Automatic Extraction of Implicit Interpretations from Modal Constructions

Jordan Sanders and **Eduardo Blanco**
Human Intelligence and Language Technologies Lab
University of North Texas
Denton, TX, 76203
jordansanders3@my.unt.edu, eduardo.blanco@unt.edu

Abstract

This paper presents an approach to extract implicit interpretations from modal constructions. Importantly, our approach uses a deterministic procedure to normalize eventualities and generate potential interpretations. An annotation effort demonstrates that these interpretations are intuitive to humans and most modal constructions convey at least one interpretation. Experimental results show that the task is challenging but can be automated.

1 Introduction

People use language to communicate not only facts, but also intentions, uncertain information and points of view. Modality can be broadly defined as a grammatical phenomenon used to express the speaker's opinion or attitude towards a proposition (Lyons, 1977). Modality has also been defined as "the category of meaning used to talk about possibilities and necessities, essentially, states of affairs beyond the actual." (Hacquard, 2011). Within computational linguistics, processing modality has proven useful for, among others, recognizing textual entailment (Snow et al., 2006; MacCartney et al., 2006), machine translation (Murata et al., 2005; Baker et al., 2012), and sentiment analysis (Wiebe et al., 2005).

In the absence of modality markers, it is understood that the author of a proposition agrees with it (Hengeveld and Mackenzie, 2008). Adding a modality marker—also referred to as cue—casts doubt on the truth of the proposition, e.g., *Mary got a new job last week* vs. *Mary likely got a new job last week*. Modality is surprisingly common (Morante and Sporleder, 2012), and notoriously difficult to annotate and process automatically (Rubinstein et al., 2013; Vincze et al., 2011). In MEDLINE, 11% of sentences contain speculative language (Light et al., 2004) and in biomedical abstracts, 18% (Vincze et al., 2008). Rubin (2006) reports that 59% of statements in 80 New York Times articles include epistemic modality. Despite modality being ubiquitous, there is not an agreed upon annotation schema.

In this paper, we extract implicit interpretations intuitively understood by humans when reading modal constructions. We do not follow any specific theory of modality. Instead, we manipulate modal constructions to automatically generate potential interpretations, and then assign factuality scores to them. Consider statement (1) below:

1. John likely contracted the disease when a mouse bit him in the Adirondacks.

Even though *likely* syntactically attaches to *contracted*, a natural reading suggests that *John contracted the disease* is factual; the only bit of uncertain information is how (or when) he contracted the disease. In other words, assuming that the author of statement (1) is truthful, event *contracted* occurred with AGENT *John* and THEME *the disease*, but the MANNER (or TIME) may not have been *when a mouse bit him in the Adirondacks*.

A key feature of the work presented in this paper is that the interpretations extracted from modal constructions are not tied to any syntactic or semantic representation. Given modal constructions in plain text, we extract implicit interpretations in plain text, and these interpretations can be processed with any existing NLP pipeline. The main contributions of

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1098–1107,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

this paper are: (1) procedure to automatically generate potential interpretations from modal constructions; (2) annotations assessing the factuality of potential interpretations generated from OntoNotes;[1] and (3) experimental results using several features.

2 Previous Work

Theoretical works in philosophy and linguistics have studied modality for decades (Palmer, 2001; Jespersen, 1992). Morante and Sporleder (2012) summarize some of these works and related phenomena, e.g., evidentiality, certainty, factuality, subjectivity. There are several expressions that have modal meanings (Fintel, 2006), including auxiliaries (must, should, etc.), adverbs (perhaps, possibly, etc.) nouns (possibility, chance, etc.) adjectives (necessary, possible, etc.) and conditionals (e.g., If the light is on, Sandy is home). Most previous works in computational linguistics target modal adverbs (Rubinstein et al., 2013; Carretero and Zamorano-Mansilla, 2013; de Waard and Maat, 2012), and some also target other modal triggers such as reporting verbs (e.g., The evidence *suggests* that he caused the fire), references, or all verbs (Diab et al., 2009). Following these previous works, we focus on modal adverbs.

Beyond theoretical works, there are many proposals to annotate modality. Doing so has proven challenging: following different annotations schemas on the same source text yields little overlap (Vincze et al., 2011), and Carretero and Zamorano-Mansilla (2013) present an analysis of disagreements when targeting modal adverbs. Annotation schemas typically include 3 tasks: identifying modality triggers, their scopes, and sources (Quaresma et al., 2014; Sánchez and Vogel, 2015). Many also classify the modality into several types (epistemic, circumstantial, ability, deontic, etc.) or a fine-grained taxonomy (Rubinstein et al., 2013; Nissim et al., 2013). In this paper, we are not concerned with modeling modality per se, or classifying instances of modality into predefined classes or hierarchies. Instead, we extract implicit interpretations from modal constructions in order to mirror intuitive readings.

FactBank is probably the best-known corpus for event factuality (Saurí and Pustejovsky, 2009). It was created following carefully crafted annotation guidelines and examples comprising 34 pages.[2] The guidelines detail a manual normalization step to "identify the full event that needs to be assessed in terms of its factuality" (p. 12), and the annotation process includes identifying the sources that are assessing factuality (p. 15). de Marneffe et al. (2012) reannotate a subset of FactBank with factuality values from the reader's perspective—they call it veridicality—using crowdsourcing. Both FactBank and de Marneffe et al. (2012), rely on manual normalization to identify the eventuality whose factuality is being annotated. Instead, we present an automated approach: we manipulate semantic roles and syntactic dependencies deterministically to generate several potential interpretations per modal construction, and then assess their factuality.

Many other efforts expand on FactBank using crowdsourced annotations, different annotation schemas (usually simpler) or other domains. Prabhakaran et al. (2012) use crowdsourcing to classify propositions into 5 modalities: ability, effort, intention, success and want. Soni et al. (2014) target the factuality of quotes (direct and indirect) in Twitter. Lee et al. (2015) detect events and assess factuality using easy-to-understand short instructions to crowdsource annotations. Unlike us, they annotate factuality at the individual token level, where annotated tokens are deemed events by annotators. Prabhakaran et al. (2015) define and annotate propositional heads with four categories: (1) non-belief propositions, or (2) committed, non-committed or reported belief. Instead of assessing factuality only for propositional heads (usually verbs, one assessment per proposition), we do so for potential interpretations automatically generated by manipulating verbs and their arguments deterministically.

All works cited in the previous two paragraphs either manually normalize text prior to assessing factuality—making automation from plain text impossible—or assess factuality for tokens deemed events (ordered, delay, agreed, etc.) or full propositions (a verb and all its arguments). Unlike them, we automatically generate potential interpretations from a single modal construction—or, equivalently, automatically generate several normalizations—and then assess their factuality.

[1] Available at www.sanders.tech

[2] https://catalog.ldc.upenn.edu/docs/LDC2009T23/annotationGuidelines.pdf

3 Terminology and Background

We use the term *modal construction* to refer to verb-argument structures modified by a modal adverb (possibly, probably, etc.). We use the term *implicit interpretation*, or *interpretation* to save space, to refer to meaning intuitively understood by humans when reading a modal construction. *Potential interpretations* are interpretations automatically generated whose factuality has yet to be determined. The *factuality* of an interpretation is a score indicating its likelihood—whether it is true, false or unknown given the modal construction.

We work on top of OntoNotes (Hovy et al., 2006) because it includes text from several genres (news, broadcast and telephone conversations, weblogs, etc.) and includes part-of-speech tags, parse trees, PropBank-style semantic roles and other linguistic information.[3] Very briefly, PropBank (Palmer et al., 2005) has two kinds of semantic roles: numbered roles (ARG_0, ARG_1, etc.), which are defined in verb-specific framesets, and argument modifiers (ARGM-TMP, ARGM-LOC, etc.), we refer the reader to the aforementioned reference, and the guidelines and framesets[4] for more details. We transformed the parse trees in OntoNotes into syntactic dependencies using Stanford CoreNLP (Manning et al., 2014).

4 Corpus Creation

We define a two-step procedure to create a corpus of modal constructions and the implicit interpretations intuitively understood by humans when reading them. First, we automatically generate potential interpretations from modal constructions by manipulating syntactic dependencies and semantic roles. Second, we manually score potential interpretations according to their likelihood. These interpretations and scores are later used to learn how to score potential interpretations automatically (Section 6).

4.1 Generating Potential Interpretations

Selecting Modal Constructions. OntoNotes is a large corpus containing 63,918 sentences. Creating a corpus of interpretations for all modal constructions is outside the scope of this paper. In order to alleviate the annotation effort, we focus on selected modal constructions. Specifically, we select verb-argument structures that have one ARGM-ADV or ARGM-MNR role, and that role is one of the following modal adverbs: certainly, clearly, definitely, likely, obviously, possibly, probably, surely, or unlikely. These adverbs are the most frequent that satisfy the above filter. Additionally, we discard verb-argument structures with *to be* as the main verb. These rules retrieve 324 modal constructions.

Automatic Normalization. Modal constructions often occur in long multi-clause sentences. In order to identify the eventuality from which potential interpretations should be generated, we automatically normalize the original sentence. Normalizing consists of a battery of deterministic steps implemented using syntactic dependencies and semantic roles. In contrast with previous work (Section 2), our normalization is fully automated. Hereafter, we use *verb* to refer to the main verb in the modal construction, *adverb* to the modal adverb, and *sem_roles* to all semantic roles in the modal construction.

1. Remove *adverb*.
2. Convert negated verb-argument structures into their positive counterparts. We follow 3 steps inspired by the rules to form negation proposed by (Huddleston and Pullum, 2002):
   (a) Remove the negation mark by deleting the token whose syntactic dependency is *neg*.
   (b) Remove auxiliaries, expand contractions, and fix third-person singular and past tense. For example (before: after), *doesn't go*: *goes*, *didn't go*: *went*, *won't go*: *will go*. To implement this step, we loop through tokens whose head is the negated verb with dependency *aux*, and use a list of irregular verbs[5] and grammar rules to convert to third-person singular and past tense based on orthographic patterns.
   (c) Rewrite negatively-oriented polarity-sensitive items. For example (before: after), *anyone*: *someone*, *any longer*: *still*, *yet*: *already*. *at all*: *somewhat*. We use the correspondences between negatively-oriented and positively-

[3] We use the CoNLL-2011 Shared Task distribution (Pradhan et al., 2011), http://conll.cemantix.org/2011/

[4] http://propbank.github.io/

[5] https://en.wikipedia.org/wiki/English_irregular_verbs

Sent. 1: The danger is [probably]$_{\text{ARGM-ADV}}$ [he]$_{\text{ARG}_0}$ [can]$_{\text{ARGM-MOD}}$ [not]$_{\text{ARGM-NEG}}$ [deliver]$_{verb}$ [the promises that he made during the campaign.]$_{\text{ARG}_1}$		
Normalization	Step	Output
	1	The danger is he cannot deliver the promises that he made during the campaign.
	2	The danger is he can deliver the promises that he made during the campaign.
	3	The danger is he will deliver the promises that he made during the campaign.
	4	He will deliver the promises that he made during the campaign.
	5	He will deliver the promises that he made during the campaign.
Sent. 2: [...] I wouldn't define victory as simply not raising taxes—although [I]$_{\text{ARG}_0,\,v_1,\,v_2}$ [definitely]$_{\text{ARGM-ADV},\,v_1}$ [would]$_{\text{ARGM-MOD},\,v_1}$ [like]$_{v_1}$ [to [defer]$_{v_2}$ [raising taxes]$_{\text{ARG}_1,\,v_2}$ [as long as prudently possible.]$_{\text{ARG}_2,\,v_2}$]$_{\text{ARG}_1,\,v_1}$		
Normalization	Step	Output
	1	I wouldn't define [...] although I would like to defer raising taxes as long as prudently possible.
	2, 3	I would define [...] although I will like to defer raising taxes as long as prudently possible.
	4	I will like to defer raising taxes as long as prudently possible.
	5	Normalization 1: I will like to defer raising taxes as long as prudently possible. Normalization 2: I will defer raising taxes as long as prudently possible.
Interpretations	From	Potential Interpretation
	norm. 1	$\{\text{ARG}_0\}$ will like to defer raising taxes as long as prudently possible.
		I will like $\{$to $\text{ARG}_1\}$.
	norm. 2	$\{\text{ARG}_0\}$ will defer raising taxes as long as prudently possible.
		I will defer $\{\text{ARG}_1\}$ as long as prudently possible.
		I will defer raising taxes $\{\text{ARG}_2\}$.
		I will defer $\{\text{ARG}_1\}$ $\{\text{ARG}_2\}$.

Table 1: Step-by-step execution of the procedure to automatically normalize modal constructions (Sentences 1 and 2) and generate potential interpretations (Sentence 2).

oriented polarity-sensitive items by (Huddleston and Pullum, 2002, pp. 831).

3. Fix modal verbs and tense. If a modal verb (can, could, may, would, should, must, etc.) has as syntactic head *verb*, we transform the modal construction into past or future depending on the modal and tense of *verb*. For example: *could go: went*, *can go: will go*, *should have gone: went*. We use the same grammar rules and list of irregular verbs as in Step (2b).
4. Select relevant tokens. We remove all tokens in the original sentence except *verb* and tokens belonging to the roles in *sem_roles*. Additionally, we fix phrasal verbs by adding tokens with the part-of-speech tag RP whose syntactic head is *verb* and dependency type *prt* (semantic roles in OntoNotes are annotated for verb tokens, missing the preposition when *verb* is a phrasal verb would inadvertently change meaning). We also add all tokens to the left of *verb* until we find the first token whose part-of-speech tag does not start with VB, MD, RB or EX (verbs, modals, adverbs and existential *there*).
5. Generate additional normalizations. If *verb* is followed by TO + *verb*$_2$ (e.g., want to go, like to play, intend to pass), we generate an additional normalization for *verb*$_2$ after merging the semantic roles of *verb* and *verb*$_2$.

Table 1 exemplifies the automatic normalization step by step with 2 modal constructions.

Generating Potential Interpretations in Plain Text. Inspired by the rules Blanco and Sarabi (2016) used to generate interpretations from negation, we generate potential interpretations from modal constructions by toggling off combinations of roles in *sem_roles*. We consider numbered roles (ARG$_0$–ARG$_5$), and argument modifiers (ARGM-) ending in LOC, TMP, MNR, PRP, CAU, EXT, PRD or DIR.

Table 1 lists some potential interpretations generated from a sample modal construction. The total number of potential interpretations for the 324 selected modal construction is 1,756 (average: 5.4).

We recognize that our procedure to generate implicit interpretations is unable to generate some useful interpretations. For example, from *This is [a person who]*$_{\text{ARG}_1}$ *[likely]*$_{\text{ARGM-ADV}}$ *[died]*$_{verb}$ *[on impact versus perhaps freezing to death]*$_{\text{ARGM-MNR}}$, we generate *This is a person who died* {ARGM-MNR}, which is factual: the only uncertain information is the manner in which the person died. Since we toggle off semantic roles of *verb*, our procedure is unable to generate *A person died on impact* and *A person died freezing to death*; the former interpretation would receive a higher factuality score than the latter. We argue that automation is preferable, and reserve for future work generating interpretations that require splitting semantic roles.

4.2 Scoring Potential Interpretations

After automatically generating potential interpretations, we collected manual annotations to determine their factuality. The annotation interface showed the original sentence containing the modal construction, the previous and next sentences as context, and no additional information. Following previous work (Saurí and Pustejovsky, 2009; de Marneffe et al., 2012), we found it useful not to restrict answers to *yes* or *no*, but to allow for degrees of certainty. Specifically, we asked “Given the 3 sentences above, do you believe that the statement [potential interpretation] below is true?”. Answers are a score ranging from -5 to 5, where -5 indicates *Certainly no*, 5 indicates *Certainly yes*, and the scores in between indicate a continuum of certainty (0 indicates *unknown*).

After pilot annotations, we examined disagreements and defined the following simple guidelines:

1. Context (previous sentence, target sentence, and next sentence) is taken into account.
2. World knowledge available at the time the original sentence was authored—not new knowledge available after—is taken into account.
3. Semantic roles toggled off are replaced with a semantically related substitute (Turney and Pantel, 2010) for the original role, e.g., give: take, customer: sales associate.

5 Corpus Analysis

The total number of modal constructions selected is 324 and the number of potential interpretations automatically generated in 1,756 (average: 5.4 interpretation per modal construction). 39.4% of interpretations are scored with a high degree of certainty. We define *high certainty* as a score below -3 (interpretation is false) or larger than 3 (interpretation is true). Importantly, on overage, modal constructions have 2.13 interpretations scored with high certainty, and 1.23 scored 3 or higher. In other words, on average, our procedure generates over 2 interpretation that are either true or false, and over 1 interpretation that is true per modal construction.

# roles toggled off	#	%$\neq$ 0	Mean score	
			> 0	< 0
0	345	87.25	3.96	-3.94
1	800	48.50	3.67	-3.90
2	479	20.46	3.55	-4.03
3	120	5.83	3.50	-3.00

Table 2: Number of interpretations generated by toggling off 0, 1, 2 or 3 roles (#), percentage of interpretations not scored zero (% $\neq$ 0), and mean scores of interpretations with positive and negative scores.

Role	#	% $\neq$ 0	Mean score	
			> 0	< 0
None	345	87.25	3.96	-3.94
ARG$_1$	671	30.40	3.60	-3.92
ARG$_0$	604	25.50	3.72	-3.94
ARG$_2$	140	28.57	3.85	-3.84
ARGM-MNR	271	32.84	3.40	-3.85
ARGM-TMP	231	28.57	3.71	-3.84
ARGM-LOC	82	23.17	3.43	-4.60
Other	290	20.00	3.38	-3.87

Table 3: Number of interpretations generated by toggling off each semantic role (#), percentage of interpretations not scored zero (% $\neq$ 0), and mean score of interpretations with positive and negative scores.

Tables 2 and 3 present basic corpus statistics. The percentage of interpretations annotated with a score different than 0 depends greatly on the number of roles toggled off (Table 2): 0: 87.25%, 1: 48.50%, 2: 20.46%, 3: 5.83%. Note that the number of roles toggled off does not significantly affect the mean score of interpretations not scored 0 (Table 2, last 2 columns). Most interpretations have either ARG$_0$ or ARG$_1$ toggled off (Table 3), and the percentages of interpretations not scored zero range from 20% to 32.84% depending on the semantic role. Note that the average score of interpretations scored positively and negatively, however, does not depend on whether a semantic role is toggled off.

	Original sentence and sample of automatically generated potential interpretations	Score
1	Context, previous sentence: *The last thing we want to do is react to every wild statement that they make.* Original sentence: *[But]*$_{\text{ARGM-DIS}}$ *[they]*$_{\text{ARG}_0}$ *[certainly]*$_{\text{ARGM-ADV}}$ *[chose]*$_{\textit{verb}}$ *[that]*$_{\text{ARG}_1}$ *[to get our attention and that of the international community.]*$_{\text{ARGM-PRP}}$ Context, next sentence: *Uh but what they've got to realize is there is no magic bullet here.*	
	- Interpretation 1.1: But they chose that to get our attention and that of the international community.	5
	- Interpretation 1.2: But they chose $\{\text{ARG}_1\}$ to get our attention and that of the international community.	-5
2	Context, previous sentence: *Saddam Hussein (interrupting): Before you offer me your rotten goods, I ask you did you find weapons of mass destruction in Iraq or not?* Original sentence: *Rumsfeld (disconcerted): We haven't found them yet, but [we]*$_{\text{ARG}_0}$ *[will]*$_{\text{ARGM-MOD}}$ *[surely]*$_{\text{ARGM-ADV}}$ *[find]*$_{\textit{verb}}$ *[them]*$_{\text{ARG}_1}$ *[one day]*$_{\text{ARGM-TMP}}$. Context, next sentence: *Do you deny that you had intentions to manufacture a nuclear bomb?*	
	- Interpretation 2.1: We will find them one day.	4
	- Interpretation 2.2: We will find them $\{\text{ARGM-TMP}\}$.	-3
3	*"This is a rare case of [a company with a big majority holder which]*$_{\text{ARG}_0}$ *[will]*$_{\text{ARGM-MOD}}$ *[probably]*$_{\text{ARGM-ADV}}$ *[act]*$_{\textit{verb}}$ *[in the interests of the minority holders]*$_{\text{ARG}_1}$*", one investor says.*	
	- Interpretation 3.1: $\{\text{ARG}_0\}$ will act in the interests of the minority holders.	4
	- Interpretation 3.2: A company with a big majority holder will act $\{\text{ARG}_1\}$.	4
4	*I wouldn't define victory as simply not raising taxes—although [I]*$_{\text{ARG}_0,\, v_1,\, v_2}$ *[definitely]*$_{\text{ARGM-ADV},\, v_1}$ *[would]*$_{\text{ARGM-MOD},\, v_1}$ *[like]*$_{v_1}$ *[to [defer]*$_{v_2}$ *[raising taxes]*$_{\text{ARG}_1,\, v_2}$ *[as long as prudently possible.]*$_{\text{ARG}_2,\, v_2}$*]*$_{\text{ARG}_1,\, v_1}$	
	- Interpretation 4.1: I will like to defer raising taxes as long as prudently possible.	5
	- Interpretation 4.2: I will defer raising taxes as long as prudently possible.	1

Table 4: Annotation Examples. For each example, we show the original sentence containing the modal construction, context if helpful to determine scores, and 2 selected interpretations and their scores. Square brackets indicate semantic roles.

5.1 Annotation Quality

The annotation guidelines (Section 4.2) to score potential interpretations were defined after examining disagreements in pilot annotations. After defining the guidelines, inter-annotator agreement was 0.92 on 18% of randomly selected interpretations.[6] Agreement measures designed for categorical labels are unsuitable, as not all disagreements are equal, e.g., 4 vs. 5, -2 vs. 5. Because of the high agreement and following previous work (Agirre et al., 2012), the rest of interpretations were annotated once.

5.2 Annotation Examples

Table 4 presents annotation examples. For each example, we include the original sentence containing a selected modal construction, its context (previous and next sentence) if helpful for scoring, and 2 automatically generated potential interpretations with their annotated scores.

Example (1) shows that context helps in determining the factuality of potential interpretations (item (1) in the guidelines). After reading the three sentences, it is clear that *they* are making *wild statements*, and are hoping to get *attention* for it. Interpretation 1.1 removes adverb *certainly* and receives the highest score, 5. Interpretation 1.2 is obtained after toggling off ARG_1, and receives the lowest score, -5. This low score is justified by item (3) in our annotation guidelines: replacing *wild statements* with a semantically (different but) related substitute, e.g., *But they chose reasonable statements / good manners to get our attention and that of the international community*, yields an unlikely interpretation.

The interpretations in Example (2) show again the importance of context, and also exemplify item (2) in the annotation guidelines. Interpretation 2.1, *We will find them one day* receives a high score ($4/5$), as given the context (and assuming that Rumsfeld is truthful), it is very likely that they will find the weapons of mass destruction, but it is not guaranteed. Note that annotators are not allowed to use the fact that the weapons were never found (item (2) in the guidelines). In Interpretation 2.2, *one day* could be replaced with *never / at no time* or similar constructions, and doing so yields the opposite of the intended meaning (score: -3). A possible descrip-

[6] We set an internal deadline of 3 days after agreeing on the guidelines, and we could annotate 18% of instances in that time.

Type	Feature	Description
baseline	adverb	Word form of adverb
	adverb_pos	Part-of-speech of adverb
	verb	Word form of verb
	verb_pos	Part-of-speech of verb
	distance	Number of tokens between adverb and verb
	direction	Whether adverb occurs before or after verb
adverb and verb	adverb_rel_pos	Part-of-speech tags of the parent, and left and right siblings of adverb
	adverb_subcat	Concatenation of part-of-speech tags of all siblings of adverb
	verb_rel_pos	Part-of-speech tags of the parent, and left and right siblings of verb
	adverb_subcat	Concatenation of part-of-speech tags of all siblings of verb
	path, path_l	Syntactic path between adverb and verb, and length of the path
	ancestor	POS tag of the lowest common ancestor between verb and adverb
	has_sem_role	Flags indicating whether a semantic role is in the modal construction
interpretation	num_roles_int	Number of roles toggled off in the potential interpretation
	sem_roles_int	Flags indicating which roles are toggled off in the interpretation
	roles_distance	Number of tokens between each semantic role and verb
	roles_direction	Whether each semantic role occurs before or after verb
	roles_path	Syntactic path between each role and verb
	roles_path_l	Length of syntactic path between each role and verb

Table 5: Features used to predict factuality scores to automatically generated potential interpretations. Features extracted from semantic role are extracted for ARG_0–ARG_5 and modifiers (ARGM-) ending in LOC, TMP, MNR, PRP, CAU, EXT and PRD.

tion of these scores could be "almost certainly true" (4 out of 5), and "most probably false" (-3 out of -5). We see scores as a continuum of certainty, but textual description may help understand the examples.

Example (3) demonstrates the usefulness of the normalization process—specifically, Step 4, selecting relevant tokens—and the importance of replacing roles with semantically related substitutes (item (3) in the guidelines). In interpretation 3.1, $\{\text{ARG}_0\}$ *will act in the interests of the minority holders*, ARG_0 can be replaced with *a company with several minority holders*, yielding a valid interpretation scored 4 (out of 5). Similarly, in interpretation 3.2, *A company with a big majority holder will act* $\{\text{ARG}_1\}$, ARG_1 can be replaced with *in the interests of the big majority holder*, yielding another valid interpretation also scored 4 (out of 5).

Finally, Example (4) shows Step 5 in the automatic normalization procedure (Section 4). By creating an additional verb-argument structure, we are able to differentiate between liking to do something (Interpretation 4.1, score 5/5) and actually doing that something (Interpretation 4.2, score 1/5).

6 Learning to Score Potential Interpretations

In order to automatically score potential interpretations, we follow a standard supervised machine learning approach. Each potential interpretation becomes an instance, and we split modal constructions (and their potential interpretations) into training (80%) and test (20%). When splitting, we make sure that the amount of modal constructions for each adverb in each split is proportional, i.e., 80% of modal constructions with each adverb are in the train split and the rest in the test split. Splitting instances randomly would assign interpretations generated from the same modal construction to the train and test splits, and bias the results.

We trained a Support Vector Machine (SVM) for regression with RBF kernel using scikit-learn (Pedregosa et al., 2011), which uses LIBSVM (Chang and Lin, 2011). The SVM parameters (C and γ) were tuned using 10-fold cross-validation with the training set, and we report results using the test split.

Features	Pearson
baseline	-0.029
adverb and verb	0.025
interpretation	0.494
baseline + adverb and verb	-0.013
baseline + interpretation	0.463
adverb and verb + interpretation	0.465
baseline + adverb and verb + interpretation	0.468

Table 6: Pearson correlations obtained with test instances and several feature combinations.

6.1 Feature Selection

The full set of features is detailed in Table 5. *Baseline* features are simple features characterizing *adverb* and *verb* and we do not elaborate on them. *Adverb and verb* features are extracted from the modal construction (constituent tree and semantic roles) and provide additional information about the modal construction. *Interpretation* features characterize the potential interpretation whose factuality is being scored, and are also derived from the constituent tree and semantic roles.

Most *adverb and verb* features are standard in semantic role labeling (Gildea and Jurafsky, 2002). We include the part-of-speech tags of the parent, and left and right siblings of *adverb* and *verb*, as well as their subcategorization, i.e., the concatenation of the sibling's part-of-speech tags. We also include syntactic path between *adverb* and *verb*, and its length. Additionally, we include the common ancestor, i.e., the syntactic node of the lowest common node that is an ancestor of both *adverb* and *verb*, and use binary features to indicate whether each semantic role is present in the modal construction.

Finally, *interpretation* features characterize the semantic roles toggled off to generate the potential interpretation. We include the number of roles toggled off to generate the potential interpretation, and binary flags indicating which roles. Additionally, for each role toggled off, we include the distance from the verb (number of tokens), whether it occurs before or after the verb, the syntactic path to the verb and the length of the path.

7 Experimental Results

Table 6 details results obtained with test instances using several feature combinations derived from gold linguistic information (POS tags, parse trees, semantic roles, etc.). *Baseline* and *adverb and verb* features, which characterize the modal construction from which potential interpretation are extracted, are virtually useless. They yield Pearson correlations of -0.029 and 0.025 individually, and -0.013 combined. These results suggest that the verb and adverb in the modal construction (word forms, syntactic paths, etc.) are insufficient to rank potential interpretations generated from the modal construction.

Interpretation features, which capture differences between potential interpretations being scored (number of roles toggled off, roles toggled off, etc.), obtain a modest Pearson correlation of 0.494. Combining *interpretation* features with other features proved detrimental, Pearson correlations are between 0.463 and 0.468.

8 Conclusions

Modality is a pervasive phenomenon used to talk about what is not factual. In this paper, we have presented a methodology to extract implicit interpretations from modal constructions. First, we automatically generate potential interpretations using syntactic dependencies and semantic roles, and then assign to them a factuality score.

The most important conclusion of the work presented here is that several interpretations automatically generated from a single modal construction often receive scores indicating high certainty. Indeed, on average, modal constructions have 2.13 interpretations scored lower or equal than -3, or higher or equal than 3. This contrast with previous work, which only assess factuality of one normalization per proposition.

Experimental results using supervised machine learning and relatively simple features show that the task is challenging but can be automated. We believe better results could be obtained by incorporating features capturing knowledge in the context of the modal construction, including other clauses in the same sentence, and the previous and next sentences. Another extension of the current work is to investigate a similar approach for other modality markers such as nouns (e.g., possibility, chance), adjectives (e.g.necessary, probable,) and certain verbs (e.g., claim, suggests).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of Modality and Negation in Semantically-Informed Syntactic MT. *Comput. Linguist.*, 38(2):411–438, June.

Eduardo Blanco and Zahra Sarabi. 2016. Automatic generation and scoring of positive interpretations from negated statements. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1431–1441, San Diego, California, June. Association for Computational Linguistics.

Marta Carretero and Juan Rafael Zamorano-Mansilla. 2013. An analysis of disagreement-provoking factors in the analysis of epistemic modality and evidentiality: the case of english adverbials. In *Proceedings of IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 16–23, Potsdam, Germany, March.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May.

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

Anita de Waard and Henk Pander Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, ACL '12, pages 47–55, Stroudsburg, PA, USA.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73, Suntec, Singapore, August.

Kai Von Fintel. 2006. Modality and language. In D. Borchert, editor, *Encyclopedia of Philosophy*, pages 20–27. Macmillan Reference.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, September.

Valentine Hacquard. 2011. Modality. In C. Maienborn, K. von Heusinger, and P. Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 1484–1515. Mouton de Gruyter.

Kees Hengeveld and J. Lachlan Mackenzie. 2008. *Functional Discourse Grammar: A Typologically-Based Theory of Language Structure*. Oxford University Press.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% Solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60, Morristown, NJ, USA.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, April.

Otto Jespersen. 1992. *The philosophy of grammar*. University of Chicago Press, Chicago.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal, September.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA, May 6.

John Lyons. 1977. *Semantics*. Cambridge University Press. Cambridge Books Online.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 41–48, Stroudsburg, PA, USA.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

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

Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Hitoshi Isahara, and Qing Ma. 2005. Correction of errors in a verb modality corpus for machine translation with a machine-learning method. 4(1):18–37, March.

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

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

F. R. Palmer. 2001. *Mood and Modality*. Cambridge University Press, second edition. Cambridge Books Online.

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.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado, June.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June.

P. Quaresma, A. Mendes, I. Hendrickx, and T. Gon?alves. 2014. Automatic tagging of modality: identifying triggers and modal value. In *Proceedings of the 10th Joint ACL SIGSEM - ISO Workshop on Interoperable Semantic Annotation*.

Victoria L. Rubin. 2006. *Identifying certainty in texts*. Ph.D. thesis, Syracuse University, Syracuse, NY.

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

Liliana Mamani Sánchez and Carl Vogel. 2015. A hedging annotation scheme focused on epistemic phrases for informal language. In *Proceedings of the Workshop on Models for Modality Annotation*.

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

Rion Snow, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 33–40, Stroudsburg, PA, USA.

Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland, June.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Veronika Vincze, György Szarvas, Richard Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9((Suppl 11)):S9.

Veronika Vincze, György Szarvas, György Móra, Tomoko Ohta, and Richárd Farkas. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the bioscope and genia event corpora. *Journal of Biomedical Semantics*, 2(5):1–11.

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