**Machine classification of modal meanings: An empirical study and some consequences**

Aynat Rubinstein, Valentina Pyatkin, Shoval Sadde, Reut Tsarfaty, and Paul Portner

**Introduction.** We discuss the linguistic relevance of a computational study on modality (Authors 2021) which sets out to detect modals in texts without assuming they come from a closed class of lexical items, to classify their meaning in terms of modality type (or "modal flavor"), and to identify the eventuality they modalize. Building on a linguistically motivated annotation of modal meaning in news text (Rubinstein et al. 2013), we show that, while the detection of modal auxiliaries is trivial, detection and classification of a more open, semantically defined class is difficult. We also show that jointly performing the tasks of classifying the modality type and identifying the modalized eventualities produces superior results to doing either separately. We also discuss the distribution of modality types across our typology and the learnability of the subtypes. Our results suggest that the standard typology due to Kratzer (1981, 1991) should be restructured by grouping together epistemics and certain circumstantials as a "facts and knowledge" class.

**The study.** We use the **taxonomy** in Table 1 for classifying senses. It is based on a *coarse-grained* split between Priority modality (Portner 2009) and Plausibility modality and six *finer-grained* sub-types adapted by Rubinstein et al. (2013) in their annotated corpus. The taxonomy unifies and harmonizes the different modal senses offered by previous computational studies (Ruppenhofer & Rehbein 2012; Marasović & Frank 2016; Baker et al. 2012; Mendes et al. 2016). Examples with modals of a variety of parts of speech (POSs) are shown.

| **Priority** | |
|---|---|
| Norms and Rules | *the ballot which* **must** *be held by the end of March* |
| Desires and Wishes | *extend our full* **support** *to the George W. Bush administration* |
| 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: Proposed Taxonomy with Examples from GME.

| | Baseline | | RoBERTa | |
|---|---|---|---|---|
| | Aux V | All | Aux V | 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 2: F1 on Auxiliary Verbs (*can, could, may, must, should, shall*) vs. All triggers, Majority Vote Baseline vs. RoBERTa.

| Rules | Intentions | Knowledge | World | Agent |
|---|---|---|---|---|
| 60.42 (50.94) | 46.1 (39.11) | 59.27 (50.95) | 54.64 (52.58) | 72.72 (67.39) |

Table 3: F1 results RoBERTa (vs. Baseline) for fine-grained senses. *Wishes/Goals* unified due to data sparsity.

For training and testing our models, we 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). We processed the corpus by extracting modal triggers and their prejacents into a CoNLL-formatted file. We added lemmas, POS tags, and dependencies using spaCy (Honnibal et al. 2020). As opposed to previous work, which trained and evaluated only on sentences known to contain modals, we use the entire dataset. We also accommodate sentences that contain multiple modals with different senses. We experiment with three **tasks**: (i) classifiying the sense of words specified by fiat as modal, (ii) detecting modal words and classifying their sense, and (iii) identifying also the modalized event. The results for the second task are shown in Table 2 (all results will be discussed in the talk), comparing a majority vote baseline to a fine-tuned RoBERTa-based classifier (Liu et al. 2019). The results show that detecting modality at the fine-grained level beyond the small set of modal auxiliary verbs is not trivial, with RoBERTa performing significantly better and far better than chance. The breakdown of RoBERTa's F1 scores is given in

Table 3. The largest label-wise gain in absolute points in comparison to the baseline is for *Rules* ($\sim$10) and *Knowledge* ($\sim$8), and the smallest is for *World* ($\sim$2).

**Consequencs for semantics.** Rubinstein et al.'s (2013) annotation effort had already noted that the distinction between *Knowledge* (epistemic) and *World* (circumstantial) modality is often very unclear. An example from the corpus is given in (1):

(1) That will facilitate their **possible** convergence later with the international system.

On the *World* reading, (1) is based on some event taken as evidence, and the accessible worlds are the ones where that event is the same in relevant respects. The assertion of (1), on this reading, is that the circumstances of the US following certain standards of the Kyoto Protocol make it possible that its policies will converge at a later date. On the *Knowledge* reading, the same kind of evidence is relevant, but in addition, the mental state of the author plays a crucial role. In other words, the evidence isn't enough, and we need to include private knowledge of the author (i.e. that US officials intend to work towards covergence once the political situation changes) to understand why (1) is a justified assertion. Thus, the two readings differ in whether the speaker's mental state is involved *in addition to* an evidential event, not *instead of* it.

We see evidence for this view in the experiment in cases where the GME Corpus and the model differ in that one assigns *Knowledge* and the other *World*. There are 36 such cases, of which we judge both annotation and model to be correct in 17 of them (i.e., true examples of ambiguity). We judge only the model to be correct in 7 examples, only the annotators to be correct in 8, and 4 where neither was correct. Overall, the model errs on the side of annotation as *World* over *Knowledge*; this may be partially due to the fact that the model did not use extrasentential context. We also note that most of the confusion occurs with particular high frequency lexemes, in particular *would* (n=13), *could* (n=5), and *possible* (n=3), with idiosyncratic confusion around *clear(ly)* (n=3).

When a modal is embedded under an epistemic (or doxastic) operator, it is typically forced to take into account some knowledge (Hacquard 2006; Yalcin 2007). In (2), the modal background from *would* involves both relevant circumstances and the judgment of Mr. Carmona or other individuals at the company he represents. Annotators were correct in this case, whereas the model did not detect the *Knowledge* signal given by the embedding verb.

(2) Mr. Carmona said that operations **would** return to normal at the oil company.

The idea of collapsing epistemic modality with some cases of circumstantial modality is not new (Hacquard 2010, Kratzer 2012, p. 24), but our computational study sheds new light on the issues. We find that the model makes the smallest gains over baseline for the class of non-ability circumstantial modals (even setting aside cases which humans annotated as ambiguous between epistemic and circumstantial). We believe that collapsing circumstantial modality (perhaps not including ability modals) with epistemics would lead to more reliable classification, and we suggest that this change would reflect the linguistic reality that "epistemic modal" is not the class we thought it was. Examples like (1) and (2), and the comparison of the annotation and computational model, suggest that epistemic modality should be understood of as a sub-type of circumstantial modality.

**Summary.** We have shown that state-of-the-art NLP models can extract a significant amount of detailed information on the meanings of modal elements from annotated text. Perhaps more interestingly, they reveal patterns that fail to align with our standard theoretical assumptions, but which ultimately may be vindicated by a reassessment of the relevant categories. We have made this point with regard to the categories of epistemic/circumstantial modality, and in the presentation we will expand upon it regarding the split between modals and attitude verbs and the distinction between bouletic and teleological modality.