Semantically Enriched Models for Modal Sense Classification

Mengfei Zhou¹ Anette Frank¹,² Annemarie Friedrich³ Alexis Palmer¹

¹Department of Computational Linguistics, Heidelberg University, Germany
²Research Training Group AIPHES, Dept. of Computational Linguistics, Heidelberg University
³Department of Computational Linguistics, Universität des Saarlandes, Germany

{zhou,frank,palmer}@cl.uni-heidelberg.de
afried@coli.uni-saarland.de

Abstract

Modal verbs have different interpretations depending on their context. Previous approaches to modal sense classification achieve relatively high performance using shallow lexical and syntactic features. In this work we uncover the difficulty of particular modal sense distinctions by eliminating both distributional bias and sparsity of existing small-scale annotated corpora used in prior work. We build a semantically enriched model for modal sense classification by novelly applying features that relate to lexical, proposition-level, and discourse-level semantic factors. Besides improved classification performance, especially for difficult sense distinctions, closer examination of interpretable feature sets allows us to obtain a better understanding of relevant semantic and contextual factors in modal sense classification.

1 Introduction

Factuality recognition (de Marneffe et al., 2011) is an important subtask in information extraction. Beyond bare filtering aspects of veridicality recognition, classification of modal senses plays an important role in text understanding, plan recognition, and the emerging field of argumentation mining. Communication revolves about hypothetical, planned, apprehended or desired states of affairs. Such ‘extrapositional’ meanings are often linguistically marked using modal verbs, adverbs, or attitude verbs, as in (1) for hypothetical situations.

(1) a. He must’ve hurt himself.
   b. He has certainly found the place by now.
   c. We anticipate that no one will leave.

Following Kratzer (1991)’s seminal work in formal semantics, recent computational approaches such as Ruppenhofer and Rehbein (2012) distinguish different modal ‘senses’, most prominently, epistemic (2.a), deontic/bouletic (2.b) and circumstantial/dynamic (2.c) modality.

(2) a. Geez, Buddha must be so annoyed!
   (epistemic – possibility)
   b. We must have clear European standards.
   (deontic – permission/request)
   c. She can’t even read them.
   (dynamic – ability)

Modal sense tagging is typically framed as a supervised classification task, as in Ruppenhofer and Rehbein (2012), who manually annotated the modal verbs must, may, can, could, shall and should in the MPQA corpus of Wiebe et al. (2005). The obtained data set comprises 1340 instances. Maximum entropy classifiers trained on this data yield accuracies from 68.7 to 93.5 for the different lexical classifier models. While these accuracies seem high, we note a strong distributional bias in their data set. Due to the small data set size (200-600 instances per modal verb) and its distributional bias, classifiers trained on this corpus are prone to overfitting and hardly beat the majority baseline. Indeed, none of the classification models in Ruppenhofer and Rehbein (2012) (henceforth R&R) is able to beat the baseline with uniform settings across all modal verb types.

Of particular concern in our work are specific sense ambiguities that are difficult to discriminate, such as dynamic vs. deontic readings of can (3.a), epistemic vs. dynamic readings of could (3.b) or epistemic vs. deontic readings of should (3.c).

(3) a. You can do this, if you want.
   ability (dy) vs. permission (de)
   b. He could have arrived in time.
   possibility (ep) vs. ability (dy)
   c. He should be aware of the issue.
   possibility (ep) vs. obligation (de)
In this paper we reexamine prior work on modal sense classification and show that specific distinctions are difficult for state-of-the-art models. We show that modal sense classification is a challenging problem that profits from lexical, proposition-level and discourse-level semantic information.

**Our goals and contributions** are as follows:

(i) We investigate the impact of **semantic and discourse-related factors for modal sense classification**, looking in particular at difficult modal sense distinctions. Accordingly, we define a range of semantically inspired linguistic feature classes. The feature groups are related to lexical and propositional semantics, as well as discourse-level semantics, ranging from tense and aspect to speaker/hearer orientation.

As an example, one of our hypotheses is that aspectual event types play a decisive role in deontic vs. epistemic sense disambiguation for modal verbs such as **must**. Our intuition is that events are more likely to co-occur with the deontic sense of **must** (4.a,b), whereas statives are more likely to co-occur with the epistemic sense (4.c).

(4)  
   a. The prisoners **must** return their weapons.  
   b. Prisoners of war **must** be returned to their home countries.  
   c. They **must** be so scared.

(ii) As a precondition for the aims of this work, we construct a large corpus that is balanced for modal sense distribution and less prone to overfitting compared to prior work. To this end we apply a **paraphrase-driven cross-lingual modal sense projection approach** using parallel corpora. We show that this automatic acquisition method yields modal sense annotations of very high accuracy.

(iii) Using this corpus as training data, we devise a **novel, semantically enriched model for modal sense classification**. We assess the impact of diverse feature groups for modal sense classification in unbiased classification settings and analyze to what extent they contribute to solving difficult disambiguation problems.

**Overview.** We review related work in Section 2. Section 3 outlines an automatic modal sense projection approach using parallel corpora. We apply this method to bilingual corpora and evaluate the quality of the obtained data set. Section 4 motivates and describes semantic and discourse-oriented features for modal sense classification. These are examined in classification experiments in Section 5. We reconstruct the modal sense classifier of Ruppenhofer and Rehbein (2012) to compare against prior work. We evaluate the performance of different models in unbiased classification experiments, using the harvested sense-labeled corpora for training. We analyze the impact of different feature groups on disambiguation performance and relate them to specific difficult disambiguation classes. Section 6 concludes.

2 Related Work

Most relevant to our work is the state of the art in modal sense classification in Ruppenhofer and Rehbein (2012). They manually annotated modal verbs in the MPQA corpus of Wiebe et al. (2005). Their annotation scheme departs from both the earlier setting in Baker et al. (2010) and a more recent proposal in Nissim et al. (2013). Baker et al. (2010) distinguish 8 categories. Next to **requirement, permissive, want and ability**, they include **success, effort, intention and belief**. They measured precision in automatic tagging of 86.3% by examining 249 modality-tagged sentences. Nissim et al. (2013) propose a fine-grained hierarchical modality annotation scheme that can be applied cross-linguistically. It includes (subtypes) of factuality, as well as speaker attitude. To our knowledge their annotation scheme has not been used for computational tagging.

Ruppenhofer and Rehbein (2012) apply the well-established modal sense categories of Kratzer (1991): **epistemic, deontic/bouletic and circumstantial/dynamic** modality. They add the categories: **concessive, conditional and optative**. Their annotation scheme proves reliable both in inter-annotator agreement, which ranges from \( K = 0.6 \) to 0.84 for the different modal verbs, and classification performance, which yields accuracies between 68.7 and 93.5, depending on the verb. However, the sense distributions of their data set are heavily biased (cf. Table 2, Section 5), and as a consequence, the majority sense baselines are hard to beat. The classification model of Ruppenhofer and Rehbein (2012) employs a mixture of target and contextual features, taking into account surface, lemma and PoS information, as well as syntactic labels and path features linking targets to their surrounding words and constituents. These features are able to capture very diverse contextual factors, but it is difficult to interpret their impact for distinguishing modal senses.
3 Paraphrase-driven Sense Projection

Given the sparsity and distributional bias in existing modal sense annotated corpora such as the MPQA, we propose a method for cross-lingual sense projection to alleviate the manual annotation bottleneck. Our approach exploits the paraphrasing behaviour of modal senses, which holds across modal verbs, modal adverbs and certain attitude verbs. As illustrated in (5) and (6), this paraphrasing behaviour is applicable across languages.

(5) a. *He may* be home by now. (possibility)
    b. *You may* enter this building. (permission)
    c. *May* you live 100 years. (wish)

(6) a. *Vielleicht* ist er schon zu Hause.
    *Maybe is he already at home.*
    b. *Es ist gestattet, das Gebäude zu betreten.*
    *It is permitted the building to enter*
    c. *Hoffentlich* werden Sie 100 Jahre.
    *Hopefully become you 100 years*

Capitalizing on the paraphrasing capacity of such expressions, we apply a semi-supervised cross-lingual projection approach, similar to prior work in annotation projection (Yarowsky and Ngai, 2001; Diab and Resnik, 2002):

(i) we select a seed set of cross-lingual sense indicating paraphrases,
(ii) we extract modal verbs in context that are in direct alignment with one of the seed expressions in word-aligned parallel corpora, and
(iii) we project the label of the sense-indicating paraphrase to the aligned modal verb.

Experimental setup and annotation scheme.

German is our source language, and we project into English. We adopt R&R’s annotation scheme, which is grounded in Kratzer’s modal senses epistemic, deontic and dynamic. While R&R add the novel categories conditional, concessive and optative,\(^1\) we subsume the former two as cases of epistemic and optative as a subtype of deontic.

Seed selection. The seeds were manually selected from PPDB (Ganitkevitch et al., 2013) and parallel corpora from OPUS (Tiedemann, 2012). The major criterion, besides frequency of occurrence, was non-ambiguity regarding the modal sense. We chose 30 seed phrases. Examples are adverbs like *wahrscheinlich* (probably – epistemic), *hoffentlich* (hopefully – deontic), adjectives like *erforderlich* (necessary – deontic), verbs like *gelingen* (succeed – dynamic), *erlauben* (admit – deontic) or affixes such as -bar (-able) as in *lesbar* (readable) – dynamic). For projection we employed the word-aligned Europarl (Koehn, 2005) and OpenSubtitles parallel corpora.

Projection and validation. We extracted 11,610 instances with direct alignment of modal sense paraphrase and modal verb. 80.6% were labeled epistemic, 8.2% deontic, 11.2% dynamic.

In order to assess the quality of the heuristically sense-labeled modal verbs we performed manual annotation on a balanced subset of the acquired data consisting of 420 sentences. We established annotation guidelines that ask the annotators to consider four paraphrasing possibilities for modal verbs: *possibility (epistemic), request (deontic), permission (deontic)*\(^2\) and *ability (dynamic)*. We performed annotation by two linguistically trained experts. They also annotated a balanced subset of 103 instances from R&R’s MPQA data set, in order to calibrate our annotation quality against the MPQA gold standard.

On the automatically acquired data (from Europarl and Open Subtitles) we obtain high annotator agreement at $\kappa=0.87$.\(^3\) Evaluating projected sense labels against ground truth, we observe high accuracy of .92. Agreement for MPQA is lower. There we achieve moderate agreement: $\kappa$ of 0.66 and 0.77 against the gold standard and 0.78 between annotators. In R&R, agreement averaged over the different modal verbs was 0.67. Our annotation reliability is largely comparable.

4 Semantic Features for Modal Sense Classification

In our work we expand the feature inventory used for modal sense classification to incorporate semantic factors at various levels. An overview of our semantic features is given in Table 1. We define specific feature groups for focused experimental investigation in Section 5. Feature extraction is performed using Stanford’s CoreNLP (Manning et al., 2014) and Stanford parser (Klein and Manning, 2002) to obtain syntactic dependencies.

\(^1\)Examples: “Should anyone call, please take a message” (conditional), “But, fool though he may be, he is powerful” (concessive), and “Long may she live!” (optative). (R&R)

\(^2\)We split permission and request to make the task more accessible and merged them to deontic later.

\(^3\)Cohen’s Kappa, Cohen (1960)
**VB: Lexical features of the embedded verb.**
The *embedded verb* in the scope of the modal plays an important role in determining modal sense. For instance, with the embedded verb *fly* in (7.a), we prefer a dynamic reading of *can*, whereas with *eat* in (7.b) we find a deontic reading.

(7) a. The children can fly (if they just believe, says Peter Pan!)
   b. The children can eat (ice cream) now.

We extract the *lemma* of the embedded verb and its *part-of-speech* tag in the sentence. We also extract whether the verb has a *particle* (e.g. *the plane could take off*), and if yes, which.

**SBJ: Subject-related features.** These features capture syntactic and semantic properties of the subject of the modal construction. In (8) a non-animate, abstract subject favors an epistemic reading for *could*, whereas with an animate subject, a dynamic reading is preferred. Other factors involve speaker/hearer/third party distinctions (9).

(8) (The conflict | He) could now move to a next stage. (ep | dy)
(9) a. I must be home by noon. (deontic only)
   b. He must be home by noon. (de or ep)

We extract *person* and *number* of the subject and the *noun_type* (common, proper, pronoun). Person is identified via personal pronoun features, and the other features are extracted from POS tags.

The *countability* of the noun is obtained from the Celex database (Baayen et al., 1996).

Lexical semantic features for the subject NP are extracted from WordNet (Fellbaum, 1999). Following Reiter and Frank (2010), we take the most frequent sense of the noun in WN *(subject_sense0)*, add the direct hypernym of this sense, the direct hypernym of that hypernym, etc., resulting in features *subject_sense[1-3]*. We also extract the top sense in the WN hierarchy *subject_sense_top* (e.g. *entity*) and the WN *lexical_filename* (e.g. *person*).

**TVA: Tense/voice/grammatical aspect features.**
These features capture tense and grammatical aspect of the embedded verb complex. LA below notes how grammatical aspect influences modal sense. At the same time, tense is an important factor for modal sense disambiguation. (10) clearly favors an epistemic reading, as the event is located in the past, whereas deontic sense is favored with future events in indicative mood as in (4.a).

We restrict the *tense* feature to the values {past, present}, determined via patterns of POS tags. We capture grammatical aspect features using sequences of POS tags of the verbal complex, following Loaiciga et al. (2014). The boolean features *perfect* and *progressive* indicate the respective grammatical aspect; *voice* indicates active or passive voice.

**LA: Lexical aspectual class.** Verbs can be used in a *dynamic* or *stative* sense, e.g. *I ate an apple* vs. *I like apples* (Vendler, 1957). The lexical aspect of a verb in context influences modal sense in some cases. In contrast to (4.a), for example, where the eventive verb *return* triggers the deontic sense, perfect aspect in (10) coerces the clause to stative, triggering the epistemic sense of *must*.

(10) The prisoners must have returned their weapons.

We label the lexical aspectual class of the embedded verb following Friedrich and Palmer (2014), who make use of both syntactic-semantic contextual features and linguistic indicators (Siegel and McKeown, 2000), which are patterns of usage for verb types estimated over a large

---

**Table 1: Individual features and feature groups.**

| Feature groups | Attributes |
|----------------|------------|
| **VB** | lemma, part-of-speech, POS of head, up, off, on,... |
| **TVA** | tense, progressive, perfect, voice, present / past, true / false, true / false, active / passive |
| **LA** | lexical aspect, dynamic / stative |
| **NEG** | negation true / false |
| **WNV** | WN sense [0 – 2] WN senseTop top sense in hypernym hierarchy |
| **SBJ** | number, sg, pl, person 1, 2, 3, countability from Celex, e.g. count, noun type common, proper, pronoun, WN sense [0 – 2] WN senses (head+hypernyms), WN senseTop top sense in hypernym hierarchy, WN lex. fn. person, artifact, event, ... |
| **Sentence structure** | S conjunct clause true / false, adjunct clause true / false, relative clause true / false, temporal mod. true / false |

---

47
parsed but otherwise unlabeled corpus. Accuracy for this prediction task is reported as around 84%.

**NEG: Negation.** Negation is a semantic feature at the proposition level that can have reflections in modal sense selection. *Should*, e.g., seems to favor a deontic meaning when negated in (11.a). Also, negation can interact with disambiguation of epistemic vs. deontic readings depending on propositional or discourse context. In (11.b), the favored reading is deontic in the negative sentence.

(11) a. He *should* (not) have returned.
    (ep/de (pos) vs. de (neg))

    b. He *may* (not) drink more gin tonight.
    (ep/de (pos) vs. de (neg))

The negation feature captures the presence or absence of negation in the modal construction. We use the dependency label `NEG` to identify negation.

**WNV: Lexical semantic features of the embedded verb.** This feature group encourages semantic generalization for lexical features of the embedded verb. It can play a role in interaction with other features, such as lexical and grammatical aspect and proposition-level features such as negation or the combined lexical semantic features described below (WN). The features in this group are parallel to the WordNet features described for the SBJ feature group above (WN). The intuition is that certain subject-predicate combinations may have a preference for certain modal senses. In (13), for example, *can* appears with a proposition that is subject to specific prescriptions or “laws”: soldiers are subject to restrictions with respect to consuming alcohol.

(13) a. Soldiers *can* drink when off duty.

**TVA/LA: Features of the verb complex.** Finally, this feature group uses both lexical aspect (LA) and tense, voice, and grammatical aspect (TVA) features. The goal is to investigate whether these two views of the verb complex are more effective separately or in combination.

5 Experiments & Results

Our experiments have several objectives:

(i.) We aim to show that modal sense classification, especially difficult sense distinctions, can profit from semantic and discourse-oriented features. To this end we construct contrasting classifier models with different feature sets:

- R&R’s shallow lexical and syntactic path features \( F_{R&R} \), a feature set consisting of only our newly designed semantic features \( F_{Sem} \), and a combined set \( F_{all} \) consisting of both \( F_{R&R} \) and \( F_{Sem} \).

- However, any classifier trained only on the highly unbalanced MPQA data set will have difficulty separating the effect of distributional bias in the training data from the predictive force of its feature set. A classifier that follows the majority class in the training data will neutralize the potential impact of its feature set. In order to counterbalance the distributional bias and also the sparsity inherent in the data, we evaluate the different classifier models in different classification settings:

(ii.) We extend the training set using heuristically labeled instances obtained from modal sense projection (cf. Section 3), thereby eliminating sparsity and reducing distributional bias.

(iii.) We further evaluate classifiers trained on perfectly balanced data. This eliminates the distributional bias in training and will allow us to carve out the impact of the different feature sets.
(iv.) Finally we measure the impact of individual feature groups via ablation (Section 5.3).

A note on notation: Subscripts on classifier names indicate the source of the training data. \( CL_M \) denotes a classifier trained only on MPQA data; \( CL_{MH} \) combines MPQA and heuristically-tagged data; \( CL_H \) is a classifier trained only on heuristically-tagged data. Superscripted +b or −b indicates a balanced vs. unbalanced training set.

### 5.1 Experimental settings

**Replicating R&R’s modal sense classifier.** We replicate R&R’s classifier by re-implementing their feature set,\(^4\) a mixture of target and contextual features that take into account surface, lemma and PoS information, as well as syntactic labels and path features linking targets to surrounding words and constituents (cf. R&R, Table 5).

We train one classifier per modal verb, using R&R’s best feature setting (context feature window=3 tokens left and right of target, target-specific features). Averaged accuracies for the replicated classifiers appear in Table 4 as \( CL_{M}^{-b} \) (feature set \( F_{R&R} \)). Our scores are very similar to their published results, which appear in the same table in the column headed “R&R”.\(^5\)

**Extending and balancing training data sets.** From the 11,610 heuristically sense tagged instances (Section 3), we construct balanced (+b) training corpora for each modal verb. The composition of this data is shown in Table 2. To alleviate training data sparsity, we add this data to the (unbalanced) MPQA data; this configuration results in \( CL_{M}^{-b} \). Finally, we re-balance both \( CL_M \) and \( CL_{MH} \) by under- and oversampling.\(^6\)

**Classification setup and test data.** Training on balanced data reduces distributional bias, but evaluating performance on an unbalanced, naturally-distributed data set gives us a more realistic picture. To this end, and in order to compare to prior work, our test data is drawn exclusively from MPQA. For \( CL_{H}^{+b} \), we evaluate on R&R’s full data set; the composition of the test set appears in the right-hand side of Table 2. The other two models (\( CL_M \) and \( CL_{MH} \)) are evaluated in a 5-fold CV setting, with testing on the naturally distributed MPQA instances. For each CV setting, only the training section is adapted, by addition of heuristic instances, and/or balancing. Table 3 exemplifies one run of our cross-validation setting. First, we split MPQA into 80% train (\( CL_{M}^{+b} \)) and 20% test, then we add the heuristically-tagged data (\( CL_{MH}^{-b} \)) and re-balance (to produce \( CL_{M}^{+b} \) and \( CL_{MH}^{+b} \)).

**Baselines.** For unbalanced classifiers, we compare to the MFS baseline (BL\(_{Maj,M} \)), taking the most frequent sense for each modal verb from the MPQA training data. For balanced classifiers, we compare to the random baseline (BL\(_{Ran} \)), determined by the (evenly distributed) number of class labels seen in training for each modal verb.

### 5.2 Comparative performance evaluation

Table 4 compares accuracy of classifiers trained on ±balanced data, from different sources, and with different feature sets. We report results for individual classifiers (per modal verb) and macro- and micro-average across all verbs. The two bold-faced numbers per table row indicate the best models for unbalanced and for balanced data. For the balanced classifiers, where we find more interesting differences, we test significance using McNemar’s test (\( p<0.05 \)) (McNemar, 1947). Within a row (for +b classifiers and micro-averages), a superscript on a number indicates which classifier is significantly outperformed by the result. Across feature sets, we compare micro-averages and mark significance by subscripts (\( R=F_{R&R} \), \( S=F_{Sem} \)).

We first discuss the classifiers trained on unbalanced data. With \( F_{R&R} \), \( CL_{M}^{-b} \) yields performance comparable to R&R’s results, at 84.44% accuracy, 1.02pp below the majority baseline. Individual lexical classifiers also approach R&R’s performance, though never beating the baseline.\(^7\)

| Feature Set | CL\(_H^{+b}\) train ep de dy | Full MPQA test ep de dy |
|-------------|-------------------------------|--------------------------|
| must        | 800 800 0                     | 11 183 0                 |
| may         | 950 950 0                     | 130 9 0                  |
| can         | 150 150 150                   | 2 115 271                |
| could       | 40 40 40                      | 156 17 67                |
| should      | 150 150 0                     | 26 248 0                 |
| shall       | 0 5 5                         | 0 11 2                   |

Table 2: Heuristic (+b) training data and MPQA (-b) training and test data

\(^4\)Following R&R we use the Stanford parser for processing and induce maximum entropy models using OpenNLP with default parameter settings.

\(^5\)R&R performed 10-fold cross-validation (CV) for evaluation. We perform 5-fold cross-validation instead.

\(^6\)When doing oversampling, we generally perform a mixture of over- and undersampling, targeting about half the size of the larger class. The data sets are available at http://projects.cl.uni-heidelberg.de/modals.

\(^7\)We report individual results, while R&R aggregated
Changing from $F_{R&R}$ to $F_{Sem}$ and $F_{Alt}$, classifier $CL_{M}^{b}$ for *could* is now able to beat the baseline. The effect is stronger for $F_{Sem}$, which reflects the impact of the semantic features. Interestingly, accuracy of $F_{Sem}$ is comparable to $F_{R&R}$, even though the classifiers learn *only* on the basis of semantic features. Combining the two feature sets ($F_{Alt}$) produces minimal differences for $CL_{M}^{b}$, but yields stronger gains for $CL_{M,H}^{b}$.

The addition of heuristically-tagged data in $CL_{M,H}^{b}$ helps for some verbs, but hurts for others. Despite the larger training set size, individual classifier performances tend to drop, meaning they do not profit much from the reduced training bias.

For classifiers trained on balanced data, the picture changes. Accuracies on balanced data are lower, reflecting the lack of distributional bias. But all results are well above the random BL. 8

---

### Table 3: Cross-validation, one run: representative class distributions of training and test data.

|            | $CL_{M}^{b}$ train | $CL_{M,H}^{b}$ train | $CL_{M}^{b}$ train | $CL_{M,H}^{b}$ train | MPQA test |
|------------|---------------------|----------------------|---------------------|----------------------|-----------|
| must       | 6 149 0             | 806 949 0            | 70 70 0             | 870 870 0            | 5 34 0    |
| may        | 105 6 0             | 1055 956 0           | 50 50 0             | 999 1000 0           | 25 3 0    |
| can        | 1 98 212            | 151 248 362          | 100 100 100         | 250 250 250          | 1 17 60   |
| could      | 120 15 57           | 160 55 97            | 54 54 54            | 94 94 94            | 36 2 10   |
| should     | 21 196 0            | 171 355 0            | 100 100 0           | 250 250 0            | 5 52 0    |
| shall      | 0 9 1               | 0 14 6               | 0 10 10             | 0 15 15             | 0 2 1     |

Table 4: Classifier accuracy for various training data and feature sets. See text for details.

| $F_{R&R}$ | R&R | $CL_{M}^{b}$ | $BL_{M_{M},M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $BL_{Run}$ |
|-----------|-----|-------------|-------------------|--------------|------------|--------------|------------|--------------|-----------|
| must      | 93.50 | 94.32 | 94.32 | 82.00 | **76.25** | 73.24 | 71.65 | 50.00 |
| may       | 81.43 | 93.57 | 93.57 | 90.71 | 79.29 | **88.57** | 90.71 | 50.00 |
| night     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| can       | 68.70 | 66.56 | 69.92 | 64.25 | 49.86 | 53.19 | **57.84** | 33.33 |
| could     | 62.50 | 65.00 | 90.81 | 59.17 | 41.25 | 44.17 | **49.17** | 33.33 |
| should    | 91.29 | 90.77 | 90.81 | 90.77 | 80.21 | **85.83** | 76.33 | 50.00 |
| shall     | 83.33 | 84.61 | 90.00 | 70.00 | 90.00 | 53.85 | 50.00 |
| macro-avg.| 83.73 | 84.44 | **85.46** | 82.41 | 70.98 | **76.43** | 71.36 | 52.38 |
| micro-avg.| 78.71 | **80.22** | 75.22 | 62.59 | **66.24** | 66.08 | 41.54 |

| $F_{Sem}$ | R&R | $CL_{M}^{b}$ | $BL_{M_{M},M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $BL_{Run}$ |
|-----------|-----|-------------|-------------------|--------------|------------|--------------|------------|--------------|-----------|
| must      | 93.50 | 93.28 | **94.32** | 88.11 | 85.48 | 87.07 | 86.08 | 50.00 |
| may       | 81.43 | 92.86 | 93.57 | 87.14 | 83.57 | **87.14** | 84.29 | 50.00 |
| night     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| can       | 68.70 | 65.03 | **69.92** | 61.43 | 58.38 | 58.61 | 55.78 | 33.33 |
| could     | 72.08 | 65.00 | 90.81 | 69.17 | 59.17 | 57.50 | 50.00 | 33.33 |
| should    | 91.29 | 89.71 | 90.81 | 90.79 | 82.68 | 81.97 | 79.15 | 50.00 |
| shall     | 83.33 | 84.61 | 66.67 | 66.67 | 66.67 | 66.67 | 66.67 | 50.00 |
| macro-avg.| 83.73 | 85.18 | **85.46** | 80.47 | 77.99 | 76.99 | 71.64 | 52.38 |
| micro-avg.| 79.59 | **80.22** | 75.22 | 62.59 | **66.24** | 66.08 | 41.54 |

| $F_{Alt}$ | R&R | $CL_{M}^{b}$ | $BL_{M_{M},M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $CL_{M}^{b}$ | $CL_{M,H}^{b}$ | $BL_{Run}$ |
|-----------|-----|-------------|-------------------|--------------|------------|--------------|------------|--------------|-----------|
| must      | 93.50 | **94.32** | **94.32** | 92.27 | 86.02 | **90.72** | 88.66 | 50.00 |
| may       | 81.43 | **93.57** | **93.57** | 92.14 | 87.86 | **92.14** | **92.14** | 50.00 |
| night     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| can       | 68.70 | 65.28 | **69.92** | 65.27 | 54.50 | 58.60 | **63.50** | 33.33 |
| could     | **66.67** | 65.00 | 90.81 | 65.42 | **63.33** | 59.58 | 54.17 | 33.33 |
| should    | 91.29 | 90.77 | **90.81** | 90.77 | 84.09 | **90.79** | 84.09 | 50.00 |
| shall     | 83.33 | 84.61 | **90.00** | 83.33 | 90.00 | 53.85 | 50.00 |
| macro-avg.| 83.73 | 84.85 | **85.46** | 85.12 | 79.88 | **83.12** | 76.63 | 52.38 |
| micro-avg.| 79.11 | **80.22** | **78.47** | 71.73 | **75.06** | 73.31 | 41.54 | 50.00 |
Compared to CL^{+b}M and CL^{+b}H, we observe the best results for CL^{+b}MH, which mixes MPQA and out-of-domain data. Here, the best performance is obtained with F_{All}. In fact, CL^{+b}MH with 83.12% on balanced mixed data closely approaches the performance of the classifiers trained on biased training data and their majority baseline, with about 2pp difference, and being almost identical to R&R’s published results.

Looking at individual modal classifiers, we see even more interesting results. can and could, both with 3-fold sense distinctions and lowest performance overall, suffer the greatest loss in the balanced setting, in ranges of 41-57% for F_{R&R}. These verbs are hard to classify, and here we see a marked performance rise as the training data changes (from CL^{+b}M to CL^{+b}H), though these differences are not significant. Comparing F_{Sem} to F_{R&R}, we obtain better results overall, always above 50% accuracy. With F_{All} we reach a range of 54-63%, achieving strong gains of more than +20pp for could, and about +5pp for can. We also note an almost continuous rise for should with a final +5pp gain over F_{R&R}. Across different feature sets, CL^{+b}MH performs best, that is, combining MPQA and out-of-domain data is effective.

To summarize, with increasingly refined models and a tendency of CL_{MH} and CL_{H} outperforming CL_{M}, we obtain a coherent picture: semantic features contribute important information and reach their best performance with a mixture of training sets. We also note that F_{Sem} and F_{All} jointly yield significant gains over F_{R&R} for could, must, should, can and may.9

### 5.3 Impact of feature groups

A confusion analysis of the predictions made by CL^{+b}H using F_{R&R} yields some insight into the most difficult sense distinctions for specific modal verbs. Table 5 highlights the most prominent misclassification classes: for instance, deontic can is misclassified as dynamic in 106 cases; epistemic could is misclassified as dynamic in 53 cases, etc.

For a deeper analysis of the impact of our semantic features, particularly on specific sense distinctions, we conducted a quantitative and qualitative evaluation by ablating individual feature groups (FGs) from the full feature sets F_{Sem} and F_{All}, for all balanced classifiers.

It turns out that precisely for the modal verbs that exhibit prominent confusion classes in Table 5 we observe a significant performance drop when omitting individual feature groups (FGs): Table 6 reports all configurations where omitting a particular FG yielded a significant accuracy loss. In the following we analyze these cases in more detail.

### Analysis

**Gains** (or rescues) due to FGx are cases in which including FGx turns a wrong classification into a correct one, compared to a model that ablates FGx. **Losses** record the opposite: a correct classification made without FGx becomes incorrect when FGx is active.

Overall, for both models F_{Sem} and F_{All} we observe more gains than losses due to the FGs SBJ, NEG, TVA/LA and WN: 140 vs. 41 (29% losses) for F_{Sem} and 195 vs. 42 (22% losses) for F_{All}. For must there are only gains and no losses at all.

We observe different performance for correction of misclassifications for the different modal verbs, and we see clearly distinct contribution of FGs for the individual modal verb classifiers.

The most clear-cut positive effects are obtained for must, with the highest number of gains (62/81 for F_{Sem}/F_{All}) and no losses. Here, exclusively the FGs TVA and TVA/LA are effective, leading to a majority of rescues of deontic readings that otherwise would be misclassified as epistemic. 5 rescues in the other direction occur, only with F_{Sem}.

For must through FG TVA/LA all meet the assumption that dynamic event readings of the verb go along with deontic sense (14.a), while stative readings (14.b) go along with epistemic sense.

(14) a. “Everything **must be done** by everyone to bring about de-escalation” […]

b. And as all **must now know** […] Mugabe has no chance of winning any ballot […]

A particularly strong effect is seen for TVA, which avoids misclassification of up to 12% of all

| can | ep | de | dy | could | ep | de | dy |
|-----|----|----|----|-------|----|----|----|
| ep  | 1  | 0  | 1  | ep    | 92 | 11 | 53 |
| de  | 8  | 1  | 106| de    | 6  | 2  | 9  |
| dy  | 28 | 21 | 223| dy    | 30 | 6  | 31 |

Table 5: Confusion analysis: CL^{+b}H using F_{R&R}
instances of must as epistemic. All cases follow
the pattern in (15.a): the verb is not in past tense,
and we prefer a deontic interpretation, whereas
past tense in (15.b) indicates epistemic usage.

(15) a. [..] whoever is on the other side is the
evil that must be destroyed [..]
b. The event must have rocked the halls of
power [..]

should displays similar sense ambiguities and
confusion patterns, but here the picture is less
clear: as with must we obtain rescues of deontic
readings, but here the WN features are most effec-
tive, jointly with SBJ. In contrast to must, we ob-
serve a mixture of gains (30/13) and losses (11/7)
due mostly to over-correction. While for the other
modal verbs, the gains/losses ratio is best for the
F_all model, should performs best with F_Sem.

For could, with a 3-way ambiguity, a different
feature set is active: SBJ and NEG. Most rescues
to epistemic are due to including SBJ fea-
tures, and a strong effect is also seen for NEG. For
both FGs we also observe gains of dynamic
readings from epistemic misclassifications, while
this effect is stronger for NEG, also in avoiding over-
correction. On the losses side, we observe 32% of
losses as opposed to gains for F_all.

SBJ features apparently capture a preference for
inanimate, abstract subjects for epistemic as opposed
to deontic (or dynamic) readings, as with the message
or propositional anaphora in (16.a,b). The same
pattern is observed with should (16.c).

(16) a. “the message could not be clearer.”
b. [..] officials said this could prompt indus-
tries to change behavior . . .
c. [..] if that should prove necessary, De
Winne will [..] pilot the space ship.

For NEG we see a clear effect that could, if negated,
is correctly analyzed as dynamic, while non-negated instances are classified as epistemic.

(17) a. Baghdad insisted [..] it could not be a
threat to the United States.
b. Two basic principles could still, perhaps,
make it possible.

Finally, can is our most difficult case. We ob-
tain moderate gains (15) by rescues of dynamic
readings from epistemic/deontic, through the SBJ
feature. As we see no gains with F_Sem, this means
we are still lacking precise features that can differ-
entiate epistemic and dynamic readings.

| verb | FG | comp. | impact |
|------|----|-------|--------|
| can  | SBJ | F_all | 2.83*  |
| could| SBJ | F_Sem | 12.50**|
|      |     | F_all | 6.25*  |
|      |     | F_Sem | 4.58*  |
|      |     | F_all | 6.25** |
| must | TVA | F_Sem | 5.69*  |
|      |     | F_all | 10.32**|
|      |     | F_Sem | 9.79** |
|      |     | F_all | 10.32**|
|      |     | F_Sem | 6.21** |
|      |     | F_all | 10.31**|
|      |     | F_Sem | 3.09*  |
|      |     | F_all | 10.32**|
| should| SBJ | F_Sem | 10.60* |
|      |     | F_all | 5.64** |
|      |     | F_Sem | 6.01*  |
|      |     | F_all | 12.37**|

Table 6: Accuracy loss by FG omission. 3rd column specifies from which feature set we ablate.

6 Conclusion

We show that difficult problems in modal sense
disambiguation can be addressed with seman-
tically enriched classification models that draw
upon lexical, propositional and discourse-level
semantic information. Our model obtains significant
improvements, especially for difficult sense dis-
tinctions, in balanced training setups. This will
prove advantageous when applying the classifiers
to documents with sense distributions that differ
from training. We further presented a method for
automatic induction of training corpora that helps
to alleviate sparsity and can be used to tailor train-
data to specific genres and domains.

The insights we gain from analyzing the im-
 pact of feature groups indicate avenues for future
work: The sensitivity of modal senses to seman-
tic properties of the subject calls for integration of
antecedent information with pronominal sub-
jects. The dependence on temporal information
calls for temporal resolution. Our current model
offers only a simple approximation of proposi-
tional semantics. We expect further improvements
with a more effective representation of proposi-
tional content and addition of more training data.

Acknowledgements This work has been partially
funded through the Leibniz ScienceCampus Empirical Linguistics
and Computational Language Modeling, supported by the
Leibniz Association under grant no. SAS-2015-IDS-LWC
and by the Ministry of Science, Research, and Art (MWK)
of the state of Baden-Württemberg. The third author is sup-
pported in part by the MMCI Cluster of Excellence. We thank
the anonymous reviewers for helpful comments.
References

Baayen, H. R., Piepenbrock, R., and Gulikers, L. (1996). CELEX2. Philadelphia: Linguistic Data Consortium.

Baker, K., Bloodgood, M., Dorr, B. J., Filardo, N. W., Levin, L., and Piatko, C. (2010). A Modality Lexicon and its use in Automatic Tagging. In Proceedings of LREC, pages 1402–1407.

Cohen, J. (1960). A coefficient for agreement for nominal scales. Education and Psychological Measurement, (20):37–46.

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2011). Veridicality and Utterance Understanding. 2011 IEEE Fifth International Conference on Semantic Computing, pages 430–437.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In Proceedings of ACL 2002, pages 255–262, Philadelphia, Pennsylvania, USA.

Fellbaum, C. (1999). WordNet. Wiley Online Library.

Friedrich, A. and Palmer, A. (2014). Automatic prediction of aspectual class of verbs in context. In Proceedings of the ACL 2014.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The Paraphrase Database. In Proceedings of the ACL-HLT 2013, pages 758–764, Atlanta, Georgia.

Klein, D. and Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. In Advances in neural information processing systems, pages 3–10.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of Machine Translation Summit X, pages 79–86.

Kratzer, A. (1991). Modality. In von Stechow, A. and Wunderlic, D., editors, Semantics: An International Handbook of Contemporary Research, pages 639–650. Berlin: de Gruyter.

Loaiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-French Verb Phrase Alignment in Europarl. In Proceedings of LREC 2014.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of ACL 2014: System Demonstrations, pages 55–60.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika.

Nissim, M., Pietrandrea, P., Sanso, A., and Mauri, C. (2013). Cross-linguistic annotation of modality: a data-driven hierarchical model. In Proceedings of the 9th Joint ISO - ACL SIGSE Workshop on Interoperable Semantic Annotation, pages 7–14, Potsdam, Germany.

Reiter, N. and Frank, A. (2010). Identifying Generic Noun Phrases. In Proceedings of the ACL 2010, pages 40–49, Uppsala, Sweden.

Ruppenhofer, J. and Rehbein, I. (2012). Yes we can!? Annotating the senses of English modal verbs. In Proceedings of the LREC 2012, pages 1538–1545.

Siegel, E. V. and McKeown, K. R. (2000). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. Computational Linguistics, 26(4):595–628.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, Proceedings of LREC-2012, pages 2214–2218, Istanbul, Turkey.

Vendler, Z. (1957). Linguistics in Philosophy, chapter Verbs and Times, pages 97–121. Cornell University Press, Ithaca, New York.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2-3):165 – 210.

Yarowsky, D. and Ngai, G. (2001). Inducing Multilingual POS Taggers and NP Brackets via Robust Projection Across Aligned Corpora. In Proceedings of the Second Meeting of ACL 2001, pages 200–207.