A Compositional Interpretation of Biomedical Event Factuality

Halil Kilicoglu, Graciela Rosemblat, Michael J. Cairelli, Thomas C. Rindflesch
National Library of Medicine
National Institutes of Health
Bethesda, MD, 20894
{kilicogluh,grosemblat,mike.cairelli,trindflesch}@mail.nih.gov

Abstract

We propose a compositional method to assess the factuality of biomedical events extracted from the literature. The composition procedure relies on the notion of semantic embedding and a fine-grained classification of extra-propositional phenomena, including modality and valence shifting, and a dictionary based on this classification. The event factuality is computed as a product of the extra-propositional operators that have scope over the event. We evaluate our approach on the GENIA event corpus enriched with certainty level and polarity annotations. The results indicate that our approach is effective in identifying the certainty level component of factuality and is less successful in recognizing the other element, negative polarity.

1 Introduction

The scientific literature is rich in extra-propositional phenomena, such as speculations, opinions, and beliefs, due to the fact that the scientific method involves hypothesis generation, experimentation, and reasoning to reach, often tentative, conclusions (Hyland, 1998). Biomedical literature is a case in point: Light et al. (2004) estimate that 11% of sentences in MEDLINE abstracts contain speculations and argue that speculations are more important than established facts for researchers interested in current trends and future directions. Such statements may also have an effect on the reliability of the underlying scientific claim. Despite the prevalence and importance of such statements, natural language processing systems in the biomedical domain have largely focused on more foundational tasks, including named entity recognition (e.g., disorders, drugs) and relation extraction (e.g., biological events, gene-disease associations), the former task addressing the *conceptual* level of meaning and the latter addressing the *propositional* level.

The last decade has seen significant research activity focusing on some extra-propositional aspects of meaning. The main concern of the studies that focused on the biomedical literature has been to distinguish facts from speculative, tentative knowledge (Light et al., 2004). The studies focusing on the clinical domain, on the other hand, have mainly aimed to identify whether findings, diseases, symptoms, or other concepts mentioned in clinical reports are present, absent, or uncertain (Uzuner et al., 2010). Various corpora have been annotated for relevant phenomena, including hedges (Medlock and Briscoe, 2007) and speculation/negation (Vincze et al., 2008; Kim et al., 2008). Several shared task challenges with subtasks focusing on these phenomena have been organized (Kim et al., 2009; Kim et al., 2012). Supervised machine learning and rule-based approaches have been proposed for these tasks. In general, these studies have been presented as extensions to named entity recognition or relation extraction systems, and they often settle for assigning discrete values to propositional meaning elements (e.g., assessing the *certainty* of an *event*).

Kilicoglu (2012) has proposed a unified framework for extra-propositional meaning, encompassing phenomena discussed above as well as discourse level relations, such as Contrast and Elab-

oration, generally ignored in the studies of extra-propositional meaning (Morante and Sporleder, 2012). The framework uses *semantic embedding* as the core notion, *predication* as the representational means, and *semantic composition* as the methodology. It relies on a fine-grained linguistic characterization of extra-propositional meaning, including modality, valence shifters, and discourse connectives. In the current work, we present a case study of applying this framework to the task of assessing biomedical event *factuality* (whether an event is characterized as a fact, a counter-fact, or merely a possibility), an important step in determining current trends and future directions in scientific research. For evaluation, we rely on the meta-knowledge corpus (Thompson et al., 2011), in which biological events from the GENIA event corpus (Kim et al., 2008) have been annotated with several extra-propositional phenomena, including certainty level, polarity, and source. We discuss in this paper how two of these phenomena relevant to factuality (certainty and polarity) can be inferred from the semantic representations extracted by the framework. Our results demonstrate that certainty levels can be captured correctly to a large extent with our method and indicate that more research is needed for correct polarity assessment.

2 Related Work

Modality and negation are the two linguistic phenomena that are often considered in computational treatments of extra-propositional meaning. Morante and Sporleder (2012) provide a comprehensive overview of these phenomena from both theoretical and computational linguistics perspectives. In the FactBank corpus (Saurí and Pustejovsky, 2009), events from news articles are annotated with their factuality values, which are modeled as the interaction of *epistemic modality* and *polarity* and consist of eight values: FACT, PROBABLE, POSSIBLE, COUNTER-FACT, NOT PROBABLE, NOT CERTAIN, CERTAIN BUT UNKNOWN, and UNKNOWN. Saurí and Pustejovsky (2012) propose a factuality profiler that computes these values in a top-down manner using lexical and syntactic information. They capture the interaction between different factuality markers scoping over the same event. de Marneffe et al. (2012) investigate *veridicality* as the pragmatic component of factuality. Based on an annotation study that uses FactBank and MechanicalTurk subjects, they argue that veridicality judgments should be modeled as probability distributions. They show that context and world knowledge play an important role in assessing veridicality, in addition to lexical and semantic properties of individual markers, and use supervised machine learning to model veridicality. Szarvas et al. (2012) draw from previous categorizations and annotation studies to introduce a unified subcategorization of semantic uncertainty, with EPISTEMIC and HYPOTHETICAL as the top level categories. Re-annotating three corpora with this subcategorization and analyzing type distributions, they show that out-of-domain data can be gainfully exploited in assessing certainty using domain adaptation techniques, despite the domain- and genre-dependent nature of the problem.

In the biomedical domain, several corpora have been annotated for extra-propositional phenomena, in particular, negation and speculation. The GENIA event corpus (Kim et al., 2008) contains biological events from MEDLINE abstracts annotated with their certainty level (CERTAIN, PROBABLE, DOUBTFUL) and assertion status (EXIST, NON-EXIST). The BioScope corpus (Vincze et al., 2008) consists of abstracts and full-text articles as well as clinical text annotated with negation and speculation markers and their scopes. While they clearly address similar linguistic phenomena, the representations used in these corpora are significantly different (cue-scope representation vs. tagged events), and there have been attempts at reconciling these representations (Kilicoglu and Bergler, 2010; Stenetorp et al., 2012). BioNLP shared tasks on event extraction (Kim et al., 2009; Kim et al., 2012) and CoNLL 2010 shared task on hedge detection (Farkas et al., 2010) have focused on GENIA and BioScope negation/speculation annotations, respectively. Supervised machine learning techniques (Morante et al., 2010; Björne et al., 2012) as well as rule-based methods (Kilicoglu and Bergler, 2011) have been attempted in extracting these phenomena and their scopes. Wilbur et al. (2006) propose a more fine-grained annotation scheme with multi-valued qualitative dimensions to characterize scientific sentence fragments: *certainty* (complete uncertainty to com-

plete certainty), *evidence* (from no evidence to explicit evidence), *polarity* (positive or negative), and *trend/direction* (increase/decrease, high/low). In a similar vein, Thompson et al. (2011) annotate each event in the GENIA event corpus with five *meta-knowledge* elements: Knowledge Type (Investigation, Observation, Analysis, Method, Fact, Other), Certainty Level (considerable speculation, some speculation, and certainty), Polarity (negative and positive), Manner (high, low, neutral), and Source (Current, Other). Their annotations are more semantically precise as they are applied to events, rather than somewhat arbitrary sentence fragments used by Wilbur et al. (2006). Miwa et al. (2012) use a machine learning-based approach to assign meta-knowledge categories to events. They cast the task as a classification problem and use syntactic (dependency paths), semantic (event structure), and discourse features (location of the sentence within the abstract). They apply their system to BioNLP shared task data, as well, overall slightly outperforming the state-of-the-art systems.

3 Methods

We provide a brief summary of the framework here, mainly focusing on *predication* representation, embedding predicate categorization, and the compositional algorithm.

3.1 Predications

The framework uses the *predication* construct to represent all levels of relational meaning. A predication consists of a predicate *P* and *n* logical arguments (logical subject, logical object, adjuncts). They can be nested; in other words, they can take other predications as arguments. We call such constructs *embedding predications* to distinguish them from *atomic predications* that can only take atomic terms as arguments. While some embedding predications operate at the basic propositional level, extra-propositional meaning is exclusively captured by embedding predications. We use the notion of *semantic scope* to characterize the structural relationships between predications. A predication Pr_1 is said to *embed* a predication Pr_2 if Pr_2 is an argument of Pr_1. Similarly, a predication Pr_2 is is said to be within the *semantic scope* of a predication Pr_1, if a) Pr_1 embeds Pr_2, or b) there is a predication Pr_3, such that Pr_1 embeds Pr_3 and Pr_2 is within the semantic scope of or shares an argument with Pr_3. Scope relations play an important role in the composition procedure. A predication also encodes the *source* (*S*) and *scalar modality value* of the predication (MV_{Sc}). A formal definition of predication, then, is:

$$Pr := [P, S, MV_{Sc}, Arg_{1..n}], n >= 1$$

By default, the source of a predication is the writer of the text (*WR*). The source may also indicate a term or predication that refers to the source (i.e., who said what is described by the predication? what is the evidence for the predication?). The scalar modality value of the predication is a value in the [0,1] range on a relevant modality scale (*Sc*), which is assigned according to lexical properties of the predicate *P* and modified by its discourse context. By default, an unmarked, declarative statement has the scalar modality value of 1 on the EPISTEMIC scale (denoted as $1_{epistemic}$), corresponding to a fact.

3.2 Categorization

With the embedding categorization, we aim to provide a fine-grained characterization of the kinds of extra-propositional meanings contributed by predicates that indicate embedding. A synthesis of various linguistic typologies and classifications, the categorization is similar to the certainty subcategorization proposed by Szarvas et al. (2012); however, it not only targets certainty-related phenomena, but is rather a more general categorization of embedding predicates that indicate extra-propositional meaning. We distinguish four main classes of embedding predicates: MODAL, RELATIONAL, VALENCE_SHIFTER and PROPOSITIONAL; each class is further divided into subcategories. For the purposes of this paper, MODAL and VALENCE_SHIFTER categories are most relevant (illustrated in Figure 1).

A MODAL predicate associates its embedded predication with a modality value on a scale determined by the semantic category of the modal predicate (e.g., EPISTEMIC scale, DEONTIC scale). The scalar modality value (MV_{Sc}) indicates how strongly the embedded predication is associated with the scale *Sc*, 1 indicating strongest positive association and 0 negative association. VALENCE_SHIFTER

predicates do not introduce new scales but trigger a scalar shift of the embedded predication on the associated scale.

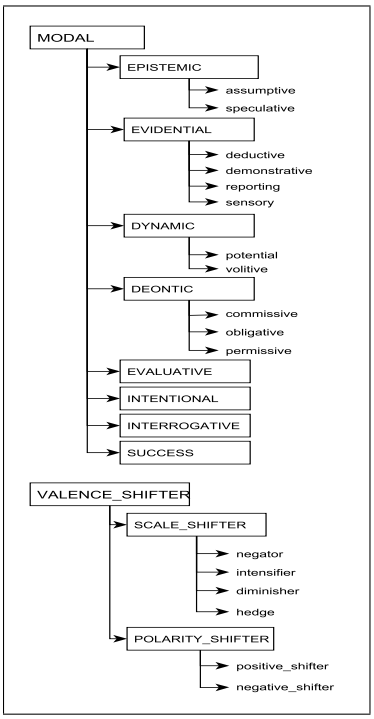

Figure 1. Embedding predicate types.

The MODAL subcategories relevant for factuality computation and examples of predicates belonging to these categories are as follows:

- EPISTEMIC predicates indicate a judgement about the factual status of the embedded predication (e.g., *may, possible*).
- EVIDENTIAL predicates indicate the type of evidence (observation, inference, etc.) for the embedded predication (e.g., *demonstrate*, *suggest*).
- DYNAMIC predicates indicate ability or willingness towards an event (e.g., *able, want*).
- INTENTIONAL predicates indicate effort of an agent to perform an event (e.g., *aim*).
- INTERROGATIVE predicates indicate questioning or inquiry towards the embedded event (e.g., *investigate*).
- SUCCESS predicates indicate degree of success associated with the embedded predication (e.g., *manage, fail*).

Each subcategory is associated with its own modality scale, except the EVIDENTIAL category, which is associated with the EPISTEMIC scale. The categories listed above also have secondary epistemic readings, in addition to their primary scale; for example, INTERROGATIVE predicates can indicate uncertainty. The EPISTEMIC scale is the most relevant scale to investigate factuality. Our model of this scale and how modal auxiliaries correspond to it is illustrated in Figure 2. It is similar to the characterization of factuality values by Saurí and Pustejovsky (2012), although numerical epistemic values are assigned to predications ($\text{MV}_{epistemic}$), rather than discrete values like Probable or Fact. In this, the characterization follows that of Nirenburg and Raskin (2004), which lends itself more readily to the type of operations proposed for scalar modality values.

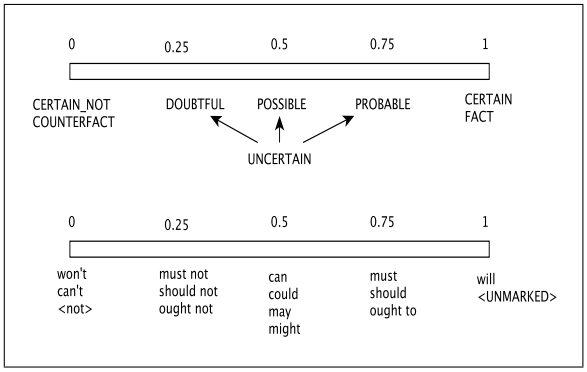

Figure 2. The epistemic scale with characteristic values and corresponding modal auxiliaries.

The SCALE_SHIFTER subcategory of valence shifters also plays a role in factuality assessment. Predicates belonging to this category change the scalar modality value of the predications in their scope. The subtypes of this category are NEGATOR, INTENSIFIER, DIMINISHER, and HEDGE. A DIMINISHER predicate (e.g., *hardly*) lowers the modality value, while an INTENSIFIER increases it (e.g., *strongly*). On the other hand, a negation marker belonging to the NEGATOR category (e.g., *no* in *no indication*) inverts the modality value of the embedded predication. The HEDGE category contains *attribute hedges* (e.g., *mostly*, *in general*) (Hyland,

1998), whose effect is to make the embedded predication more vague. We model this by decreasing, increasing or leaving unchanged the modality value depending on the position of the embedded predication on the scale.

Lexical and semantic knowledge about predicates belonging to embedding categories are encoded in a dictionary, which currently consists of 987 predicates, 544 of them belonging to MODAL and 95 to SCALE_SHIFTER categories. A very preliminary version of this dictionary was introduced in Kilicoglu and Bergler (2008). It was later extended and refined using several corpora and linguistic classifications (including Saurí (2008) and Nirenburg and Raskin (2004)). Since predicates collected from external resources do not neatly fit into embedding categories and we target deeper levels of meaning distinctions, the dictionary construction involved a fair amount of manual refinement. The dictionary encodes the lemma and part-of-speech of the predicate as well as its extra-propositional meaning senses. Each sense consists of five elements:

1. *Embedding category*, such as ASSUMPTIVE.
2. *Prior scalar modality value* (if any).
3. *Embedding relation classes* indicate the semantic dependencies used to identify the logical object argument of the predicate.
4. *Scope type* indicates whether the predicate allows a wide or narrow scope reading (for example, in *I don't think that P*, because *think* allows narrow scope reading, the negation is transferred to its complement (*I think that not P*)).
5. *Argument inversion* (true/false) determines whether the object and subject arguments should be switched in semantic interpretation.

The entry in Table 1 indicates that the modal auxiliary *may* is associated with two modal senses (i.e., it is ambiguous) with differing scalar modality values. It also indicates that a predication embedded by SPECULATIVE *may* will be assigned the epistemic value of 0.5 initially. *Scope type* and *argument inversion* attributes are not explicitly given, indicating default values for each.

Lemma	*may*	
POS	*MD* (modal)	
Sense.01	Category	SPECULATIVE
	Scalar modality value	0.5
	Embedding rel. classes	*AUX*
Sense.02	Category	PERMISSIVE
	Scalar modality value	0.6
	Embedding rel. classes	*AUX*

Table 1. Dictionary entry for *may*.

3.3 Composition

Semantic composition is the procedure of bottom-up predication construction using the knowledge encoded in the dictionary and syntactic information in the form of dependency relations. Dependency relations are extracted using the Stanford CoreNLP toolkit (Manning et al., 2014). We use the Stanford collapsed dependency format (de Marneffe et al., 2006) for dependency relations. We illustrate the salient steps of this procedure on a sentence from the GENIA event corpus (sentence 9 from PMID 10089566 shown in row (1) in Table 2). For brevity, the simplified version of the sentence is given in row (2), in which textual spans are substituted with the corresponding event annotations.

As the first step in the procedure, the syntactic dependency graphs of sentences of a document are combined and transformed into a semantically enriched, directed, acyclic semantic document graph through a series of dependency transformations. The nodes of the semantic graph correspond to textual units of the document and the direction of the arcs reflects the direction of the *semantic dependency* between its endpoints. The transformation is guided by a set of rules, illustrated on row (3). For example, the first three transformations are due to the Verb Complex Transformation rule, which reorders the dependencies that a verb is involved in such that semantic scope relations with the auxiliaries and other verbal modifiers are made explicit. The resulting semantic dependencies on the right indicate that *involve* is within the scope of *not*, which in turn is in the scope of *may*, and the entire verb complex *may not involve* is within the scope of *thus*, which indicates a discourse relation.

The next steps of the compositional algorithm, *argument identification* and *scalar modality value*

(1)	*Thus HIV-1 gp41-induced IL-10 up-regulation in monocytes may not involve NF-kappaB, MAPK, or PI3-kinase activation, but rather may operate through activation of adenylate cyclase and pertussis-toxin-sensitive Gi/Go protein to effect p70(S6)-kinase activation.*	
(2)	*Thus* E_{27} *may not* E_{32}, E_{33}, *or* E_{34}, *but rather may* E_{39} *and* E_{40}.	
(3)	*advmod(involve,thus)* *aux(involve,may)* *neg(involve,not)* *prep_of(activation,cyclase)*	*ADVMOD(thus,may)* *AUX(may,not)* *NEG(not,involve)* *PREP_OF(activation,adenylate cyclase)*
(4)	*involve*:CORRELATION*(*E_{32}*,WR,0.5*$_{epistemic}$*,*E_{27}*,* E_{28}*)* *not*:NEGATOR*(*EM_{53}*,WR,0.5*$_{epistemic}$*,*E_{32}*)* *may*:SPECULATIVE*(*EM_{57}*,WR,1.0*$_{epistemic}$*,*EM_{53}*)* *operate*:REGULATION*(*E_{39}*,WR,0.5*$_{epistemic}$*,*E_{38}*,*E_{27}*)* *may*:SPECULATIVE*(*EM_{58}*,WR,1.0*$_{epistemic}$*,*E_{39}*)*	

Table 2. Composition example for 10089566:S9.

composition, play a role in factuality assessment[1]. *Argument identification* is the process of determining the logical arguments of a predication, based on the bottom-up traversal of the semantic graph. It is guided by *argument identification rules*, each of which defines a mapping from a lexical category and an embedding class to a logical argument type. Such a rule applies to a predicate specified in the dictionary that belongs to the lexical category and serves as the head of a semantic dependency labeled with the embedding relation class. With argument identification rules, we determine, for example, that the second instance of *may* in the example, has as its logical object, the predication indicated by *operate*, since there is an *AUX* embedding relation between *may* and *operate*, which satisfies the constraint defined in the embedding dictionary (Table 1).

Scalar modality value composition is the procedure of determining the relevant scale for a predication and its modality value on this scale. The following principles are applied:

1. Initially, every predication is assigned to EPISTEMIC scale with the value of 1 (i.e., a fact).

2. A MODAL predicate places its logical object on the relevant MODAL scale and assigns to it its prior scalar modality value, specified in the dictionary.

3. A SCALE_SHIFTER predicate does not introduce a new scale but changes the existing scalar modality value of its logical object.

4. The scalar influence of an embedding predicate (*P*) extends beyond the predications it embeds to another predication in its scope (Pr_e), if one of the following constraints is met:

   - *P* is associated with the epistemic scale and the intermediate predications (Pr_i) are either of SCALE_SHIFTER type or are associated with epistemic scale
   - *P* is of SCALE_SHIFTER type and at most one intermediate predication is of MODAL type
   - *P* is of a non-epistemic MODAL type and Pr_i all belong to SCALE_SHIFTER type

Assuming that we have a predicate *P* which indicates an embedding predication *Pr* and a predication (Pr_e) under its scalar influence, the scalar modality value of Pr_e is updated differently, based on whether the predicate *P* is a MODAL or a SCALE_SHIFTER predicate. All update operations used for MODAL predicates are given in Table 3 and those for SCALE_SHIFTER predicates in Table 4. For MODAL predicates, the composition is modeled as the interaction of the prior scalar modality value of the embedding predicate ($\mathrm{MV}_{Sc}(\mathrm{P}_{modal})$) in the first column and the current scalar modality value associated with the embedded predication ($\mathrm{MV}_{Sc}(\mathrm{Pr}_e)$) in the second column, resulting in the value shown in

[1]The compositional steps that we do not discuss here are *source propagation* and *argument propagation*.

the third column ($MV_{Sc}(Pr_e)'$). When *P* is a scale-shifting predicate, the update procedure is guided by its type, as illustrated in Table 4. *X* and *Y* represent arbitrary values in the range of [0,1].

	$MV_{Sc}(P_{modal})$	$MV_{Sc}(Pr_e)$	$MV_{Sc}(Pr_e)'$
(1)	= X	= 1.0	X
(2)	= X	= 0.0	1-X
(3)	$>$ Y	$>$ 0.5 $\wedge$ = Y	min(0.9, Y+0.2)
(4)	$<$ Y $\wedge >=$ 0.5	$>$ 0.5 $\wedge$ = Y	min(0.5,Y-0.2)
(5)	$<$ 0.5	$>$ 0.5 $\wedge$ = Y	1-Y
(6)	$>=$ 0.5	$<$ 0.5 $\wedge$ = Y	Y
(7)	$<$ 0.5	$<$ 0.5 $\wedge$ = Y	1- Y

Table 3. The composition of scalar modality values in MODAL contexts.

For the example shown in Table 2, the computation in row (1) of Table 3 applies when we encounter the SPECULATIVE *may* node dominating the *operate* node in the semantic graph: since $MV_{epistemic}$(*may*)=0.5 and *operate* at the time of composition has epistemic value of 1, its scalar modality value gets updated to 0.5.

	Type	$MV_{Sc}(Pr_e)$	$MV_{Sc}(Pr_e)'$
(1)	NEGATOR	= 0.0	0.5
(2)	NEGATOR	$>$ 0.0 $\wedge$ = Y	1-Y
(3)	INTENSIFIER	(= 0.0 $\vee$ = 1.0) $\wedge$ = Y	Y
(4)	INTENSIFIER	$>=$ 0.5 $\wedge$ = Y	min(0.9,Y+0.2)
(5)	INTENSIFIER	$<$ 0.5 $\wedge$ = Y	max(0.1,Y-0.2)
(6)	DIMINISHER	(= 0.0 $\vee$ = 1.0) $\wedge$ = Y	Y
(7)	DIMINISHER	$>=$ 0.5 $\wedge$ = Y	max(0.5,Y-0.2)
(8)	DIMINISHER	$<$ 0.5 $\wedge$ = Y	max(0.4,Y+0.2)
(9)	HEDGE	= 0.0	0.2
(10)	HEDGE	= 1.0	0.8
(11)	HEDGE	= Y	Y

Table 4. The composition of scalar modality values in SCALE_SHIFTER contexts.

When *not*, a NEGATOR, is encountered in composition, the scalar modality value of its embedded predication (*involve*) is updated to 0, due to row (2) in Table 4 (1-1=0). In the next step of composition, when the first instance of SPECULATIVE *may* is encountered, the nodes in its scope, *not* and *involve*, have epistemic values of 1 and 0, respectively. The scalar modality value of *not* gets updated to *min(0.5,1-0.2)=0.5* (row (4) in Table 3). Row (2) in Table 3 applies to *involve*, resulting in 0.5 as its new epistemic value (1-0.5).

Row (4) in Table 2 shows the annotations generated by the system. The system takes as input GENIA event annotations (e.g., CORRELATION and REGULATION), which we expand with scalar modality values and sources. For example, E_{32}..E_{34}, three events triggered by *involve* and annotated as CORRELATION events in GENIA, have epistemic value of 0.5 and WR as the source (only one of the events, E_{32}, is shown for brevity). The system also generates other embedding predications (indicated with *EM*) corresponding to fine-grained extra-propositional meaning. To clarify, the content of first three predications (first an event and the latter two extra-propositional) are expressed in natural language below:

- E_{32}: Correlation between gp41-induced IL-10 upregulation and NF-kappaB activation is POSSIBLE according to the author.
- EM_{53}: That there is no correlation between gp41-induced IL-10 upregulation and NF-kappaB activation is POSSIBLE according to the author.
- EM_{57}: That it is possible there no correlation between gp41-induced IL-10 upregulation and NF-kappaB activation is a FACT according to the author.

3.4 Data and Evaluation

We assessed our methodology on the meta-knowledge corpus (Thompson et al., 2011), in which GENIA events are annotated with certainty levels (CL) and polarity. This corpus consists of 1000 MEDLINE abstracts and contains 34,368 event annotations. Uncertainty is only annotated in this dataset for events with Analysis knowledge type. Such events correspond to 17.6% of the entire corpus. Of all Analysis events, 33.6% are annotated with L2 (high confidence), 11.4% with L1 (low confidence), and 55% with L3 (certain) CL values. Polarity, on the other hand, is annotated for all events (6.1% negative).

Factuality values are often modeled as discrete categories (e.g., PROBABLE, FACT). Thus, to evaluate our approach, we converted the scalar modality values associated with predications (MV_{Sc}) to discrete CL and polarity values using mapping rules, shown in Table 5. The rules were based on the analysis of 100 abstracts that we used for training.

Condition	Annotation
$\text{MV}_{epistemic} = 0 \vee \text{MV}_{epistemic} = 1$	L3
$\text{MV}_{epistemic} > 0.6 \wedge \text{MV}_{epistemic} < 1$	L2
$\text{MV}_{epistemic} > 0 \wedge \text{MV}_{epistemic} <= 0.6$	L1
$\text{MV}_{potential} > 0$	L2
$\text{MV}_{interrogative} = 1 \vee \text{MV}_{intentional} = 1$	L1
$\text{MV}_{epistemic} = 0 \vee \text{MV}_{potential} = 0 \vee \text{MV}_{success} = 0$	Negative

Table 5. Mapping scalar modality values to event certainty and polarity.

We evaluated CL mappings in two ways: a) we restricted it only to Analysis type events, the only ones annotated with L1 and L2 values, and b) we evaluated them on the entire corpus. For polarity, we only considered the entire corpus. As evaluation metrics, we calculated precision, recall, and F_1 score as well as accuracy on the discrete values we obtained by the mapping.

Another evaluation focused more directly on factuality. We represented the gold CL-polarity pairs as numerical values and calculated the average distance between these values and those generated by the system. The lower the distance, the better the system can be considered. In this evaluation scheme, annotating a considerably speculative (L1) event as somewhat speculative (L2) is penalized less than annotating it as certain (L3). We mapped the gold annotations to the numerical values as follows: L3-Positive $\rightarrow$ 1, L2-Positive $\rightarrow$ 0.8, L1-Positive $\rightarrow$ 0.6, L1-Negative $\rightarrow$ 0.4, L2-Negative $\rightarrow$ 0.2, L3-Negative $\rightarrow$ 0.

4 Results and Discussion

The results of mapping the system annotations to discrete values annotated in the meta-knowledge corpus are provided in Table 6.

When the CL evaluation is limited to Analysis events, we obtain an accuracy of approximately 82%. The baseline considered by Miwa et al. (2012) is the majority class, which would yield an accuracy of 55% for these events. Their CL evaluation is not limited to Analysis events, and they report F_1 scores of 97.6%, 66.5%, and 74.9% for L3, L2, and L1 levels, respectively, on the test set. Restricting the system to Analysis events, we obtain the results shown at the top of the table (86.2%, 73.3%, and 79.6%). Lifting the Analysis restriction, we obtain the results shown at the bottom (97.5%, 66.6%, and 68.1% for L3, L2, and L1, respectively). The results are very similar for L3 and L2 levels, while our system somewhat underperformed on L1. With respect to negative polarity, our system performed poorly (49.4% vs. Miwa et al.'s 63.4%), while the difference was minor for positive polarity (97.5% vs. 97.7%).

Type	Precision	Recall	F_1	Accuracy
CL evaluation limited to Analysis events				
CL				81.75
L3	78.43	95.57	86.15	
L2	90.65	61.46	73.25	
L1	83.22	76.28	79.60	
Evaluation on the entire test set				
CL				95.13
L3	97.27	97.74	97.51	
L2	73.08	61.55	66.61	
L1	61.42	76.28	68.05	
Polarity				95.32
Positive	95.99	99.15	97.54	
Negative	74.17	37.04	49.41	

Table 6. Evaluation results.

In comparing to Miwa et al.'s results, several points need to be kept in mind. First, in contrast to their study, we have not performed any training on the corpus data, except determining the mapping rules shown in Table 5. Secondly, knowing whether an event is an Analysis event or not is a significant factor in determining the CL value and their machine learning features are likely to have exploited this fact, whereas we did not attempt to identify the knowledge type of the event. Thirdly, L1 and L2 values appear only for Analysis events, therefore the evaluation scenario that only considers Analysis events is likely to overestimate the performance of our system on L1 and L2 and underestimate it on L3.

While our system performed similarly to Miwa et al.'s with regards to positive polarity, our mappings for negative polarity were less successful, which suggests that modeling negative polarity as the lower end of several modal scales (the last row of Table 5) may not be sufficient for correctly capturing the polarity values. Our preliminary analysis of the results indicate that scope relationships between predications could play a more significant role. In other words, whether an event is in the scope of a predication trigger by a NEGATOR predicate may be a better predictor of negative polarity.

With the evaluation scheme that is based on average distance, we obtained a distance score of 0.12. For the majority class baseline, this score would be 0.21. Our score shows clear improvement over the baseline; however, it is not directly comparable to Miwa et al.'s results. This evaluation scheme, to our knowledge, has not previously been used to evaluate factuality and we believe it is better suited to the gradable nature of factuality.

Analyzing the results, we note that many errors are due to problems in dependency relations and transformations that rely on them. Errors in dependency relations are common due to complexity of the language under consideration, and these errors are further compounded by hand-crafted transformation rules that can at times be inadequate in capturing semantic dependencies correctly. In the following example, the prepositional phrase attachment error caused by syntactic parsing (*to suppress...* is attached to the main verb *result*, instead of to *ability*) prevents the system from identifying the semantic dependendency between *ability* and *suppress*, causing a L2 recall error. While the system uses a transformation rule to correct some prepositional phrase attachment problems, this particular case was missed.

- *The reduction in gene expression resulted from the ability of IL-10 to suppress IFN-induced assembly of signal transducer ...*
- *prep_to(result,suppress)* vs. *prep_to(ability,suppress)*

Prior scalar modality values in the dictionary have been manually determined and are fixed. They are able to capture the meaning subtleties to a large extent and the composition procedure attempts to capture the meaning changes due to markers in context. However, some uncertainty markers are clearly more ambiguous than others, leading to different certainty level annotations in similar contexts and our method may miss these differences due to the fixed value in the dictionary. For example, the adjective *potential* has been almost equally annotated as an L1 and L2 cue in the meta-knowledge corpus. This also seems to confirm the finding of de Marneffe et al. (2012) that world knowledge and context have an effect on the interpretation of factuality.

We also noted what seem like annotation errors in the corpus. For example, in the sentence *L-1beta stimulation of epithelial cells did not generate any ROIs*, the event expressed with *generation of ROIs* seems to have negative polarity, even though it is not annotated as such in the corpus.

5 Conclusion

We presented a rule-based compositional method for assessing factuality of biological events. The method is linguistically motivated and emphasizes generality over corpus-specific optimizations, and without making much use of the corpus for training, we were able to obtain results that are comparable to the performance of the state-of-the-art systems for certainty level assignments. The method was less successful with respect to polarity assessment, suggesting that the hypothesis that negative polarity can be modeled as corresponding to the lower end of the modal scales may be inadequate. In future work, we plan to develop a more nuanced approach to negative polarity.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13 Suppl 11:S4.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Pro-*

ceedings of the 5th International Conference on Language Resources and Evaluation, pages 449–454.

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

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 — Shared Task*, pages 1–12.

Ken Hyland. 1998. *Hedging in scientific research articles*. John Benjamins B.V., Amsterdam, Netherlands.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9 Suppl 11:s10.

Halil Kilicoglu and Sabine Bergler. 2010. A High-Precision Approach to Detecting Hedges and Their Scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 70–77.

Halil Kilicoglu and Sabine Bergler. 2011. Effective Bio-Event Extraction using Trigger Words and Syntactic Dependencies. *Computational Intelligence*, 27(4):583–609.

Halil Kilicoglu. 2012. *Embedding Predications*. Ph.D. thesis, Concordia University.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13 Suppl 11:S1.

Marc Light, Xin Y. Qiu, and Padmini Srinivasan. 2004. The language of bioscience: facts, speculations, and statements in between. In *BioLINK 2004: Linking Biological Literature, Ontologies and Databases*, pages 17–24.

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.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics*, pages 992–999.

Makoto Miwa, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, 13:108.

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

Roser Morante, Vincent van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 40–47.

Sergei Nirenburg and Victor Raskin. 2004. *Ontological Semantics*. The MIT Press, Cambridge, MA.

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

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.

Roser Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

Pontus Stenetorp, Sampo Pyysalo, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Bridging the gap between scope-based and event-based negation/speculation annotations: A bridge not too far. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 47–56.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *JAMIA*, 17(5):514–518.

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

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356.