The Possible, the Plausible, and the Desirable: Event-Based Modality Detection for Language Processing

Valentina Pyatkin* Bar Ilan University pyatkiv@biu.ac.il

Shoval Sadde* Bar Ilan University shovatz@gmail.com

Aynat Rubinstein Hebrew University of Jerusalem aynat.rubinstein@mail.huji.ac.il

Paul Portner Georgetown University paul.portner@georgetown.edu

Reut Tsarfaty Bar Ilan University reut.tsarfaty@biu.ac.il

Abstract

Modality is the linguistic ability to describe events with added information such as how *desirable, plausible*, or *feasible* they are. Modality is important for many NLP downstream tasks such as the detection of hedging, uncertainty, speculation, and more. Previous studies that address modality detection in NLP often restrict modal expressions to a closed *syntactic class*, and the modal *sense labels* are vastly different across different studies, lacking an accepted standard. Furthermore, these senses are often analyzed independently of the events that they modify. This work builds on the theoretical foundations of the *Georgetown Gradable Modal Expressions* (GME) work by Rubinstein et al. (2013) to propose an *event-based modality* detection task where modal expressions can be words of any syntactic class and sense labels are drawn from a comprehensive taxonomy which harmonizes the modal concepts contributed by the different studies. We present experiments on the GME corpus aiming to detect and classify fine-grained modal concepts and associate them with their modified events. We show that detecting and classifying modal expressions is not only feasible, but also improves the detection of modal events in their own right.

1 Introduction

Modality refers to the linguistic ability to describe alternative ways the world could be.[1] Modal expressions aim to identify wishes, rules, beliefs, or norms in texts (Kratzer, 1981; Portner, 2009), which is a crucial part of *Natural Language Understanding* (NLU) (Morante and Sporleder, 2012).

Concretely, events in natural language are often reported in a manner that emphasizes *non-actual* perspectives on them, rather than their actual *propositional* content. Consider examples (1a)–(1b):

(1) a. *We presented a paper at ACL'19.*
b. *We did not present a paper at ACL'20.*

The propositional content p ="present a paper at ACL'X" can be easily verified for sentences (1a)-(1b) by looking up the proceedings of the conference to (dis)prove the existence of the relevant publication. The same proposition p is still referred to in sentences (2a)–(2d), but now in each one, p is described from a different perspective:

(2) a. *We* **aim** *to present a paper at ACL'21.*
b. *We* **want** *to present a paper at ACL'21.*
c. *We* **ought** *to present a paper at ACL'21.*
d. *We are* **likely** *to present a paper at ACL'21.*

These sentences cannot be verified or falsified simply by examining whether p actually came or will come to pass, and in fact, such verification is not the goal of this way of reporting. Rather, speakers describe such events in order to indicate PLANS (2a), DESIRES (2b), NORMS (2c), or the assessed PLAUSIBILITY (2d) of the associated propositional content p. Investigating how to classify these perspectives on events has been the focus of extensive research on modality in theoretical linguistics (Kratzer, 1981; Palmer, 1986; Portner, 2009).

In terms of NLP technology, modal concepts as expressed in (2) are relevant to many downstream tasks, such as the automatic detection of hedging and speculation (Vincze et al., 2008; Malhotra et al., 2013), uncertainty (Vincze et al., 2008; Miwa et al., 2012; Zerva et al., 2017; Prieto et al., 2020), opinion (Wiebe et al., 2005; Rubin, 2010; Miwa et al., 2012), and factuality (Saurí and Pustejovsky, 2009; Rudinger et al., 2018). Although these tasks rely on modality features, so far there is no accepted standard for modal concepts and labels, which aligns with the semantic space of modal senses that linguists identify. Consequently, modality features are

*Equal contribution

[1] In formal semantics, these alternatives are referred to as *possible worlds* or *situations* (Kripke, 1959; Lewis, 1973; Barwise and Perry, 1981; Kratzer, 2010).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 953–965
August 1–6, 2021.

either treated idiosyncratically or are absent from semantic frameworks (Donatelli et al., 2018, §4.6).

In support of such downstream tasks, a different type of NLP investigations targets modality annotation and detection in its own right (Ruppenhofer and Rehbein (2012); Baker et al. (2012); Zhou et al. (2015); Marasović and Frank (2016); Hendrickx et al. (2012); Nissim et al. (2013); Ghia et al. (2016); Mendes et al. (2016); Lavid et al. (2016), and others). However, each of these studies creates its own scheme, and none of these schemes has been picked up as an accepted standard by the community. Moreover, different endeavors suffer from one (or more) of the following types of deficiencies with respect to their expressivity and coverage.

First, many studies limit the modal *triggers*, i.e., the expressions that trigger the modal meaning, to a closed class of auxiliary verbs (e.g., *can, might, should, must* in English (Ruppenhofer and Rehbein, 2012; Marasović et al., 2016; Quaresma et al., 2014)). However, as acknowledged by linguists (Kratzer, 1981) and NLP researchers (Rubin, 2010; Baker et al., 2012; Nissim et al., 2013), words of any Part-of-Speech (POS) can trigger modality. Consider, for instance, the following triggers: *We* **should** *remain calm* (AUX); *We have a* **plan** *to reduce the costs* (NOUN); *Our agency* **prefers** *this equipment* (VERB); *Marx is* **probably** *patriotic* (ADV); *Devaluation has been* **necessary** (ADJ).

Second, the modal *senses*, i.e., the labels that indicate the modal perspectives, differ from one study to another, with no accepted standard. Some studies focus only on a particular sense, such as epistemic modality (Rubin, 2010; Ghia et al., 2016). Others use labels that mix modal senses with orthogonal notions (e.g., *force*, distinguishing permission from requirement as in Baker et al. (2012)), thereby making their deployment into existing annotations and tasks less transparent. In general, there is no single annotation standard that covers the full spectrum of modal senses attested in the data and confirmed by the latest linguistic theories, as portrayed by Portner (2009).

Finally, *modality detection* in NLP has often been cast as a word-sense disambiguation (WSD) task (Ruppenhofer and Rehbein, 2012) or as a sentence-classification task (Marasović and Frank, 2016). Both perspectives are insufficient for any practical use. The latter is too coarse-grained, as a sentence may contain multiple events, each of which potentially carries a different modal sense. The former is uninformative, because the modal trigger is not explicitly associated with the event being modified. Ghia et al. (2016) take a step in the right direction, offering to annotate modal sense *constructions*.

The current work proposes to address all of the aforementioned deficiencies as follows. We define a prediction task that we term *event-based modality detection*, where, given a sentence as input, we aim to return all of its modal *triggers*, their associated modal *senses*, and, for each trigger, the respective *event* being modified. Crucially, the modal triggers can be from any syntactic class. The modal senses are drawn from a single taxonomy that we motivate based on linguistic research and which harmonizes the different modal concepts contributed in previous studies (§3). Finally, we propose to view modal triggers as semantic modifiers of eventive heads in event-based (a.k.a., Neo-Davidsonian; Parsons (1990)) semantics. This is motivated by practical concerns – when extracting events from texts to benefit downstream tasks, one would want easy access to the features that indicate the perspective on each event, above and beyond its participants.

The accompanying annotation standard we assume for the task is based on the *Georgetown Gradable Modal Expressions* (GME) framework (Rubinstein et al., 2013), with two simplifications that are designed to allow for more consistent annotations and increased ease-of-use by non-experts. First, we change the modal sense labels to be intuitive and self-explanatory. Second, instead of the event span (a.k.a., *prejacent*) in the GME, we mark the *head* of the event being modified.

To assess the feasibility of the proposed task, we use the GME corpus (Rubinstein et al., 2013) to train and test the automatic detection of modal *triggers*, their *senses*, and associated *events*. Our experiments show that while identifying a closed set of auxiliary verbs as modal triggers is straightforward, expanding the set of triggers to any syntactic class indeed makes it a harder task. Notwithstanding this difficulty, we show that a model based on large pre-trained contextualized embeddings (Liu et al., 2019) obtains substantial improvements over our baseline on the full task. Moreover, we show that detecting modalized events in fact improves with the availability of information about the modal triggers. All in all, we contribute a new task, a new standard and a set of strong baselines for the event-based modality task we defined.

2 Linguistic Background

Modal expressions allow language users to discuss alternative realities. For example, the sentence *She can reach the ceiling* is modal because it describes the event of her reaching the ceiling as feasible, but potentially non-actual. Similarly, *She hopefully will reach the ceiling* is modal because it describes such an event as desirable, and likewise potentially non-actual. A sentence like *She was reported to reach the ceiling* describes the event of her reaching the ceiling as potentially actual, according to one's state of knowledge, yet implying that in reality it could have been otherwise.

Over the last 40 years linguists have achieved an increasingly refined understanding of how to classify modal senses. The most traditional and fundamental distinction is between *epistemic* modals and *non-epistemic* modals (also called *root* modals). Epistemic modals have to do with knowledge and plausibility of the event actually happening. Non-epistemic modals have to do with agent actions and motivations underlying the events.[2]

Epistemic modality is not a unified class. Some modals express a perspective on the event that is based on knowledge, while others express a perspective related to the objective chance of the event happening (a.k.a., *circumstantial* modality in Kratzer (1981)). Furthermore, linguists posit two types of non-epistemic modal senses: one which focuses on the *objective* abilities and dynamic unfolding of events (Palmer, 1986), and another which focuses on *subjective* reasons to prioritise one event over another (Portner, 2009). Within the latter subtype there are further subdivisions according to whether the event is prioritised in terms of norms (*deontic*), desires/preferences (*bouletic*), or goals/plans (*teleological*) (Kratzer, 1981; Portner, 2009; Rubinstein, 2012; Matthewson and Truckenbrodt, 2018).

The traditional three-way classification of modal senses into *deontic, epistemic*, and *dynamic*, which has been used in previous NLP work (e.g., Ruppenhofer and Rehbein (2012); Marasović et al. (2016)), did not attend to these subdivisions, which are nonetheless expected to be important for reasoning and other tasks that require deep understanding. Baker et al. (2012) make finer-grained distinctions in the non-epistemic case, distinguishing between requirements, permissions, wants, and intentions, but not all of these in fact track distinct modal senses. For example, their "require" modality conflates both rule-based obligations and goal-oriented preferences.

Most importantly, the discussion of modality in NLP often resorts to linguistic regimes that are not understandable by non-linguists and non-expert practitioners, making the output of these systems essentially unusable for NLP engineers and designers of downstream tasks. This paper aims to bridge this gap, offering a single task and annotation standard that cover the rich space of concepts, while being intuitively understandable and easy-to-use.

A Note on Modality vs. Factuality. A related but different line of work in NLP investigates the automatic identification and classification of the *factual* status of events (Saurí and Pustejovsky, 2009; Rudinger et al., 2018). That is, the factuality classification task has to do with automatically detecting whether, in actuality, a reported event has *happened* or has *not happened*.[3]

It is important to note that *factuality* and *modality* are distinct and completely orthogonal notions (see, e.g., Ghia et al. 2016). For example, the sentences *The WSJ announced that she reached the shore* and *She was able to reach the shore* share the propositional content of $p =$ '*she reached the shore*' and its implied factuality status (happened), but differ in the manner of reporting the event p. The former is based on knowledge, while the latter puts emphasis on the ability of the agent in p. It is precisely this change of perspective that is missing in the realm of NLU and related downstream tasks.

The upshot of Rudinger et al.'s (2018) work is the claim that factuality is determined at event level, and that expressions contributing to factuality may be of any syntactic class. We likewise propose to relate modal triggers to an event being modified, and we similarly adopt an inclusive view of the syntactic classes that express modality. In contrast to event-based factuality detection, as proposed by Rudinger et al. (2018) and others, which classifies which events came to pass, event-based modality detection as proposed here classifies an orthogonal dimension of meaning related to semantic properties of events that *may* be non-actual, providing information about *why* they are portrayed as such.

[2]The same split is motivated also on syntactic grounds: epistemic modals appear in high positions in the syntactic structure, in particular above tense and aspect, while root modals appear lower in the structure, closer to the verb phrase (see Hacquard (2010) for an overview).

[3]Rudinger et al. (2018) define factuality status on a scale of {+3,-3}. 0 indicates an event with unclear factuality status.

3 Event-Based Modality Detection: Proposed Task Definition

We propose an *event-based modality* detection task that rests upon three assumptions: (i) the set of possible *modal triggers* is open-ended, and may be of any POS tag, (ii) the associated *modal senses* are fine-grained and form an hierarchical taxonomy, and (iii) each trigger is associated with an *event*.

Consider, for instance, the following examples:

(3) a. He was **reported**$_i$ to *be*$_i$ in custody.
b. It is **believed**$_j$ that the glass will *make*$_j$ it **possible**$_k$ to *see*$_k$ the satellite at night.

In these examples, the words in bold indicate the modal expression, which we call a *trigger*. The co-indexed items in italics mark the head of the event for which the modal perspective is ascribed. In (3a), '**reported**' triggers a modal perspective on the event of '*being (in custody)*'. In (3b), '**believed**' triggers a modal perspective on the '*making*' event, and '**possible**' indicates a modal perspective on the '*seeing (the satellite)*' event.

Clearly, the modal perspectives on these events, i.e., the modal senses, are of different types. How should we label these fine-grained modal senses?

A Hierarchical Taxonomy of Modal Senses Having established that a given expression serves as a modal trigger, we are interested in classifying the particular sense, or perspective, that it assigns to the modal event. Figure 1 presents the complete taxonomy that we propose for modal sense classification in NLP. It is based on the modal senses proposed and justified by Rubinstein et al. (2013), with a few simplifications that make it intuitive and easy-to-use by NLP practitioners and non-linguists.[4]

The highest level of the hierarchy tracks the distinction between events whose PLAUSIBILITY is being assessed, and events whose PRIORITY is stated. More specifically, plausibility has to do with events that are expected to happen or not happen, given a relevant set of assumptions which are made explicit. Plausibility can be assessed based on our state of knowledge ("I **heard**$_i$ she *got married*$_i$"), based on what is objectively probable due to facts about the world ("The ice cream will **definitely**$_i$ *melt*$_i$ in the sun"), or based on inherent (physical) abilities of an agent ("I **can**$_i$ easily *swim*$_i$ 10 km").

[4]Cf. Manning's Law, item 5 https://en.wikipedia.org/wiki/Manning's_Law

Priority	
Norms and Rules	*the ballot which* **must** *be held by the end of March*
Desires and Wishes	*we do* **support** *certain limitations on the villains*
Plans and Goals	*a* **necessity** *emerged to enter the Pilgrim's House*
Plausibility	
State of Knowledge	*The ship is* **believed** *to carry illegal immigrants*
State of the World	*The disease* **can** *be contracted if a person is bitten*
State of the Agent	*They are* **able** *to do whatever they want*

Table 1: Modal-Sense Examples

In contrast, the PRIORITY branch marks a perspective where events are prioritized, or considered "good" by the speaker (or more generally, by a relevant attitude holder) (Portner, 2009). Events can be preferred because they are normatively obliged or commendable ("You **should**$_i$ n't *drink and drive*$_i$), because they realize a goal ("The **plan**$_i$ to *reduce*$_i$ costs in Q2"), or because they are otherwise desirable, as a matter of personal taste or preference ("I will **preferably**$_i$ *meet*$_i$ them over lunch").

To make these notions accessible, we assign intuitive labels to these fine-grained concepts. On the PLAUSIBILITY side, we distinguish plausibility based on the state of KNOWLEDGE (previously, *epistemic*), plausibility based on a state of the WORLD (*circumstantial*), and plausibility based on the objective abilities of the AGENT (*dynamic*). On the PRIORITY side, we distinguish priorities based on RULES AND NORMS (*deontic*), priorities based on DESIRES AND WISHES (*bouletic*), and priorities based on PLANS AND GOALS (*teleological*). As illustrated in Table 2, modal triggers on both sides of the sense hierarchy may be of any POS tag.

The proposed taxonomy unifies and harmonizes the different modal senses offered by previous studies. Importantly, we enrich the *epistemic-deontic-dynamic* classification used in previous NLP research (Ruppenhofer and Rehbein, 2012; Marasović and Frank, 2016) with the finer-grained notions introduced by Rubinstein et al. (2013) and refer to the various labels in work by Baker et al. (2012); Mendes et al. (2016). More concretely, in GME and in our taxonomy, what in previous annotations was a monolithic *deontic* class (Ruppenhofer and Rehbein, 2012; Marasović and Frank, 2016) now corresponds to the PRIORITY node, with three linguistically-motivated sub-classes (Portner,

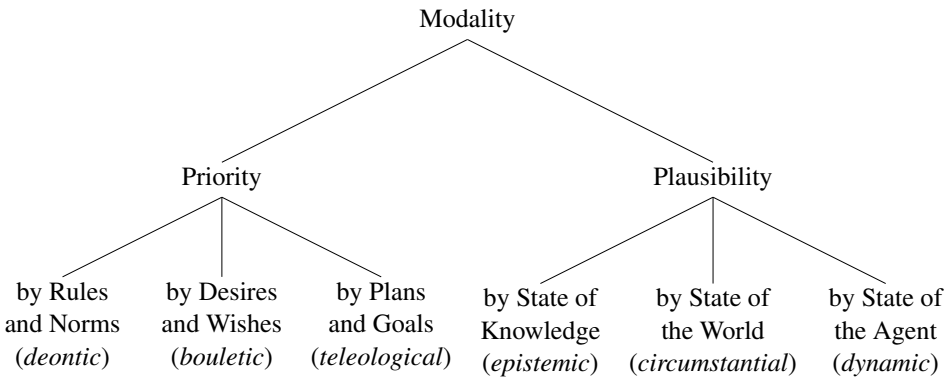

Figure 1: The Proposed Hierarchical Taxonomy of Modal Senses

	Priority	Plausibility
Aux	*We* **should** *remain calm*	*there is little I* **can** *do*
Verb	*Our agency seriously* **needs** *equipment*	*powers that* **enable** *him to defend the rights*
Noun	*a* **plan** *to reduce carbon-dioxide emissions*	*their* **incapacity** *to put crime under control*
Adverb	*Marx is* **sufficiently** *patriotic*	*President Mugabe* **easily** *won Zimbabwe's election*
Adjective	*devaluation was* **necessary**	*this complex decision was not* **easy** *for him*

Table 2: Modal Triggers with Diverse Parts-of-Speech Tags: Sentence Excerpts from the GME corpus.

2009): a RULES-AND-NORMS class, a DESIRES-AND-WISHES class, and PLANS-AND-GOALS.

Among modal events that do not involve priorities or norms, the sub-class which concerns the state of an AGENT corresponds to *dynamic* modality in previous studies (Ruppenhofer and Rehbein, 2012; Marasović et al., 2016). The two other sub-classes of plausibility modality, state of WORLD and state of KNOWLEDGE taken together, correspond to *epistemic* in these previous works.

To justify our fine-grained distinction, consider how the latter two senses, state of the WORLD and the state of KNOWLEDGE, correspond to interesting applications in the BioNLP literature, where it is vital to distinguish *fact* from *analysis* (Miwa et al., 2012). The difference is seen in the interpretations of **may** in the following examples from the BioScope corpus (Vincze et al., 2008):

(4) a. Symptoms **may** *include* fever, cough or itches.

b. The presence of urothelial thickening and mild dilatation of the left ureter suggest that the patient **may** *have* continued vesicoureteral reflux.

In (4a), we classify **may** to the plausibility branch with a state of the WORLD sub-class. In Miwa et al.'s terms this would be referred to as *fact*. In (4b), we classify **may** to the plausibility branch with a state of KNOWLEDGE sub-class. In Miwa et al.'s terms this would be referred to as *analysis*.

4 Experimental Setup

Goal We set out to assess the feasibility of our proposed event-based modality task. Concretely, we would like to gauge how well we can learn to detect and classify the different levels of modal senses afforded by our taxonomy (§3) and to identify the events modified by the triggers.

Data Our experiments use the Georgetown Gradable Modal Expressions Corpus (GME; Rubinstein et al. (2013)), a corpus obtained by expert annotations of the MPQA Opinion Corpus (Wiebe et al., 2005). The MPQA corpus is a 301,090-token corpus of news articles, which, following Ruppenhofer and Rehbein (2012), has become a benchmark for the annotation of modality.

The GME corpus annotates various properties of modal expressions, including their sense in context, the proposition they apply to, the polarity of their environment, and whether or not they are qualified by a degree expression.[5] Rubinstein et al. (2013) claim inter-annotator agreement scores as follows: Krippendorf's $\alpha = 0.89$ for a 2-way distinction corresponding to Priority versus Plausibility, $\alpha = 0.49$ for their finest-grained sense classification, and $\alpha = 0.65$ for prejacent span detection.

We processed the corpus by extracting the modal triggers and their corresponding proposi-

[5] See Rubinstein et al. (2013) for details about the annotation process and the full scheme of annotated features.

tional spans (*propositional argument* in GME) into a CoNLL-formatted file. Using spaCy (Honnibal et al., 2020), we obtained the lemmas, POS tags, and dependencies. The topmost head of the propositional span is considered the head of the event being modified. We transformed the spans of modal propositions into BIO-tags, as shown in Table 3.

We shuffled and split the data into 90% training and validation sets, and a 10% test set. The training and validation set was then split into 5 folds, and in each fold, 20% of the sentences were randomly assigned to validation, 80% to training. As opposed to Marasović and Frank (2016), who trained and evaluated only on sentences already known to contain modal triggers, we use the entire dataset, including sentences with no modality.[6]

Corpus Statistics The GME corpus, containing 11K sentences, shows that modality is a pervasive phenomenon (modal triggers were found in 96% of the documents and in 48% of the sentences). We find in the corpus 8318 modal triggers which correspond to 1502 unique types.

Aside from verbs, nouns (e.g., *rights, possibility*) and adjectives (e.g., *fair, important*) are among the most frequently used modal expressions, with verbs making up 37% of the modals in the corpus, adjectives 30%, and nouns 20%. The remaining modals are either adverbials, auxiliaries, or particles. While most verbal triggers are modal verbs (e.g., *could, must, should*; MV henceforth), 38% have other POS tags. 736 triggers appear only once in the entire corpus with a modal meaning.[7]

About 25% of modal triggers are ambiguous in terms of their modal sense (Plausibility vs. Priority), posing an additional classification challenge on top of the varied distribution of trigger POS tags. Modal triggers can also be multi-word expressions, with about 200 such instances in the corpus (e.g., *have to*).

The modal-triggers' sense-labels are rather balanced: 48% of the triggers in the corpus belong to 'Plausibility' and 52% to 'Priority'. For the finer-grained senses, the most common and least common classes make up 33% and 7% of the corpus, respectively.

The Proposed Tasks We experiment with three tasks, with an increasing level of complexity:

[6]The processed data is available at https://github.com/OnlpLab/Modality-Corpus.

[7]Words like *can* and *right* have non-modal meanings in addition to modal meanings.

1. MODAL SENSE CLASSIFICATION. Here we aim to classify the modal sense of a trigger, assuming a modal trigger is already known. Specifically, we examine the contribution of the context to the lemma. We perform sense classification with the following variations: (i) *Vote*: a majority vote, (ii) *Token*: out of context token-based classification where the trigger token is encoded using GloVe (Pennington et al., 2014)), (iii) *Context*: Token-in-context classification, given the whole sentence encoded with RoBERTa (Liu et al., 2019) as input, with a marked trigger position, (iv) *Masked*: given the sentence encoded with RoBERTa but with the trigger masked, (v) *Trigger+Head*: only the trigger word and event head are given, encoded with RoBERTa, and finally, (vi) *Full+Head*: the full sentence is encoded using RoBERTa with both the trigger and the event head marked.

2. MODALITY DETECTION AND CLASSIFICATION. This is a realistic scenario, where we do not assume the trigger is known. We aim to both identify the trigger and label its sense. We model this as a tagging task. Every token in the corpus is assigned a BIOSE tag if it belongs to a modal trigger, which is appended with a suffix indicating its modal sense. We additionally perform variations of this task by including the head of the event as a feature (with either gold or predicted heads). Table 3 shows an example of the BIOSE tagging of modal triggers, with and without the event.

3. MODAL-EVENT DETECTION. Detecting and classifying modal triggers in isolation is insufficient for applications, as it is crucial to detect the event being modified. Here we predict a modal event and aim to relate it to its trigger and modal sense. We model this as sequence labeling, with the different tagging schemes to indicate the event being modified. First, we aim to detect *only* the event. In (i), we predict BIO tags for the propositional spans. In (ii), we predict a HEAD label for the event head. Next, we aim to *jointly* predict the modal triggers and their modified events. To this end, in (iii) we predict BIOSE-{E|T} for the event span, concatenating the related modal trigger. That is, within a single event span marked with BIO, E marks the propositional content and T marks the trigger. We experiment with and without the *modal sense* appended to the trigger. Finally, in (iv) we predict BIOSE-{sense} tags that indicate the modal trigger along with a HEAD tag for the event head.

Text	BIOSE	Event Head	Event Span
Japan	O	O	O
has	O	O	O
taken	O	O	O
a	O	O	O
leading	O	O	O
role	O	O	O
in	O	O	O
the	O	O	O
international	O	O	O
drive	S-GOALS	S-GOALS	B-T
to	O	O	I-E
rebuild	O	H	I-E
Afghanistan	O	O	I-E
...	O	O	O

Table 3: Representing Event-Based Modality Using a BIO Tagging Scheme. On the left, the BIOSE-label tags are used to label the modal triggers. In the middle column BIO tags track the modal triggers, and H indicates the event head. On the right, the BIO tags track the event span, with the T and E labeling the trigger and event span respectively.

	Vote	Token	Mask	Context	Head	Head +Trigger
Coarse	89.1	88.7	78.0	**90.7**	90.5	90.1
Fine	72.0	72.4	58.3	**76.4**	76.2	75.1

Table 4: Modal Sense Classification with Oracle Triggers.

The labels that indicate modal sense are drawn from the proposed hierarchy, and we experiment with multiple levels of granularity: *Modal/Not Modal*: a binary distinction, indicating if the token is a modal trigger or not. *Coarse-grained*: a 3-way distinction, indicating if the token is a modal trigger, and if so, what coarse-grained sense it has (Plausibility vs. Priority). *Fine-Grained*: indicating if the token is a modal trigger, and if so, which one of the senses at the lowest level of the hierarchy it has. We conflated *Desires/Wishes* and *Plans/Goals* into a single type called *Intentions*, since both these senses are under-represented in our corpus. See appendix A for the complete label distribution in our data.

Evaluation Metrics We report for all experiments BIOSE-chunk Precision, Recall, and (Macro) F1, calculated with the official *ConllEval* script (Sang and Buchholz, 2000). When evaluating span tagging for event-based modality we report labeled and unlabeled scores. When we report *unlabeled* F1 for trigger classification, we check whether the token has been correctly identified as modal vs. not-modal, regardless of its sense.

Models Our baseline for modal trigger detection is a simple majority vote baseline where each token in the test set is tagged with its most frequent label in the training set. For detecting modal triggers as well as for event detection, we experiment by fine-tuning a ROBERTA-based classifier (Liu et al., 2019).[8] The encoded sequence is fed through a linear layer with a softmax function predicting the appropriate tag for a given token. For the shorter spans (modal triggers) we predict the tag for every token-in-context. For the longer spans (events spans or events+trigger spans) we perform CRF decoding. The models we used are AllenNLP (Gardner et al., 2018) implementations. Whenever we use the trigger or the event as features to the model, we add special tokens to the input, marking their respective spans in the sentence. The hyperparameters of the models are as follows: we use ROBERTA$_{BASE}$ and fine-tune it for 6 epochs with a batch-size of 8, a learning rate of $1e^{-5}$ and the adam optimizer.[9]

	Baseline		RoBERTa	
	MV	ALL	MV	ALL
Modal/Not	99.04	68.24	99.9	73.2
Coarse-Grained	93.29	63.94	93.3	68.9
Fine-Grained	73.48	55.23	78.5	58.14

Table 5: The Diversity of Modal Triggers: F1 of MV triggers vs. All triggers, Majority Vote Baseline vs. RoBERTa

5 Results

Setting the Stage Before evaluating our models on the proposed tasks, we first assess the empirical challenge of our *event-based modality detection* task relative to the modal sense *sentence classification* (SC) setup of Marasović and Frank (2016). Their work focuses on 6 modal auxiliary verbs (*can, could, may, must, should*, and *shall*) and modal senses from a restricted set of three labels (*deontic, dynamic, epistemic*). Note that their proposed setup is not designed to separate modal sentences from non-modal ones, as the Marasović and Frank (2016) dataset contains *only* modal sentences. Second, it cannot directly indicate that a sentence contains multiple modal triggers with different senses.

[8] We also experimented with a PyTorch-based sequence tagging model (NCRF++ by Yang and Zhang (2018)) with GoogleNews-vectors-negative300 embeddings (https://code.google.com/archive/p/word2vec/), but this setting did not outperform our majority vote baseline (and certainly under-performed the model based on contextualized representations), and we didn't pursue this direction further.

[9] The code for data processing, configuration files and training are available at https://github.com/OnlpLab/Modality.

		Modal/Not Modal			Coarse-Grained			Fine-Grained		
		P	R	F1	P	R	F1	P	R	F1
Unlabeled	Baseline	75.81	62.07	68.24	75.81	62.07	68.24	75.81	62.07	68.24
	RoBERTa	70.05	76.68	73.2	72.07	76.17	74.04	74.01	74.41	74.2
Labeled	Baseline	NA	NA	NA	71.36	57.92	63.94	58.68	45.56	51.29
	RoBERTa	NA	NA	NA	67.03	70.89	68.89	57.98	58.32	58.14

Table 6: Precision, Recall, and F1 for Baseline and RoBERTa. In *labeled* the model tagged each token for modal/not modal and classified the identified modal tokens. In *unlabeled* the labels are given, but not counted beyond the modal/not-modal distinction.

Dataset - Triggers	Sentence Sense Accuracy
Marasovic̀ - MV	79
GME - MV	73
GME - ALL	69

Table 7: Replicating the Setup of Marasović and Frank (2016) on the GME Data. Results drop for GME when using only sentences with modal verbs (MV), and even further when using all of GME's sentences (namely with all modal triggers).

We trained and tested a CNN compatible to theirs[10] on their data as well as our data (GME), using their proposed settings. We mapped our *Priority*, *Agent*, and *Knowledge* to their *deontic*, *dynamic*, and *epistemic*, respectively, and ignored our *State of the World (circumstantial)*. Here, we report the same sentence-based accuracy metrics as they do. Table 7 shows the results on the two datasets, theirs and GME. We see that accuracy on the SC task drops when switching from their data to ours, and that it drops further when moving from a closed set of POS (Modal Verbs) to all targets. All in all, sentence classification is not sufficient to reflect the richness of *event-based* modality annotation, and we conjecture that the SC setup would be too restrictive for real-world applications.

Modal Sense Classification Next we report results for the first task we define, labeling the modal sense of an oracle trigger, as shown in Table 4. The majority vote baseline is high, which is partly due to the trigger lemma overlap between train and dev/test (between 73%-79% depending on the split). Additionally we found only 25% of the trigger lemmas in the corpus to be ambiguous between Plausibility and Priority. Exposing the context, either by means of the full sentence or only the event head, improves results, and the improvement is more substantial for the fine-grained distinctions. Removing the lemma and using *only* context (Masked) harms the results, but it is still impressive and shows that the environment has non-negligible contribution to sense disambiguation. Finally, the sense classification is surprisingly effective also in cases where different modal events in the same sentence are intertwined. An interesting example is the following sentence, with modal triggers in **bold** (sense in brackets): "How **can**(Plausibility), under such circumstances, America **allow**(Priority) itself to express an **opinion**(Plausibility) over the issue of human **rights**(Priority) in other countries." Even when masking the triggers, the fine-tuned language model was able to correctly identify this alternating pattern of Plausibility and Priority.

Modal Triggers Detection Table 5 shows the modal trigger detection results when applied only to the six modal verbs (MVs), as opposed to modal triggers of unrestricted POS tags (ALL). We see that when targeting only MVs, detecting modal elements is almost trivial for both the baseline and RoBERTa. Both models are also quite proficient (F1=93) at separating the different high-level modal senses (Priority vs. Plausibility) of the modal types that we defined. Once we switch to 'All triggers', results substantially drop. Also, when switching to finer-grained categories we observe an expected drop for both the baseline and RoBERTa, with RoBERTa performing significantly better.

Table 6 presents the breakdown of the scores, labeled and unlabeled, for the different levels of granularity by the different models. In all cases RoBERTa shows at least 5 absolute points consistent increase in F1 scores over the baseline, for all levels of granularity. Furthermore, our unlabeled scores demonstrate that predicting the fine-grained categories by RoBERTa actually helps to determine the modal/non-modal decision boundary, with an F1 improvement of about 1 absolute point at all levels. For the labeled accuracy, we observe an expected drop in the F1 scores when taking into account fine-grained labels. Yet, the performance is better than a majority vote baseline and is far better than chance for these nuanced distinctions.

[10] Some dependencies in the Marasović and Frank (2016) code are deprecated, so we use a simple off-the-shelf CNN model of AllenNLP (Gardner et al., 2018).

F1	No-Head	Head Gold	Head Predict	Joint
Modal / Not-Modal	73.2	87.6	69.4	73.3
Coarse-Grained	68.9	79.8	63.2	67.3
Fine-Grained	58.14	66.7	52.1	56.0

Table 8: Modal Trigger Tagging Results, F1 on Detected Spans, with and without Event Head Information.

	F1	No Trigger	Trigger Gold	Trigger Predict	Joint Joint
Span	Modal / Not	51.1	71.13	53.55	50.05
	Coarse-Grained	51.1	70.91	53.56	49.85
	Fine-Grained	51.1	70.38	53.09	48.24
Head	Modal / Not	56.3	72.3	55.8	56.9
	Coarse-Grained	56.3	71.6	56.0	60.7
	Fine-Grained	56.3	70.9	55.2	55.3

Table 9: Event Detection Results, F1 on Detected Spans, with and without Modal Trigger Information.

In the Fine-Grained Labeled RoBERTa setting the breakdown of the F1 performance by label is: *agent*: 72.7, *world*: 54.7, *rules/norms*: 60.4, *knowledge*: 59.3, *intentional*: 46.1. These scores do not correlate with the frequency of each sense in the training data, e.g. *agent* is the least frequent sense, but the model performed best at tagging it. Looking at ambiguous lemmas, i.e., lemmas that can have different modal senses depending on context, one can see that *agent* and *rules/norms* are the least ambiguous senses, which explains their higher performance scores. Breaking down the performance by coarse grained POS tag shows that VERBS are easiest to tag (66.5), followed by ADVERBS (59.7), then ADJECTIVES (55.9) and lastly, NOUNS, which, with a score of 43.8, seem to be the hardest to tag. Interestingly, ADJECTIVES are more ambiguous than NOUNS; we thus do not have a satisfying explanation for why it is harder to classify the modality of NOUN triggers.

Table 8 shows the effect of event heads on modal trigger identification and classification, considering whether to model them separately or jointly in realistic scenarios, where the trigger is not known in advance. Gold event information as a feature for modal trigger tagging is helpful, but when this information is predicted, propagated errors decrease performance. Jointly predicting both triggers and event heads only very slightly decreases performance for the more fine-grained sense categories, making it a viable option for classification.

Event Detection Based on Modal Triggers Table 9 shows that event-span detection is a harder task than merely locating the triggers (cf. Table 6). Interestingly, predicting the span *given* information about the trigger (Trigger as Feature) works better than predicting the span with no such information (No-trigger). This holds both when the triggering event is provided by an Oracle ('Gold'), or whether it is predicted by RoBERTa ('Predict'). Improving modal trigger prediction is thus expected to further contribute to the accurate identification of events, and to event-span boundary detection. In general, head prediction shows better results than span prediction, partly due to the F1 score on spans being a restrictive metric in cases of partial overlap.

Error Analysis To qualitatively assess the usability of RoBERTa's output, two trained human experts manually inspected the errors in 112 modal triggers in the dev set. Out of 36 false negatives (FN), 6 (16% of the FN) are in fact correct (incorrectly tagged by the annotators as modal), and out of 27 false positives, 21 (78% of the FP) are in fact correct (modals missed by the annotators). This leads to the conclusion that the gold annotation by the experts, while being precise, has incomplete coverage and lower recall. It implies that RoBERTa's precision is in actuality *higher*, with a larger share of its predictions being correct.

6 Conclusion

We propose an *event-based modality detection* task which is based on solid theoretical foundations yet is adapted to fit the needs of NLP practitioners. The task has three facets: modal triggers can be of any syntactic type, sense labels are drawn from a unified taxonomy we propose, and modal triggers are associated with their modified events. We propose this task and standard as a potential extension for standard semantic representations (AMR, SDG, UCCA, etc.) towards easy incorporation of modal events as features in downstream tasks.

Acknowledgements

We thank Yoav Goldberg, Ido Dagan, Noah Smith, Graham Katz, Elena Herburger, and members of the BIU-NLP Seminar for thoughtful feedback and fruitful discussion. We also thank 3 anonymous reviewers for their insightful remarks. This research is supported by an ERC-StG grant of the European Research Council (no. 677352), the Israel Science Foundation (grant no. 1739/26 and grant no. 2299/19), and the National Science Foundation (BCS-1053038), for which we are grateful.

References

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics*, 38(2):411–438.

Jon Barwise and John Perry. 1981. Situations and attitudes. *The Journal of Philosophy*, 78(11):668–691.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Elisa Ghia, Lennart Kloppenburg, Malvina Nissim, and Paola Pietrandrea. 2016. A construction-centered approach to the annotation of modality. In *Twelfth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 67–74, Portorož, Slovenia.

Valentine Hacquard. 2010. On the event relativity of modal auxiliaries. *Natural Language Semantics*, 18:79–114.

Iris Hendrickx, Amália Mendes, and Silvia Mencarelli. 2012. Modality in text: a proposal for corpus annotation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Angelika Kratzer. 1981. The notional category of modality. In Hans-Jürgen Eikmeyer and Hannes Rieser, editors, *Words, Worlds, and Contexts*, pages 38–74. Walter de Gruyter, Berlin. Reprinted in *Formal Semantics: The Essential Readings*, ed. Paul Portner and Barbara H. Partee (2002), 289–323. Oxford: Blackwell.

Angelika Kratzer. 2010. Situations in natural language semantics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2010 edition. First published February 2007.

Saul A. Kripke. 1959. A completeness theorem in modal logic. *The Journal of Symbolic Logic*, 24(1):1–14.

Julia Lavid, Marta Carretrero, and Juan Rafael Zamorano-Mansilla. 2016. A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT. In *Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning*. CSLI Publications.

David Lewis. 1973. *Counterfactuals*. Harvard University Press, Cambridge, Mass.

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*.

Ashutosh Malhotra, Erfan Younesi, Harsha Gurulingappa, and Martin Hofmann-Apitius. 2013. 'HypothesisFinder:' a strategy for the detection of speculative statements in scientific text. *PLOS Computational Biology*, 9(7):e1003117.

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.

Ana Marasović, Mengfei Zou, Alexis Palmer, and Anette Frank. 2016. Modal sense classification at large. paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *LiLT (Linguistic Issues in Language Technology)*, 14.

Lisa Matthewson and Hubert Truckenbrodt. 2018. Modal flavour/modal force interactions in German: *soll*, *sollte*, *muss* and *müsste*. *Linguistische Berichte*, 255:259–312.

Amália Mendes, Iris Hendrickx, Liciana Ávila, Paulo Quaresma, Teresa Gonçalves, and João Sequeira. 2016. Modality annotation for Portuguese: from manual annotation to automatic labeling. In *Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning*. CSLI Publications.

Makoto Miwa, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, 13(108).

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Malvina Nissim, Paola Pietrandrea, Andrea Sansò, and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14, Potsdam, Germany. Association for Computational Linguistics.

Frank R. Palmer. 1986. *Mood and Modality*. Cambridge University Press, Cambridge.

Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Paul Portner. 2009. *Modality*. Oxford University Press.

Mario Prieto, Helena Deus, Anita de Waard, Erik Schultes, Beatriz García-Jiménez, and Mark D. Wilkinson. 2020. Data-driven classification of the certainty of scholarly assertions. *PeerJ 8*, 8:e8871.

Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves. 2014. Automatic tagging of modality: identifying triggers and modal value. In *The 10th Joint ACL SIGSEM - ISO Workshop on Interoperable Semantic Annotation*, pages 95–102.

Victoria L. Rubin. 2010. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46(5):533–540.

Aynat Rubinstein. 2012. *Roots of Modality*. Ph.D. thesis, University of Massachusetts Amherst.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Dan 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.

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-2012)*, pages 1538–1545.

Erik Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9:S9.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626*.

Chrysoula Zerva, Riza Batista-Navarro, Philip Day, and Sophia Ananiadou. 2017. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*, 33(23):3784–3792.

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.

A Data

A.1 GME in numbers

The GME dataset (Rubinstein et al., 2013) annotates the MPQA corpus (Wiebe et al., 2005) with information about modality. The corpus consists of 534 documents which in turn contain 11,048 sentences. 5288 sentences have modal triggers, and of them, in 1141 the modal trigger is an auxiliary verb. There are 7979 instances of modal triggers (tokens), which belong to 1141 unique words (types). 1229 of the modal triggers are modal verbs. The breakdown of the modal triggers into the different modal senses is given in Table 10.

<table>
<tr><th>Type</th><th>Quantity</th><th colspan="2">2-way ambiguity</th><th>3-way ambiguity</th></tr>
<tr><td>Rules & Norms</td><td>2316</td><td colspan="2" rowspan="2"></td><td rowspan="3">537</td></tr>
<tr><td>Desires & Wishes</td><td>142</td></tr>
<tr><td>Plans & Goals</td><td>1077</td><td colspan="2">210</td></tr>
<tr><td>Knowledge</td><td>1527</td><td rowspan="2">557</td><td rowspan="2"></td><td rowspan="3"></td></tr>
<tr><td>World</td><td>1303</td><td rowspan="2">202</td></tr>
<tr><td>Agent</td><td>447</td><td></td></tr>
</table>

Table 10: Label Counts in the GME Data

A.2 Data Pre-processing

We parsed the data using spaCy, and obtained the lemma, POS, and dependency information for all tokens in our corpus. We split the data into 5 folds, where each fold had a different split of training and validation set, but the test set is the same for all folds. Train and validation sets are of 9894 sentences (validation 1975 and training 7919), while the test set has 1096 sentences. The train and validation sets have 7160 modal triggers, while the test set has 819.

B Additional Materials

Please refer to the following github repositories for code and data:

Code Code and models and evaluation scripts used in our experiments

https://github.com/OnlpLab/Modality

Data A processed version of the GME corpus, including all annotation layers and meta-information.

https://github.com/OnlpLab/Modality-Corpus

C Experimental Setting

We had 4 GeForce GTX 1080 Ti available for training and hyper-parameter search. Our models are based on RoBERTa-base, which has 82M parameters and it takes about 45 minutes to train a single tagging model.

Tables 11 and 12 show the results of the baseline and RoBERTa respectively. On the right hand side of the tables, the scores are split by modal senses. Here too, we observe that RoBERTa obtains substantial improvements on per-label scores over the baseline.

	P/R/F1		Labels F1				
	Labeled	**Unlabeled**					
Modal vs. Not-Modal	75.81 62.07 68.24	75.81 62.07 68.24					
Priority vs. Plausibility	71.36 57.92 63.94	75.81 62.07 68.24	**Priority** 55.46		**Plausibility** 72.51		
Fine-Grained	58.68 45.56 51.29	75.81 62.07 68.24	**Rules** 50.94	**Intentions*** 39.11	**Knowledge** 50.95	**World** 52.58	**Agent** 67.39

Table 11: Classifying Modal Events: Baseline Results (ambiguities not shown). We unified wishes and goals into *intentions* for reasons of data sparsity.

	P/R/F1		Labeled F1				
	Labeled	**Unlabeled**					
Modal vs. Not-Modal	NA NA NA	70.05 76.68 73.2					
Priority vs. Plausibility	67.03 70.89 68.89	72.07 76.17 74.04	**Priority** 62.98		**Plausibility** 75.52		
Fine-Grained	57.98 58.32 58.14	74.01 74.41 74.2	**Rules** 60.42	**Intentions*** 46.1	**Knowledge** 59.27	**World** 54.64	**Agent** 72.72

Table 12: Classifying Modal Events: RoBERTa Results (ambiguities not shown). We unified wishes and goals into *intentions* for reasons of data sparsity.