Annotating Event Mentions in Text with Modality, Focus, and Source Information

Suguru Matsuyoshi[†], Megumi Eguchi[†], Chitose Sao[†], Koji Murakami[†], Kentaro Inui[‡,†], Yuji Matsumoto[†]

[†]Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

[‡]Graduate School of Information Science,
Tohoku University
6-6-05, Aoba, Aramaki, Aoba-ku, Sendai, Miyagi, 980-8579, Japan

{matuyosi, megumi-e, chitose-s, kmurakami, matsu}@is.naist.jp, inui@ecei.tohoku.ac.jp

Abstract

Many natural language processing tasks, including information extraction, question answering and recognizing textual entailment, require analysis of the polarity, focus of polarity, tense, aspect, mood and source of the event mentions in a text in addition to its predicate-argument structure analysis. We refer to modality, polarity and other associated information as extended modality. In this paper, we propose a new annotation scheme for representing the extended modality of event mentions in a sentence. Our extended modality consists of the following seven components: Source, Time, Conditional, Primary modality type, Actuality, Evaluation and Focus. We reviewed the literature about extended modality in Linguistics and Natural Language Processing (NLP) and defined appropriate labels of each component. In the proposed annotation scheme, information of extended modality of an event mention is summarized at the core predicate of the event mention for immediate use in NLP applications. We also report on the current progress of our manual annotation of a Japanese corpus of about 50,000 event mentions, showing a reasonably high ratio of inter-annotator agreement.

1. Introduction

In Natural Language Processing (NLP), development of syntactic and semantic parsers is a major task, and recently we have seen the development of precise POS taggers, dependency parsers, and predicate-argument structure analyzers. Identifying predicate-argument structure in a sentence is important but insufficient for applications such as information extraction (IE), question answering (QA) and recognizing textual entailment (RTE). These applications require analyzing the polarity, focus of polarity, tense, aspect, mood and source of the event mentions in a text in addition to predicate-argument structure analysis. For example, the verb "submitted" in sentence (1) has three arguments "Ann", "the plan" and "the comittee", which can be identified by predicate-argument structure analysis, and the verb and these arguments correspond to an event mention "Ann submitting the plan to the comittee" (underlined in sentence (1)). The modality toward the event mention shows John's assertion that the event actually happened. The verb "cause" and its arguments "mercury-based vaccines" and "autism in children" underlined in sentence (2) correspond to an event mention, toward which the modality indicates the doctor's inference that the underlined event does not happen.

(1) John claimed that Ann **submitted** the plan to the committee.

(2) The doctor speculated that mercury-based vaccines did not **cause** autism in children.

Distinguishing assertion and inference is essential for NLP applications such as IE and RTE, because information asserted is obviously much more reliable than information inferred. We refer to the modality, polarity and other associated information of an event mention in a given sentence as *extended modality*. Extended modality covers almost all the components of modality in a broad sense, as described in detail in section 3, and is important for interpreting a writer's attitude toward event mentions.

A sentence may have several event mentions, though it does not include any connectives such as "and" and "although". For example, sentence (3) has three event mentions, where the core predicates are "decide(d)", "stop" and "buy(ing)".

(3) Jim **decided** to **stop buying** that weekly magazine.

Modality in a narrow sense (hereafter referred as restricted modality) and polarity can be assigned to each event mention, even if the event mention is not the main proposition in the sentence. For example, the restricted modality of an event "Jim buying that weekly magazine" can be regarded as volition due to indirect effect of "decided" and its polarity negative due to direct effect of "stop". We would like to recognize extended modality of such event mentions as well as the main event mention. It is because they are dependent on the main event in the sentence but can be regarded as distinct events in applications such as IE and RTE. For interpreting a writer's attitude toward such event mentions, their extended modality should be analyzed.

In this paper, as a first step toward constructing an analyzer of extended modality, we propose a new annotation scheme for representing extended modality of event mentions in a sentence, and report on the current progress of manual annotation of a Japanese corpus of about 50,000 event mentions.

2. Related work

A writer's attitude toward event mentions is mainly represented with restricted modality. In English, restricted modality has two basic categories; one is Propositional modality, which consists of Epistemic modality and Evidential modality, and the other is Event modality which

	certainty	transition of certainty	evaluation	modality in a narrow sense	polarity	focus	source	time	conditional
(Light et al., 2004)	√								
(Rubin et al., 2005)	√			√			√	√	
(Saurí et al., 2006)	√	√		√	√			√	√
(Prasad et al., 2006)	√			√	√	√	√		
(Saurí and Pustejovsky, 2007)	√				√		√		
(Medlock and Briscoe, 2007)	√								
(Szarvas et al., 2008)	√				√				
(Saurí, 2008; Saurí and Pustejovsky, 2009)	√	√		√	√		√	√	√
(Hara and Inui, 2008; Inui et al., 2008)	√	√		√	√		√	√	√
(Kawazoe et al., 2009)	√				√		√		√
(Im et al., 2009)	√	√		√	√			√	√
Our work	√	√	√	√	√	√	√	√	√

Table 1: Components of extended modality considered in our work and related works.

is divided into Deontic modality and Dynamic modality (Palmer, 2001). Other important categories of modality are Future, Negative, Interrogative, Imperative (Jussive), Presupposed, Conditional, Purposive, Resultative, Desiderative, and Fears. Modal logic with possible worlds has been studied to capture the logical properties of modal expressions strictly (Portner, 2009). By using well-defined accessibility relations or conversational backgrounds, the modal logic represents various modal expressions while differentiating between ones that have almost the same meaning. Classifications of restricted modality in Linguistics and systems of modal logic are helpful for constructing a system of extended modality in our research, but they are unable to be adopted directly. We give three reasons for this. The first one is that these works mainly focus on classification of modal expressions while we focus on classification of event mentions based on their extended modality. The second reason is that the main targets in these works are modality classes of the main proposition in a sentence, and we want to deal with the extended modality of both the main proposition in a sentence and the event mentions embedded in that proposition, as mentioned in Section 1. The third reason is that classifications in these works are too fine-grained to identify automatically. We will describe our annotation scheme of extended modality in Section 3.

In recent years, there has been increasing attention paid to annotating phrases and event mentions in text with extended modality in the fields of bioinformatics and NLP. We show a list of related work with considered components of extended modality in Table 1.

Studies in bioinformatics (Light et al., 2004; Medlock and Briscoe, 2007; Szarvas et al., 2008) mainly focus on certainty toward event mentions in text, because they want to distinguish between asserted propositions and propositions with inferential expressions and hedges. For example, Szarvas et al. mark hedging expressions, such as "may" and "possible", or negations, such as "not" and "neither", and their scopes in text (Szarvas et al., 2008). They do not mark the focus of a negation, even if it is partial negation.

Related studies in NLP are roughly divided into two groups: One is for an annotation scheme of extended modality (Saurí et al., 2006; Prasad et al., 2006; Saurí, 2008; Saurí and Pustejovsky, 2009; Kawazoe et al., 2009; Im et al., 2009) and the other is for automatically analyzing extended modality in text (Rubin et al., 2005; Saurí and Pustejovsky, 2007; Hara and Inui, 2008; Inui et al., 2008). A pioneering work for an annotation scheme is Saurí et al.'s FactBank (Saurí and Pustejovsky, 2009). In FactBank, an event mention is annotated with its source, introduced by Wiebe et al. (2005), epistemic modality and polarity for representing event factuality, along with attribute values for tense, aspect, modality and polarity provided by a "MAKEINSTANCE" element in TimeML (Saurí et al., 2006). A factuality value of FactBank is represented as a combination of values at epistemic modality axis ("certain (CT)", "probable (PR)", "possible (PS)" and "underspecified (U)") and polarity axis ("positive (+)", "negative (−)" and "underspecified (u)"). For example, an event in text is labeled with "CT+" when it is certain that the event happened or will happen according to the source of the text. An event in text is labeled with "PR−" when it is propable that the event did not happen or will not happen according to the source of the text. The factuality category from FactBank is practical for applications such as IE, QA and RTE, because it summarizes information of whether a target event actually happens or not along with degree of certainty in a way that is easily accesible for applications. However, the framework of FactBank is not sufficient for extended modality of event mentions in text because FactBank relies on a "MAKEINSTANCE" element in TimeML (Saurí et al., 2006) for restricted modality except for epistemic modality. In TimeML, restricted modality is specified at "modality" attribute of the "MAKEINSTANCE" element with a surface auxiliary verb, such as "SHOULD" and "MUST". This scheme cannot be directly applied to agglutinative languages such as Korean and Japanese, because they tend to have complex systems of auxiliary verbs and sequences of auxiliary verbs. In fact, Im et al. have proposed an extended version of TimeML for the Korean language (KTimeML) (Im et al., 2009). Their KTimeML has the advantage of handling sequences of auxiliary verbs, but cannot easily be used to obtain the extended modality of each event mention in text because pieces of information about extended modality toward an event mention in a sentence spread over

several XML tags in the same sentence, and analyzing the extended modality requires extra effort.
For automatically analyzing extended modality in text, Saurí et al. have proposed a rule-based method using information of polarity particles, modality markers and epistemic predicates (Saurí and Pustejovsky, 2007). In their algorithm, values of epistemic modality and polarity toward events in a sentence are determined by the upper factuality values and rules one by one from the top of a dependency tree to the lowest level. Inui et al. (2008) have proposed a method of analyzing restricted modality and polarity of event mentions in Japanese text with a conditional random fields model. Implemention of such an analyzer for extended modality is an area of future work.

3. Proposed scheme

In this section, we propose an annotation scheme of extended modality of event mentions in a sentence. In this research, we work with English and Japanese, but we believe that our annotation scheme is applicable to other languages such as Spanish and Korean.

3.1. Definition of an event mention

We define an event mention in text as follows:

> consisting of a core predicate and its arguments (complements and adjuncts) in the sentence.

A predicate is a verb, "be" + an adjective or "be" + a noun. For example, the verb "ate" in sentence (4) is the core predicate of an event mention "Mary eating a cake with her favorite fork yesterday". In sentence (5), "is effective" is the core predicate of an event mention "xylitol being effective at preventing tooth decay".

(4) Mary **ate** a cake with her favorite fork yesterday.

(5) Xylitol **is effective** at **preventing** tooth decay.

Because nominalized verb phrases and adjective phrases should be extracted as expressions consisting of predicate argument structures in applications such as IE and RTE, we regard these phrases (with their arguments) as event mentions. For example, we regard "xylitol preventing tooth decay" as an event mention in sentence (5), where a nominalized verb "prevention" is the core predicate.

3.2. Our extended modality

For constructing a system of our extended modality of event mentions in a sentence in consideration of NLP applications, we have the following four desiderata:

1. Information of the extended modality of an event mention should be gathered into one piece, specifically at the core predicate.
2. A system of extended modality should be language-independent.
3. The polarity of event mentions should be divided into two distinct classes: polarity on actuality and polarity from the view of the source's evaluation.
4. Labels in each component of extended modality should not be too fine-grained.

Desideratum 1. facilitates use in NLP applications, because spreading of pieces of information about the extended modality toward an event mention over several points requires taking extra effort in working out the summarized extended modality by considering labels or surface forms of several modality expressions related to the event mention. Desideratum 2. means that we aim for a language-independent scheme of extended modality, such as the scheme of semantic role labeling in The CoNLL-2009 Shared Task (Hajič et al., 2009). We determine Desideratum 3. because we believe that division of polarity into the above two classes is important for explicitly capturing the factuality of an event in text. A good example for description of this is a subjunctive mood. There are two event mentions "I studying mathematics harder" and "I passing the examination" in sentence (6).

(6) If I had **studied** mathematics harder, I could have **passed** the examination.

We can see that both of the events did not occur, so polarity values on actuality of them are negative. In the meantime, it appears that the writer of sentence (6) wanted the latter event "I passing the examination" to occur and therefore he/she evaluated occurence of the event as "positive". We refer to such polarity from the view of the source's evaluation as subjective polarity. Polarity on actuality and subjective polarity should not be grouped into a single "polarity" attribute in a system of extended modality, because they should be treated in a different way in NLP applications. Desideratum 4. indicates that a scheme of extended modality should be abstract to the extent of application-independence. We take the position that each application may subdivide our fundamental classification if necessary. This desideratum also comes from the fact that classifications of restricted modality in Linguistics, e.g., (Palmer, 2001) and systems of modal logic (Portner, 2009) are too sophisticated, and it is very difficult to implement analyzers of extended modality based on them with the current level of technology in NLP.

As important components of extended modality of an event mention, we take *Source* and *Focus* information because they are deeply associated with modality and useful for applications such as IE and RTE.

We reviewed the literature about extended modality in Linguistics, such as (Palmer, 2001; *Nihongo kizyutsu bumpou kenkyukai*, 2003; *Nihongo kizyutsu bumpou kenkyukai*, 2007; Masuoka, 2007) and NLP, such as (Wiebe et al., 2005; Prasad et al., 2006; Saurí, 2008; Inui et al., 2008), and made an annotation scheme of extended modality for event mentions in a sentence. Our extended modality consists of the following seven components:

> *Source* (S), *Time* (T), *Conditional* (C),
> *Primary modality type* (P), *Actuality* (A),
> *Evaluation* (E), and *Focus* (F).

In the rest of this subsection, we describe each component of our extended modality and labels in the component[1]. Due to space limitation, the components' names are represented by their initial letters respectively in tables where labeled examples are listed.

3.2.1. Source

Source expresses an agent or an organization that takes an attitude toward an event mention in a sentence. This information helps a reader judge credibility of contents conveyed from a given sentence. We specify nested sources introduced by Wiebe et al. (2005) in this component. When a certain agent, other than the writer of a sentence, is clearly identified as the source, *Source* of the target event mention is "wr_(the agent)", as in the example with ID=1 in Table 4. In the case where the agent is represented with a pronoun, the *Source* of the event mention is "wr_ot", as in the example with ID=2 in the same table. The *Source* of a target event mention is "wr_arb" when the event mention is an overheard statement, as in the example with ID=3 in the same table. In the case of none of the above, the writer of the sentence is regarded as the only source of a target event mention, and its *Source* is "wr", as in the other examples in the same table.

3.2.2. Time

Time shows relative time of occurrence of a target event on the base of the time when the source took an attitude toward the event mention. *Time* can be "future" or "notFuture". This information, together with *Actuality* described in subsubsection 3.2.5., indicates whether factuality of the event is by nature fixed or not. For example, suppose that the writer of a sentence conjectures about occurrence of an event. In this case, "future" of *Time* means that the writer selected conjecture because factuality of the event is not fixed by nature, as in the example with ID=4 in Table 4. On the other hand, "notFuture" of *Time* means that the writer selected conjecture because factuality of the event is fixed and he/she does not know the factuality, as in the example with ID=5 in the same table. Thus, *Time* is not the tense of the core predicate of a target event mention.

3.2.3. Conditional

Conditional conveys whether a target event mention is a proposition with a condition. This information is useful for applications such as IE and RTE, because a proposition with a condition should be distinguished from one with no condition in these applications. We give *Conditional* the value "condition" for event mentions that exist in a conditional clause, as in the example[2] with ID=6 in Table 4, and "hasCondition" to event mentions that exist in the main clause of a conditional sentence, as in the example with ID=7. In the case of none of the above, *Conditional* of an event mention is "notConditional".

3.2.4. Primary modality type

Primary modality type represents a primary category that determines the fundamental meaning of a event mention. *Primary modality type* is assigned a language-independent category label derived from restricted modality, i.e. "assertion", "volition", "wish", "imperative", "permission" or "interrogative", but, unlike TimeML (Saurí et al., 2006), it is not assigned a surface auxiliary verb. Table 4 shows examples of event mentions annotated with these labels (especially, see examples with ID=2, 7, 1, 8, 9, 10, 11 and 12.). On a trial basis, "imperative" is divided into three sub-classes: "imperative-direct", "imperative-indirect" and "imperative-together".

3.2.5. Actuality

Actuality expresses degree of certainty toward an event mention in text and transition of certainty toward it. This informaiton consists of epistemic modality and polarity on actuality, and so corresponds roughly to factuality in FactBank (Saurí and Pustejovsky, 2009). We have the following five labels that represent degree of certainty: "certain+", "certain−", "probable+", "probable−" and "unknown". We use "probable+" and "probable−" in the cases of "possible+" and "possible−", which are labels in FactBank, respectively, because variety of modality expressions in Japanese makes it difficult to distinguish consistently between "probable+" and "possible+" and between "probable−" and "possible−". Table 4 shows examples of event mentions annotated with labels of *Actuality*. Adding to the above five labels, we use the following four labels for capturing aspect, especially inchoative aspect and terminate aspect, of a target event mention: "certain− → +", "certain+ → −", "probable− → +" and "probable+ → −". For example, "certain− → +" expresses transition of certainty (from "certain−" to "certain+") toward a target event mention, as in the example with ID=13 in Table 4. The example with ID=14 in the same table is annotated with "certain+ → −" that expresses transition of certainty (from "certain+" to "certain−"). Capturing transition of certainty is important for RTE, because, for example, "certain+ → −" of the *Actuality* of an event mention entails that the event happened in the past and will not happen in the future.

3.2.6. Evaluation

Evaluation indicates subjective polarity toward a target event mention in a sentence. As mentioined previously, we divide polarity of event mentions into polarity on actuality and subjective polarity. The former class is essential for NLP applications that focus on facts. Therefore, we handle polarity of this class as an element of *Actuality*. On the other hand, the latter class is essential for NLP applications that focus on opinions, such as opinion mining and sentiment analysis. We handle polarity of the class in this component of extended modality, but not in *Actuality*. *Evaluation* can be "positive", "negative" or "neutral". When the source evaluate occurence of a target event mention as positive, *Evaluation* is "positive", as in the examples with

[1] A document, written in Japanese, about description in considerable detail of our extended modality is available at the following URL. `http://cl.naist.jp/nltools/modality/manual.pdf`

[2] As an exception, in order to express degree of certainty, we determine the *Actuality* of an event mention in a conditional clause, by extracting the event mention from the clause and regarding it as a stand-alone sentence.

ID=1, 7, 8 and 15 in Table 4. When he/she evaluates it negatively, *Evaluation* is "negative", as in the examples with ID=9, 14 and 16 in the same table. In the case of absence of a source's evaluation in the sentence, *Evaluation* of the event mention is "neutral".

3.2.7. Focus

Focus represents the focus of negation, inference or interrogative. This information of an event mention in text indicates a statement entailed by the event mention. For example, the event "he staying for you" in the sentence with ID=17 in Table 4 did not happen, but the sentence entails the realization of the event "he staying". In order to express this entailment clearly, we specify "$negation$(for you)" to *Focus* of the event mention. The label means that the focus of negation is on the phrase "for you" in the event mention. Example sentences including foci of inference and interrogative are the ones with ID=5 and 12 in the same table, respectively. On a trial basis, as in the scheme (Prasad et al., 2006), we also specify focus on a connective in a sentence to *Focus*, as in the examples with ID=18, 19 and 20 in the same table.

4. Annotated corpus

4.1. Construction

We have constructed a corpus of sentences annotated with labels in our scheme of extended modality described in the previous section. We selected Japanese as a target language. Our corpus has 50,108 event mentions, which come from the following four resources:

(A) Blog posts (19,237 event mentions / 5,687 sentences)

Blog posts collected over about six months.

(B) Web Documents (4,401 event mentions / 4,401 sentences)

Documents retrieved using query phrases from the Web.

(C) The corpus from Murakami et al. (2009) (13,527 event mentions / 2,878 sentences)

Sentences extracted from the Web for the purpose of annotating pairs of sentences with semantic relations.

(D) Posts from Q&A sites (12,943 event mentions / 5,432 sentences)

Posts from Yahoo! JAPAN Q&A sites that are included in *the Balanced Corpus of Contemporary Written Japanese*[3].

We defined an event mention in text in Subsection 3.1. Unfortunately, there is no system that automatically identifies all event mentions in Japanese text with high precision. So, we took the following two strategies for annotating event mentions:

1. We identify only the core predicate of an event mention, not the arguments explicitly, as in the case of "MAKEINSTANCE" element in TimeML (Saurí et al., 2006). We assign labels of extended modality for the event mention to the core predicate.
2. We overdetect event mention candidates from text and then filter out pseudo event menions from the set of the candidates manually.

First, we analyzed each sentence in the above resources using morphological and dependency analyzers and picked out event mention candidates in the sentence using several syntactic patterns, such as a verb, a nominalized verb, an adjective, a noun + "*da*", and a verbal noun (*sahen* noun) + "*suru*". Next, we asked an annotator, who is a native speaker of Japanese, to exclude pseudo event mentions from the candidates. Pseudo event mentions mainly arise from grammaticalized verbs, metaphors, and errors of the above analyzers. Then, she annotated each event mention with our labels of extended modality, where she assigned the labels to the core predicate of the event mention. Table 4 shows examples annotated with our labels, where the core predicates of target event mentions are indicated by boldface.

Table 2 shows a distribution of labels of each component of extended modality in the corpus. In this table, every component has a major label whose frequency in the corpus is 80% or over. In analyzing extended modality, though, it is important to recognize the remaing minor labels properly for IE and opinion mining.

We examined inter-annotator agreement on annotating event mentions with labels of our extended modality. We extracted 300 annotated event mentions randomly from the corpus, and asked another annotator, who is a native speaker of Japanese, to independently annotate these event mentions without knowledge of the previous labels. Then, we worked out the κ statistics for components of extended modality which are shown in Table 3. Figures in the table, whose average is 0.71, indicate a reasonably high ratio of inter-annotator agreement for our annotation scheme.

4.2. Challenges in actual annotation

We faced the following challenges in actual annotation.

Adverbs representing frequency Event mentions including adverbs of frequency accompanied with "*nai*" (not), such as "*metta-ni nai*" (not very often), were problematic for specifying *Actuality*. We decided to neglect the degree of frequency and assign "certain+" to *Actuality* in these cases.

Tough constructions We do not have appropriate labels for an event mention in tough construction, such as "*X-nikui*" (it is tough to X) and "*X-yasui*" (it is easy to X). We assign "probable−" or "probable+" to *Actuality* of such event mention tentatively.

Potential We do not have a label "potential" in *Primary modality type* component of extended modality. Probably, "assertion" in this component should be divided into "assertion-standard" and "assertion-potential".

5. Concluding remarks

In this paper, we proposed a new annotation scheme of extended modality consisting of the following seven components: *Source*, *Time*, *Conditional*, *Primary modality type*,

[3] http://www.tokuteicorpus.jp/

Actuality, *Evaluation* and *Focus*. Based on this scheme, we constructed an annotated corpus of about 50,000 event mentions in Japanese[4]. Averaged κ statistics for components of extended modality is 0.71 that shows a reasonably high ratio of inter-annotator agreement between two annotators.

Our future work contains the following two tasks: one is to apply our annotation scheme to different kinds of resources for assessing the scheme and extending our corpus; the other is to implement an analyzer of our extended modality with high precision, using machine learning approaches, such as support vector machines and conditional random fields, and the corpus as the training data.

Acknowledgment

This work was partly supported by Japan MEXT Grant-in-Aid for Young Scientists (Start-up) (No. 20800029) and Grant-in-Aid for Scientific Research on Priority Areas (No. 21013036), and by National institute of information and Communications of Technology Japan.

6. References

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Kazuo Hara and Kentaro Inui. 2008. Factuality analysis for event extraction. In *IPSJ SIG Notes*, volume 2008-NL-183, pages 75–80. (in Japanese).

Seohyun Im, Hyunjo You, Hayun Jang, Seungho Nam, and Hyopil Shin. 2009. KTimeML: Specification of temporal and event expressions in Korean text. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 115–122.

Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. 2008. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 314–321.

Ai Kawazoe, Manabu Saito, Kiyoko Kataoka, and Daisuke Bekki. 2009. A preliminary study for constructing japanese corpus annotated for certainty and uncertainty. In *IPSJ SIG Notes*, volume 2009-NL-189, pages 77–84. (in Japanese).

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases*, pages 17–24.

Takashi Masuoka. 2007. *A study on Japanese modality*. Kurosio Publishers. (in Japanese).

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

Koji Murakami, Shouko Masuda, Suguru Matsuyoshi, Eric Nichols, Kentaro Inui, and Yuji Matsumoto. 2009. Annotating semantic relations combining facts and opinions. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 150–153.

Nihongo kizyutsu bumpou kenkyukai, editor. 2003. *Modern Japanese Grammar volume 4*. Kurosio Publishers. (in Japanese).

Nihongo kizyutsu bumpou kenkyukai, editor. 2007. *Modern Japanese Grammar volume 3*. Kurosio Publishers. (in Japanese).

Frank Robert Palmer. 2001. *Mood and Modality Second edition*. Cambridge University Press.

Paul Portner. 2009. *Modality*. Oxford University Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. Annotating attribution in the Penn discourse treebank. In *the COLING/ACL Workshop on Sentiment and Subjectivity in Text*, pages 31–38.

Victoria Rubin, Elizabeth Liddy, and Noriko Kando, 2005. *Chapter 7: Certainty Identification in Texts: Categorization Model and Manual Tagging Result*, pages 61–74. Springer-Verlag New York.

Roser Saurí and James Pustejovsky. 2007. Determining modality and factuality for text entailment. In *the International Conference on Semantic Computing*, pages 509–516.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. In *Language Resources and Evaluation*.

Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky, 2006. *TimeML Annotation Guidelines Version 1.2.1*. `http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf`.

Roser Saurí, 2008. *FactBank 1.0 Annotation Guidelines*. `http://www.cs.brandeis.edu/~roser/pubs/fb_annotGuidelines.pdf`.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation 39 issue 2-3*, pages 165–210.

[4]We plan to distribute this corpus freely at our web site.

Resource		(A)	(B)	(C)	(D)
Number of event mentions		19,237(100%)	4,401(100%)	13,527(100%)	12,943(100%)
Component	**Label**				
Source	wr	19,154(100%)	4,371(100%)	13,288(98%)	12,796(99%)
	wr_ot	72(0%)	9(0%)	18(0%)	33(0%)
	wr_arb	4(0%)	3(0%)	121(1%)	68(1%)
	wr_*STRING*	7(0%)	18(0%)	100(1%)	46(0%)
Time	future	1,500(8%)	349(8%)	2,051(15%)	1,849(14%)
	notFuture	17,737(92%)	4,052(92%)	11,476(85%)	11,094(86%)
Conditional	condition	667(3%)	142(3%)	600(5%)	986(8%)
	hasCondition	56(0%)	2(0%)	172(1%)	279(2%)
	notConditional	18,514(97%)	4,257(97%)	12,755(94%)	11,678(90%)
Primary modality type	assertion	18,194(94%)	4,110(93%)	12,735(94%)	10,462(81%)
	volition	408(2%)	111(3%)	322(2%)	265(2%)
	wish	269(1%)	25(1%)	55(1%)	288(2%)
	imperative-direct	96(1%)	20(0%)	29(0%)	436(3%)
	imperative-indirect	143(1%)	58(1%)	269(2%)	380(3%)
	imperative-together	29(0%)	17(0%)	17(0%)	10(0%)
	permission	4(0%)	0(0%)	14(0%)	10(0%)
	interrogative	94(1%)	60(2%)	86(1%)	1,092(9%)
Actuality	certain+	16,498(86%)	3,879(88%)	11,578(85%)	8,838(68%)
	certain−	1,036(5%)	126(3%)	637(5%)	894(7%)
	certain− → +	0(0%)	0(0%)	20(0%)	48(0%)
	certain+ → −	2(0%)	0(0%)	11(0%)	21(0%)
	probable+	951(5%)	187(4%)	669(5%)	722(6%)
	probable−	95(1%)	29(1%)	94(1%)	125(1%)
	probable− → +	1(0%)	0(0%)	13(0%)	6(0%)
	probable+ → −	1(0%)	0(0%)	12(0%)	2(0%)
	unknown	653(3%)	180(4%)	493(4%)	2,287(18%)
Evaluation	positive	903(5%)	192(4%)	677(5%)	1,249(10%)
	negative	78(0%)	34(1%)	101(1%)	233(2%)
	neutral	18,256(95%)	4,175(95%)	12,749(94%)	11,461(88%)
Focus	$negation$(*P1*)	0(0%)	2(0%)	7(0%)	8(0%)
	$negation$(*P1*;*P2*)	1(0%)	0(0%)	3(0%)	0(0%)
	$inference$(*P1*)	19(0%)	9(0%)	10(0%)	31(0%)
	$inference$(*P1*;*P2*)	0(0%)	0(0%)	12(0%)	0(0%)
	$interrogative$(*P1*)	20(0%)	27(1%)	44(0%)	263(2%)
	$interrogative$(*P1*;*P2*)	1(0%)	0(0%)	1(0%)	0(0%)
	no	19,196(100%)	4,363(99%)	13,450(100%)	12,641(98%)

Table 2: Distribution of labels of each component of extended modality in our corpus. [(A) Blog posts, (B) Web documents, (C) The corpus from Murakami et al. (2009), (D) Posts from Q&A sites]

Source	*Time*	*Conditional*	*Primary modality type*	*Actuality*	*Evaluation*	*Focus*
0.69	0.76	0.68	0.66	0.70	0.72	0.75

Table 3: κ statistics for components of extended modality.

ID	Sentence [The core predicate of the target event mention is indicated by boldface.] and labels
1	*hayaku ie-ni **kaeri**-tai-to Taro-ga itta.* (Taro said he wanted to **go** home soon.) S=wr_Taro, T=future, C=notConditional, P=wish, A=unknown, E=positive, F=no
2	*saisho-wa datsu-sute-no shozyo-ni **kurusin**-da-soudesu.* (He said he **suffered** from symptoms due to stopping steroid medicines at that time.) S=wr_ot, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=no
3	*touyaku-wa sanhankikan-no kino-wo antei-sa-seru-tameni **tsudukeru**-noda-souda.* (I hear that the mediciation **is continued** for regulating functions of three semicircular canals.) S=wr_arb, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=no
4	*kuron-gizyutsu-no hattatsu-niyoru jinkozoki-ga **shutsugen-suru**-no-wa jikan-no mondai-da-to omou.* (It is only a matter of time before progress of cloning technology leads to **producing** artificial organs.) S=wr, T=future, C=notConditional, P=assertion, A=probable+, E=neutral, F=no
5	*osoraku seigo-2-kagetsu-mae-kurai-kara suteroidozai-no **shiyo**-wo shi-te-ta-to omoware-masu.* (I guess he has **been** on steroids since a month or two after birth.) S=wr, T=notFuture, C=notConditional, P=assertion, A=probable+, E=neutral, F=$inference$(a month or two after birth)
6	*ashita **hare-tara**, mizumi-ni sakana-tsuri-ni iko-to omou.* (If it **is** nice out tomorrow, I will go fishing in that lake.) S=wr, T=future, C=condition, P=assertion, A=certain+, E=neutral, F=no
7	*ashita hare-tara, mizumi-ni sakana-tsuri-ni **iko**-to omou.* (If it is nice out tomorrow, I will **go** fishing in that lake.) S=wr, T=future, C=hasCondition, P=volition, A=probable+, E=positive, F=no
8	*kangensui-no motsu koka-wo **jikkanshi**-te-mi-te-kudasai!* (**Feel** for yourself the effects of restoration water!) S=wr, T=future, C=notConditional, P=imperative-direct, A=unknown, E=positive, F=no
9	*konki-ga nai kata-wa, te-wo **dasa**-nai hou-ga bunan-desu.* (A person with no patience had better not **try** it.) S=wr, T=future, C=notConditional, P=imperative-indirect, A=unknown, E=negative, F=no
10	*soba-wo tabe-ni **iki**-mase-n-ka?* (Would you like to **go** eating soba noodles with me?) S=wr, T=future, C=notConditional, P=imperative-together, A=unknown, E=positive, F=no
11	*okii tsukue-wo **tsukat**-temo-ii-desu-yo.* (You may **use** a larger desk.) S=wr, T=future, C=notConditional, P=permission, A=unknown, E=positive, F=no
12	*soredewa, kishiritoru-ni-wa donna kouka-ga **aru**-no-desho-ka?* (Then, how **is** xylitol effective?) S=wr, T=notFuture, C=notConditional, P=interrogative, A=unknown, E=neutral, F=$interrogative$(how)
13	*sorede, Taro-wa sono hamigakiko-wo **tsukai** hajime-ta-no-desu.* (So, Taro began to **use** the toothpaste.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain$-\rightarrow+$, E=neutral, F=no
14	*Jim-ha shukanshi-no **kodoku**-wo tori-yameru koto-ni-shi-ta.* (Jim decided to stop **buying** the weekly magazine.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain$+\rightarrow-$, E=negative, F=no
15	*anata-wa sono toki-ni kanojo-ni shinjitsu-wo **tsutaeru**-beki-dat-ta.* (You should have **told** the truth to her at that time.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain$-$, E=posiive, F=no
16	*kare-ga kuru-to shit-te-i-tara, pati-ni **ika**-nakat-ta-noni.* (If I had known he would come to the party, I would not have **been** there.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain+, E=negative, F=no
17	*kare-wa kimi-no-tame-ni **nokot**-ta-no-de-wa-nai.* (It was not for you that he **stayed**.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain$-$, E=neutral, F=$negation$(for you)
18	*kusuri-wo non-da-kara genki-ni **nat**-ta wake-de-wa-nai.* (It is not because he took the medicine that he **recovered**.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=$negation$(because;took)
19	*kyunyu-suteroidozai-wa koukateki-ni kyunyu-dekiru-node, hayaku kouka-ga **de**-ta-no-de-wa-nai-ka.* (I guess that inhaled steroids **worked** so quickly because they can be inhaled effectively.) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=$inference$(because;be inhaled)
20	*Taro-wa, eiyo-wo tot-ta-kara, genki-ni **nat**-ta-no-desu-ka?* (Is it because Taro received appropriate nutrition that he **recovered**?) S=wr, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=$interrogative$(because;received)

Table 4: Examples annotated with labels in our scheme of extended modality. (S: *Source*, T: *Time*, C: *Conditional*, P: *Primary modality type*, A: *Actuality*, E: *Evaluation*, F: *Focus*)