Veridicality and utterance understanding

Marie-Catherine de Marneffe, Christopher D. Manning and Christopher Potts
Linguistics Department
Stanford University
Stanford, CA 94305
{mcdm,manning,cgpotts}@stanford.edu

***Abstract*—Natural language understanding depends heavily on assessing *veridicality* – whether the speaker intends to convey that events mentioned are actual, non-actual, or uncertain. However, this property is little used in relation and event extraction systems, and the work that has been done has generally assumed that it can be captured by lexical semantic properties. Here, we show that context and world knowledge play a significant role in shaping veridicality. We extend the FactBank corpus, which contains semantically driven veridicality annotations, with pragmatically informed ones. Our annotations are more complex than the lexical assumption predicts but systematic enough to be included in computational work on textual understanding. They also indicate that veridicality judgments are not always categorical, and should therefore be modeled as distributions. We build a classifier to automatically assign event veridicality distributions based on our new annotations. The classifier relies not only on lexical features like hedges or negations, but also structural features and approximations of world knowledge, thereby providing a nuanced picture of the diverse factors that shape veridicality.**

I. Introduction

Utterance understanding depends heavily on assessing whether the speaker intends to convey that the events described are actual, non-actual, or uncertain. A declarative like "The cancer has spread" conveys firm speaker commitment, whereas qualified variants such as "There are strong indicators that the cancer has spread" or "The cancer might have spread" imbue the claim with uncertainty. These *event veridicality* judgments, and the factors that guide them, have received considerable attention in logic, linguistics, and NLP [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13].

The FactBank corpus [14] provides an extensive set of veridicality annotations, which capture the ways in which lexical meanings and local semantic interactions determine veridicality. In this paper, we extend the FactBank annotations by bringing context and world knowledge into the picture. We had a large group of linguistically naïve annotators re-annotate a portion of FactBank using only their commonsense intuitions. Together with the original FactBank annotations, the results paint a nuanced picture of the ways in which semantics and pragmatics interact to shape readers' veridicality judgments.

Our annotations also show that some uncertainty in the judgments must be captured. The newspaper says, *United Widget said that its profits were up in the fourth quarter*, but just how trustworthy is United Widget on such matters? Speakers are likely to vary in what they intend in such cases, and hearers (readers) are thus forced to operate under uncertainty when making the requisite inferences. The inherent uncertainty of pragmatic inference suggests to us that veridicality judgments are not always categorical, and thus are better modeled as distributions. We therefore treat veridicality as a distribution-prediction task. We trained a maximum entropy classifier on our annotations to assign event veridicality distributions. Our features include not only lexical items like hedges, modals, and negations, but also structural features and approximations of world knowledge. The resulting model yields insights into the complex pragmatic factors that shape readers' veridicality judgments.

II. Corpus annotation

FactBank provides veridicality annotations on 9,472 events from 208 newswire documents, relative to multiple discourse participants. The tags (summarized in Table I) annotate ⟨event, participant⟩ pairs at the sentence level. Each tag consists of a veridicality value (certain, probable, possible, underspecified) and a polarity value (positive, negative, unknown). The participant can be anyone mentioned in the sentence as well as its author. In (1), for example, the event described by *means* is assigned two values, one relative to the source *experts* and the other relative to the author of the sentence.

(1) Some experts now predict Anheuser's entry into the fray **means** near-term earnings trouble for all the industry players.
Veridicality(means, experts) = PR+
Veridicality(means, author) = Uu

Discriminatory tests [6, 230-235], informed by lexical theories [15], [16], were used to guide the annotation, which was "textual-based, that is, reflecting only what is expressed in the text and avoiding any judgment based on individual knowledge" [14, 253]. The level of inter-annotator agreement was high (κ = 0.81, a conservative figure given the partial ordering in the tags).

How do readers assess veridicality when encouraged to offer judgments about their *utterance* understanding? To gain an empirical foundation in this area, we collected ten annotations each for 642 events from the FactBank training set annotated at the author level (all the PR+, PS+, PR-, PS- items plus some randomly chosen Uu, CT+ and CT- items). Subjects were recruited via Mechanical Turk, given a brief training session to help them conceptualize the tags properly, and then asked to decide whether the bold-faced event described in the sentence did (or will) happen. 177 workers participated. As expected

978-0-7695-4492-2/11 $26.00 © 2011 IEEE
DOI 10.1109/ICSC.2011.10

TABLE I
FactBank annotation scheme

Value	Definition	Count
CT+	According to the source, it is certainly the case that X	7,749 (57.6%)
PR+	According to the source, it is probably the case that X	363 (2.7%)
PS+	According to the source, it is possibly the case that X	226 (1.7%)
CT-	According to the source, it is certainly not the case that X	433 (3.2%)
PR-	According to the source it is probably not the case that X	56 (0.4%)
PS-	According to the source it is possibly not the case that X	14 (0.1%)
CTu	The source knows whether it is the case that X or that not X	12 (0.1%)
Uu	The source does not know what the factual status of the event is, or does not commit to it	4,607 (34.2%)
		13,460

TABLE II
Inter-annotator agreement comparing FactBank annotations with MTurk annotations

	κ	p **value**
CT+	0.37	< 0.001
PR+	0.79	< 0.001
PS+	0.86	< 0.001
CT-	0.91	< 0.001
PR-	0.77	< 0.001
PS-	-0.001	$= 0.982$
Uu	0.06	$= 0.203$
Overall	0.60	< 0.001

TABLE III
Confusion matrix comparing the FactBank annotations (rows) with our annotations (columns)

Fact-				MTurk				
Bank	CT+	PR+	PS+	CT-	PR-	PS-	Uu	Total
CT+	54	2	0	0	0	0	0	56
PR+	4	63	2	0	0	0	0	69
PS+	1	1	55	0	0	0	2	59
CT-	5	0	0	146	0	0	2	153
PR-	0	0	0	0	5	0	1	6
PS-	0	0	0	0	0	0	1	1
Uu	94	18	9	12	2	0	21	156
Total	158	84	66	158	7	0	27	500

given the more unconstrained setting, the level of agreement was lower than for FactBank (Fleiss κ = 0.53; again, this is conservative because it is insensitive to the ordering of the tags), but most of the disputes were about degree (e.g., CT+ vs. PR+) rather than about the basic veridicality judgment. At least 6 out of 10 workers agreed on the same tag for 500 of the 642 sentences (78%). For 53% of the examples, at least 8 Turkers agreed with each other, and total agreement is obtained for 26% of the data (165 sentences).[1]

The new annotations largely agree with those of FactBank, but there are some systematic differences. As a first step towards understanding these differences, we first restrict attention to the 500 examples for which at least six out of ten Turkers agreed on the annotation.

Table II assesses the agreement between FactBank and the MTurk annotations on this 500-sentence subset of the data in which at least 6 of the 10 Turkers agreed on the label. We treat FactBank as one annotator and our collective Turkers as a second, with the majority label the correct one for that annotator. Our goal here is not to rigorously assess agreement, but rather only to probe for similarities and differences. The very poor value for PS- derives from the fact that, in this subset, that label was chosen only once in FactBank and not at all by our annotators.

There is modest to very high agreement for all the categories except Uu and CT+. The confusion matrix in Table III explicates these numbers. In FactBank, the Uu category is used much more often than it is used by the Turkers, and the dominant alternative choice for the Turkers is CT+. Thus, the low score for Uu also effectively drops the score for CT+. But why do Turkers go for CT+ where FactBank says Uu?

The divergence traces to the way in which lexicalist theories handle events embedded under attitude predicates like *say*, *report*, and *indicate*: any such embedded event is tagged Uu in FactBank. In contrast, in our annotations, readers do not view the veridicality of reported events as unknown. Instead, they are sensitive to markers in the embedded clause, the embedding verb, and subject. For example, even though the events in (1) and (2) are all embedded under some attitude predicate (*say*, *predict*), the events in (2a) and (2b) are assessed as certain (CT+), whereas the words *highly confident* in (2c) trigger PR+, and *may* in (2d) mainly leads to PS+.

(2) a. Magna International Inc.'s chief financial officer **resigned**, the company said. [CT+: 10]

b. In the air, U.S. Air Force fliers say they have **engaged** in "a little cat and mouse" with Iraqi warplanes. [CT+: 9, PS+: 1]

c. Merieux officials said that they are "highly confident" the offer will be **approved**. [PR+: 10]

d. U.S. commanders said 5,500 Iraqi prisoners were taken in the first hours of the ground war, though some military officials later said the total may have **climbed** above 8,000. [PS+: 7, PR+: 3]

The FactBank and MTurk annotators also treat "possible" and "probable" differently. In FactBank, markers of possibility

[1] Our annotations are available at http://christopherpotts.net/ling/data/factbank/.

or probability (*could*, *likely*) uniquely determines the corresponding tag [6, 233]. In contrast, the Turkers allow the bias created by these lexical items to be swayed by other factors; as seen in (3), *could* can appear in sentences marked possible, probable, or unknown. (In FactBank all these sentences are marked PS+.)

(3) a. They aren't being allowed to leave and could **become** hostages. [PS+: 10]

b. Iraq could start **hostilities** with Israel either through a direct attack or by attacking Jordan. [Uu: 6, PS+: 3, PR+: 1]

Similarly, in both sets of annotations, *expected* and *appeared* are often markers of PR events. However, in FactBank, the association is very tight, whereas our annotations often show a move to PS, as in (4) and (5).

(4) a. Big personal computer makers are developing 486-based machines, which are expected to **reach** the market early next year. [PR+: 10]

b. Beneath the tepid news-release jargon lies a powerful threat from the brewing giant, which last year accounted for about 41% of all U.S. beer sales and is expected to see that **grow** to 42.5% in the current year. [PS+: 6, PR+: 3, CT+: 1]

(5) a. Despite the lack of any obvious successors, the Iraqi leader's internal power base appeared to be **narrowing** even before the war began. [PR+: 7, CT+: 1, PS+: 1, PS-: 1]

b. Saddam appeared to **accept** a border demarcation treaty he had rejected in peace talks following the August 1988 cease-fire of the eight-year war with Iran. [PS+: 6, PR+: 2, CT+: 2]

For the purposes of comparing our annotations with those of FactBank, it is useful to single out the Turkers' majority-choice category, as we did above. However, we have 10 annotations for each event, which invites exploration of the full distribution of annotations, to see if the areas of stability and variation can teach us something about the nature of speakers' veridicality judgments.

Figure 1 provides a high-level summary of the reaction distributions that our sentences received. The labels on the y-axis characterize types of distribution. For example, '5/5' groups the sentences for which the annotators were evenly split between two categories (e.g., a sentence for which 5 Turkers assigned PR+ and 5 assigned PS+, or a sentence for which 5 Turkers chose PR+ and 5 chose Uu). The largest grouping, '10', pools the examples on which all the annotators were in agreement.

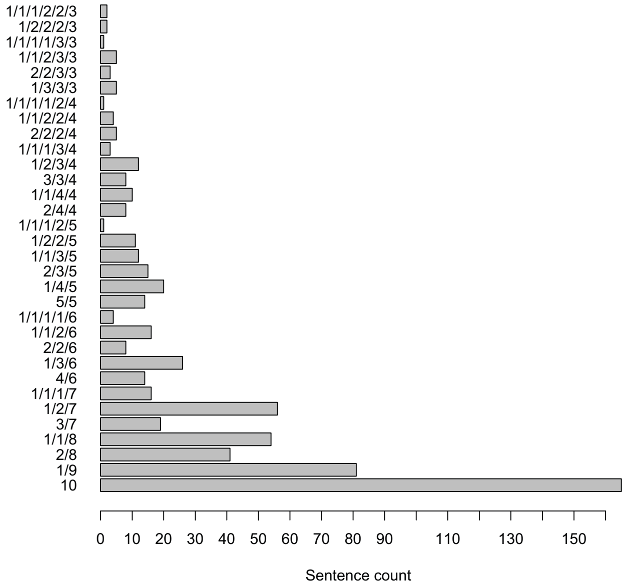


Fig. 1. Reaction distributions by type

We can safely assume that some of the variation seen in Figure 1 is due to the noisiness of the crowd-sourced annotation process. Some annotators might have been inattentive or confused, or simply lacked the expertise to make these judgments [17]. For example, the well-represented '1/9' and '1/1/8' groups probably represent examples for which veridicality assessment is straightforward but one or two of the annotators did not do a good job. If all the distributions were this skewed, we might feel secure in treating veridicality as categorical. However, there are many items for which it seems implausible to say that the variation is due to noise. For example, the '5/5' group includes sentences like (6) and (7), for which the judgments depend heavily on one's prior assumptions about the entities and concepts involved. This suggests to us that work on veridicality should embrace variation and uncertainty as being part of the characterization of veridicality, rather than trying to approximate the problem as one of basic categorization.

(6) In a statement, the White House said it would do "whatever is necessary" to ensure **compliance** with the sanctions. [Uu: 5, PR+: 5]

(7) Diplomacy appears to be making headway in **resolving** the United Nations' standoff with Iraq. [PR+: 5, PS+: 5]

III. System

To automatically assign veridicality, we built a maximum entropy classifier [18]. In all our experiments, we use the Stanford Classifier [19] with Gaussian prior $\mathcal{N}(0,1)$. For classification tasks, the dominant tradition within computational linguistics has been to adjudicate differing human judgments and to assign a single class for each item in the training data. However, in section II, we saw evidence in our annotations that veridicality is not necessarily categorical, in virtue of the uncertainty involved in making pragmatic judgments of this sort. In order to align with our theoretical conception of the problem as probabilistic, we treat each annotator judgment as a training item. Thus, each of the 642 sentences appears 10 times in our training data, with the label assigned by one Turker.

As test set, we used 130 sentences from the test items in FactBank, also annotated using the same Mechanical Turk prompt. For 112 of the 130 sentences, at least six Turkers agreed on the same value. The features were selected through 10-fold cross-validation on the training set.

a) Predicate classes: Saurí [6] defines classes of predicates that project the same veridicality value onto the events they introduce. The classes (e.g., ANNOUNCE, CONFIRM, SAY) also define the grammatical relations that need to hold between the predicate and the event it introduces. Like Saurí, we used dependency graphs produced by the Stanford parser [20], [21] to follow the path from the target event to the root of the sentence. If a predicate in the path was contained in one of the classes and the grammatical relation matched, we added features for the lemma of the predicate, its predicate class, and whether the lemma was negated.

b) World knowledge: For each verb found in the path and contained in the predicate classes, we also added the lemma of its subject. Our rationale for including the subject is that readers' interpretations differ for sentences such as "The FBI said it **received** ..." and "Bush said he **received** ...", presumably because of world knowledge they bring to bear on the judgment. To approximate such world knowledge, we obtained subject–verb bigram and subject counts from the New York Times portion of GigaWord and then included log(subject–verb counts/subject counts) as a feature. The intuition is that some of the embedded clauses carry the main point of the sentence and the subject–verb pairs then serve as evidential modifiers of the embedded clause [22], with the overall frequency of such modifiers contributing to readers' veridicality assessments.

c) General features: We used the event lemma, the lemma of the sentence root, the incoming grammatical relation to the event, and a general class feature.

d) Negation: This feature captures the presence of negative contexts. Events are considered negated if they have a negation dependency in the graph or an explicit marker of negation as dependent (e.g., negation (*not*, *never*), downward-monotone quantifiers (*no*, *any*), or restricting prepositions). Events are also considered negated if embedded into a negative context (e.g., *fail*, *cancel*).

e) Modality: We used Saurí's list of modal words. We distinguished between markers found as direct governors or children of the event under consideration, and modal words elsewhere in the sentence.

f) Conditional: Antecedents of conditionals and words clearly marking uncertainty are reliable features of the Uu category. We marked events in *if*-clauses or embedded under markers such as *call for*.

TABLE IV
LOG LIKELIHOOD VALUES FOR THE TRAINING AND TEST DATA

	Train	**Test**
lower-bound	−10813.97	−1987.86
classifier	−8021.85	−1324.41
upper-bound	−3776.30	−590.75

g) Quotation: Another reliable feature of Uu. We generated such feature if the sentence opens and ends with quotes, or if the root subject is *we*.

IV. RESULTS

Table IV gives log-likelihood values of the classifier for the training and test sets, along with the upper and lower bounds. The upper bound is the log-likelihood of the model that uses the exact distribution from the Turkers. The lower bound is the log-likelihood of a model that uses only the overall rate of each class in our annotations for the training data.

KL divergence provides a related way to assess the effectiveness of the classifier. The KL divergence of one distribution from another is an asymmetric measure of the difference between them. We use example (2d) to illustrate. For that sentence, the classifier assigns a probability of 0.64 to PS+ and 0.28 to PR+, with very low probabilities for the remaining categories. It thus closely models the gold distribution [PS+: 7, PR+: 3]. The KL-divergence is correspondingly low: 0.13. The KL-divergence for a classifier that assigned 0.94 probability to the most frequent category (i.e., CT+) and 0.01 to the remaining categories would be much higher: 5.76.

The mean KL divergence of our model is 0.95 (SD 1.13) for the training data and 0.81 (SD 0.91) for the test data. The mean KL divergence for the baseline model is 1.58 (SD 0.57) for the training data and 1.55 (SD 0.47) for the test data. To assess whether our classifier is a statistically significant improvement over the baseline, we use a paired two-sided t-test over the KL divergence values for the two models. The t-test requires that both vectors of values in the comparison have normal distributions. This is not true of the raw KL values, which have approximately gamma distributions, but it is basically true of the log of the KL values: for the model's KL divergences, the normality assumption is very good, whereas for the baseline model there is some positive skew. Nonetheless, the t-test arguably provides a fair way to contextualize and compare the KL values of the two models. By this test, our model improves significantly over the lower bound (two-sided t = -11.1983, df = 129, p-value $< 2.2e{-}16$).

We can also compute precision and recall for the subsets of the data where there is a majority vote, i.e., where six out of ten annotators agreed on the same label. This allows us to give results per veridicality tag. We take as the true veridicality value the one on which the annotators agreed. The value assigned by the classifier is the one with the highest probability. Table V reports precision, recall, and F1 scores on the training (10-fold cross-validation) and test sets, along with the number of instances in each category. PR- and PS-

TABLE V
PRECISION, RECALL AND F1 ON THE SUBSETS OF THE DATA WHERE THERE IS MAJORITY VOTE

	Train				Test			
	#	**P**	**R**	**F1**	**#**	**P**	**R**	**F1**
CT+	158	74.3	84.2	78.9	61	86.9	86.9	86.9
CT-	158	89.4	91.1	90.2	31	96.6	90.3	93.3
PR+	84	74.4	69.1	71.6	7	50.0	57.1	53.3
PS+	66	75.4	69.7	72.4	7	62.5	71.4	66.7
Uu	27	57.1	44.4	50.0	6	50.0	50.0	50.0
Macro-avg		74.1	71.7	72.6		69.2	71.1	70.0
Micro-avg		78.6	78.6	78.6		83.0	83.0	83.0

items did not appear in the test data and were very infrequent in the training data, so we left them out.

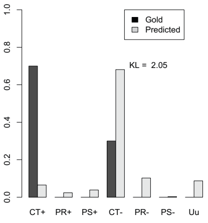
The classifier weights give insight into the interpretation of lexical markers. Some markers behave as linguistic theories predict. For example, *believe* is often a marker of probability whereas *could* and *may* are more likely to indicate possibility. But as seen in 3, world knowledge and other linguistic factors shape the veridicality of these items. The greatest departure from theoretical predictions occurs with the SAY category, which is logically non-veridical but correlates highly with certainty (CT+) in our corpus. Conversely, the class KNOW, which includes *know*, *acknowledge*, *learn*, is a marker of possibility rather than certainty. Our model thus shows that to account for how readers interpret sentences, the space of veridicality should be cut up differently than the lexicalist theories propose.

V. ANALYSIS AND DISCUSSION

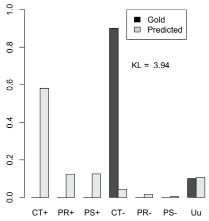
We focus on two kinds of errors. First, where there is a majority label — a label six or more of the annotators agreed on — in the annotations, we can compare that label with the one assigned the highest probability according to our model. Second, we can study cases where the the annotation distribution diverges considerably from our model's distribution (i.e., cases with a very high KL-divergence).

For the majority-label cases, errors of polarity are extremely rare; the classifier wrongly assesses the polarity of only four events, shown in (8). Most of the errors are thus in the degree of confidence (e.g., CT+ vs. PR+). The graphs next to the examples compare the gold annotation from the Turkers (the black bars) with the distribution proposed by the classifier (the gray bars). The KL divergence value is included to help convey how such values relate to these distributions.

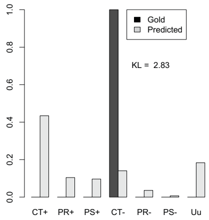
(8) a. Addressing a NATO flag-lowering ceremony at the Dutch embassy, Orban said the occasion indicated the end of the embassy's **mission** of liaison between Hungary and NATO. [CT+:7, CT-: 3]



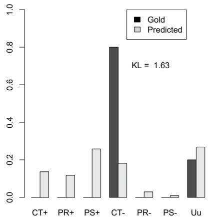
b. But never before has NATO **reached** out to its former Eastern-bloc enemies. [CT-: 9, Uu: 1]



c. Horsley **was** not a defendant in the suit, in which the Portland, Ore., jury ruled that such sites constitute threats to abortion providers. [CT-: 10]



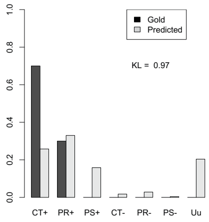
d. A total of $650,000, meanwhile, is being offered for information leading to the **arrest** of Kopp, who is charged with gunning down Dr. Barnett Slepian last fall in his home in Buffalo. [CT-: 8, Uu: 2]



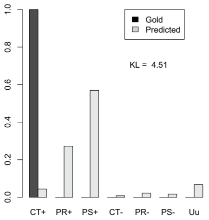
When the system missed CT- events, it failed to find an explicit negative marker, as in (8b), where (due to a parse error) *never* is treated as a dependent of the verb *have* and not of the *reaching out* event. Similarly, the system could not capture instances in which the negation was merely implicit, as in (8d), where the non-veridicality of the arresting event requires deeper interpretation that our feature-set can manage.

In (9), we give examples of CT+ events that are incorrectly tagged PR+, PS+, or Uu by the system because of the presence of a weak modal auxiliary or a verb that lowers certainty, such as *believe*. As we saw in section II, these markers correlates strongly with the PS categories.

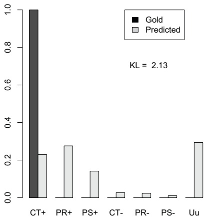
(9) a. The NATO summit, she said, would produce an **initiative** that "responds to the grave threat posed by weapons of mass destruction and their means of delivery." [CT+: 7, PR+: 3]



b. Kopp, meanwhile, may have approached the border with Mexico, but it is **unknown** whether he crossed into that country, said Freeh. [CT+: 10]

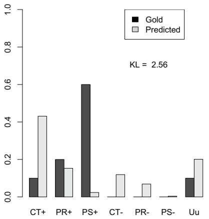


c. They believe Kopp was driven to Mexico by a female friend after the **shooting**, and have a trail of her credit card receipts leading to Mexico, the federal officials have said. [CT+: 10]

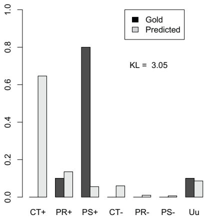


In the case of PR+ and PS+ events, all the erroneous values assigned by the system are CT+. Some explicit modality markers were not seen in the training data, such as *potential* in (10a), and thus the classifier assigned them no weight. In other cases, such as (10b), the system did not capture the modality implicit in the conditional.

(10) a. Albright also used her speech to articulate a forward-looking vision for NATO, and to defend NATO's potential **involvement** in Kosovo. [PS+: 6, PR+: 2, CT+: 1, Uu: 1]

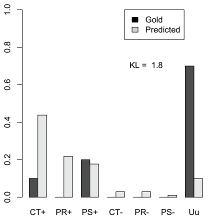


b. "And we must be resolute in spelling out the consequences of intransigence," she added, referring to the threat of NATO air **strikes** against Milosevic if he does not agree to the deployment. [PS+: 8, PR+: 1, Uu: 1]

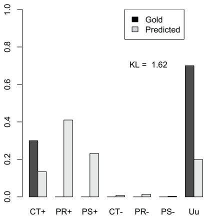


The only Uu events that the system correctly retrieved were antecedents of a conditional. For the other Uu events in (11), the system assigned CT+ or PR+. The majority of Uu events proved to be very difficult to detect automatically since complex pragmatic factors are at work, many of them only very indirectly reflected in the texts.

(11) a. Kopp's stepmother, who married Kopp's father when Kopp was in his 30s, said Thursday from her home in Irving, Texas: "I would like to see him come forward and clear his name if he's not guilty, and if he's guilty, to contact a priest and make his amends with society, face what he **did**." [Uu: 7, PS+: 2, CT+: 1]

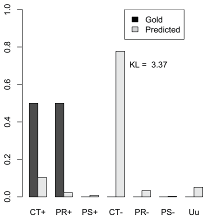


b. Indeed, one particularly virulent anti-abortion Web site lists the names of doctors it says perform abortions, or "crimes against humanity," with a code indicating whether they are "**working**," "wounded" or a "fatality." [Uu:7, CT+: 3]

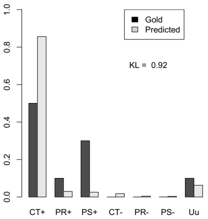


It is also instructive to look at the examples for which there is a large KL-divergence between our model's predicted distribution and the annotation distribution. Very often, this is simply the result of a divergence between the predicted and actual majority label, as discussed above. However, examples like (12) are more interesting in this regard: these are cases where there was no majority label, as in (12a), or where the model guessed the correct majority label but failed to capture other aspects of the distribution, as in (12b) and (12c).

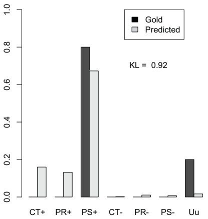
(12) a. On Tuesday, the National Abortion and Reproductive Rights Action League plans to hold a news **conference** to screen a television advertisement made last week, before Slepian died, featuring Emily Lyons, a nurse who was badly wounded earlier this year in the bombing of an abortion clinic in Alabama. [CT+: 5, PR+: 5]



b. Vacco's campaign manager, Matt Behrmann, said in a statement that Spitzer had "sunk to a new and despicable low by **attempting** to capitalize on the murder of a physician in order to garner votes." [CT+: 5, PR+: 1, PS+: 3, Uu: 1]

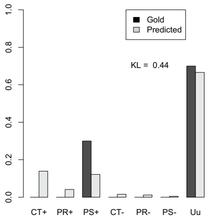


c. Since there is no federal homicide statute as such, the federal officials said Kopp could be **charged** under the recent Freedom of Access to Clinic Entrances Act, which provides for a sentence of up to life imprisonment for someone convicted of physical assaults or threats against abortion providers. [PS+: 8, Uu: 2]



In (12a), the classifier is confused by an ambiguity: it treats *hold* as a kind of negation, which leads the system to assign a 0.78 probability to CT-. In (12b), there are no features indicating possibility, but a number of SAY-related features are present, which leads to a very strong bias for CT+ (0.86) and a corresponding failure to model the rest of the distribution properly. In (12c), the classifier correctly assigns most probability to PS+, but the rest of the probability mass is distributed between CT+ and PR+. This is another manifestation of the problem, noted above, that we have very few strong indicators of Uu. The exception to that is conditional antecedents. As a result, we do well with cases like (13), where the event is in a conditional; the classifier assigns 70% of the probability to Uu and 0.15 to PS+.

(13) On Monday, Spitzer called for Vacco to revive that unit immediately, vowing that he would do so on his first day in office if **elected**. [Uu: 7, PS+: 3]



Overall the system assigns incorrect veridicality distributions in part because it misses explicit linguistic markers of veridicality, but also because contextual and pragmatic factors cannot be captured. This is instructive, though, and serves to further support our central thesis that veridicality judgments are not purely lexical, but rather involve complex pragmatic reasoning.

VI. Conclusion

We focused on the nature of event veridicality assessment. To do this, we extended FactBank [14] with veridicality annotations that are informed by context and world knowledge. While the two sets of annotations are similar in many ways, their differences highlight areas in which pragmatic factors play a leading role in shaping speakers' judgments. In addition, because each one of our sentences was judged by ten annotators, we actually have annotation *distributions* for our sentences, which allow us to identify areas of uncertainty in veridicality assessment. This uncertainty is so pervasive that the problem itself seems better modeled as one of predicting a distribution over veridicality categories, rather than trying to predict a single label. The predictive model we developed is true to this intuition, since it trains on and predicts distributions. Although automatically assigning veridicality judgments that correspond to speakers' intuitions when pragmatic factors are allowed to play a role is challenging, our classifier shows that it can be done effectively using a relatively simple feature set, and we expect performance to improve as we find ways to model richer contextual features.

These findings resonate with the notion of entailment used in the Recognizing Textual Entailment (RTE) challenges [23], where the goal is to determine, for each pair of sentences $\langle T, H\rangle$, whether T (the *text*) justifies H (the *hypothesis*). The original task definition draws on "common-sense" understanding of language [24], and focuses on how people interpret utterances naturalistically. Thus, these entailments are not calculated over just the information contained in the sentence pairs, as a more classical logical approach would have it, but rather over the full utterance meaning. As a result, they are imbued with all the uncertainty of utterance meanings [25], [26], [27]. This is strongly reminiscent of our distinction between semantic and pragmatic veridicality. For example, as a purely semantic fact, *might*(S) is non-veridical with regard to S. However, depending on the nature of S, the nature of the source, the context, and countless other factors, one might nonetheless infer S. This pragmatic complexity is one of the central lessons of our new annotations.

Acknowledgment

We thank our colleagues Eric Acton, Alex Djalali and Scott Grimm for normalizing the test set items in preparation for the annotation. This work was supported in part by ONR grant No. N00014-10-1-0109 and ARO grant No. W911NF-07-1-0216.

References

[1] J. Barwise, "Scenes and other situations," *The Journal of Philosophy*, vol. 78, no. 7, pp. 369–397, 1981.

[2] F. Zwarts, "Nonveridical contexts," *Linguistic Analysis*, vol. 25, no. 3–4, pp. 286–312, 1995.

[3] A. Giannakidou, "Affective dependencies," *Linguistics and Philosophy*, vol. 22, no. 4, pp. 367–421, 1999.

[4] V. L. Rubin, E. D. Liddy, and N. Kando, “Certainty identification in texts: Categorization model and manual tagging results,” in *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*, J. G. Shanahan, Y. Qu, and J. Wiebe, Eds. Springer-Verlag New York, Inc., 2005.

[5] V. L. Rubin, “Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements,” in *Proceedings of the NAACL-HLT 2007*, 2007.

[6] R. Saurí, “A factuality profiler for eventualities in text,” Ph.D. dissertation, Computer Science Department, Brandeis University, 2008.

[7] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, “A simple algorithm for identifying negated findings and diseases in discharge summaries,” *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[8] P. L. Elkin, S. H. Brown, B. A. Bauer, C. S. Husser, W. Carruth, L. R. Bergstrom, and D. L. Wahner-Roedler, “A controlled trial of automated classification of negation from clinical notes,” *BMC Medical Informatics and Decision Making*, vol. 5, no. 13, 2005.

[9] Y. Huang and H. J. Lowe, “A novel hybrid approach to automated negation detection in clinical radiology reports,” *Journal of the American Medical Informatics Association*, vol. 14, no. 3, pp. 304–311, 2007.

[10] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, “BioInfer: a corpus for information extraction in the biomedical domain,” in *BMC Bioinformatics*, 2007.

[11] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, “The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts,” in *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, 2008.

[12] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, “Overview of BioNLP’09 shared task on event extraction,” in *Proceedings of the Workshop on BioNLP: Shared Task*, 2009.

[13] R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas, “The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text,” in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task*, 2010.

[14] R. Saurí and J. Pustejovsky, “FactBank: A corpus annotated with event factuality,” *Language Resources and Evaluation*, vol. 43, no. 3, 2009.

[15] P. Kiparsky and C. Kiparsky, “Facts,” in *Progress in linguistics*, M. Bierwisch and K. E. Heidolph, Eds. The Hague, Paris: Mouton, 1970.

[16] L. Karttunen, “Presuppositions and compound sentences,” *Linguistic Inquiry*, vol. 4, no. 2, pp. 169–193, 1973.

[17] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng, “Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. ACL, October 2008, pp. 254–263.

[18] A. Berger, S. Della Pietra, and V. Della Pietra, “A maximum entropy approach to natural language processing,” *Computational Linguistics*, vol. 22, pp. 39–71, 1996.

[19] C. D. Manning and D. Klein, “Optimization, maxent models, and conditional estimation without magic,” in *Tutorial at HLT-NAACL 2003 and ACL 2003*, 2003, http://nlp.stanford.edu/software/classifier.shtml.

[20] D. Klein and C. D. Manning, “Accurate unlexicalized parsing,” in *Proceedings of the 41st Meeting of the Association of Computational Linguistics*, 2003.

[21] M.-C. de Marneffe, B. MacCartney, and C. D. Manning, “Generating typed dependency parses from phrase structure parses,” in *Proceedings of LREC-06*, 2006.

[22] M. Simons, “Observations on embedding verbs, evidentiality, and presupposition,” *Lingua*, vol. 117, no. 6, pp. 1034–1056, 2007.

[23] I. Dagan, O. Glickman, and B. Magnini, “The PASCAL recognising textual entailment challenge,” in *Machine Learning Challenges, Lecture Notes in Computer Science*, J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d’Alch Buc, Eds., vol. 3944. Springer-Verlag, 2006, pp. 177–190.

[24] S. Chapman, *Paul Grice: Philosopher and Linguist*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2005.

[25] A. Zaenen, L. Karttunen, and R. Crouch, “Local textual inference: Can it be defined or circumscribed?” in *ACL Workshop on Empirical Modelling of Semantic Equivalence and Entailment.* Ann Arbor, MI: Association for Computational Linguistics, 2005.

[26] C. D. Manning, “Local textual inference: it’s hard to circumscribe, but you know it when you see it – and NLP needs it,” 2006, ms., Stanford University.

[27] R. Crouch, L. Karttunen, and A. Zaenen, “Circumscribing is not excluding: A reply to Manning,” 2006, ms., Palo Alto Research Center.