from an IE viewpoint. MiST consists of 3737 annotated modal verb instances selected from texts of five scientific disciplines (henceforth *domains*), which is larger than all existing comparable datasets (see Table 1). Our **corpus analysis** reveals differences in modal use between scientific domains, and between academic and non-academic use. We perform an inter-annotator agreement study and ensure high data quality via adjudication.

Based on MiST, as well as related corpora, we conduct an extensive **computational study** on automatically classifying functions of modals, comparing CNN-based (Marasović and Frank, 2016) and BERT-based models (similar to Pyatkin et al., 2021). In contrast to prior modeling work, we circumvent modifying the transformer's input by selecting the modal's contextualized output embedding and/or the CLS embedding as input to the classifier. We find that in most cases, a model using both embeddings works best.

To sum up, our paper lays the groundwork for both corpus-linguistic and computational work on modeling functions of modal verbs in scientific text. Our contributions are as follows.

- Our new large-scale dataset annotated with functions of modals in scientific text is publicly available.[1]
- We conduct an in-depth corpus study detailing the corpus construction process, agreement, and corpus statistics, as well as a comparison with existing schemes (Sec. 3).
- Our computational experiments provide a systematic comparison of neural models for modal classification on scientific text (Sec. 4 and 5). We find that a combination of the CLS embedding and the embedding of the modal verb itself works best.
- We show that models trained on out-of-genre data do not work well on scientific text, while classifiers trained on annotated scientific text perform well on unseen scientific domains. In sum, these experimental findings underline the value of our new dataset.

2 Related work

Our work relates to several areas, which we survey in this section.

Annotated corpora. Prior annotation studies on modal verb senses carried out by expert annotators are of limited scale (see Table 1). Ruppenhofer and Rehbein (2012, henceforth RR12, **Modalia** dataset) annotate *senses* of modal verbs in the MPQA Opinion corpus, which consists of news texts. Their linguistically motivated label set includes *dynamic*, *epistemic*, and *deontic* (see Sec. 1), as well as *optative* for wishes, *concessive* if a state of affairs is taken as a given, and *conditional* for *if*-clauses and inversion constructions. On the same texts, Rubinstein et al. (2013, henceforth Rubin13) annotate modal expressions including nouns (e.g., "hope"), adjectives, adverbs, and verbs of propositional attitude (e.g., "believe"). Their annotation scheme is similar to RR12 with minor modifications. Pyatkin et al. (2021, henceforth Pyatkin21) use this dataset with six renamed categories and refer to it as the Georgetown Gradable Modal Expressions (**GME**) corpus. Marasović et al. (2016) annotate a 3-way distinction of modal senses (*dynamic*, *epistemic*, and *deontic*) on **MASC** (Ide et al., 2008), covering several domains. They also introduce the Modalia version **Modalia**$_{\mathbf{M}}$ using this 3-way scheme, mapping *conditional* and *concessive* to *epistemic* and *optative* to *deontic*. Finally, their **EPOS** dataset consists of 7693 sentences for which the same 3-way annotation has been derived via cross-linguistic projection from Europarl (Koehn, 2005) and OpenSubtitles (Tiedemann, 2012). King and Morante (2020) annotate modal verbs in the vaccination debate domain (**VCM**).

[1] github.com/boschresearch/mist_emnlp_findings2022

Dataset	# inst.	# cat.	genre	lang.
Modalia	1158	6	news	EN
Rubin13/GME*	1912	6	news	EN
MASC	1962	3	multi-genre	EN
VCM	450	6	vaccination debate	EN
EP	888	13	multi-genre	PT
CuiChi13	263	6	news	CN
MiST (ours)	3737	7	scientific papers	EN

Table 1: **Datasets manually annotated with modal verb categories.** *Rubin13: 6 base categories + 3 supertypes + 1 multi-label combination; GME is the same dataset as Rubin13 using the 6 base categories (renamed) and 2 supertypes. For Rubin13/GME, EP, and CuiChi13, we count only modal verb instances.

Several annotated datasets target modal expressions in a variety of domains, e.g., focusing on *could* (Moon et al., 2016) in English GigaWord (Parker et al., 2009), or negotiation dialogues (Lapina and Petukhova, 2017). We are also aware of a cluster of works on annotating and tagging Portuguese data (**EP**) using multi-genre data and RR12-style annotation schemes (e.g., Mendes et al., 2016; Avila and Mello, 2013; Quaresma et al., 2014b). Cui and Chi (2013) conduct a small annotation study on Chinese modals with Rubin13-style labels (**CuiChi13**). Yamazaki (2001) performs a corpus study on how American English native speakers interpret modal verbs in the chemistry domain.

Modeling. Early approaches to modal sense classification leverage a lexicon (Baker et al., 2010), or make use of "traditional" features (such as n-grams or part-of-speech tags) and maximum entropy classifiers (Ruppenhofer and Rehbein, 2012; Zhou et al., 2015) or SVMs (Quaresma et al., 2014a,b). Li et al. (2019) create context vectors for modals by computing weighted sums of the non-contextualized word embeddings of selected context words. Marasović and Frank (2016, henceforth MF16) generate a sentence embedding using a CNN, hence classifying *sentences* instead of *modal instances*. Our models are most similar to those of Pyatkin21, who encode input sentences using RoBERTa (Liu et al., 2019), with the CLS embedding as input for a linear classifier. Their model variants differ in the input: the *Context* model marks the modal trigger with special tags (`Sue <target>can</target> swim`); the *Trigger+Head* model encodes only the trigger and its dependency head without further context.

Further related work. Other related work includes research on speculation in biomedical data (Szarvas et al., 2008; Kim et al., 2011) and on event factuality (e.g., Saurí and Pustejovsky, 2009; Stanovsky et al., 2017; Rudinger et al., 2018; Pouran Ben Veyseh et al., 2019). Bijl de Vroe et al. (2021) integrate a lexicon-based method for modality detection in event extraction; using this tagger, Guillou et al. (2021) find that entailment graph construction does not profit from tagging for modality. Vigus et al. (2019) propose to annotate modal structures as dependencies. Rhetorical analysis of scientific text is often based on Argumentative Zoning (Teufel et al., 1999). Lauscher et al. (2018a,b) provide a dataset and neural methods for extracting and classifying claims from scientific text. Luan et al. (2018), Jiang et al. (2019), and Friedrich et al. (2020) present data-driven work on scientific IE. Heffernan (2021) uses modality as a feature to recognize problem-solving utterances in scientific text.

3 MiST Corpus

In this section, we describe our new dataset, including its annotation scheme and detailed corpus and inter-annotator agreement statistics. We annotate instances of *can, could, may, might, must*, and *should* in research papers from five scientific fields: computational linguistics (CL), materials science (MS), agriculture (AGR), earth science (ES), and computer science (CS). Modal usage is influenced by sociopragmatic context (Yamazaki, 2001) and, as a form of hedging, needs to be understood in its social, cultural and institutional context (Hyland, 1998), here the *global* scientific community. Hence, we do not restrict document selection to native English authors.

3.1 Document and Sentence Selection

We select modal verb occurrences as follows. In our *full-text subset* of 73 documents, the CL papers are taken from the ACL Anthology,[2] spanning the years 2013-2015. Data from the other domains stems from the OA-STM corpus,[3] with the exception of five open-access documents for MS.

Because some modal-domain combinations are rare, we additionally sample sentences from 348 documents with Creative Commons licenses such that we have at least 100 instances for each modal-domain pair. For CS, we sample papers tagged

[2] aclanthology.org

[3] elsevierlabs.github.io/OA-STM-Corpus

	CL	MS	AGR	CS	ES	Total
Complete corpus						
sent with modals	925	718	497	746	584	3470
annotated modals	1011	757	543	806	620	3737
Full-text subset						
documents	30	16	10	7	10	73
sents. with modals	693	462	195	445	258	2053
- in % of sents.*	9.3	11.6	9.0	14.2	11.0	10.8
sents. with ≥ 2 modals	61	29	24	48	19	181
avg. #tokens/sent.	27.7	26.4	32.1	27.4	31.9	28.4
annotated modals	760	492	223	497	279	2251

Table 2: **MIST corpus statistics**. *in % of total sentences of the documents. In the *full-text subset*, all modals within the documents have been annotated. The *complete corpus* contains the full-text subset and additionally sampled individual sentences with annotations.

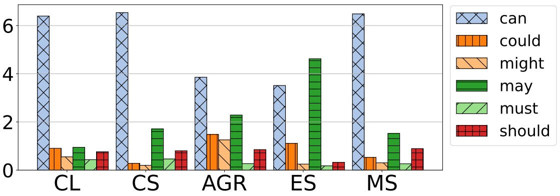


Figure 2: **MIST:** Distribution of modals by domain, computed over full-text annotation subset.

with *cs:CV* and published in 2018 from ArXiv.[4] Additional MS papers published between 2015 and 2021 were retrieved via PubMedCentral.[5] For ES and AGR, we use the DOAJ API[6] to retrieve documents matching the topics of the full-text subset. For AGR, we add articles from the Journal of Agricultural Science published 2017-2021.[7] In total, we obtain a large-scale dataset of 3737 annotated instances (see Table 2, *complete corpus*).

3.2 Annotation Scheme

Our annotation scheme comprises seven labels for functions of modals (see Table 3). Table 4 shows which labels apply for each modal, for more details see Appendix C. The labels apply if the author uses the modal verb:

feasibility: to indicate that it is possible for an external actor, e.g., a human, to do or achieve something;

capability: to convey that something has a certain intrinsic property, ability, or capacity;

inference: to state that they inferred something based on some given information;

speculation: to indicate speculations;

options: to indicate potential options;

deontic: to express a desire, or a requirement, or an obligation;

rhetorical: for conventionalized, fixed expressions.

Our inventory intends to capture the most frequent and relevant functions of modal verbs in scientific discourse. Table 3 classifies a set of utterances according to our, RR12's and Rubin13's schemes.[8] A detailed description of the commonalities and differences is provided in Appendix A. During annotation scheme design, we started out with their categories, but then tailored our scheme to the scientific domain, adding some pragmatic distinctions that are relevant in scientific writing. Annotators have access to the full documents. For labels involving inference, uncertainty or speculation, annotators are instructed to only refer to the text and not to make use of their own knowledge of whether something is the case.

3.3 Annotation Process

Our annotation scheme takes a multi-label approach in which all applicable features may be selected. For each instance of the *full-text subset*, we collect the annotations of three annotators (two for MS) using the web-based annotation systems Swan (Gühring et al., 2016) and INCEpTION (Klie et al., 2018). We ensure consistency across sub-corpora by means of an adjudication step (for all instances) performed by one author of this paper, who then also labeled the additionally sampled instances. Our total group of annotators consists of one undergraduate as well as three graduate students of CL, one undergraduate student of CS, one graduate student of MS and one physicist holding a PhD degree. While not all annotators are native speakers of English, they are either domain experts or have a strong linguistic background.

3.4 Corpus Analysis

Modal distributions. We first analyze the usage of the different modals per domain. As shown in Table 2, in the full-text subset, the ratio of sentences including modal verbs ranges from 9.0% (AGR) to 14.2% (CS). In Figure 2, we plot the distribu-

[4] kaggle.com/Cornell-University/arxiv

[5] ncbi.nlm.nih.gov/pmc

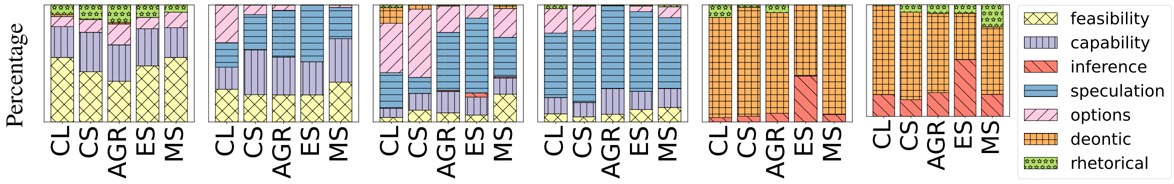
[6] doaj.org/api/v2/docs

[7] cambridge.org/core/journals/journal-of-agricultural-science

[8] According to our interpretation of their guidelines. To facilitate comparison with Pyatkin21, we also added their mapping to the Rubin13 scheme to the table.

Example	Ours	RR12	Rubin13 / Pyatkin21
Several supercapacitors **can** be integrated and connected in series.	***feasibility***	*dynamic*	*Circum. / State of the World*
The device **can** light up a red light-emitting diode and works well.	***capability***	*dynamic*	*Ability / State of the Agent*
The overlap in the ranges [...] indicates that the sample **must** be older than 50.70 Ma.	***inference***	*epistemic*	*Epistemic / State of Knowledge*
The real shielding **can** of course be different.	***options***	*deontic*	*Circum. / State of the World*
DA3 **may** therefore indicate a continuation of high nutrient surface water with an elevated freshwater input.	***speculation***	*epistemic*	*Epistemic / State of Knowledge*
Energy storage devices **should** be able to endure high-level strains.	***deontic***	*deontic*	*Bouletic / Desires and Wishes*
A GCR proton [...] **must** have at least 150 MeV to reach the station.	***deontic***	*deontic*	*Teleological / Plans and Goals*
You must leave the lab tidy.	***deontic***	*deontic*	*Deontic / Rules and Norms*
It **can** be seen in Figure 1 that...	***rhetorical***	*dynamic*	*Ability / State of the Agent*
For instance, despite graphene, the band gaps of silicone **can** be opened and tuned when exposed to an external electric field.	***feas.***, ***cap.***	*dynamic*	*Circum. / State of the World*
These results suggest that epeiric seas [...] **may** have played an important role in the driving mechanism for OAE 2.	***inf.***, ***spec.***	*epistemic*	*Epistemic / State of Knowledge*
Long ***may*** *she live!*	***deontic***	*optative*	*Bouletic / Desires and Wishes*

Table 3: **MiST annotation scheme** in comparison to those of RR12 and Rubin13/Pyatkin21.



(a) can (b) could (c) may (d) might (e) must (f) should

Figure 3: **MiST: Label distributions** by modal verb and scientific domain (adjudicated complete corpus).

tions of modals by domain. Except in the case of ES, *can* is the most frequently used modal by a large margin. In AGR and ES, *may* is also used frequently. Overall, the distributions of CL, CS and MS are somewhat similar, while AGR and ES exhibit different modal usage patterns. The distributions differ from modal usage in other genres (details for MASC and Modalia see Appendix B.2), e.g., the percentage of *can* is much higher in MiST.

Label distributions. Next, we drill down on the functions of the modals by domain. If an instance has more than one label, both labels are counted. The label distributions differ strongly by modal (see Figure 3 and Table 4), but at times also visibly between domains. Previous corpus-linguistic studies (Takimoto, 2015; Hardjanto, 2016) observe more hedging in humanities and social sciences text compared to the natural sciences. ES, which deals with earth's present features and its past evolution, has notably more ***inference*** usages of *must* and *should*. In MS, many cases of *could* are classified as ***feasibility***, as it is common to report experiments in the past tense in this domain. Also, in MS *may* is sometimes used interchangeably with *can* as in "stress–strain data may be obtained for ductile materials." The larger amount of ***rhetorical*** instances in MS is due to cases such as "We should note that."

Comparing the label distributions of MiST and those of MASC and Modalia$_\text{M}$, we also find notable differences (details in Appendix B.2). For example, *may* is used mostly in *epistemic* senses. Our annotations reveal that in AGR and ES, these are mostly ***speculation***; CL and CS texts use this modal to indicate (mostly algorithmic) ***options***. Finally, the use of *should* seems most community-specific: while it is used predominantly in a *deontic* way in MASC and Modalia$_\text{M}$, usage in MiST varies by domain. Overall, these observations support the hypothesis that modal usage depends on the socio-pragmatic context, and demonstrate the value of genre-specific data such as MiST.

Label co-occurrence. In the full-text subset and in the complete corpus 24.5% and 22.3% of instances carry more than one label, respectively. Figure 4 shows the total number of label co-occurrences in the adjudicated gold standard. Overall, ***speculation*** co-occurs most with other labels, indicating that the author likely had two reasons for using the modal, for example indicating a ***capability***, but marking at the same time that it is unclear whether it actually

	can	could	may	might	must	should
feasibility	823	161	62	52	0	0
capability	476	188	91	102	0	0
inference	0	0	*8	0	63	127
speculation	0	206	257	398	0	0
options	183	64	205	70	0	0
deontic	*13	0	25	0	444	330
rhetorical	157	0	*4	*8	24	41

Table 4: **MIST: Label counts**, all domains, adjudicated complete corpus. *Omitted from experiments.

cap.	153					
inf.	0	1				
spec.	73	208	8			
opt.	67	117	0	1		
deon.	2	2	15	0	15	
rhet.	128	3	2	5	6	50
	feas.	***cap.***	***inf.***	***spec.***	***opt.***	***deon.***

Figure 4: **MIST: Label co-occurrence counts**, all domains, adjudicated complete corpus.

holds ("The urban ecosystems could account for a significant portion of terrestrial carbon (C) storage (...)."). Often, both a ***feasibility*** and a ***capability*** reading are possible (see lower part of Table 3), as in "The above construction can be further simplified.", where simplifiability is an intrinsic property of the construction, but the simplification needs an external actor.

3.5 Inter-Annotator Agreement

Computing agreement for our dataset is not straightforward for two reasons. First, we are dealing with a multi-label scenario, for which standard agreement coefficients cannot easily be applied. Second, for some modal-domain combinations, we only have limited data. Averaging across modal verbs is not meaningful: due to the notably different label distributions, good agreement could only mean that annotators distinguish modal verbs well (Artstein et al., 2009; Artstein, 2017).

Following the idea of Krippendorff's diagnostics (Krippendorff, 1980), we evaluate (on the full-text subset) for each modal-label combination how often annotators agree on whether the label applies or not. For each pair of annotators, we compute κ (Cohen, 1960) for this binary decision for each label, mapping all respective other labels to OTHER. In Figure 5, we report the average of these κ-scores over the pairs of annotators for each valid modal-label combination. For some combinations, high agreement is reached. For infrequent labels or modals, agreement is less satisfying. Many "disagreements" occur in cases where in fact several readings are possible.

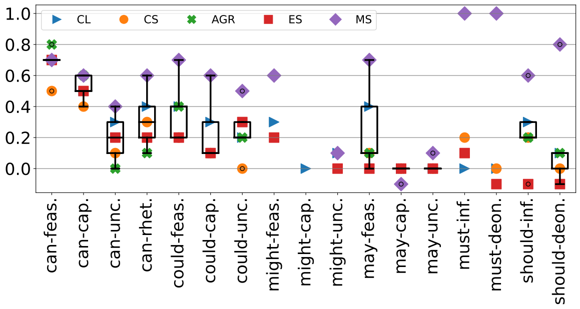


Figure 5: **MIST: Inter-annotator agreement** in terms of avg. κ for labels that have been assigned to at least 6 instances of respective modal (by any annotator), on full-text subset.

Qualitative analysis revealed that some annotators over- or under-used some labels, especially ***uncertainty***, which in the initial round of annotation described here was defined to include both ***options*** and ***speculation***. We hence decided to **ensure high quality** of our corpus through an **adjudication step**. In 62.2% of instances, the adjudicator's labels exactly match the majority vote across annotators; in 90.5%, they overlap with the majority vote labels. We further introduced the label ***options***, and two adjudicators re-labeled all instances initially labeled with ***speculation***. Out of these, both labeled 166 instances, reaching F1-agreements of 72.7/81.3/83.5/86.9 for ***capability***, ***feasibility***, ***options*** and ***speculation***, respectively. In the remainder of this paper, we perform experiments based on the adjudicators' labels.

4 Computational Modeling

We now describe our neural models for classifying functions of modal verbs. We assume that targets have been pre-defined, e.g., using a part-of-speech tagger. Our models are based on a pre-trained transformer that provides embeddings for sentences and contextualized token embeddings. We fine-tune SciBERT (SB, Beltagy et al., 2019), which has the same architecture as BERT (Devlin et al., 2019), but has been trained on large volumes of scientific text. On top, we use **multiple classification heads**, i.e., one per modal, as the label distributions vary substantially by modal. The largest version of our models is trained jointly on multiple datasets and therefore has the aforementioned output heads for each dataset (see Figure 6). The output dimension of these heads varies according to the labelset size of the respective dataset.

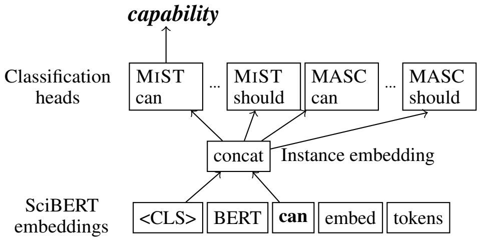


Figure 6: **Model architecture.** SB$_{\text{CLS,modal}}$ model.

We test the following model variants:

SB$_{\text{CLS}}$**.** We feed the CLS embedding of an input sentence into a linear layer with softmax (for single-label classification) or sigmoid (for multi-label classification) activation. This model uses the same decision basis for all modal verbs within a single sentence.

SB$_{\text{modal}}$**.** We select the embedding of the word-piece token corresponding to the modal to be classified (*modal embedding*),[9] and feed this embedding into the linear layer as above. We expect this model to be able to distinguish different modal verbs in the same sentence. The model primarily reflects local context, but to some extent also dependency context (Tenney et al., 2019).

SB$_{\text{CLS,modal}}$**.** We concatenate the CLS embedding with the modal embedding before feeding it into the linear layer. This model should distinguish modal verbs in the same sentence, at the same time leveraging the CLS embedding that intends to cover the entire sentence.

5 Experiments

In this section, we report our experimental results.

5.1 Evaluation Metrics

As majority classifiers are known to provide a strong baseline for modal sense classification (see Rubin13, MF16), we report F$_1$ scores in order to evaluate how well a classifier performs across labels. We compute macro-average F$_1$ (**mF$_1$**) as the average of the per-label F$_1$ scores for the set of labels with which the modal is labeled at least once in the entire corpus and which are not omitted from the experiments due to extreme sparsity (see Table 4). We also report accuracy; we compute it globally across samples and labels, i.e., we simply count for each label how often the classifier (in)correctly did (not) assign it. For hyperparameter tuning and early stopping, we use the macro-average of weighted F$_1$ scores for each modal-domain combination. These weighted F$_1$ scores are computed by weighting per-label F$_1$ scores by the label's support in the validation set. For computing all metrics, we use TorchMetrics.[10]

5.2 Baselines

We report results for the following baselines: **Maj** always predicts the label most frequent in training. We also re-implement MF16's **CNN** with 300-dimensional GloVe embeddings (Pennington et al., 2014) and filter region sizes of 3, 4, and 5 with 100 filters each. Replicating MF16's Table 4 (with their hyperparameters and training a separate model for each modal), we find that our CNN implementation is comparable to theirs, with 77.6 % accuracy on all verbs compared to MF16's 76.5%. On MiST, we use only one model with per-modal heads. **SB**$_{\text{CLS-mark}}$ is our re-implementation of Pyatkin21's *Context* model (their most accurate model), but using SciBERT and per-modal heads. We also investigate whether the genre-specific pre-training is beneficial, replacing SciBERT with BERT (**BERT**$_{\text{CLS,modal}}$), and how the model size affects performance, comparing to **BERT-large**$_{\text{CLS,modal}}$ (to date, there is no SciBERT-large).

5.3 Experimental Settings

We randomly split MiST into a training and a test set of complete documents, aiming at covering approximately 25% of each domain's modal instances in the test set, with real test set sizes ranging from 22.8% to 27%. In our *CV training setting*, we split the training set into 5 folds of complete documents, and train 5 models on 4 folds each, using the respective fifth fold for model selection. We train for at most 100 epochs, performing early stopping with a patience of 10 epochs. We then run each of these five models on the unseen test set, reporting average scores along with standard deviations. Hyperparameters are reported in Appendix D.1.

5.4 Experimental Results on MiST

Here, we evaluate the neural architectures described above on MiST, and investigate performance in the absence of in-domain training data.

Comparing Model Architectures. Table 5 reports the mF$_1$ scores of the various neural models

[9] SciBERT and BERT both tokenize all modals in MiST into a single word piece. (Otherwise, one could use the embedding of the modal's first word-piece token.)

[10] github.com/PyTorchLightning/metrics

	can	**could**	**may**	**might**	**must**	**should**
#inst. train	987	343	369	366	397	342
#inst. test	340	105	141	117	105	119
Maj	18.9 ±0.0	15.5 ±0.4	12.8 ±0.0	23.0 ±0.0	30.2 ±0.0	28.9 ±0.0
CNN	58.8 ±5.5	55.2 ±7.1	40.2 ±5.4	37.8 ±4.4	41.1 ±4.2	64.2 ±10.0
SB$_{\text{CLS}}$	74.8 ±2.1	71.9 ±4.0	50.1 ±3.1	64.1 ±4.3	78.2 ±4.3	82.5 ±2.7
SB$_{\text{CLS-mark}}$	76.6 ±1.7	63.7 ±1.4	49.1 ±1.8	61.5 ±1.5	73.7 ±4.0	85.5 ±2.3
SB$_{\text{modal}}$	76.7 ±2.5	71.3 ±1.8	**50.2** ±4.3	**65.3** ±4.7	76.9 ±2.7	84.5 ±1.4
SB$_{\text{CLS, modal}}$	77.4 ±1.0	**73.7** ±3.8	47.2 ±1.1	64.5 ±2.7	**78.4** ±1.1	**85.7** ±0.5
BERT$_{\text{CLS,modal}}$	74.9 ±2.2	73.3 ±2.0	47.9 ±1.6	64.1 ±1.4	73.8 ±3.4	85.4 ±1.0
BERT-large$_{\text{CLS,modal}}$	**77.7** ±1.3	68.9 ±2.5	46.2 ±3.0	61.4 ±3.2	76.0 ±2.0	84.9 ±1.1

Table 5: **Macro F$_1$ (mF$_1$) on test set of MIST.** #inst. train refers to the entire training set.

Macro F$_1$	**CL**	**CS**	**Agr**	**ES**	**MS**
+	**57.0** ±4.6	**53.2** ±5.6	**61.7** ±8.1	**58.7** ±3.1	**59.7** ±1.1
-	54.2 ±1.7	50.2 ±7.6	60.4 ±7.0	54.8 ±1.8	58.8 ±5.1
Accuracy	**CL**	**CS**	**Agr**	**ES**	**MS**
+	**92.3** ±1.0	**92.7** ±1.4	**93.5** ±1.3	**93.5** ±0.8	**93.4** ±0.5
-	91.5 ±1.3	91.9 ±0.9	**93.5** ±1.6	92.5 ±0.9	92.5 ±1.4

Table 6: Results for 6-fold CV on MIST by domain when **training with (+) and without (-) in-domain data**, averaged over modals. Cross-validated averages and standard deviations of averages of per-modal scores.

on MIST. The magnitude of these scores differs by modal verb. The CNN learns more than Maj., but is always outperformed by the SciBERT-based models. SB$_{\text{modal}}$ is better than SB$_{\text{CLS}}$ on *can*, *might*, and *should*, but worse on *must*, where using an additional sentence-wide embedding is beneficial. For most of the verbs, SB$_{\text{CLS,modal}}$ is the best SciBERT-based model, but SB$_{\text{modal}}$ is better on *may* and *might*. In general, SB$_{\text{CLS, modal}}$ tends to have smaller standard deviations across CV training configurations than the other SciBERT-based models. On *could* and *must*, SB$_{\text{CLS,modal}}$ is better than SB$_{\text{CLS-mark}}$, suggesting that directly using the modal's embedding instead of modifying the input is more effective.

On most verbs, SciBERT and BERT perform comparably, but the domain specificity of SciBERT leads to clear improvements on *can* and *must*. Interestingly, increasing the model size for BERT is beneficial on the very same verbs; at the same time, however, it hurts performance on the other verbs, with an especially sharp loss on *could*.

With the exception of *can*, SB$_{\text{CLS,modal}}$ is also the most accurate model (scores in Appendix D.3). For this model, during development, we experimented with using only one classifier head for all modals (not reported in tables). Compared to per-modal heads, we observed either no difference or slightly worse (by around 1 point mF$_1$ on average) performance for all modals except *must*, where mF$_1$ increased by around 15 points. These gains were due to similar ***rhetorical*** instances, e.g., "We must note that..." and "We should note that...".

Cross-Domain Results on MIST. We conduct a cross-domain experiment on MIST to determine the extent to which in-domain training data is necessary for classifying modal verbs in different scientific communities. Since some modal-domain combinations have rather little data, in this experiment, we split MIST into six folds and use each fold once for testing. We use four of the remaining five folds for training and one for early stopping.

Table 6 reports the cross-validated averages and standard deviations of averages of per-modal mF$_1$ and accuracy to show the overall effect of in-domain data. Models trained on other (scientific) domains work well on unseen domains, as the performance does not decrease substantially when training without in-domain data. As one would expect, domain-specific data usually leads to improvements, especially for domains in which a specific modal has a visibly different label distribution (see Figure 3), e.g., cross-validated mF$_1$ for *could* on CL increased by around 18 points. For other modal-domain combinations, gains were less distinct or sometimes non-existent, and cross-validated scores had a high variance. On average, standard deviation of accuracy was 2.5 and 2.7 for with and without in-domain data, respectively. For mF$_1$, standard deviation was 10.7 when training with in-domain data and 11.3 when training without in-domain data.

In sum, we expect classifiers trained on MIST to also generalize to new scientific domains to some extent. For optimal performance, adding in-domain data is beneficial in most cases.

5.5 Transfer from GME to MIST

In this experiment, we show that functions of modal verbs in scientific text cannot be determined simply using existing datasets. We train a model only on an out-of-genre resource (GME in the version

feasibility, ***options***	*State of the World*
capability, ***rhetorical***	*State of the Agent*
speculation, ***inference***	*State of the Knowledge*
deontic	*Priority* (*Desires*+*Wishes*, *Plans*+*Goals*, *Rules*+*Norms*)

Table 7: **Transfer experiment: Mapping** between GME (Pyatkin21) and MIST schemes.

	can	could	may	might	must	should
Maj$_{\text{GME}_\text{T}}$	33.5 ±0.0	19.7 ±0.1	15.9 ±0.0	30.7 ±0.0	30.2 ±0.0	28.9 ±0.0
SB$_{\text{CLS, modal; GME}_\text{T}}$	56.1 ±7.7	43.7 ±6.5	19.8 ±4.4	30.2 ±4.0	33.9 ±7.7	39.8 ±6.2
SB$_{\text{CLS, modal; MIST-small}}$	82.9 ±0.8	78.1 ±1.7	**43.7** ±1.2	63.4 ±5.0	69.2 ±6.3	76.9 ±11.6
SB$_{\text{CLS, modal; MIST}}$	**84.3** ±0.8	**79.6** ±1.8	43.1 ±1.1	**69.9** ±1.1	**74.6** ±3.8	**84.1** ±2.4

Table 8: **Transfer experiment: Macro F$_1$** on *mapped* test set of MIST.

Macro F$_1$	**can**	**could**	**may**	**might**	**must**	**should**
Maj$_{\text{GME}_\text{T}}$	33.5 ±0.0	19.7 ±0.1	15.9 ±0.0	30.7 ±0.0	30.2 ±0.0	28.9 ±0.0
SB$_{\text{CLS, modal;GME}_\text{T}}$	56.1 ±7.7	43.7 ±6.5	19.8 ±4.4	30.2 ±4.0	33.9 ±7.7	39.8 ±6.2
SB$_{\text{CLS, modal;MIST-small}}$	82.9 ±0.8	78.1 ±1.7	**43.7** ±1.2	63.4 ±5.0	69.2 ±6.3	76.9 ±11.6
SB$_{\text{CLS, modal;MIST}}$	**84.3** ±0.8	**79.6** ±1.8	43.1 ±1.1	**69.9** ±1.1	**74.6** ±3.8	**84.1** ±2.4
Accuracy						
Maj$_{\text{GME}_\text{T}}$	68.7 ±0.0	63.4 ±0.2	67.0 ±0.0	85.7 ±0.0	88.6 ±0.0	85.9 ±0.0
SB$_{\text{CLS, modal;GME}_\text{T}}$	76.1 ±2.6	70.1 ±2.7	68.9 ±2.1	74.7 ±10.5	88.9 ±1.7	85.6 ±2.9
SB$_{\text{CLS, modal;MIST-small}}$	90.1 ±0.3	86.1 ±1.1	**79.6** ±1.0	85.9 ±1.1	92.1 ±0.7	92.5 ±1.2
SB$_{\text{CLS, modal;MIST}}$	**91.0** ±0.5	87.0 ±1.2	**79.6** ±0.9	**88.4** ±0.7	**93.4** ±0.7	**94.3** ±0.6

Table 9: **Transfer experiment:** on *mapped* test set of MIST.

published by Pyatkin21).[11] We train on **GME$_\text{T}$**, i.e., all instances from GME (including the test set) that cover MIST's set of modal verbs using mapped labels as shown Table 7. Resolving GME's *State of Knowledge* into ***inference*** and ***speculation*** and *State of the World* into ***feasibility*** and ***options*** would require a manual re-annotation. We map ***deontic*** to Pyatkin21's supertype *Priority*.

GME$_\text{T}$ consists of 1276 instances, of which 370/238/139/61/196/272 are instances of *can*, *could*, *may*, *might*, *must*, and *should*, respectively. We train and evaluate all models in this experiment using the mapped annotation scheme, using the SB$_{\text{CLS,modal}}$ architecture with sigmoid heads. For hyperparameter tuning and evaluation, we perform the steps as described in Appendix D on GME$_\text{T}$ with five randomly induced folds (**SB$_{\text{CLS,modal; GME}_\text{T}}$**). **SB$_{\text{CLS,modal; MIST}}$** is SB$_{\text{CLS,modal}}$ trained on MIST with *mapped labels*. **SB$_{\text{CLS,modal; MIST-small}}$** is trained on a randomly downsampled version of MIST to account for the notably larger size of MIST compared to GME$_\text{T}$. We approximately proportionally randomly downsample each MIST fold (with mapped labels) to get **MIST-small**, which has exactly the same number of instances as GME$_\text{T}$.

Table 9 shows the results of our transfer experiment. SB$_{\text{CLS,modal; GME}_\text{T}}$ learns more than just the majority baseline **Maj$_{\text{GME}_\text{T}}$** (except for *might*, for which GME contains little data), but clearly lags behind the models trained on MIST in both mF$_1$ and accuracy, with average mF$_1$ being between 23.9 and 37.1 points lower than models trained on MIST-small. A related experiment (reported in Appendix D.4) using prior corpora of annotated modals in multi-task objectives confirmed the limited amount of transferability. Hence, genre-specific data is clearly required for classifying functions of modal verbs in scientific discourse, demonstrating the value of MIST.

6 Conclusion and Outlook

In this paper, we have introduced a new large-scale dataset of scientific text annotated for functions of modal verbs. Our corpus and computational studies reveal differences and similarities in modal usage across genres and domains. We have shown that neural classification is robust across scientific domains, but also that annotated scientific text is essential for good performance. To sum up, our paper lays the groundwork for informed IE from sentences containing modals in scientific texts, e.g., distinguishing speculations from capabilities attributed to a method or device.

Future work. Our experiments on MIST point to various next steps, e.g., identifying domain-adaptation methods that more effectively leverage annotations across domains, genres, or even languages, or developing data-augmentation techniques targeted to scientific text. Another next step is to integrate our methods for generating metadata for facts into IE systems. Existing open IE systems do not handle the meaning of modal verbs adequately. As an outlook, in Appendix E, we outline how this could be improved using a classifier trained on MIST.

[11] We thank the anonymous reviewers for proposing this interesting experiment.

Acknowledgements

We thank Alex Fraser, Alexis Palmer, Jannik Strötgen, Heike Adel, and the anonymous reviewers for their valuable comments. Thanks also go to Daria Stepanova for a very helpful discussion on knowledge representation. We thank our annotators Sherry Tan, Prisca Piccirilli, Johannes Hingerl, Federico Tomazic, and Anika Maruscyk for their dedication to the project and insightful discussions. We also thank Josef Ruppenhofer for answering our questions on the Modalia annotation scheme, and Valentina Pyatkin and Shoval Sadde for answering our clarification questions on their models and experiments.

Limitations

Closed class of targets. Our work is limited to a closed class of linguistic expressions (modal verbs). Such approaches are sometimes seen as “too narrow” to be of interest to the NLP community. However, we argue that examining components of language understanding in detail will ultimately point to relevant research directions. In addition, as we have shown, modal verbs are a very common phenomenon, occurring in about every tenth sentence in scientific text. Nevertheless, we admit that a limitation of our study is the focus on a closed set of verbs in the English language. Future work might generalize our ideas to a more open class of targets (yet, it is a challenge to come up with a well-defined selection).

Limited data for minority classes. For some categories, data is limited due to the difficulty of data collection (we can only sample for modal verbs, not for labels). We have already enriched the dataset by a second annotation round, further data collection is unfortunately infeasible in the context of our project.

Applications. Our study provides the first steps (an annotated dataset, a corpus-linguistic study and NLP models) of research into the computational modeling of modal verbs in scientific text. Our distinctions intuitively should be of high relevance to processing and mining scientific text. Besides the case study on why the distinctions matter for open information extraction (IE) and practical suggestions for incorporating them into existing Open IE systems in Appendix E, demonstrating the usefulness of our work on existing scientific relation extraction datasets (which unfortunately do not commonly mark the “information status” of the annotated relations) is beyond the scope of this paper (but planned future work).

Ethics Statement

The corpus described in this work consists of open-access scientific articles annotated with several categories. The annotators involved in the project gave their explicit consent to the publication of their annotations and were compensated substantially above the minimum wage in our country.

References

Ron Artstein. 2017. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. Semi-Formal Evaluation of Conversational Characters. In *Languages: From Formal to Natural: Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, page 22–35, Berlin, Heidelberg. Springer.

Luciana Beatriz Avila and Heliana Mello. 2013. Challenges in modality annotation in a Brazilian Portuguese spontaneous speech corpus. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 1–6, Potsdam, Germany. Association for Computational Linguistics.

Kathryn Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Anita de Waard, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert M. Patton, and Michal Shmueli-Scheuer, editors. 2020. *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*, pages 113–120.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Yanyan Cui and Ting Chi. 2013. Annotating modal expressions in the Chinese treebank. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 24–32, Potsdam, Germany. Association for Computational Linguistics.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2019. *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2020. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. MinIE: Minimizing facts in open information extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640, Copenhagen, Denmark. Association for Computational Linguistics.

Kanyarat Getkham. 2011. Hedging devices in applied linguistics research articles. *Interdisciplinary discourses in language and communication*, pages 141–154.

Liane Guillou, Sander Bijl de Vroe, Mark Johnson, and Mark Steedman. 2021. Blindness to Modality Helps Entailment Graph Mining. *arXiv preprint arXiv:2109.10227*.

Timo Gühring, Nicklas Linz, Rafael Theis, and Annemarie Friedrich. 2016. SWAN: an easy-to-use web-based annotation system. In *Proceedings of the Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS)*, pages 103–113, Hildesheim, Germany.

Edith AS Hanania and Karima Akhtar. 1985. Verb form and rhetorical function in science writing: A study of MS theses in biology, chemistry, and physics. *The ESP Journal*, 4(1):49–58.

Tofan Dwi Hardjanto. 2016. Hedging through the use of modal auxiliaries in English academic discourse. *Humaniora*, 28(1):37–50.

Kevin Heffernan. 2021. *Problem-solving recognition in scientific text*. Ph.D. thesis, University of Cambridge.

Kevin Heffernan and Simone Teufel. 2018. Identifying Problems and Solutions in Scientific Text. *Scientometrics*, 116(2):1367–1382.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the manually annotated sub-corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V. Chawla, and Meng Jiang. 2019. The role of "condition": A novel scientific knowledge graph representation and construction model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, page 1634–1642, New York, NY, USA. Association for Computing Machinery.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA. Association for Computational Linguistics.

Liza King and Roser Morante. 2020. Must children be vaccinated or not? annotating modal verbs in the vaccination debate. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5730–5738, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Petr Knoth, Christopher Stahl, Bikash Gyawali, David Pride, Suchetha N. Kunnath, and Drahomira Herrmannova, editors. 2020. *Proceedings of the 8th International Workshop on Mining Scientific Publications*. Association for Computational Linguistics, Wuhan, China.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online. Association for Computational Linguistics.

Angelika Kratzer. 1981. The notional category of modality. *Words, worlds, and contexts*, 38:74.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology.* Sage, Beverly Hills, CA.

Valeria Lapina and Volha Petukhova. 2017. Classification of modal meaning in negotiation dialogues. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018a. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018b. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3326–3338, Brussels, Belgium. Association for Computational Linguistics.

Bo Li, Mathieu Dehouck, and Pascal Denis. 2019. Modal sense classification with task-specific context embeddings. In *ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium. hal-02143762.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ana Marasović and Anette Frank. 2016. Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 111–120, Berlin, Germany. Association for Computational Linguistics.

Ana Marasović, Mengfei Zhou, Alexis Palmer, and Anette Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. In *Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning*. CSLI Publications.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Amália Mendes, Iris Hendrickx, Luciana Ávila, Paulo Quaresma, Teresa Gonçalves, and João Sequeira. 2016. Modality annotation for Portuguese: from manual annotation to automatic labeling. *Language Issues in Language Technology (LiLT)*, 14.

Lori Moon, Patricija Kirvaitis, and Noreen Madden. 2016. Selective annotation of modal readings: Delving into the difficult data. In *Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning*. CSLI Publications.

Vivi Nastase, Benjamin Roth, Laura Dietz, and Andrew McCallum, editors. 2019. *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. Association for Computational Linguistics, Minneapolis, Minnesota.

Harinder Pal and Mausam. 2016. Demonyms and compound relational nouns in nominal open IE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39, San Diego, CA. Association for Computational Linguistics.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge University Press.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English GigaWord, fourth edition. *LDC2009T13*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Paul Portner. 2009. *Modality*, volume 1. Oxford University Press.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 953–965, Online. Association for Computational Linguistics.

Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves. 2014a. Automatic Tagging of Modality: identifying triggers and modal values. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 95–101. European Language Resources Association.

Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves. 2014b. Tagging and labelling Portuguese modal verbs. In *International Conference on Computational Processing of the Portuguese Language*, pages 70–81. Springer.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 38–46, Potsdam, Germany. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. 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 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1538–1545, Istanbul, Turkey. European Language Resources Association (ELRA).

Ute Römer. 2004. A corpus-driven approach to modal auxiliaries and their didactics. *How to Use Corpora in Language Teaching, Studies in Corpus Linguistics*, pages 185—-199.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357, Vancouver, Canada. 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.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Masahiro Takimoto. 2015. A corpus-based analysis of hedges and boosters in English academic articles. *Indonesian Journal of Applied Linguistics*, 5(1):95–105.

PyTorchLightning Team. 2020. Torchmetrics: Machine learning metrics for distributed, scalable PyTorch applications. `https://github.com/PyTorchLightning/metrics`.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Meagan Vigus, Jens E. L. Van Gysel, and William Croft. 2019. A dependency structure annotation for modality. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 182–198, Florence, Italy. Association for Computational Linguistics.

Kai Von Fintel. 2006. Modality and language. In Donald M. Borchert, editor, *Encyclopedia of Philosophy - Second Edition*. Macmillan Reference USA.

Sander Bijl de Vroe, Liane Guillou, Miloš Stanojević, Nick McKenna, and Mark Steedman. 2021. Modality and negation in event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 31–42, Online. Association for Computational Linguistics.

Astuko K Yamazaki. 2001. The pragmatic Function of Modal Verbs in Scientific Papers. Technical Report 36, Tokyo University of Fisheries.

Mengfei Zhou, Anette Frank, Annemarie Friedrich, and Alexis Palmer. 2015. Semantically enriched models for modal sense classification. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 44–53, Lisbon, Portugal. Association for Computational Linguistics.

Supplementary Material

A Comparison to Existing Annotation Schemes for Modal Senses

Table 3 classifies a set of utterances according to our, RR12's and Rubin13's schemes (according to our interpretation of their guidelines).[12] These two works inspired ours, but with the aim of knowledge graph construction in mind, we tailored an annotation scheme making explicit the various pragmatic and rhetorical reasons for using modals in scientific writing. We thereby follow Moon et al. (2016), who argue that "not everything about modal auxiliary meaning can be represented at once," and that "it is important to focus on the parts of modal auxiliary meaning that most directly impact an automated learner." While we fully agree with the linguistic classification of the examples by RR12 and Rubin13, we found certain sub-distinctions to be essential for understanding modal usage in the scientific context, and designed our annotation scheme for *functions* of modals accordingly, intentionally conflating what is traditionally treated separately as *modal sense disambiguation* and *veridicity* (Karttunen and Zaenen, 2005) from the author's point of view.

The definition of Rubin13's label *Circumstantial*, focusing less on dispositions rather than on abilities in particular circumstances (Von Fintel, 2006), is closer to our ***feasibility***, which could be interpreted as an ability of the actor given the circumstances (but sometimes overlaps with internal properties of the object under discussion). Conversely, we do not distinguish personal wishes and goals as in Rubin13. The label *deontic* for *can* of RR12 falls under our label ***options*** if options are introduced, and maps to our ***deontic*** otherwise. Within the *epistemic* notion, we further distinguish whether a statement is derived from other facts (***inference***) or whether an author ***speculates*** (both labels may apply at the same time). As some usages of modal verbs in scientific writing are rather conventional, we introduce the label ***rhetorical***.

B Further Corpus Statistics

B.1 Impact of Negation

Analyzing all negated modal verb constructions, we found only two instances where negation affects the modality label. For example, "Submarine volcanism alone cannot be the sole driving mechanism for OAEs" is labeled with ***capability*** when ignoring the negation. Otherwise, this becomes an ***inference***.

B.2 Comparison of Label Distributions of MiST, MASC, and Modalia$_{\text{M}}$

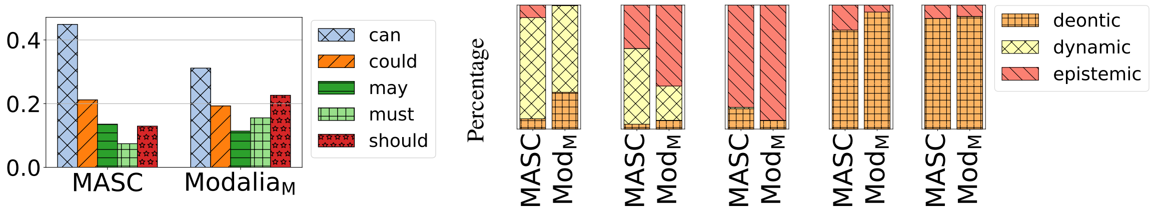
The distribution of modal functions and sense differs between corpora and genres (academic writing vs. news). Comparing Figure 3 and Figure 7, we note several differences. The most frequent modal in all genres is *can*, but it is much more frequent in CL, CS, and MS. For *can* and *could*, *dynamic*/***feasibility***/***capability*** uses are predominant, with the exception of Modalia$_{\text{M}}$, where the majority class of *could* is *epistemic*. *Can* and *could* are not used in the *deontic* sense in MiST; their *epistemic* uses are all related to ***speculation***.

C Annotation Guidelines

In this section, we describe our annotation guidelines for marking up modal verbs in scientific publications with regard to whether they are used for particular rhetorical, semantic or pragmatic reasons *as they were presented to the annotators*. Depending on the context, modal verbs can modify a sentence's propositional content such that uncertainty about the truth of the proposition is implied (e.g., "X *is* the cause for Y" vs. "X *may* be the cause for Y"), but in other circumstances, they simply indicate properties or capabilities (e.g., "X *can* float"). Our goal is to provide information about the functions of modal verbs in our corpus that then can be used in a preprocessing step for information extraction. For example, when disregarding a modal's contribution to the discourse, when processing "X *can* float", the relation `float(X)` may be extracted, but it should be flagged somehow as the sentence does not state that X is currently floating or that it always floats. In contrast, adding has_capability(X, float) to our knowledge base is desirable.

We consider *can, could, must, should, may, might* as well as their negated forms for annotation. They are pre-marked in the corpus to ensure that no modal verb is overlooked. Our annotation scheme is based on the observation that it is not always possible to assign exactly one type to every instance. We decided to follow a feature-based annotation approach in which a modal verb is represented by features that do or do not apply. Our feature sets reflects the range of functions a modal verb can

[12] To facilitate comparison with Pyatkin21, we also added their mapping to the Rubin13 scheme to the table.



(a) Distributions of modals (b) can (c) could (d) may (e) must (f) should

Figure 7: **MASC and Modalia$_M$**: Modal distributions and label distributions by modal verb.

have, i.e., the meaning that it adds to the sentence (e.g., capability), or the rhetorical or pragmatic reason for using it (e.g., uncertainty). To determine which features apply, annotators are asked to think of the sentence without the modal verb first, and then observe how the meaning has changed when adding the modal. Selecting the features accordingly means determining the reason(s) for which the author uses the respective modal verb. The next section explains the set of set of functions in our scheme.

It is important to note that we annotate *the author's* intentions and understanding, not the reader's (which might be based on additional context or knowledge). However, for making the judgment of why an author uses a modal verb in a particular context, annotators are of course asked to consider the broader context.

C.1 Functions of Modal Verbs

feasibility: We use this feature when it is possible for an external actor to do or achieve something and indicating this is the reason why the modal verb is used. ***feasibility*** can be seen as general possibility involving some external actor, e.g., a human agent.

Example 1. *Several supercapacitors **can** be integrated and connected in series.*

It is possible for somebody to integrate and connect supercapacitors, it needs a human agent to do it. The focus is not on an internal capability of the supercapacitors here.

capability: ***capability*** is annotated if the modal verb is used to express that something has a certain property, ability or capacity. Ruppenhofer and Rehbein (2012) have a corresponding category named *dynamic*. We mark up ***capability*** only in cases where the modal verb is used to convey information about an intrinsic property of an entity.

Example 2. *The device **can** light up a redlight-emitting diode and works well.*

The device has the ability to light, the device is able to light up, being able to light up is an inherent property of the device.

Example 3. *They hope the government **can** introduce a new law.*

The government is able to introduce a new law. Therefore ***capability*** is marked up. Note that even if the sense of the utterance is "it is desirable that the government introduces a new law," ***deontic*** doesn't apply here as the desire is expressed by "hope" and not by the clause containing the modal verb.

inference: This feature covers cases in which an author states that she inferred something based on some given information. ***inference*** corresponds to the category *epistemic* in previous work, as it also applies if the author draws a conclusion based on some information. ***inference*** applies especially when the author predicts something, e.g., based on computational results, experimental outcomes, or empirical knowledge. In order to correctly identify this feature, usually a broader context needs to be taken into consideration and domain knowledge is sometimes crucial.

Example 4. *The maximum power density was measured to 0.350 mW cm^2. Therefore, it **must** be the case that the open-circuit voltage reaches at least 1 V.*

Based on the measurement of the power density, the author infers that the open-circuit voltage is 1 V.

Example 5. *(According to these calculations...) The three lowest-energy isomers of C60O3 **should** exist in equilibrium at room temperature by using a modified and extended Hűckel method.*

The author predicts that these isomers exist in equilibrium at room temperature, based on some calculations.

speculation*: *speculation is used when the truth value of an utterance is not clear according to the author. Note that we annotate this feature only in cases where ***feasibility*** or ***capability*** are not clearly the predominant readings, and use both features only if a speculation reading is really predominant.

Example 6. *This problem **might** be mitigated by using better semantic-based retrieval model.*

Here, we label both ***feasibility*** and ***speculation***. Consider replacing *might* with *can*: then, the ***feasibility*** is clear, but no ***speculation*** is involved, which is the author's reason for choosing *might* instead.

options*: *options is marked up when the author uses the modal verb to enumerate some potential options.

Example 7. *The real shielding **can** of course be different.*

A different shielding may be used; potentially the shielding is different, but it can also stay the same. Note that "being different" is not a property, hence ***capability*** or ***feasibility*** wouldn't fit here.

Example 8. *Grounding this in our example, w1 **may** represent breakfast, w2 pancakes, and w4 hashbrowns.*

Breakfast, pancakes and hashbrowns are options for w1, w2 and w4.

Example 9. *This process **can** last from several hours to a few days depending on the applied temperature.*

The reason for using the modal verb here is mostly to convey the uncertainty about the duration, it does not describe a capability of the process.

Example 10. *We showed that combining a model based on minimal units with phrase-based decoding **can** improve both search accuracy and translation quality.*

In this case, we label both ***capability*** and ***options***, as the sentence both indicates a capability of the combination method, but at the same time could be read as a hedging term (i.e., improvements occur only in certain circumstances).

deontic*: *deontic is selected if the author uses the modal verb to express a desire, i.e., how the world should be like, to express a requirement for something, e.g., an experiment, or to state an obligation.

Example 11. *Our daily life requires matchable energy storage devices, which **should** have the capability to endure high-level strains.*

It is desirable that energy storage devices have the capability to endure high-level strains.

Example 12. *A GCR proton at the maximum latitude of the ISS **must** have at least about 150 MeV to reach the Station.*

It is required that a GCR proton has at least about 150 MeV to reach the station.

Example 13. *Temperature **should** be treated as a concept.*

The author prescribes that temperature is treated as a concept, it is necessary that temperature is treated as a concept.

rhetorical*:** In some contexts, modal verbs are used because of conventions and there is no substantial semantic need for doing so.[13] We annotate this cases with ***rhetorical.

Example 14. *It **can** be seen in Figure 1 that...*

Can simply be stated at 'In Figure 1 you see ..."

If annotators feel that ***feasibility*** or ***capability*** are also strongly present in such a case, they may select these features in addition.

Example 15. *Value: <first part of the definition> The value **can** also be described via <second part of the definition>.*

A value is defined as (first part of the definition) and also as (second part of the definition). (In this example, ***speculation*** and ***feasibility*** are also applicable.)

Other: This label is used if none of the above features apply. Please extract those sentences and explain why you couldn't decide for a predefined feature. Also think about whether you have a tendency towards one or more features but there is something that has to be captured in our scheme in additional. *This was used during annotation scheme development.*

C.2 Additional examples: *feasibility* vs. *capability*

As stated above, features are not mutually exclusive. Sometimes, multiple readings/interpretations may be possible. Under certain circumstances annotators are asked to select multiple features. In this section, we show some not-so-clear-cut examples to complement the above guidelines, which work with mostly clear examples.

[13] In several cases that we observed, we also felt that they corresponded to over-use of modal verbs by non-native English speakers (though they are not syntactically wrong, just unnecessary to some extent).

If we want to annotate ***feasibility*** or ***capability*** but it is hard to decide which of both features apply, we follow the following guidelines.

We annotate both features when there is an external actor (e.g., a human) involved, but if it can also be interpreted as describing a particular internal property of the referent of the subject. The referent has this property already before an external actor is involved.

Example 16. *This simpler distribution Q* ***can*** *be viewed as an approximation to P.*
feasibility: A human agent views Q as an approximation to P.
capability: Without a human agent viewing Q, Q is still an approximation to P, P has the property of being an approximation to P in general without somebody actually viewing it.

Example 17. *For instance, despite graphene, the band gaps of silicone* ***can*** *be opened and tuned when exposed to an external electric field.*
feasibility:A human agent opens the band gaps of silicone.
capability: some materials have the property of having openable band gaps, it is always possible to open band gaps of silicone under this circumstances.

Whenever there is a human actor involved, we mark up ***feasibility*** even if the sentence includes a passive construction which could indicate a ***capability***. ***capability*** and ***feasibility*** are only used at the same time if the modal verb is used to signal an intrinsic property (band gaps *of* silicone can be opened) and an external actor is involved. We do not mark up ***capability*** if ***feasibility*** applies but there isn't a general property. In this case an external actor has to do something first. As a consequence, some entity has a capability.

Example 18. *The resulting expression combines similarity terms which* ***can*** *be divided into two groups.*
feasibility: An human actor is needed to divide the terms into groups. Being dividable is not an intrinsic, common property of these terms. ***feasibility*** is the strongest modal function in this utterance. The modal verb is not used to convey an information about a ***capability***, as it is not an intrinsic property of similarity terms that they can be divided (we consider this to be an artifact of their being grouped).

Hyperparameter	CNN	SB
Learning rate	$1e-3, 5e-3, 1e-4, 5e-4, 3e-5, 5e-5$	$5e-4, 3e-5, 5e-5$
# warm-up epochs	N/A	$1, 2$
Batch size	$8, 16, 32$	$8, 16, 32$
Dropout	$0.1, 0.5$	N/A

Table 10: **Hyperparameter values** searched during hyperparameter selection for CNN and SB.

Example 19. *Similar symmetry* ***can*** *be achieved with the following factorization.*
feasibility: it needs a human agent to achieve something.
The modal verb is not used to indicate an intrinsic property of being achievable. The sense of the utterance is that somebody, i.e., the author, achieves similar symmetry with a certain formula that they mention. Even if no human actor is explicitly mentioned in the text due to a passive construction, only ***feasibility*** may apply.

Example 20. *Word vectors* ***can*** *be trained directly on a new corpus.*
feasibility: It is possible for somebody to train some word vectors on a new corpus.
Word vectors cannot be trained directly on a new corpus in general, not all word vectors are trainable on a new corpus, so we don't annotate ***capability***.

When it is clearly possible for an entity to have a property but this doesn't apply in general, we still use ***capability***, but possibly ***speculation*** in addition.

Example 21. *graphene aerogels with ...* ***can*** *present superelasticity.*
capability: some of these aerogels have this property, it is possible for aerogels to have this property.
speculation: It is uncertain whether each aerogel has this property, only some of them may present superelasticity, or aerogels have this property only under particular circumstances.

D Experimental Studies

D.1 Hyperparameters

This section describes the hyperparameter tuning for our main experiments. For CNN and SB, we tune learning rates, batch sizes, dropout probabilities (only CNN) and learning rate warm-up lengths (only SB) using grid search on the values shown in Table 10 as follows: Similar to cross validation (CV), for each hyperparameter configuration, we train five models on 4 folds each for 10 epochs and

	can	**could**	**may**	**might**	**must**	**should**
Maj	85.0	78.1	80.7	91.5	93.5	92.0
CNN	91.3	85.8	84.5	91.2	93.6	93.3
SB_{CLS}	93.7	90.1	86.7	92.1	**96.7**	96.2
SB_{modal}	94.2	90.2	86.6	**92.6**	96.6	96.7
$\text{SB}_{\text{CLS-mark}}$	**94.4**	89.3	86.6	91.9	96.2	96.8
$\text{SB}_{\text{CLS, modal}}$	94.3	**90.8**	**87.0**	**92.6**	**96.7**	**96.9**
$\text{BERT}_{\text{CLS,modal}}$	93.7	90.7	86.5	92.1	96.0	**96.9**
$\text{BERT-large}_{\text{CLS,modal}}$	94.3	89.6	86.0	92.3	96.5	96.8

Table 11: **Accuracy on test set of MiST**. Standard deviations are rather small, between 0 and 1.4.

use the respective remaining fold (*validation fold*) for model selection. For each of the five models, we average weighted F_1 scores (see Sec. 5.1) on the validation fold across modal verbs. We then choose the hyperparameter setting that performs best on average across the different models. The tuned batch sizes and learning rates are 32 and 5^{-3} (CNN), and 8 and 3^{-5} (SB). SB is warmed up for 2 epochs. We use a dropout probability of 0.1 in the output heads, and the Adam optimizer (Kingma and Ba, 2014) with a weight decay of 1^{-3} (CNN) respectively 0 (SB).

D.2 Training Details, Model Size, etc.

All experiments were performed on a single Nvidia Tesla V100 GPU. Training and testing the $\text{SB}_{\text{CLS,modal}}$ models in the 5-fold CV training setting used in the model architecture comparison experiment (cf. Table 5) took 1.2 hours (for the entire experiment).

SciBERT has the same number of parameters as BERT-base, i.e., 110M. The linear layer we add on top of SciBERT in the $\text{SB}_{\text{CLS,modal}}$ has less than 11k parameters.

D.3 Further Experiment Results

This section provides further experimental results, elaborating on the study described in Sec. 5.4.

Table 11 provides accuracy scores for the models whose F_1 scores are reported in Table 5.

D.4 Cross-Genre Multi-Tasking Experiment

We investigate whether we can improve classification on MiST by using existing modal sense classification datasets as auxiliary tasks in training. Table 12 shows the results of co-training with $\text{Modalia}_{\text{M}}$, MASC, EPOS, and GME (see Sec. 2), and the first three at once. On GME, we follow Pyatkin21's experiments and collapse *Desires+Wishes* and *Plans+Goals* to a *Intentional*.

Train	**can**	**could**	**may**	**might**	**must**	**should**
MiST	77.4 ±1.0	73.7 ±3.8	47.2 ±1.1	**64.5** ±2.7	**78.4** ±1.1	85.7 ±0.5
+ EPOS	**78.4** ±1.5	69.6 ±3.9	49.4 ±2.2	64.0 ±2.4	74.9 ±2.4	86.1 ±2.5
+ MASC	**78.4** ±1.6	72.2 ±1.3	**51.5** ±1.5	63.2 ±3.2	75.8 ±1.9	84.4 ±1.0
+ $\text{Modalia}_{\text{M}}$	76.6 ±1.5	71.5 ±1.9	49.4 ±3.3	59.9 ±3.4	75.6 ±4.3	**86.4** ±1.4
+ GME	76.7 ±2.1	70.5 ±4.2	47.0 ±2.4	62.3 ±6.0	70.8 ±6.1	84.1 ±1.6
+ E/M/Mo	77.0 ±1.6	**73.9** ±1.7	50.0 ±2.4	62.8 ±3.3	77.2 ±1.9	85.6 ±0.6

Table 12: **Multi-task setup: Macro F_1** on test set of MiST when co-training with other corpora. E/M/Mo: EPOS, MASC and $\text{Modalia}_{\text{M}}$ together.

The only verb where co-training leads to clear improvements is *may*. Here, it increases per-label F_1 scores (not reported in tables) for ***spec.***, ***opt.***, ***feas.*** (for the latter two except for GME), and ***cap.*** (except for $\text{Modalia}_{\text{M}}$). For the other verbs, classification performance is similar (e.g., *should*) or decreased (e.g., *might*, which is only covered by GME, but just with few instances). Thus, in line with the findings from the pure transfer experiment, using modal sense information from out-of-genre datasets for classifying modal verbs in scientific text is non-trivial.

E Case Study: Treatment of Modality in Open Information Extraction

We now discuss how handling modal verbs in Open Information Extraction (OIE) systems may be improved using our classification scheme by adding interpretations instead of just pin-pointing modal verbs. The same principles can be applied to relation extraction settings with predefined schemas when relations are rooted in a verbal argument structure.

E.1 Analysis of Existing OIE Systems

We run four popular recent OIE systems on sentences from MiST and perform a qualitative analysis of the results. We find that the examined systems either have no specific mechanism for handling modality, or include modality information only in rather rudimentary ways. **OpenIE4**[14] (Christensen et al., 2011; Pal and Mausam, 2016) and **OpenIE6**[15] (Kolluru et al., 2020) extract information from sentences in the form of standard subject–relation–object triples, simply considering modals part of the predicate, e.g., a sentence such

[14] knowitall.github.io/openie/

[15] github.com/dair-iitd/openie6

as "X may influence Y" yields the extraction (X; may influence; Y). **RnnOIE**[16] (Stanovsky et al., 2018) generates a representation resembling Semantic Role Labeling (SRL), in which spans within the sentence are annotated to indicate the relation-evoking verb and its respective arguments, e.g., `[ARG0: X] [ARGM-MOD: may] [V: influence] [ARG1: Y].` Within this representation, modal verbs are treated as a simple modifier of the relation verb (ARGM-MOD). In sum, modals are extracted by all of these OIE systems, but their classification and interpretation is left to the downstream system.

MinIE[17] (Gashteovski et al., 2017) includes a notion of modality by adding a binary modality value (*certainty*/*possibility*) to each extracted triple. In practice, we observe that the occurrence of virtually any modal in the input sentence results in the triple being assigned the *possibility* label. This means that sentences such as "X can influence Y," "X should influence Y," "X must influence Y," or "X may influence Y" are in effect all being treated as paraphrases. In sum, existing state-of-the-art OIE systems do not handle the meaning of modal verbs in a way that could inform downstream use.

E.2 Discussion: Modality-informed Open IE

In light of the weaknesses of existing systems, we now sketch an approach by which OIE systems could be extended to incorporate modality information, which could be generated by a classifier (as described in Sec. 4). As motivated by Figure 1, we posit that there are two main ways in which modality information should be incorporated into extractions. (For an overview, see also Table 13.) First, we propose specific relation templates for the ***capability*** and ***deontic*** modalities: *hasCapabilityTo_** for the former and *isRequiredTo_** and *isAllowedTo_** for the latter. In a given extracted triple, these relation templates would be instantiated with the main verb of the extraction, e.g., "X can influence Y" (***capability***) would yield *(X, hasCapabilityTo_influence, Y).*[18]

Second, to cover cases modifying not only the relation but the entire fact, we propose the meta-property *hasFactualityRating* (see also Figure 1). This property could take the values *speculation* (for ***speculation***), *possible* (for ***options*** and ***feasibility***), *inferred* (for ***inference***), and *true* (for ***rhetorical*** and as the default value of the property). For example, the sentence "X might influence Y" (***speculation***) would yield *(X, influence, Y)* with *hasFactualityRating(speculation)*, whereas "These sandstones may contain reworked material." (***options***), would lead to *(sandstones, contain, reworked_material)* with *hasFactualityRating(possible)*. Similar approaches to handling veridicality of utterances have for instance been proposed by de Marneffe et al. (2012).

Modal function	**IE extraction(s)**
capability	*hasCapabilityTo_**
deontic	*isRequiredTo_** (must, should) / *isAllowedTo_** (other modals)
feasibility	*hasFactualityRating(possible)*
inference	*hasFactualityRating(inferred)*
speculation	*hasFactualityRating(speculation)*
options	*hasFactualityRating(possible)*
rhetorical	*hasFactualityRating(true)*

Table 13: **Mapping modal functions to Open IE extractions**, *=modified main verb.

We argue that such an approach would constitute an improvement over existing ways of handling modality in OIE. Enabling the identification across surface representations is one aim of OIE systems. Looking further ahead, explicitly disambiguating modal verbs as well as other constructions expressing the same meaning will result in a uniform representation. For example, *X can (**capability**) influence Y)* and *X is able to influence Y* would both be retrieved by searching for *hasCapability*, and *X must (**deontic**) Y* and *X has to Y* would be retrieved when searching for *isRequiredTo*. In addition, *hasFactualityRating* properties of extracted triples will immediately clarify their factuality status, avoiding, e.g., erroneously taking speculation as fact. Taken together, we have outlined a way to take OIE systems to the next level with regard to the treatment of modal verbs.

[16] demo.allennlp.org/open-information-extraction

[17] github.com/uma-pi1/minie

[18] In an OWL-like ontology, these concretely instantiated predicates may then be considered subproperties of generic *hasCapabilityTo* / *isRequiredTo* / *isAllowedTo* properties.