A New Dataset and Evaluation for Belief/Factuality

Vinodkumar Prabhakaran[1], Tomas By[2], Julia Hirschberg[1], Owen Rambow[3]*, Samira Shaikh[4], Tomek Strzalkowski[4], Jennifer Tracey[5], Michael Arrigo[5], Rupayan Basu[6], Micah Clark[2], Adam Dalton[2], Mona Diab[7], Louise Guthrie[2], Anna Prokofieva[1], Stephanie Strassel[5], Gregory Werner[7], Janyce Wiebe[8], Yorick Wilks[2]

[1]Department of Computer Science, Columbia University, New York, NY, USA
[2]Florida Institute for Human and Machine Cognition (IHMC), FL, USA
[3]Center for Computational Learning Systems, Columbia University, New York, NY, USA
[4]State University of New York - University of Albany, NY, USA
[5]Linguistic Data Consortium (LDC), University of Pennsylvania, PA, USA
[6]Amazon.com Inc., CA, USA
[7]George Washington University, DC, USA
[8]University of Pittsburgh, PA, USA

*corresponding author; email address: rambow@ccls.columbia.edu

Abstract

The terms "belief" and "factuality" both refer to the intention of the writer to present the propositional content of an utterance as firmly believed by the writer, not firmly believed, or having some other status. This paper presents an ongoing annotation effort and an associated evaluation.

1 Introduction

This paper presents an ongoing project aimed at developing a community-wide evaluation of expressed belief, also known as "factuality". Belief and factuality are closely related to hedging, veridicality, and modality. The project has grown out of the DARPA DEFT project; participants include the Linguistic Data Consortium (LDC) and three performer sites: Columbia University/George Washington University, the Florida Institute for Human and Machine Cognition, and the University of Albany. The goal of our research project is not linguistic annotation, but the identification of meaning which is expressed in a non-linguistic manner. Such a meaning representation is useful for many applications; in our project we are specifically interested in knowledge base population. A different part of the DEFT program is concerned with the representation of propositional meaning, following the tradition of the ACE program in representing entities, relations and events (ERE) (Doddington et al., 2004). The work presented here is concerned with the attitude of agents towards propositional content: do the agents express a committed belief or a non-committed belief in the propositional content? Our work has several characteristics that set it apart from other work: we are interested in annotation which can be done fairly quickly; we are not interested in annotating linguistic elements (such as trigger words); and we are planning an integration with sentiment annotation.

The structure of the paper is as follows: we start out by situating our notion of "belief" with respect to other notions of extra-propositional meaning (Section 2); we then present our annotation in some detail, with a special comparison to FactBank (Saurí and Pustejovsky, 2012). While the goal of this paper is not to talk about computational systems that were run as part of the evaluation (different publications will be available for that purpose), we quickly summarize their main characteristics so that the evaluation results can be interpreted. We then turn to

the pilot evaluation we have performed, presenting first the evaluation with respect to propositions (Section 5) and then a qualitative evaluation. We conclude with a discussion of plans for the upcoming open evaluation, scheduled for December 2015.

2 Terminology and Related Work

In this section, we identify how we define different terms. Different papers may have different and conflicting definitions of these terms, but for lack of space we do not provide an overview over all definitions.

While at first the terms "belief" and "factuality" appear to relate to rather different things (a subjective state versus truth), in the NLP community they in fact refer to the same phenomenon, while having rather different connotations. The phenomenon is the communicative intention of a writer[1] to present propositional content as something that she firmly believes is true, weakly believes is true, or has some other attitude towards, namely a wish or a reported belief. The term "belief" here describes the cognitive state of the writer (Diab et al., 2009), and comes from artificial intelligence and cognitive science, as in the Belief-Desire-Intention model of Bratman (1999 1987). The term "factuality" describes the communicative intention of the writer (Saurí and Pustejovsky, 2012, p. 263) (our emphasis):

> The fact that an eventuality is depicted as holding or not does not mean that this is the case in the world, but that this is how it is characterized by its informant. Similarly, it does not mean that this is the real knowledge that informant has (his true cognitive state regarding that event) but what he *wants us to believe* it is.

We would like to emphasize that the terms "belief" and "factuality" do not refer to the underlying truth of propositions, only to the intention of the writer to present them as, in her view, true. Thus, we as researchers cannot determine what is true from an analysis of factuality (or of belief). The term "factuality" is often misunderstood in this respect, which is one of the reasons we prefer not to use it. In order to understand the relation between belief/factuality and truth, we need to distinguish two assumptions. First, we may assume that the writer is not lying (assumption of truthfulness). In this paper, we make this first assumption. Second, we could assume that the writer knows what is true (assumption of truth). In this paper, we do not make this second assumption. We discuss these two assumptions in turn.

We start with the assumption of truthfulness. In the quote above, Saurí and Pustejovsky (2012) (apart from distinguishing factuality from truth) also make the point that the writer's communicative intention of making the reader believe she has a specific belief state does not mean that she actually has that cognitive state, since she may be lying. Lying is clearly an important phenomenon that researchers have looked into (Mihalcea and Strapparava, 2009; Ott et al., 2011).[2] However, we (as linguists interested in understanding how language enables communication) feel that assuming the writer is truthful is a standard assumption about communication which we should in general make. This is because if we do not make this assumption, we cannot explain why communication is possible at all, since discourse participants would have no motivation to ever adopt another discourse participant's belief as their own. We therefore do claim that we can infer belief from utterances, while assuming that the writer is not lying, and knowing that this assumption may be false in certain cases.

We now turn to the second assumption, the assumption of truth. Even if we assume that the writer is not lying, the assumption of truth is not required for communication to succeed; this is because the writer may be wrong, and this has no effect on the communication. For example, Ptolemy successfully made many people believe that the sun rotates around the earth, as was his (presumably) honest communicative intention. Therefore, to us as researchers interested in describing how language

[1] For brevity, we will assume a female writer as the source of utterances in this paper. Everything we say applies equally to spoken and written communication, and equally to male and female communicators.

[2] Sarcasm and irony differ from lying in that the communicative intention and the cognitive state are aligned, but they do not align with the standard interpretation of the utterance. Here, the intention is that the reader recognizes that the form of the utterance does not literally express the cognitive state. We leave aside sarcasm and irony in this paper; for current computational work on sarcasm detection, see for example (González-Ibáñez et al., 2011).

is used to communicate, it does not matter that astronomers now believe that Ptolemy was wrong, it does not change our account of communication and it does not change the communication that happened two millennia ago. And since we do not need to make the assumption that the writer knows what she is talking about, we choose not to make this assumption. In the case of Ptolemy, we leave this determination – what is actually true – to astronomers. In other cases, we typically have models of trustworthiness: if a writer sends her spouse a text message saying she is hungry, the spouse has no reason to believe she is wrong. We leave this issue aside in this paper.

The term "hedge" refers to words or phrases that add ambiguity or uncertainty (Propositional Hedges) or show the speakers lack of commitment to a proposition (Relational Hedges). For example, *The ball is* **sort of** *blue* contains a Relational Hedge (*sort of*) and *I* **think** *the ball is blue* includes a propositional hedge (*think*). Propositional hedges indicate non-committed belief. There has been a major effort to annotate texts with hedging information (Farkas et al., 2010), with an open evaluation. While belief and hedging are closely related, we see the belief/factuality annotation as more general than hedging (since it does not only include non-committed belief), and also more semantic (since we are not identifying language use but underlying meaning).

The term "modality" is used in formal semantics as well as in descriptive linguistics. Many semanticists (e.g. (Kratzer, 1991; Kaufmann et al., 2006)) define modality as quantification over possible worlds. Modality can be of two types: epistemic, which qualifies the speaker's commitment, and deontic, which concerns freedom to act. Belief/factuality falls under epistemic modality. Another view of modality relates more to a speaker's attitude toward a proposition (e.g. (McShane et al., 2004; Baker et al., 2010; Prabhakaran et al., 2012)), which is closer to the way we model belief.

We interpret the term "veridical" as referring to a property of certain words (usually verbs), namely to mark the proposition expressed by their syntactic complement clause as firmly believed (committed belief) by the writer (Kiparsky and Kiparsky, 1970). Veridicality as a property of lexical or lexico-syntactic elements is thus a way of relating belief/factuality to linguistic means of expressing them, but we take the notion of belief/factuality as being the underlying notion.

3 Annotation

3.1 Annotation Manual

The purpose of this annotation is to capture the commitment of the writer's belief in the propositions expressed in the text. The annotation for this project marks beliefs held by the writer only. We exhaustively annotate all (clausal) propositions in each document with a four-way belief type distinction, with the following categories.

Committed belief (CB) – the writer strongly believes that the proposition is true. Examples:

(1) a. The sun will **rise** tomorrow.
b. I know John and Katie **went** to Paris last year.

Non-committed belief (NCB) – the writer believes that the proposition is possibly or probably true, but is not certain. Examples:

(2) a. It could **rain** tomorrow.
b. I think John and Katie **went** to Paris last year.

Reported belief (ROB) – the writer attributes belief (either committed or non-committed) to another person or group. Note that this label is only applied when the writer's own belief in the proposition is unclear. Examples:

(3) a. Channel 6 said it could **rain** tomorrow.
b. Sarah said that John and Katie **went** to Paris last year.

Non-belief propositions (NA) – the writer expressed some other cognitive attitude toward the proposition, such as desire or intention, or expressly states that s/he has no belief about the proposition (e.g., by asking a question). Examples:

(4) a. Is it going to **rain** tomorrow?
b. I hope John and Katie **went** to Paris last year.

We do not make any effort to evaluate the truth value of the propositions, only the expressed level of belief in them held by the writer. Thus a strongly held false belief would not appear any different from

a strongly held true belief. Similarly, lying, sarcasm, irony, and other cases where the writer's internal belief may differ from the expressed belief are not captured. That is, we take all expressed beliefs at "face value". We also do not capture any cognitive attitudes expressed about a proposition other than belief. An NA tag signifies just that there is no belief expressed about the proposition; it does not signify that there is another cognitive attitude expressed (e.g., 4a). Similarly, a proposition tagged as CB may also have other cognitive attitudes expressed about them (e.g., in "John managed to go to Paris last week", the author is expressing CB towards the proposition *go*, but also the *success* modality (Prabhakaran et al., 2012)); we do not capture them.

Annotators are not required to identify the full text span of the proposition. Instead, we take advantage of the close relationship between the semantics of the proposition and the syntactic structure of the clause by marking only the head of the structural unit containing the proposition (propositional head). For each proposition, annotators mark a head word and tag it with one of the four belief types. Note that the syntactic head word (perhaps lemmatized) can serve as a convenient name for the proposition, so for the examples above, we can talk about the belief in the 'rain' proposition and in the 'go' proposition. When a sentence has a single clause containing only one proposition, there will be only one head word to identify (usually a verb, but see details below on identifying heads of propositions). Many sentences contain multiple propositions, and the annotation guidelines provide detailed instructions on identifying head words. Note that the (b) examples above contain an additional proposition which is not marked; a full markup for example (3b) is below.

(5) Sarah said/CB that John and Katie went/ROB to Paris last year.

This is equivalent to the following span-based annotation:

(6) [CB Sarah said [ROB that John and Katie went to Paris last year.]]

The general principles of head word selection for each proposition can be summarized as follows:

1. Annotate the lexical verb of the clause expressing the proposition, if there is one.
2. If the verb of the clause is a copula, annotate the head of the predicate that follows the copula (noun for NP, preposition for PP, etc.).
3. Deontic modal auxiliaries, which signal a complex proposition, are annotated in addition to the lexical verb, as a separate belief.

All annotations are applied to a single whitespace-delimited word. In cases where the head of a proposition is a multiword expression (MWE), the head of the MWE is selected. In cases of noun phrases where no head is apparent (e.g. *bok choy*), the last word of the MWE is selected.

3.2 Comparison with FactBank

As already explained (Section 2), we take the terms "belief" and "factuality" to refer to the same phenomenon underlyingly (with perhaps different emphases). Therefore, the FactBank annotation is basically compatible with ours. Our annotation is much simpler than that of FactBank in order to allow for a quicker annotation. We summarize the main points of simplification here.

- We have taken the source always to be the writer. As we will discuss in Section 7.1, we will adopt the FactBank annotation in the next iteration of our annotation.
- We do not distinguish between possible and probable; this distinction may be hard to annotate and not too valuable.
- We ignore negation. If present, we simply assume it is part of the proposition which is the target.

Werner et al. (2015) study the relation between belief and factuality in more detail. They provide an automatic way of mapping the annotations in FactBank to the 4-way distinction of speaker/writer's belief that we present in this paper.

3.3 Corpus and Annotation Results

The annotation effort for this phase of belief annotation for DEFT produced a training corpus of 852,836 words and an evaluation corpus of 100,037 words. All annotated data consisted of English text from discussion forum threads. The discussion forum

threads were originally collected for the DARPA BOLT program, and were harvested from a wide variety of sites. Discussion forum sites were chosen for harvesting in BOLT based on human judgement that the site was likely to contain many threads discussing either current events or personal anecdotes. For details on the BOLT collection, see Garland et al. (2012). Threads longer than 1000 words were truncated to produce documents consisting of one or more consecutive posts from a single thread. Long threads may generate multiple documents consisting of non-overlapping sections of the same thread (e.g., document 1 contains posts 1-5, while document 2 contains posts 6-12, etc.). The distribution of the four belief types in the training and evaluation corpora can be seen in Table 1.

Annotations	CB	NCB	ROB	NA
Training Corpus				
143240	79995 (56%)	3890 (3%)	7150 (5%)	52205 (36%)
Evaluation Corpus				
17553	8730 (50%)	583 (3%)	941 (5%)	7299 (42%)

Table 1: Annotation Statistics

The source data pool, annotation procedures, and annotators were the same for both the training and evaluation datasets, with the exception of the fact that the evaluation annotations received a full second pass over the annotation by a senior annotator (not the same as the first pass annotator) to increase consistency and reduce annotator errors. The training annotations were produced with a single annotation pass, and quality control was conducted through a second pass by a senior annotator on a sample of approximately 15% of the data. Inter-annotator agreement on headword selection was 93% and agreement on belief type labeling was 84%. Overall observed agreement, combining headword selection and belief type label, was 78% (Kappa score .60). Agreement on belief type was least reliable on the categories of ROB and NCB, both of which were sometime erroneously marked as CB. Both of these categories, in addition to being less frequent in the corpus, have difficult edge cases in which the annotator must make a judgment based on the context of the document (for example, deciding whether the writer clearly shares a belief attributed to another person for ROB).

4 Evaluation Systems

We conducted a multi-site pilot evaluation for the task of identifying beliefs expressed in text. Three performer sites took part in this evaluation. In this section, we briefly describe the systems built at these performer sites. The first two systems are rule-based systems, whereas the third system is a supervised learning system. We limit the discussion of these systems to a high level, postponing the detailed system descriptions to separate future publications.

4.1 System A

System A is adapted from a Sentiment Slot Filling system which participated in the 2014 TAC KBP SSF Evaluation (Shaikh et al., 2014). This system uses the Stanford Parser to create a syntactic dependency structure for every sentence in a given document. Using the dependency tree, it extracts the belief targets, which are usually the subjects of the sentence. In addition, the system extracts belief relations – a unary or binary predicate – typically a verb, an adjective or a noun. The focus of this version of System A is to identify propositional heads that express belief of any type. Each relation so extracted was initially marked as CB. A few heuristics were then applied to distinguish CBs from NCBs - such as presence of hedge words (*maybe*, *guess*). In addition, a few heuristics were added to tag relations as NAs, for example when the predicates appear in a question. The current version of System A does not account for ROB tags.

4.2 System B

System B uses the dependency tree and part-of-speech tags from the Stanford NLP tools, together with a custom verb lexicon to recognize belief expressions. The tree is processed to convert objects and complements to a single format, and then transformed into one or more belief triples (subject, verb, object). The system maintains a database of nested belief context, as in 'X believes Y believes Z', but we did not notice many instances of this phenomenon in the data. Partly because our System B

	System A			System B			System C		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
CB	35.9	39.9	37.8	42.1	36.8	39.3	68.9	77.9	73.1
NCB	13.3	8.8	10.6	4.6	7.4	5.7	52.9	29.7	38.0
ROB	0.0	0.0	0.0	1.3	0.9	1.0	43.8	15.6	23.0
NA	40.3	5.9	10.2	35.8	4.8	8.4	80.1	62.0	69.9
Overall	35.5	22.5	27.6	34.4	20.6	25.8	72.0	66.4	69.1

Table 2: Results obtained for System A, System B, and System C on the final Evaluation dataset.

recognizes reported beliefs (ROB) independently of the distinction between committed/non-committed belief in the annotations, the heuristic rules (mainly based on the presence of modal auxiliaries) that we added for the purpose of classifying the beliefs (CB, NCB, ROB, NA) did not work reliably in all cases.

4.3 System C

System C uses a supervised learning approach to identify tokens denoting the heads of propositions that denote author's expressed beliefs. It approaches this problem as a 5-way (CB, NCB, ROB, NA, *nil*) multi-class classification task at the word level. System C is adapted from a previous system which uses an earlier, simpler definition and annotation of belief (Prabhakaran et al., 2010). The system uses lexical and syntactic features for this task, which are extracted using the part-of-speech tags and dependency parses obtained from the Stanford CoreNLP system. In addition to the features described in (Prabhakaran et al., 2010), System C uses a set of new features including features based on a dictionary of hedge-words (Prokofieva and Hirschberg, 2014). The hedge features improved the NCB F-measure by around 2.2 percentage points (an overall F-measure improvement of 0.25 percentage points) in experiments conducted on a separate development set. It uses a quadratic kernel SVM to build the model, which outperformed the linear kernel in experiments conducted on the development set.

5 Proposition-Oriented Evaluation

We now describe the results obtained on a proposition-oriented quantitative evaluation of these systems. We focus on a system's ability to correctly identify the propositional heads of each type of belief (CB, NCB, ROB, NA). Only the words denoting heads of propositions will get one of these tags, and hence the majority of words in our data will not have any tags. We expect the system to find the propositional heads and to correctly assign their belief tags.

We use the entire Evaluation dataset described in Section 3 for this evaluation (entirely unseen during the development of the systems). We report precision, recall and F-measure for each belief type. We also report their micro-averages as the overall result. We compute F-measure as the harmonic mean between precision and recall. The best results obtained by each system described in Section 4 are presented in Table 2.

For System A, four different configurations were run for the evaluation, in which the NCB and NA tagging was either enabled or disabled. (The current version does not account for ROB tags.) In Table 2, Columns 2-4, we show the performance of System A while all 3 tags (CB, NCB and NA) are enabled. The results of other three configurations are comparable. Any sentence where the belief target could not be located, either due to parsing error or due to missing coreference (as supplied by ERE), was discarded. This resulted in a relatively lower recall in the evaluation, but produced high precision in a target-driven pilot evaluation (Section 6). The results obtained by System B in the evaluation are shown in Table 2: Columns 5-7. The results of System B, when ignoring the belief categories (i.e., on identifying heads of propositions), were 83.6% precision and 50% recall. Table 2: Columns 8-10 shows results obtained by System C trained on 80% of the training dataset (the rest of the corpus was used as a development set).

The supervised learning approach obtained over-

all better performance than rule based approach in our evaluation. ROB and NCB were the most difficult classes to predict for all three systems (e.g., highest recall posted for ROB is only 15.6%). CB was relatively easier to predict. NA was difficult to predict using the rule based approach, but supervised learning approach obtained reasonable performance of 69.9 F-measure.

6 An Entity-Focused Evaluation: Preliminary Investigation

In this section, we describe an initial investigation towards an entity-focused evaluation. An entity-focused evaluation tests a different kind of question about beliefs: given an entity e, what beliefs does the writer have about e? This entity-focused evaluation draws its parallels from TAC KBP Sentiment Slot Filling Evaluation (SSF) task. In the SSF, the task is to determine a target entity given a source entity and a sentiment between them. The goal is to populate a knowledge base with information regarding entities and the sentiment relations between them. In the same vein, an entity-focused belief task would provide knowledge about the salient belief relations between entities. For this purpose, we needed to define what is meant by "having a belief about an entity" and agreed on the following preliminary rules. The rules are entirely syntactic. In the following examples, the target entity is Mary, and the statement after the arrow shows what the beliefs are about her (and what the level of commitment by the writer is).

Adjunct clause case 1. If the target entity is contained in a clause (lets call it the "core clause") but NOT in an adjunct clause which modifies the core clause, we omit the adjunct clause (even though the adjunct clause in some sense pertains to the core clause but by virtue of being an adjunct, it is omissable).

(7) While John was in/CB Paris, Mary left/CB Paul
$\longrightarrow$ CB: Mary left Paul

Adjunct clause case 2. If the target is in an adjunct clause to a core clause where the target is not mentioned, we retain both the adjunct clause as a standalone belief, and the combination of the adjunct and core (i.e., we have two beliefs about the entity).

(8) John was happy/CB when Mary left/CB Paul
$\longrightarrow$ CB: John was happy when Mary left Paul ; CB: Mary left Paul

We devised similar rules for complement clauses, we omit them here.

For the actual evaluation, we used files which also had been hand-annotated for ACE entities. However, we did not have a gold annotation for entity-focused belief, as this study is still contributing towards a definition of this notion. Only two systems participated, System A and System C. System A as described in Section 4.1 already takes the notion of entity into account. For System C, we used the parse to determine the span associated with the annotated headword, and counted a proposition whose span included an entity to be about that entity. In order to understand how these two ways of determining entity-focused belief relate to each other, we compared the two systems to each other. We obtained an F-measure of 52%. We also hand evaluated the positive claims of System C, obtaining an accuracy of 48% on the positive claims. The errors are due to parse errors, the presence of the entity in adjuncts which do not appear germane (contradicting adjunct clause case 2), the presence of irrelevant adjunct clauses (counter to adjunct clause case 1), and to a lack of clarity in the annotation standard. As an example of the lack of clarity, consider the following sentence from our evaluation corpus, with *two kids* as target entity:

(9) I didn't see these two kids (sic) names on the news

two kids is a possessor of the direct object, and fell into the span of the annotated *see* for System C, but System A deemed the 'see' belief not to be about it. We conclude that this purely syntactic definition of "belief about an entity" is not satisfactory. The definition of "belief about an entity" remains an open question and we return to it in Section 7.3.

7 Plans for Next Round

7.1 Adding the Source

Currently, we are only annotating and evaluating the writer's beliefs. Beliefs attributed by the writer to other sources are marked ROB. We intend to annotate the source for all beliefs, using the method of

nested attribution pioneered by MPQA (Wiebe et al., 2005) and adopted by FactBank (Saurí and Pustejovsky, 2012). Consider the following sentence.

(10) John believes Mary knows that the clock was stolen

In the nested attribution approach, according to the writer, according to John, Mary firmly believes (CB) the 'steal' proposition. According to the writer, John firmly believes the 'know' proposition and the 'steal' proposition (as indicated by the veridical verb *know*). The writer herself firmly believes (CB) the 'believe' proposition, does not express an opinion on the 'know' proposition (ROB), and also firmly believes (CB) the 'steal' proposition (again, the reader infers this from the use of *know*). We intend to annotate all these levels of belief.

7.2 Defining the Target Proposition

In our work to date, we have assumed that the target of a belief is a proposition, and we have represented the proposition by the syntactic head word of the clause which describes the proposition (which is equivalent to a text span under syntactic projection). We are investigating extending this in several manners. First, we are considering including the heads of event noun phrases (*the sudden collapse of the building*). Second, we are looking at using a semantic representation for the proposition (as opposed the syntactic head of the text passage describing the proposition). We do not propose to develop our own semantic representation, but instead we will look to using existing relation and event representations based on the ACE program (Doddington et al., 2004). These have the advantage that there are off-the-shelf computational tools available for detecting ACE relations and events; they have the disadvantage that they do not cover all propositions we may be interested in. An alternative would be the use of a shallower semantic representation such as PropBank (Kingsbury et al., 2002), FrameNet (Baker et al., 1998), or AMR (Banarescu et al., 2013).

7.3 Entities as Targets

In Section 6, we discussed an initial evaluation of a belief being about an entity. In this section we discuss further guidelines for identifying belief targets, i.e., when one can say that someone's belief is about a certain entity.

In general, the notion of belief "aboutness" is fairly fuzzy and it may be difficult to circumscribe precisely without some additional constraints. Suppose then that one of the ultimate objectives of belief extraction is to populate a knowledge base with beliefs held about specific entities: individuals, groups, artifacts, etc., which adds this constraint that the extracted belief is knowledge-base-worthy, or reportable. Some initial guidelines may go as suggested below. The objective is to provide guidance for a human assessor — not to propose a solution. We should note that these guidelines generally transcend any syntactic or structural considerations and appeal directly to the annotators' judgment. Furthermore, we note that these guidelines are not about effects relating to information structure – in one sense of "being about", the same sentence may be referred to as being "about" different things in different contexts. We are aiming for a lexical-semantic, not a pragmatic notion of aboutness.

A belief whose target is proposition p is **about** an entity T if one of the following clauses holds:

1. p describes a property of T, where the property is considered semi-permanent but not necessarily limited to physical or mental characteristics (e.g., red, long, brainy) and may also include behavioral properties (smart, slow) as well as characteristics bestowed on by others (beloved).

2. T is an agent of p, i.e., T is said to be performing some activity, physical or mental: drive a car, send a letter, etc.

3. T is directly involved in (or affected by) p but is not an agent: this includes situations where T's involvement may be passive but is nonetheless required for p to be performed, e.g., receive a letter, win a prize, etc.

We make no claim that the above list is exhaustive or that there would not be exceptions to these rules. For this reason we may also attempt to describe conditions under which a belief is not about T. For example: a belief target p is not about entity T even though T may be mentioned within the scope of p if:

4. T appears uninvolved in p and is apparently unaffected by its execution, e.g., reading about, waiting for, etc.

We intend to explore whether we can define this notion of belief aboutness sufficiently well to obtain good inter-annotator agreement.

7.4 Combining with Sentiment

We are planning on working on an annotation and an evaluation that combines belief with sentiment. The motivation for this is that belief and sentiment are similar types of meaning: they are attitudes towards propositions or entities which are expressed directly or indirectly. The similarity can also be seen in the fact that FactBank took the notion of nested source from MPQA, which is a sentiment-annotated corpus. Furthermore, many lexical items express both a belief and a sentiment at once:

(11) I hope Bertha enjoys the oysters

Here, the writer expresses a positive sentiment towards the 'enjoy' proposition, and at the same time she is expressing a lack of certainty (NCB) in the 'enjoy' proposition.

7.5 Adding Spanish and Chinese

We will be extending our annotation (including some of the extensions mentioned above) to Spanish and Chinese.

8 Conclusion

We have presented an ongoing annotation effort related to belief/factuality and an initial evaluation based on that annotation effort. To our knowledge, the annotated corpus is by far the largest corpus annotated in terms of belief/factuality. We have presented several proposed extensions to the annotation. The linguistic resources described in this paper will be published in the LDC catalog, making them available to the broader research community. The materials will be used in an open evaluation in late 2015 or early 2016. The evaluation will cover both belief/factuality and sentiment.

Acknowledgments

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We thank several anonymous reviewers for their constructive feedback.

References

Collin F. Baker, J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montréal.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Nathaniel W. Filardo, Lori S. Levin, and Christine D. Piatko. 2010. A modality lexicon and its use in automatic tagging. In *LREC*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Michael E. Bratman. 1999 [1987]. *Intention, Plans, and Practical Reason*. CSLI Publications.

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. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program–tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. 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*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA, June. Association for Computational Linguistics.

Stefan Kaufmann, Cleo Condoravdi, and Valentina Harizanov, 2006. *Formal Approaches to Modality*, pages 72–106. Mouton de Gruyter.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.

Paul Kiparsky and Carol Kiparsky. 1970. Facts. In Manfred Bierwisch and Karl Erich Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, The Hague, Paris.

Angelika Kratzer. 1991. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*. Walter de Gruyter, Berlin.

Marjorie McShane, Sergei Nirenburg, and Ron Zacharsky. 2004. Mood and modality: Out of the theory and into the fray. *Natural Language Engineering*, 19(1):57–89.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore, August. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August. Coling 2010 Organizing Committee.

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. Association for Computational Linguistics.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*, Reykjavik, Iceland.

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

Samira Shaikh, Rob Giarrusso, Veena Ravishankar, and Tomek Strzalkowski. 2014. The SUNY Albany Sentiment Slot Filling System. In *Proceedings of the 2014 TAC KBP Sentiment Slot Filling Evaluation*, Gaithersburg, Maryland, USA. NIST.

Gregory J. Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Denver, USA, June. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language ann. *Language Resources and Evaluation*, 39(2/3):164–210.