Probing BERT's ability to encode sentence modality and modal verb sense across varieties of English

Jonas Wagner and **Sina Zarrieß**
Bielefeld University
Faculty for Linguistics and Literary Studies
{jonas.wagner,sina.zarriess}@uni-bielefeld.de

Abstract

In this research, we investigate whether BERT can differentiate between modal verb senses and sentence modalities and whether it performs equally well on different varieties of English. We fit probing classifiers under two conditions: contextualised embeddings of modal verbs and sentence embeddings. We also investigate BERT's ability to predict masked modal verbs. Additionally, we classify separately for each modal verb to investigate whether BERT encodes different representations of senses for each individual verb. Lastly, we employ classifiers on data from different varieties of English to determine whether non-American English data is an additional hurdle. Results indicate that BERT has different representations for distinct senses for each modal verb, but does not represent modal sense independently from modal verbs. We also show that performance in different varieties of English is not equal, pointing to a necessary shift in the way we train large language models towards more linguistic diversity. We make our annotated dataset of modal sense in different varieties of English available at https://github.com/wagner-jonas/VEM.

1 Introduction

Work on contextualised embeddings learned by large bidirectional language models such as BERT (Devlin et al., 2019) indicates that they may capture *senses* of lexical items (Loureiro et al., 2021). This has the potential to greatly accelerate variationist research, for example by finding community-specific senses of words (Lucy and Bamman, 2021) or tracing contact-induced semantic shifts (Miletic et al., 2021). Modal sense[1] variation has been an area of interest for variationist researchers (see, e.g. Collins et al., 2014, Hansen, 2018, or Loureiro-Porto, 2019), but has, so far, received comparably little attention in NLP. In this paper, we investigate to what extent modal sense is encoded in BERT embeddings across varieties of English, and if so, at which layer(s) and in what form.

Modality is generally analysed on sentence level (Portner, 2009, 2–6) and is primarily expressed in English by the use of modal verbs (Portner, 2009, 4), with each verb potentially evoking different senses. Consider *must* in the following two sentences: "You *must* complete all tasks for course credit" and "You *must* be tired after the long journey". In the first sentence, *must* has deontic sense, i.e. it is used to express orders or recommendations, which can also be expressed by e.g. *should*. In the second sentence, *must* has epistemic sense, i.e. a qualification of certainty. This can be expressed by many modal verbs, such as *may*, *can*, *could*, or *might*. In addition to these two, there are also concessive (granting or denying permission, e.g. *may* and *can*) and dynamic (expressing ability, e.g. *can*) sense.[2] The modal verb therefore affects the interpretation of the sentence as a whole: swapping *must* and *may* in "You *must*/*may* leave now" clearly affects more than only the meaning of the modal verbs themselves.

How often each modal verb expresses which sense is prone to variation and at times starkly differs between varieties of English. This has been researched in-depth. For example, Collins et al. (2014) investigate domain-specific variation of modal verb sense distribution in Philippine English and compare it to American and British English. Hansen (2018) provides what is probably the most comprehensive treatment of modal verb sense in varieties of English, finding that e.g. British and Indian English have high incidences of epistemic

[1] Linguists often differentiate between modality, which is analysed on sentence level, and modal verb sense for individual modal verbs. As we investigate both, we use the term "modal sense" where we refer to both.

[2] Other senses exist, but will not be discussed in this work; see also Ruppenhofer and Rehbein (2012).

Proceedings of the 15th International Conference on Computational Semantics pages 28–38
June 21–23, 2023. ©2023 Association for Computational Linguistics

must, while Hong Kong and Singapore English have higher incidences of deontic *must*.

Most of these studies remain small in scale. They investigate only a small number of varieties, a small number of modal verbs, or small corpora. This is unsurprising, as modal verb sense annotation is largely done manually. Large-scale computational investigations in this area would be a valuable contribution, but these different distributions of modal senses may pose a challenge for pre-trained language models, which are often not trained on diverse data and may struggle with other varieties' different modal verbs being usage.

These simple facts about modal sense raise questions regarding BERT's potential to capture modal sense which have not been addressed in recent work on probing BERT's abilities to encode lexical semantics, cf. among others, Aina et al. (2019); Pilehvar and Camacho-Collados (2019); Vulić et al. (2020); Garí Soler and Apidianaki (2021). Ideally, BERT would capture modal sense both at sentence (in the [CLS] token) and word level (in modal verbs' embeddings). The latter needs more differentiation: representation could be independent from the individual verbs (e.g. epistemic *must* and epistemic *may* share encoded epistemic sense) or different modal senses are only represented for each individual verb (epistemic and deontic *must* encode different senses, but these would not be shared by epistemic and deontic *should*). Further, it should show systematic encodings of lexical and sentential modal sense across different layers, in light of other work showing linguistic systematicity in processing different aspects of linguistic knowledge across layers (Aina et al., 2019; Pilehvar and Camacho-Collados, 2019; Vulić et al., 2020; Garí Soler and Apidianaki, 2021). And, last but not least, it should encode modal sense in a way that is robust to distributional differences of modal senses and verbs across varieties of English.

Beyond accelerating variationist research, correct classification of modal sense also has relevance for NLP tasks. Modal sense classification has been used in connection with sentiment analysis (Liu et al., 2014), hedging and detection of hypotheses and speculation (Morante and Daelemans, 2009; Vincze et al., 2008; Malhotra et al., 2013),[3] and factuality detection (Saurí and Pustejovsky, 2012), among others. These are key tasks that, ideally, should function in different varieties of English – not just majority varieties.

We conduct a series of experiments to investigate if, and how, BERT encodes modal sense. We train probing classifiers on annotated datasets (see Section 3 for our data) and classify modal sense. We do this for modal verbs' embeddings as well as sentence embeddings (experiment 1, Section 4). We also train separate classifiers for each modal verb (experiment 2, Section 5); we extend this methodology to data from several different varieties of English (experiment 4, Section 7). We also test whether BERT can predict masked modal verbs, even if it cannot classify modal sense (experiment 3, Section 6).

2 Background

2.1 Semantic knowledge encoded in BERT

While BERT (Devlin et al., 2019) has been used to investigate many facets of the semantic meanings of words (e.g. Wiedemann et al., 2019; Vulić et al., 2020; Zhang et al., 2020; Bhardwaj et al., 2021; Garí Soler and Apidianaki, 2021; Lucy and Bamman, 2021; Miletic et al., 2021; Apidianaki, 2023 among others), some aspects of meaning cannot be captured by BERT embeddings. Ettinger (2020) found that BERT does not appear to process negation at all: both *a robin is a* and *a robin is not a* are predicted to most likely end with *bird* or *robin*. Therefore, more research into the kinds of meaning contained in BERT embeddings is necessary.

Simultaneously, previous research on classifying modal senses with static embeddings indicates that contextualised word embeddings may be useful to improve modal sense classification. Li et al. (2019) use static embeddings for modal sense classification, but adjust each embedding's weight based on distance from the modal verb and POS-tag, which improves results. Marasović et al. (2016) present one of the most comprehensive studies on modal sense classification to date, and point to the importance of lexical features of embedded verbs and the subject in the sentence as giving cues to the modal verbs' meanings. Their experiments also analyze the effect of variation in the distribution of modal senses in different datasets and genres. In more recent work, Pyatkin et al. (2021) go beyond Marasović et al. (2016)'s setup that is restricted to modal verbs and propose a more complex modality detection task involving a broader set of modality triggers and the detection of events associated with

[3] While Vincze et al. (2008) do not explicitly mention modal sense, they do point to the importance of modal auxiliaries in uncertainty detection.

them. As our work aims for a controlled analysis of the representation of modal verb sense across varieties of English, we follow Marasović et al. (2016) and leave the exploration of further modality triggers to future work.

2.2 Variationist NLP research

There has been some NLP research into variation within English. For example, Lucy and Bamman (2021) successfully use contextualised BERT embeddings to find community-specific meanings of words like *python*, which may refer to a programming language or a fictional spaceship. Similarly, Miletic et al. (2021) use contextualised BERT embeddings to find contact-induced semantic shift in English in Quebec. These studies demonstrate that BERT can be used to study variation within English, even between different varieties. But we see two issues with them. Firstly, much World Englishes research focuses types of sense variation other than homonymy, such as the different distributions of modal senses. Secondly, by using BERT to investigate variation, the authors inherently assume that BERT can capture such variation. While their results support this assumption, this does not mean that BERT is an adequate tool to measure all kinds of differences between varieties of English.

The exact nature of BERT's training data is opaque, but Devlin et al. (2019) mention that they use two sources of data for pre-training. These are the 800 million word BooksCorpus (Zhu et al., 2015), consisting of 11,038 unpublished books, and a large 2.5 billion token corpus of Wikipedia entries. While the exact makeup of who wrote those texts is unknown, some reasonable guesses can be made regarding the larger Wikipedia sample. Wikipedia publishes data on the demographic makeup of its contributors,[4] which indicates that a plurality of edits are made from the United States, followed by the United Kingdom and Canada. This is not a perfect method – just because a user is accessing Wikipedia from the United States does not mean that they also speak American English – but it still provides a basis for the assumption that most of BERT's training data comes from the so-called "Inner Circle" (Kachru, 1985), i.e. those countries where English is spoken as a first language by most of the population. This suggests that BERT's training data lacks diversity with regards to varieties of English, which may adversely affect its ability to process English produced by speakers of those under- or unrepresented varieties.

3 Data

We use two existing datasets to test whether BERT can differentiate modal verb senses and construct a new one, taking data from the International Corpus of English (ICE). The first is a portion of the Multi-Perspective Question Answering (MPQA) dataset that has been annotated for modal sense (Ruppenhofer and Rehbein, 2012). This consists of 1,201 sentences taken from news articles dated June 2001 to May 2002 (Wiebe et al., 2005; Ruppenhofer and Rehbein, 2012). The second is the heuristically tagged EPOS-E dataset (Marasović et al., 2019), based on the EUROPARL and OpenSubtitles-English datasets, consisting of data from the European Parliament and film subtitles, comprising 2,453 sentences. Modal sense is annotated for each sentence in both datasets. For comparability, we do not report results for *ought*, as it is not evaluated in previous publications either (Marasović et al., 2016). We also remove *might* and *shall* from our results due to their low frequency.

The main difference between the two datasets (besides the annotation methodology) is size, with EPOS-E being almost twice the size of MPQA. They also draw their data from different sources, which is important given the genre effects found by Marasović et al. (2016). MPQA includes more senses than EPOS-E, which we discard in our analysis to maintain comparability. The balancing for the different senses for each modal verb also varies between them: the most common sense for *must* makes up 92% of instances in MPQA, but only 60% in EPOS-E; for *may*, this is 74% and 87%; for *can* 67% and 84%; for *could* 65% and 43%; and for *should* 92% and 94%. This, naturally, may impact classification results. As in previous research, we only investigate the modal verbs that are annotated in the dataset in cases where there is more than one modal verb per sentence.

For the last experiment, in which we test a classifier trained on EPOS-E on data from different varieties of English, we use the written components of eight sub-corpora from the International Corpus of English (ICE; `https://www.ice-corpora.uzh.ch/en.html`), a comparative corpus of varieties of English. For each variety, the same kinds of documents (like

[4] `https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia` and `https://stats.wikimedia.org/wikimedia/squids/SquidReportPageEditsPerLanguageBreakdown.htm`

student writings or fiction) are used to compile subcorpora of about 400,000 written tokens for each variety (see http://ice-corpora.net/ice/index.html for more information). We investigate Philippine (PH), Canadian (CA), Irish (IR), Hong Kong (HK), Sri Lankan (SL), Jamaican (JA), Nigerian (NI), and Indian (IN) English. For each modal verb in each variety, we randomly extract 20 sample sentences that contain the modal verb for a total of 800 sample sentences. Three annotators independently annotate these. We discard all instances where no two annotators agree on one sense, where the sense is unclear (e.g. due to missing context), and false positives (e.g. *must* as a noun instead of a modal verb). This leaves 782 sentences for analysis. Agreement between the first and second annotator is highest (83.75%), followed by agreement between the second and third annotator (79.88%), and between the first and third annotator (78.00%). We use the majority labels as gold labels. We call this corpus VEM – the **V**arieties of **E**nglish **m**odal sense corpus – and make it available at https://github.com/wagner-jonas/VEM.

4 Experiment 1

4.1 Methods

In the first experiment, we investigate whether modal verb sense classification is successful using the modal verbs' contextualised embeddings and sentence embeddings in the form of [CLS] tokens. We use a logistic regression classifier (from scikit-learn, version 1.0.2; Pedregosa et al. (2011)) with elasticnet penalty and the L1-ratio set to 0.5. We only train one classifier each for the modal verbs' embeddings and [CLS] token, but report the results split by modal sense and modal verb.

We replicate the setup from Marasović et al. (2016): first, we randomly split MPQA into training (80%) and test (20%) sets. We then train a logistic regression classifier on the training set and predict modal senses in the test set. Then, we add the data from EPOS-E to the MPQA training set – we borrow the name CL^{-b}_{ME} for this from Marasović et al. (2016) – and predict modal senses in the same test set.

We report accuracy for each layer and sense. There, the baseline is the sum of frequencies of modal verbs for which that sense is the most frequent one. That is, if *must* and *shall* are both most frequently deontic, we add up their frequencies to determine the baseline. Split by modal verb, we do not report for each layer, as we only use the 12th layer for classification using modal verb embeddings and the 7th layer for classification using the [CLS] token, since they showed the strongest overall performance (see also Figures 1 and 2). Here, the baseline is the frequency of the verb's most frequent sense.

4.2 Results

Classifying modal verb embeddings, we reach overall accuracies between 0.70 (*can* in MPQA) and 1.0 (*must* in MPQA). We beat our baseline (the most common sense for each modal verb) for *could* and *must* in both datasets, and additionally for *can* in CL^{-b}_{ME} and *may* in MPQA. We only dip below our baseline for *should*. Marasović et al. (2016) beat their respective baseline for *should* and *must* only. Taking the mean accuracy for all senses any individual verb can express, accuracies vary between 0.25 (*may* in CL^{-b}_{ME}) and 1.0 (*must* in MPQA). In this case, we beat our baseline for *could* and *must* in MPQA.

Classifying with the [CLS] token instead, we reach overall accuracies between 0.02 (*should* in CL^{-b}_{ME}) and 0.73 (*may* in CL^{-b}_{ME} and *could* in MPQA). Here, we only beat our baseline once - for *could* in MPQA. In general, precision and recall are lower than accuracy and results are stronger for MPQA than for CL^{-b}_{ME} - drastically so when classifying using the [CLS] token. For all results, see Table 1.

Separating the results by sense, MPQA performs better than CL^{-b}_{ME} (see Figure 1). Deontic sense is the only sense which (semi-)consistently performs above baseline; other senses hardly, if ever, exceed their baseline. Classifying the [CLS] token (Figure 2), no sense consistently performs above baseline. Accuracy in CL^{-b}_{ME} fluctuates between layers, with one sense usually reaching perfect accuracy and others at zero accuracy.

4.3 Interpretation

Results of the first experiment suggest that there is no single layer of the BERT model that captures modal sense (see Figures 1 – 2). Deontic sense appears easiest to classify (as it is the only sense with accuracies above baseline). Overall, modal sense classification is not successful. It also appears that no modality information is encoded at sentence level, at least in the [CLS] token, given the wild fluctuations between layers (see Figure

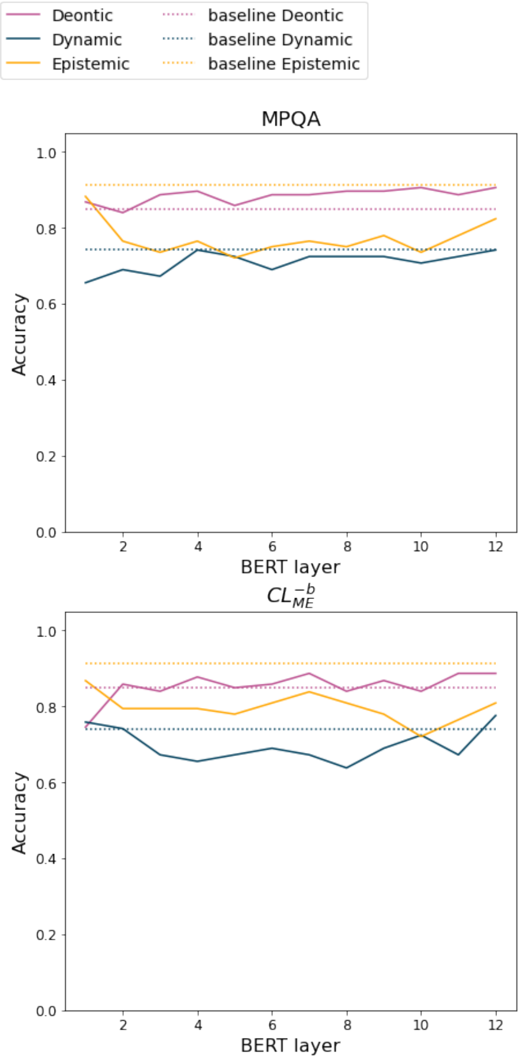

Figure 1: Accuracy of classification of modal verb embeddings per layer, split by dataset and sense.

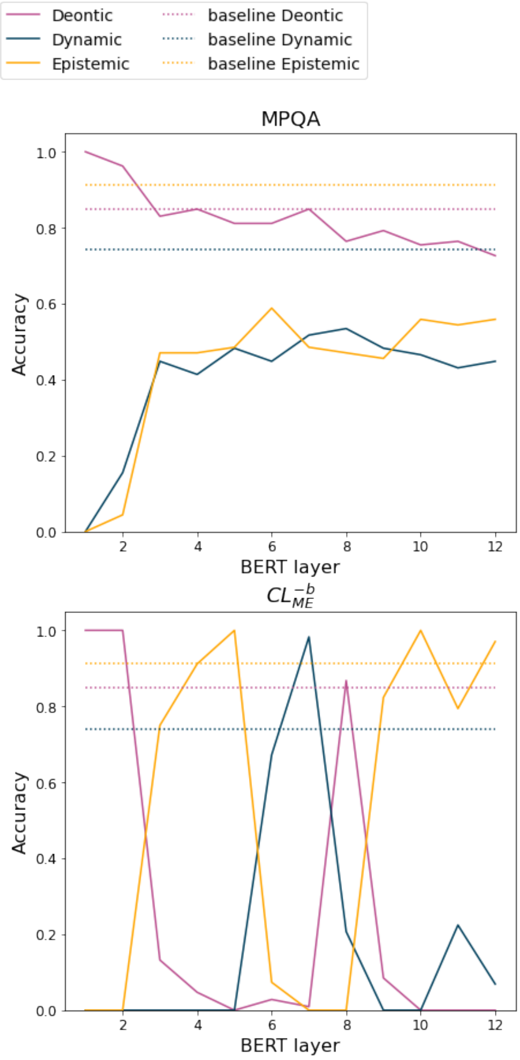

Figure 2: Accuracy of classification of [CLS] token embeddings per layer, split by dataset and sense.

2) – though this might be caused by our choice of classifier.

Viewing individual modal verbs paints a more interesting picture. Some modal verbs appear to be easier to classify, like *could* and *must*. This cannot (just) be due to lower baselines (i.e. a more balanced nature), as *could* has a baseline of 0.67 while *must*'s baseline lies at 0.91. Classification is a lot less successful using the [CLS] token rather than modal verbs' embeddings, which serves to re-affirm the notion that no modality information is encoded on sentence level.

But human speakers (and annotators) do not process modal sense in isolation - they take whichever modal verb is present into account. Thus, it may be that modal sense is not encoded as its own category, but that differences between senses for each individual modal verb (e.g. epistemic and deontic *must*) are. In the next experiment, we therefore train classifiers for each individual modal verb.

5 Experiment 2

5.1 Methods

In this experiment, we train logistic regression classifiers for each modal verb separately, using embeddings from the 12th BERT layer. We use the same parameters as in the first experiment. Note that this does not use the same train and test data as before; as we train separate classifiers for each modal verb, we split data for each modal verb into train and test sets separately. This means that we randomly split data from EPOS-E into 80% training and 20% test data and do the same for MPQA. Note also

Modal verb	*could*		*should*		*can*		*must*		*may*	
Instances	45		57		61		33		33	
Baseline	0.67		0.96		0.70		0.91		0.79	
Training data	MPQA	CL^{-b}_{ME}	MPQA	CL^{-b}_{ME}	MPQA	CL^{-b}_{ME}	MPQA	CL^{-b}_{ME}	MPQA	CL^{-b}_{ME}
Modal verb embedding										
Mean precision per sense	0.28	0.22	0.44	0.1	0.23	0.23	1.0	0.42	0.22	0.25
Mean recall per sense	0.42	0.33	0.44	0.11	0.33	0.33	1.0	0.5	0.25	0.25
Mean accuracy per sense	**0.68**	0.67	0.65	0.31	0.46	0.46	**1.0**	0.5	0.44	0.25
Overall accuracy	**0.71**	**0.69**	0.93	0.89	0.70	**0.72**	**1.0**	**0.94**	**0.88**	0.79
[CLS] embedding										
Mean precision per sense	0.09	0.55	0.09	0.33	0.09	0.34	0.16	0.33	0.1	0.12
Mean recall per sense	0.33	0.75	0.11	0.33	0.22	0.5	0.28	0.33	0.12	0.12
Mean accuracy per sense	0.28	0.6	0.28	0.33	0.27	0.34	0.41	0.33	0.2	0.23
Overall accuracy	0.33	**0.73**	0.82	0.02	0.46	0.07	0.70	0.06	0.64	0.73
Marasović et al (2016)										
Overall accuracy	**0.72**	**0.68**	**0.93**	**0.92**	0.66	0.63	0.94	0.87	0.93	0.90
Baseline	0.65		0.91		0.70		0.94		0.94	

Table 1: Results of modal classification of modal verb/[CLS] token embeddings per modal verb. Mean precision, recall, and accuracy per sense. **Bolded** accuracies are above respective baseline(s) (most frequent sense for each verb). Results from (Marasović et al., 2016) use their semantic features (F_{Sem}), which generally performed best.

that this is a novel methodology and not directly comparable to previous research. And since we train and test on data from MPQA and EPOS-E, respectively, results may skew somewhat positive as we avoid some of the genre effects that Marasović et al. (2016) observe.

5.2 Results

For MPQA, classification of sense for each modal verb shows accuracy between 0.64 (*could*) and 0.96 (*should*). We reach the lowest precision and recall for *may* (precision = 0.29; recall = 0.35); the highest for *must* (precision = 0.83; recall = 0.94). See Table 2 for more results.

For EPOS-E, nearly all metrics are higher than for MPQA. We reach the highest accuracy for *may* at 0.98, the lowest for *could* at 0.84. The lowest precision and recall are reached for *can* (precision = 0.33; recall = 0.31). We reach the highest precision and recall for *may* (precision = 0.95; recall = 0.97).

Accuracy beats the baseline (the frequency of each verb's most common sense) for *could* and *must* in both datasets, additionally for *should* and *can* in MPQA and *may* in EPOS-E. Mean accuracies for each modal verb's potential senses exceed the baseline for *could*, *must*, and *may* in EPOS-E.

5.3 Interpretation

The much improved classification results obtained in this experiment as opposed to the first, where we used a classifier trained on all modal verbs rather than a different one for each modal verb, point to BERT encoding modal verb sense separately for each modal verb. Classification accuracy in both datasets meets or beats the baseline of its most common sense for all verbs. For modal verbs that are dominated by one sense (like *should*), we only rarely exceed the baseline, which is expected, but we do not dip below it, either. The mean accuracy across a modal's possible senses beat the baseline for *could*, *must*, and *may*. These all share a comparatively low baseline, meaning their senses are more balanced than for other modal verbs (though note that *can*, *could*, and *may* in MPQA share this lower baseline but classification is less successful, indicating that it is not the only factor).

In EPOS-E, *must* and *may* see particular success, both reaching precision, recall, and accuracy exceeding 0.93 with baselines of 0.63 and 0.86, respectively. Clearly, BERT does not simply assign one sense to each of these modal verbs. These results suggest that representations for e.g. deontic and epistemic *must* are different, but that there is no overall representation for any one sense.

Lastly, BERT was trained to predict masked tokens. The final test to ascertain BERT's ability to recognise modal sense is therefore masked prediction: can BERT predict masked modal verbs?

6 Experiment 3

6.1 Methods

We mask modal verbs from MPQA and EPOS-E and let BERT predict them, using the *pipeline*

Modal verb	*could*		*should*		*can*		*must*		*may*	
Data set	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E
Instances	45	19	53	30	75	34	38	218	29	213
Mean precision per sense	0.17	0.52	0.62	0.5	0.21	0.33	0.75	0.47	0.15	0.47
Mean recall per sense	0.28	0.61	0.75	0.5	0.33	0.33	0.83	0.5	0.2	0.5
Mean accuracy per sense	0.44	**0.75**	0.75	0.5	0.42	0.33	0.83	**0.94**	0.29	**0.95**
Overall accuracy	**0.64**	**0.84**	**0.96**	0.97	**0.69**	0.91	**0.95**	**0.94**	0.72	**0.98**
Baseline	0.62	0.58	0.92	0.97	0.68	0.91	0.87	0.63	0.72	0.86

Table 2: Modal sense classification results, separate training of classifiers for each modal verb. Overall and mean results by senses. Accuracies that meet or exceed baseline in **boldface**.

function from huggingface's *transformers* library (version 4.23.1; Wolf et al. 2020).

6.2 Results

Success of masked modal verb prediction depends on the modal verb. In both datasets (see Table 3), *should* is predicted correctly most commonly, with an accuracy of 0.44 in MPQA for the top prediction and 0.80 for the top three predictions. In EPOS-E, this rises to 0.52 and 0.83, respectively. *Could* and *must* also are frequently predicted correctly in both datasets, though they switch places: *must* is predicted correctly more often than *could* in EPOS-E, but the reverse is true in MPQA. *May* is predicted correctly least often in all layers. Words other than modal verbs are only predicted rarely: accuracies lie between 0.87 (EPOS-E, first prediction only) and 0.98 (MPQA, top 3 predictions) of predictions are modal verbs.

6.3 Interpretation

This experiment indicates that, as expected from results of per-modal-verb classification in the second experiment (see Section 5), BERT succeeds at predicting modal verbs where it failed at classifying modal verb sense. *May* appears most difficult to predict. *Should*, despite not being the most common modal verb, especially in EPOS-E, where *must* occurs over seven times as often, is correctly predicted most frequently. These results are strong considering the relatively minute semantic differences between modal verbs – syntactically, any of them would be an acceptable prediction.

This partially confirms the observations made in the first and second experiment (Sections 4 and 5). There, too, classification of *must* is overall most successful. Combining this with the results from the second experiment (Section 5), it appears that the relatively strong prediction performance may be unrelated to an overarching representation of modal sense.

Lastly, the question remains whether BERT embeddings encode modal verb sense equally well in different varieties of English.

7 Experiment 4

7.1 Methods

For each modal verb in the varieties of English modal sense corpus (VEM, see Section 3), we train a logistic regression classifier on that verb's instances in the EPOS-E dataset, mirroring the methodology from the second experiment (Section 5). We then predict modal verb senses for that verb in the Varieties of English modal sense corpus (VEM) and compare overall accuracy, precision, and recall for each modal verb and for each variety.

We do not train a separate classifier on each variety of English. While this will undoubtedly lead to diminished success for some (or all) varieties, we believe that this reflects real-world scenarios. By its nature, the amount of data for minority varieties of English will be lower than for the varieties present in EPOS-E. As the point of this experiment is to see whether automatic modal sense classification for other varieties of English is viable, we therefore use the large pre-existing EPOS-E dataset.

7.2 Results

Classification of modal verbs' senses (see Table 4) is most successful for *must* (overall accuracy = 0.90; mean accuracy = 0.86). We reach the lowest overal accuracy for *could* (0.70) and the lowest mean accuracy for each verb's possible senses (0.32), mean precision (0.33), and mean recall (0.30) for *can*.

Modal verb	*could*		*should*		*can*		*must*		*may*		modal rate	
Data set	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E	MPQA	EPOS-E
Instances	216	88	254	139	355	154	173	1054	136	1009		
acc@1	0.34	0.43	0.44	0.52	0.43	0.60	0.38	0.48	0.35	0.44	0.89	0.87
acc@2	0.59	0.61	0.67	0.72	0.59	0.71	0.58	0.68	0.54	0.61	0.95	0.93
acc@3	0.78	0.75	0.80	0.83	0.69	0.74	0.72	0.78	0.65	0.71	0.98	0.95

Table 3: Results of masked modal verb prediction. The last column shows the rate of modal verbs in the top predictions.

Modal verb	*could*	*should*	*can*	*must*	*may*
Instances	156	156	154	158	158
Mean precision per sense	0.23	0.50	0.11	0.43	0.25
Mean recall per sense	0.33	0.50	0.22	0.50	0.25
Mean accuracy per sense	**0.70**	0.50	0.32	**0.86**	0.50
Overall accuracy	**0.70**	0.90	0.73	**0.90**	0.88
Baseline	0.52	0.90	0.79	0.75	0.89

Table 4: Results of modal verb sense classification on varieties of English. **Bolded** accuracies are above respective baseline(s) (most frequent sense for each verb).

Variety	PH	HK	NI	IN	SL	JA	IR	CA
Instances	99	98	99	96	97	98	96	99
Mean precision per modal verb	0.59	0.56	0.57	0.66	0.60	0.59	0.58	0.65
Mean recall per modal verb	0.55	0.54	0.60	0.69	0.60	0.58	0.53	0.61
Mean accuracy per modal verb	0.78	0.76	0.79	0.85	0.85	0.87	0.77	0.91
Overall accuracy	0.78	0.77	0.79	0.85	0.85	0.87	0.77	0.91

Table 5: Results of modal verb sense classification on varieties of English: mean metrics for each variety. Note: we do not report a baseline since, without separating by modal verbs, this would be meaningless.

We reach the highest overall accuracy for Canadian English (0.91), followed by Jamaican (0.87), Sri Lankan, and Indian English (both 0.85). We reach the lowest overall accuracy for Hong Kong English and Irish English (both 0.77) For more results, see Table 5.

We choose the Nigerian English results for a brief example. In Sentence (1), *may* is predicted to have deontic sense, when annotators agreed it should be epistemic. Note the lack of space between *i'm* and *wrong* as well as the (subjectively) non-standard use of *wonder*:

(1) I wondered at one point that you may have forgotten us, but your mail now makes me think i'mwrong

Conversely, in Example (2), epistemic *may* was classified correctly:

(2) Chieftains from the 55 local councils may be lending moral and financial support to their counterparts in the two Ibeju-Lekki councils, sources said

In all correct classifications of *may* in the Nigerian English sample, *be* occurs in the vicinity of *may* – at times negated. The reason for incorrect classification can not be as simple as non-occurrence of *be*, as *be* also occurs in 5 of 15 instances of misclassified *may*, such as in Example (3):

(3) He may be very poor, poorer than a church rat

The instance of *may* in Example (3) was also classified incorrectly as deontic. Note that this sentence appears much less non-standard than the previous example. It must be kept in mind that classification in the second experiment (see Section 5) was also not perfect, meaning that (at least some) wrong classification despite no discernible presence of non-standard language may be caused by general model errors rather than meaning variation. Genre variation and register may also play a role: Example (1) is taken from a social letter, Example (3) from a novel; Example (2), in which modal sense was classified correctly, is taken from press coverage, which may be more similar to the parliament proceedings used in EPOS-E.

7.3 Interpretation

Some of the classification performance differences between modal verbs are mirrored in the second experiment (see Section 5), though nearly all performance metrics are lower compared to the second experiment. This may be due to the different register of the texts: while EPOS-E is comprised of European Parliament proceedings and subtitles, the ICE corpora consist of various kinds of writings, none of which include parliamentary writings or subtitles. This does not account for differences between the varieties, however.

Sense classification being most successful in Canadian English is not surprising, as BERT's training materials are likely predominantly comprised of American English, to which Canadian English bears the greatest similarity (Schneider, 2006; Kytö, 2019). The strong performance reached for Sri Lankan English may be due to the later collection date in the 2010s as opposed to the majority of the ICE corpora, which were collected in the 1990s. Thus, "colonial lag" (Hundt, 2009) may be causing this data to be more similar to the EPOS-E data, though the concept is disputed. The strong performances on Jamaican and Indian English (as Outer Circle varieties; Kachru, 1985) and the poor performance on Irish English (as an Inner Circle variety), are more surprising and warrant further investigation. While the difference between varieties is not enormous (overall accuracies range from 0.77 to 0.91), they are not negligible, either.

8 Conclusion and outlook

Our experiments have demonstrated that BERT does not appear to have any representations of modal sense as its own category. Classification did not show satisfactory results for either modal verb sense the embeddings of modal verbs or modality in the [CLS] token. However, BERT showed some ability to predict masked modal verbs, though its success depends greatly on which modal verb has been masked, making it unclear whether this is truly an ability to predict specific modal verbs or rather prediction of *any* modal verb. Modality does not appear to be encoded in the [CLS] token at all, calling into question whether sentence-level encodings of modality exist in BERT. However, different classifiers may yield different results, and representations of sentence meaning other than the [CLS] token (such as summing up embeddings) may yet encode modality. Further research is thus necessary to come to a complete conclusion.

Classification was most successful when done separately for each individual modal verb. This indicates that, while BERT may not have representations of modal verb sense as its own category, it does appear to encode sense differences for each modal verb. Thus, it can differentiate between *must* in sentences like "You *must* complete all tasks for course credit" and "You *must* be tired after the long journey", but it also views the deontic modal verbs *must* and *should* in a sentence like "You *must*/*should* do your homework" as different. This has some intuitive appeal - clearly, the actual meanings of the sentence change quite considerably with the strength of deontic obligation expressed by *must* and *should*, respectively.

The results of the last experiment demonstrate that the difference in modal verb sense use across different varieties of English may negatively impact this performance. Some varieties (Canadian, Sri Lankan, Indian, and Jamaican English) reach comparatively good performance, while modal verb sense classification in Irish English proves difficult. It is clear that more focus must be put on linguistic diversity for language models to be more useful for such (often marginalised) varieties.

Further research into BERT's representations of modal sense may focus on non-categorical representations of modal sense. Those who have annotated modal sense can attest that it is not always very clear-cut, and often, more than one interpretation of modal sense can be perceived as valid. As BERT embeddings are continuous - only our classification forces them into categories - researchers may want to investigate whether non-continuous BERT embeddings of modal verbs also match human annotators' certainties or disagreements.

Acknowledgements

We would like to thank our student assistants Hannes Gröner and Bastian Bunzeck for their diligent work annotating the modal verb senses in VEM. Without them, this work would not have been possible. We also thank the comments and suggestion of the anonymous reviewers – they have contributed to improving our work. Part of this research was carried out as part of the SAIL project which is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant no NW21-059A.

References

Laura Aina, Kristina Gulordava, and Gemma Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.

Marianna Apidianaki. 2023. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, pages 1–59.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognit. Comput.*, 13(4):1008–1018.

Peter Collins, Ariane Macalinga Borlongan, and Xinyue Yao. 2014. Modality in Philippine English: A diachronic study. *Journal of English Linguistics*, 42(1):68–88.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Beke Hansen. 2018. *Corpus Linguistics and Sociolinguistics: A Study of Variation and Change in the Modal Systems of World Englishes*. Brill, Leiden, The Netherlands.

Marianne Hundt. 2009. Colonial lag, colonial innovation or simply language change? In Günter Rohdenburg and Julia Schlüter, editors, *One language, two grammars?*, pages 13–37. Cambridge University Press.

Braj B. Kachru. 1985. Standards, codification and sociolinguistic realism. The English language in the Outer Circle. In Randolph Quirk and H.G. Widdowson, editors, *English in the World. Teaching and Learning the Language and Literatures*, pages 11–30. Cambridge University Press.

Merja Kytö. 2019. English in North America. In Daniel Schreier, Marianne Hundt, and Edgar W.Editors Schneider, editors, *The Cambridge Handbook of World Englishes*, Cambridge Handbooks in Language and Linguistics, pages 160–184. Cambridge University Press.

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*, pages 1–6, Bruges, Belgium.

Yang Liu, Xiaohui Yu, Bing Liu, and Zhongshuai Chen. 2014. Sentence-level sentiment analysis in the presence of modalities. In *Computational Linguistics and Intelligent Text Processing*, pages 1–16, Berlin, Heidelberg. Springer Berlin Heidelberg.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Lucia Loureiro-Porto. 2019. Grammaticalization of semi-modals of necessity in Asian Englishes. *English World-Wide*, 40(2):115–143.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

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, 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.

Ana Marasović, Mengfei Zhou, and Anette Frank. 2019. *The MSC Data Set*. heiDATA.

Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy. 2021. Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10852–10865, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28–36, Boulder, Colorado. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

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).

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Edgar W. Schneider. 2006. English in North America. In *The Handbook of World Englishes*, pages 58–77. Blackwell Publishing Ltd.

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(S11).

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

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

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. *Proc. Conf. AAAI Artif. Intell.*, 34(05):9628–9635.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the International Conference on Computer Vision (ICVV)*, pages 19–27.