Semantic Annotation of Japanese Functional Expressions and its Impact on Factuality Analysis

Yudai Kamioka[1] **Kazuya Narita**[1] **Junta Mizuno**[2] **Miwa Kanno**[1] **Kentaro Inui**[1,3]
[1] Graduate School of Information Sciences, Tohoku University / Miyagi, Japan
[2] Resilient ICT Research Center, NICT / Miyagi, Japan [3] JST, RISTEX
{yudai.k, narita}@ecei.tohoku.ac.jp junta-m@nict.go.jp {meihe, inui}@ecei.tohoku.ac.jp

Abstract

Recognizing the meaning of functional expressions is essential for natural language understanding. This is a difficult task, owing to the lack of a sufficient corpus for machine learning and evaluation. In this study, we design a new annotation scheme and construct a corpus containing 2,327 Japanese sentences and 8,775 functional expressions. Our scheme achieves high inter-annotator agreement with kappa score of 0.85. In the experiments, we confirmed that machine learning-based functional expression analysis contributes to factuality analysis.

1 Introduction

In natural language, many expressions are used to convey information beyond the propositional content of the sentence, such as modality and polarity. Understanding such information is essential for natural language understanding.

The extra-propositional aspects of meaning are often expressed by function words and their combinations. For example, consider the following sentence:

(1) パソコンが壊れ てしまったかもしれない。
(*My computer may have been broken.*)

Three expressions are used to add extra information to the propositional content 壊れ (*break*): function words てしまっ (means it is unintentional), た (*have been*) and かもしれない (*may*) mean UNINTENTIONAL, COMPLETION, and UNCERTAIN, respectively.

Some function words such as た are used alone, and others are combined to express their meaning, such as てしまっ and かもしれない. We call the former a "function word," and the latter a "compound functional expression (CFE)." These are collectively called "functional expressions" (FEs) in this paper. Recognizing the meaning of FEs is useful for various natural language processing tasks, such as factuality analysis, machine translation, and question answering. However, two main issues cause difficulties in FE analysis. First, because FEs are usually expressed with multiple tokens, we must resolve the chunking problem. Second, FEs indicating different meanings can have the same surface form. For example, ている is used to indicate CONTINUOUS in 食べ ている ところです (*now eating*) and used to indicate HABIT in いつも歌っている (*always sing*).

In Japanese, there is no corpus large enough for machine learning and evaluation. Matsuyoshi et al. (2006) first built a dictionary of Japanese FEs named *Tsutsuji*. Imamura et al. (2011) reported that this dictionary lacks many expressions. Therefore, we designed a new scheme for annotating FE meanings, and constructed a corpus containing 2,327 sentences and 8,775 FEs. In this scheme, we reorganize a dictionary of FEs on the basis of *Tsutsuji*. Our scheme and corpus are especially compatible with factuality analysis. We selected factuality analysis as our application, because it provides verifiable evidence to confirm the importance of FEs. Using the annotations of actual text, we investigate the problems associated with FE annotation. We also verified the effect of our corpus and FE analysis on factuality analysis. Our contributions are three fold:

(1) we introduce a new annotation scheme for Japanese FEs;

(2) we build a Japanese FE corpus with high inter-

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 52–61,
Denver, Colorado, June 5, 2015.

annotator agreement;

(3) we demonstrate that improvements in FE analysis contribute to factuality analysis.

2 Related Work

Previous research efforts have addressed the problem of disambiguating functional and content usage. Tsuchiya et al. (2005) reported that more than 50% of the most frequent 180 CFEs contain ambiguities between functional and content usage. Tsuchiya et al. (2006) and Utsuro et al. (2007) used support vector machines (SVMs) for chunking, and showed that a machine learning model had advantages over a rule-based model. Suzuki et al. (2012) disambiguated functional and content usage using an example-based system.

Surprisingly, NLP research has paid insufficient attention to recognizing the meaning of FEs. Tsuchiya et al. (2005) constructed a Japanese CFE corpus. However the corpus focused only on a restricted range of expressions and is insufficient for machine learning. Matsuyoshi et al. (2006) organized a hierarchical Japanese FE dictionary, named *Tsutsuji*. *Tsutsuji* contains more than 16,000 FEs, which are categorized into 89 classes based on linguistic dictionaries. While *Tsutsuji* covers a wide range of FEs and their derivations, Imamura et al. (2011) reported that some expressions are not included. Some CFEs are contained in a dictionary of multiword expressions (Shudo et al., 2011). For example, とはいえ is included as "however."

In English, some research efforts have addressed the problem of modality and factuality. Saurí and Pustejovsky (2012) defined a list of modal words such as *perhaps* and *probably* for the factuality analysis. Szarvas et al. (2008) produced the BioScope corpus, which consists of biomedical texts annotated with negation and uncertainty, and their scopes. Diab et al. (2009) classified the writer's belief into three categories (committed belief, non committed belief, or not applicable). Diab et al. manually annotated the 10,000 words covering different domains and genres, and achieved high inter-annotator agreement of 95%. de Marneffe et al. (2012) used list of modal words and linguistic markers of negative contexts such as *no* and *any*, to automatically distribute event veridicality. Incorporating information about modality and negation has been shown to be useful for a wide range of applications. For example, Harabagiu et al. (2006) used negative markers such as *n't* as classifier features to recognize contradictions between two texts. Baker et al. (2010) showed the structure-based modality tagger improved the machine translation.

3 Annotation Scheme Design

3.1 Aims of Annotation

With the aim of creating a corpus for FE analysis, we designed an annotation scheme. The goal was to annotate its meanings to Japanese FEs. Because we are planning to use annotated labels in application tasks such as factuality analysis and FE analysis, the annotation scheme should be compatible with many applications.

3.2 Design Procedure

In the linguistics field, the meanings of FEs have been extensively researched. For example, Morita and Matsuki (1989) collected and categorized CFEs and provided explanations using an abundance of examples. As for the field of NLP, Matsuyoshi et al. (2006) provided an electronically-processable dictionary of Japanese FEs named *Tsutsuji*. *Tsutsuji* was composed according to linguistic dictionaries. There are many expressions that *Tsutsuji* lacks, because it has not been annotated for any actual texts.

We designed our annotation scheme by beginning with the semantic type categories defined in *Tsutsuji* and improving each category and entry where necessary. To be more precise, we added FEs that were not included in *Tsutsuji* but should have been. We also added and segmentalized some categories that were not appropriate for the application tasks. We used 1,627 sentences as development data, and alternated designing our scheme and annotating the corpus. A series of process was repeated several times while we carefully analyzed the feedback from the factuality analyzer described in Section 6. The following sections describe the problems encountered during the scheme's design phases, and how they were addressed.

3.3 Functional Expressions

Because different research efforts have adopted slightly different definitions of the term functional expression, we now clarify our definition. In this research, we define FEs as functional words and their combinations. Function words are non-content words; in terms of parts-of-speech (POS), they are

categorized as particles and auxiliary verbs in the Japanese POS Tagset[1]. In the phrase 読みたい (*want to read*), for example, たい is categorized as an auxiliary verb and means WISH. These are the counterparts of the modal verbs (i.e., *might, will*) and verbs in English. Treating these words as FEs is common in linguistics research.

We define some FEs as compound functional expressions (CFEs), which are expressions whose meaning cannot be derived from their components. For example, かもしれない contains three words and means UNCERTAIN. The meaning of UNCERTAIN comes only after three words are combined; however none of the three words have the meaning of UNCERTAIN. We define such multiword expressions whose meanings are clear only after their components are combined, as CFEs.

Some CFEs are composed only of function words, and some contain content words. For example, ではない is composed from three function words で, は and ない. This expression means NEGATION when its components are combined. In another case, かもしれない is composed of function words かも and ない, and contentive しれ (*know*). However, the verb しれ (*know*) in かもしれない has no meaning as a verb, and the complete expression means UNCERTAIN. We consider these expressions to be a type of FE, even if some of the components are categorized in contentive. Function words and CFEs are collectively called functional expressions (FEs) in this paper.

3.4 Category Redesign

We categorized the meanings of FEs by referring to *Tsutsuji*. Because some categories were not compatible with application tasks, we added and segmentalized some of them. For example, かもしれない (*possibly*) and だろう (*probably*) are categorized as SPECULATION in *Tsutsuji*. However, these are actually different in the following aspects: 食べるだろう (*probably eat*) has more certainty than 食べるかもしれない (*possibly eat*). This fact is useful when determining the author's degree of conviction. Thus, we segmentalized these categories into different categories.

In another example, ている is categorized only as CONTINUOUS in *Tsutsuji*. This expression actually means continuation, however it sometimes means past experience: 歩いている (*be walking*) means continuation of 歩い (*to walk*), 指摘している (*pointed out*) means past experience. This fact will have an effect on the task of temporal relation analysis. Therefore, we introduced some new categories such as EXPERIENCE to annotate appropriate labels to these expressions. As a result, meanings of Japanese FEs are classified into 72 categories in our annotation scheme. Note that the number of categories is less than that of *Tsutsuji* because we left some FEs in *Tsutsuji* out of consideration in our scheme. Some FEs, such as が and を, have no information that is useful to us, as they are related more closely to predicate-argument structure.

4 Corpus Annotated with FE

We constructed a Japanese corpus annotated with the semantic labels of FEs based on the annotation scheme we developed. All labels were annotated using the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[2]. We selected texts categorized in Yahoo! Answers in terms of usefulness, and because they were annotated with Extended Modality Tags (Matsuyoshi et al., 2010). The Extended Modality Tags contain *Actuality*, which can be used as a gold standard for factuality analysis. At this time, 2,327 out of 6,323 sentences in BCCWJ have been annotated. The guideline and corpus are available on `http://tinyurl.com/ja-fe-corpus`.

4.1 Labels

Labels are annotated at the token level. To annotate CFEs, we employed the IOB2 format (Sang, 2000) to express the range of FEs, and we used the label P for predicates. An example is shown in Table 1.

Label	Description
P	Predicates
B	Head of FE
I	Inner of FE
O	Otherwise

Token	Label
壊れ	P
て	B-UNINTENTIONAL
しまっ	I-UNINTENTIONAL
た	B-COMPLETION
かも	B-UNCERTAIN
しれ	I-UNCERTAIN
ない	I-UNCERTAIN

Table 1: Labels used in the corpus. (Chunk labels (left) and an example of actual labels (right))

4.2 Annotation

Our corpus is composed of a development set and test set. The development set contains 1,627 sen-

[1] `http://sourceforge.jp/projects/ipadic/docs/ipadic-2.7.0-manual-en.pdf/en/1/`

[2] `http://www.ninjal.ac.jp/corpus_center/bccwj/`

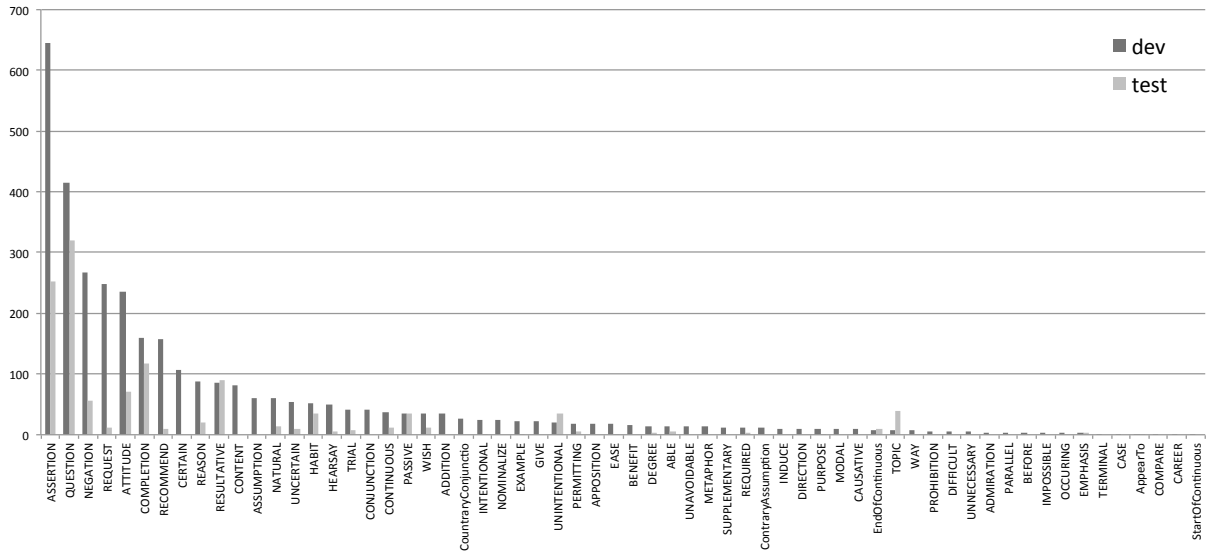

Figure 1: Distribution of semantic labels in the head clauses of the development set and the test set. (The development set contains all FEs in the sentence, while the test set contains FEs only in the head clauses. Only the labels agreed between two annotators were counted in the test set.)

tences randomly sampled from 6,323 sentences. We then labeled the 4,696 sentences using the development set as training data, and then sampled 700 sentences that contained more than three FE tokens in the head clause. We use these 700 sentences as a test set, in order to evaluate our annotation scheme and to conduct an open test.

The development set was annotated by a native Japanese speaker. For the development set, the annotator encountered issues with the original guidelines; after discussing the problems, we updated our guidelines accordingly and used the resulting guidelines for our development set. We outsourced the annotation of our test set to two other annotators, who were provided with our latest annotation guidelines and the list of FEs from the development set. To reduce time and cost, only the expressions in the head clauses were annotated in the test set, while all expressions in the development set were annotated.

The annotation procedure is as follows: i) The annotator is given the token sequence of a sentence. ii) The annotator selects a predicate that precedes an FE and annotates it with the predicate label (P). iii) On each FE, chunking labels are annotated for both head and inner chunks. iv) The most appropriate semantic label in the annotation scheme is annotated. Note that the annotator used only predicates and FEs to determine the semantic label. If the annotator could not locate an appropriate label in our scheme, the closest label was annotated and a comment was provided. The entire procedure was conducted using a spreadsheet, and we constructed a corpus from the resulting data.

4.3 Corpus Statistics

Statistical information of the constructed corpus is shown in Table 2. The results show that the percentage of CFEs in the development set was 74%, and 67% in both test sets. These percentages were significantly higher than our expectations; and extracting CFEs correctly is a crucial problem that we must resolve. Note that the number of FEs is much lower than that of *Tsutsuji*; this is because some of the FEs listed in *Tsutsuji* are infrequent and thus not found in the corpus.

	dev		test	test
	all	head	by worker1	by worker2
Sent.	1,627		700	700
FE	5,993 (584)	3,407 (317)	1,378 (179)	1,404 (187)
CFE	1,861 (437)	577 (216)	697 (121)	710 (126)
Label	67	62	37	34

Table 2: Corpus Statistics. (FEs and CFEs are noted in brackets)

As for the labels that appeared in the corpus, the development set contained 62 labels in the head clauses, and the test set contained roughly half of

that. This is because the development set contained infrequent semantic labels which did not appear in the test set. Fig. 1 shows the distribution of FE semantic labels in the head clauses of development and test set. Some labels, such as TOPIC and UNINTENTIONAL, were frequent only in the test set. This was caused by the difference of annotator between the development set and the test set. More precise guidelines will reduce these differences. Conversely, labels such as NEGATION and REQUEST were infrequent in the test set although they appeared frequently in the development set. This is because we sampled sentences for the test set, depending on the number of FE tokens in the head clauses. Some FEs consist of less than three FE tokens did not appear in the test set.

In the development set, 106 of 584 FEs contain semantic ambiguity. These expressions are likely to be labeled with more than two types of labels, and they should be disambiguated during FE analysis. The number of newly added FEs is 485. Because we only added FEs which appeared in the corpus, some FEs and their derivations must still be added.

4.4 Reported Issues

The development set was annotated according to the conditions in Section 4.2, and all FEs were annotated completely. However, the test set annotators reported some issues with annotation. The following section describes two of them.

4.4.1 Variation of functional expressions

One of the biggest problems is that many FEs have a variety of derivations. During the annotation of the development set, we treated these derived expressions the same as base forms, and we indicated this to the test set annotators. While we thought it would be easy for native Japanese speakers to identify the derivations, the annotator reported some confusion when attempting to determine whether two expressions were the same. For example, one annotator reported that he was confused when attempting to determine whether ばよろしい had the same meaning as ばよい. In reality, these were slightly different in their degree of politeness; however, the proposed scheme could not identify the difference. It was confusing because the scheme indicated that they were the same; however, ばよろしい is not in the known lists. To resolve this problem, we should complement the list or create more precise guidelines with additional derivation pattern definitions.

4.4.2 Difference between contentives and functional expressions

The second problem involves difficulties in deciding whether a token is a predicate. As we mentioned previously, some contentives lose their original meaning and can be components of compounds. For example, in したら良いのでしょうか, there are two content words: し and 良い. Our annotation scheme defines たら良い as one of the FEs; therefore, し should be annotated as a predicate and 良い is the inner FE. However, it was difficult for our annotator because 良い slightly maintains its meaning as a contentive. This example shows that our definition of the differences between contentives and FEs was not specific enough.

4.5 Annotator Agreement

To evaluate our annotation scheme, we tested two types of inter annotator agreements using the data in the test set; two outsourced annotators were employed for the evaluation. Note that the annotations in the test set were only performed on the head clauses. To evaluate the inter-annotator agreements, we employed kappa statistics and calculated three different agreements: predicate agreement, chunk agreement, and semantic label agreement. Predicate agreement shows whether two different annotators agree on the location of a head clause. Chunk agreement shows whether they agree on the beginning and ending locations of the expressions, and is calculated according to the predicate location agreed upon by both annotators. If the predicate and chunk locations are agreed to by both annotators, we then calculate the semantic label agreement according to the choice of FE semantic type. Table 4 lists the kappa results, which show very high values for all three agreements. While detecting predicate position is a difficult problem, we achieved very high agreement because of the restricted annotation range. Because we are planning to create a corpus in which all predicates and FEs are annotated, predicate agreement should be calculated once again after all instances are annotated. Table 4 also shows a kappa score of .97 for chunking. This suggests extremely high agreement. Once predicate positions are given, it should be relatively easy for a human annotator to detect the beginnings and ends of FEs.

Label	Precision	Recall	F
QUESTION	93.67(296/316)	94.59(297/314)	94.13
ASSERTION	92.86(247/266)	95.37(247/259)	94.10
COMPLETION	80.85(114/141)	93.44(114/122)	86.69
RESULTATIVE	54.32(88/162)	74.79(89/119)	62.93
HABIT	89.47(34/ 38)	40.00(34/ 85)	55.38
ATTITUDE	90.79(69/ 76)	88.46(69/ 78)	89.61
NEGATION	80.00(52/ 65)	70.27(52/ 74)	74.82
PASSIVE	94.87(37/ 39)	92.86(39/ 42)	93.85
CONTINUOUS	71.43(10/ 14)	25.64(10/ 39)	37.74
TOPIC	100.00(38/ 38)	97.44(38/ 39)	98.70
UNINTENTIONAL	82.93(34/ 41)	100.00(34/ 34)	90.67
RECOMMEND	76.92(10/ 13)	34.48(10/ 29)	47.62
REASON	100.00(21/ 21)	91.30(21/ 23)	95.45
WISH	100.00(12/ 12)	85.71(12/ 14)	92.31
NATURAL	92.31(12/ 13)	78.57(11/ 14)	84.89
REQUEST	66.67(10/ 15)	100.00(11/ 11)	80.00

Label	Precision	Recall	F
UNCERTAIN	100.00(10/ 10)	100.00(10/ 10)	100.0
EndOfContinuous	36.84(7/ 19)	100.00(9/ 9)	53.85
PERMITTING	100.00(5/ 5)	83.33(5/ 6)	90.91
TRIAL	100.00(6/ 6)	100.00(6/ 6)	100.0
ABLE	71.43(5/ 7)	100.00(5/ 5)	83.33
HEARSAY	100.00(5/ 5)	100.00(5/ 5)	100.0
REQUIRED	60.00(3/ 5)	75.00(3/ 4)	66.67
MANNER	100.00(4/ 4)	100.00(4/ 4)	100.00
AppearTo	50.00(2/ 4)	100.00(2/ 2)	66.67
NOMINALIZE	0.00(0/ 2)	0.00(0/ 2)	0.00
INTENTIONAL	100.00(2/ 2)	100.00(2/ 2)	100.00
CONTENT	0.00(0/ 1)	0.00(0/ 2)	0.00
PURPOSE	100.00(1/ 1)	50.00(1/ 2)	66.67
EXAMPLE	100.00(1/ 1)	100.00(1/ 1)	100.00
EASE	100.00(1/ 1)	100.00(1/ 1)	100.00
All labels	84.66(1142/1349)	83.31(1148/1378)	83.98

Table 3: Label-specific Inter Annotator Agreement. (Precision, Recall, and F-measure assuming worker 1 produces "gold data" and worker 2 produces system output. More details on each semantic label can be found in the annotation guidelines on the web site.)

	kappa
Predicate agreement	0.8508
Chunk agreement	0.9708
Semantic label agreement	0.8514

Table 4: Inter-Annotator Agreement (kappa)

To evaluate the semantic label agreements, we calculated the inter-annotator agreement in more detail; we treated one annotator's annotation results as "gold data," and the other annotator's results as system estimation, and evaluated F-measure. Table 3 shows the results of precision, recall, and F-measure calculations. Note that each annotator annotates different semantic label sets, and the resulting agreements differ depending on which annotator we treat as "gold." Because the differences between these two result sets are relatively miner, Table 3 shows only one of them. Semantic label-specific agreement shows that the label of CONTINUOUS and HABIT labels achieved the lowest scores. These labels contain ambiguity: each label was annotated to the same functional expression ている. These results show that determining such ambiguous labels is still difficult for native Japanese speakers.

5 FE Analysis

We evaluated our FE analysis system and verified how useful our scheme will be for actual tasks. We consider FE analysis as a sequence labeling problem. In our evaluation, we used the conditional random fields (CRF) method (Lafferty et al., 2001) because it is commonly applied to solve sequence labeling problems. We used CRFSuite (Okazaki, 2007) to implement the CRF model.

Dataset The closed test experiments were performed using 10-fold cross validation on the development set; the open tests were performed using the test set, with development set as training data.

Features The unigram and bigram features that were used included tokens, POS, and base forms. Note that POS is subdivided into four stages: we used each of them for unigrams, and only the first two stages for bigrams.

We used the longest match principle as a baseline when using the dictionary. The baseline uses the constraints for the preceding token's POS. Dictionary entries and constraints were collected from the development set. Furthermore, the system outputs the most frequent label in the development set if the expression takes more than one label.

We employed the standard evaluation metrics of precision, recall, and F-measures. Each metric was calculated by considering FEs as a unit. In other words, we accepted only the expressions in which a chunking labels (B and I) sequence matched correctly. Furthermore, we only evaluated BI sequences, because recognizing the compounds is one of the main problems in FE analysis. The entire experiment focused on only FEs, while contentives were disregarded.

The results are shown in Table 5. Every result indicates that the CRF model provides better results than the baseline. The table also shows that

Table 5: Results of FE analysis evaluation

		Method	Precision	Recall	F
Closed	Chunk	Baseline	94.91 (5257/5539)	86.50 (5184/5993)	90.51
		CRF	95.39 (5851/6134)	95.93 (5749/5993)	95.66
	Semanti label	Baseline	76.44(4234/5539)	70.43(4221/5993)	73.31
		CRF	79.83 (4897/6134)	81.18 (4865/5993)	80.50
	Chunk (only head clause)	Baseline	95.00 (2339/2462)	82.85 (2299/2775)	88.51
		CRF	93.96 (2689/2862)	94.77 (2630/2775)	94.36
	Semantic label (only head clause)	Baseline	79.37 (1954/2462)	70.09 (1945/2775)	74.44
		CRF	80.61 (2307/2862)	82.05 (2277/2775)	81.32
Open	Chunk	Baseline	83.42 (815/ 977)	58.49 (672/1149)	68.76
		CRF	91.49 (1053/1151)	92.08 (1058/1149)	91.78
	SemLabel	Baseline	53.33(521/ 977)	45.52(523/1149)	49.11
		CRF	77.32 (890/1151)	79.11 (909/1149)	**78.21**

CRF achieved a high score on chunking F-measure. These results show that it is easier than expected to detect compounds from an FE sequence. Conversely, the F-measure of semantic label estimation exceeded 80%. We analyzed outputs from the closed test to determine why the F-score was low.

(2) いつも読んでいる雑誌でもかまわない。
(*Magazines that you read all the time is okay.*)
(Gold: HABIT System: RESULTATIVE)

(3) 両親とも働いているのが条件です。
(*Working of both parents is required.*)
(Gold:CONTINUOUS, System:RESULTATIVE)

(4) 感情の高ぶりがよく描かれている。
(*The novel portrayed heightened emotion well.*)
(Gold, System: RESULTATIVE)

In (2) and (3), RESULTATIVE was labeled incorrectly; the answer shold have been HABIT and CONTINUOUS. (4) shows an example of FE correctly labeled as RESULTATIVE. These examples were ambiguous, and caused lower inter-annotator agreement. Therefore, we should improve our corpus to include more precise guidelines.

6 Factuality Analysis

To verify the practical effectiveness of our corpus for factuality analysis, we used a rule-based factuality analyzer based on FE semantic labels. We applied our factuality analyzer to 1,475 events to which FEs were attached in the head clauses of 1,627 sentences for the closed test, and to 650 events in the head clauses of 700 sentences for the open test[3]. We only selected events in head clauses because the factuality in subordinate clauses is determined not only by FEs, but also by other factors such as predicates.

In our corpus, extended modality is also annotated for each event mentioned by Matsuyoshi et al. (2010). The *actuality* of extended modality denotes the author's degree of certainty and corresponds to factuality. In this paper, by comparing the results of factuality analysis based on each of the four FE types, we show that annotating events with both FEs and factuality leads to some quantitative investigations such as i) how much effect our FE redesign has on factuality analysis, ii) how much does FE disambiguation contribute to factuality analysis, and iii) for how many events can we analyze factuality based on FEs.

FE I and II are the results of the longest matches using POS-attachment rules by *Tsutsuji* (Matsuyoshi et al., 2006) and using our dictionary. We investigate the effect of our label redesign by comparing the results based on FE I and II. In our corpus, FEs and their semantic labels are added to the dictionary *Tsutsuji*. We make comparisons based on gold data from *Tsutsuji* and our dictionary to investigate the strict effects of our label redesign. However, *Tsutsuji* does not provide gold data; therefore, we approximate the results of the longest matches. FE I and II cannot determine one semantic label for ambiguous FEs. Therefore, FE I and II allow ambiguous FEs for multiple semantic label such as "HEARSAY, UNCERTAIN, METAPHOR;" in the factuality analysis step, all effects of semantic labels are applied.

FE III is the result of the CRF shown in section 5. We investigate the contribution of FE disambiguation by comparing the results based on FE II and III.

[3]FE annotation and extended modality annotation have different criteria for judging events. 650 of 700 events in head clauses were judged as events in extended modality corpus, so we use 650 events for factuality analysis.

Table 6: The distribution of factuality values

	CT+	PR+	PR−	CT−	U	Total
Closed	476	215	51	107	626	1,475
Open	283	18	0	50	299	650

FE IV is gold annotation data. We investigate incorrect events based only on the FEs of the results based on FE IV. For the open test set, we conducted experiments using the gold data from two annotators.

6.1 Model

We use factuality values generated by combining certainty and polarity, as per Narita et al. (2013). They classify events into five factuality classes: CT+ (fact), PR+ (probable), PR− (not probable), CT− (counterfact), U (unknown or uncommitted), with reference to Saurí and Pustejovsky (2012). Table 6 shows the distribution of factuality values in our experiment. In the extended modality corpus, CT+ constitutes 68% of the total events (Matsuyoshi et al., 2010); however, in our experiment, U has the highest rate because events with FEs are selected.

Our analyzer determines event factuality by attaching FEs. For example, a NEGATION FE switches factuality to negative if it is positive, and vice versa. We constructed the following update rules and corresponding FE semantic labels for each rule:

A. polarity: +→ −, − →+
(NEGATION, IMPOSSIBLE, POINTLESS, UNNECESSARY, DIFFICULTY)

B. certainty: CT→PR
(UNCERTAIN, HEARSAY, INTENTIONAL, EASE, MODAL)

C. certainty: CT→U, PR→U
(QUESTION, REQUEST, WISH, RECOMMEND, INDUCE)

First, the factuality is set to CT+ as an initial value. Then, the analyzer identifies the attached FEs. If FEs that have update rules are found, the factuality is updated according to the rule. Rules of all attached FEs determine the factuality of the event.

Figure 2 shows the example of our model. The factuality of the event 進め (*work out*) is classified as PR− by the NEGATION FE たい and the UNCERTAIN FE みたい.

6.2 Discussion

Table 7 shows the evaluation results on different FE analysis. The open test shows higher performance

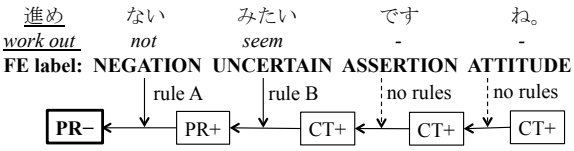

Figure 2: Applying our model to the sentence 進めないみたいですね (*It does not seem that you work out.*)

than the closed test because the open test set contains simpler events, as the frequency of PR− shows. We investigated the effect of our label redesign by comparing i) the longest-match results produced by our dictionary and *Tsutsuji*, ii) the contribution of FE disambiguation, which is obtained by comparing the CRF-based results and our dictionary, and iii) the incorrect events based only on FEs resulting from the gold data.

First, to investigate the effects of our label redesign, we compared the results of FE I and II. Table 7 shows that our label redesign improves factuality analysis.

(5) その方がより 分かり[やすい] です。
(*It is easier to understand.*)
(FE I: CT+, FE II: PR+, Gold factuality: PR+)

(5) is an improved example from our dictionary. No items for やすい are in *Tsutsuji*; therefore, adding EASE as the semantic label of the FE やすい provides an improvement.

Second, to investigate the contribution of FE disambiguation, we compared the results of FE II and III. Table 7 shows that FE disambiguation improves factuality analysis.

(6) 5階から落ちて 助かっ たんでした [よね]。
(*He survived a fall from the 5th floor.*)
(FE II: U, FE III: CT+, Gold factuality: CT+)

(6) is an improved example produced by CRF. The factuality of the event 助かっ (*survive*) is misclassified as U by our dictionary, because the FE よね is labeled as QUESTION. In contrast, CRF labels the FE よね as ATTITUDE based on context such as the COMPLETION FE た and period; therefore, so the factuality of the event 助かっ (*survive*) is correctly classified as CT+.

Finally, to investigate incorrect events based only on FEs, we evaluate the results based on FE IV. In the closed test set, approximately 40% of the events are incorrect despite the use of gold FEs. It shows that improvements in FE analysis are necessary, but

Table 7: Results of factuality analysis evaluation

	FE	Accuracy	Macro-Average		
			Precision	Recall	F_1
Closed	FE I: longest match by *Tsutsuji*	44.00 (649/1,475)	36.46	33.53	32.13
	FE II: longest match by our dictionary	54.51 (804/1,475)	50.70	44.28	46.56
	FE III: CRF	57.90 (854/1,475)	55.70	48.38	50.42
	FE IV: Gold data	61.90 (913/1,475)	56.71	54.58	54.59
Open	FE I: longest match by *Tsutsuji*	52.00 (338/650)	38.04	54.89	29.57
	FE II: longest match by our dictionary	66.46 (432/650)	50.34	61.96	50.86
	FE III: CRF	92.62 (602/650)	94.93	86.29	89.54
	FE IV: Gold data by annotator 1	94.62 (615/650)	97.14	92.83	94.76
	FE IV: Gold data by annotator 2	94.46 (614/650)	97.02	93.57	95.15

Table 8: Error type distribution

		output	
		CT+	others
FE	granularity of semantic labels	10	21
	annotation error of FEs	6	4
factuality	update rule of FEs: insufficient/misapply	9	2
	equivalent predicate of FEs	9	2
	preceding adverb/particle	4	5
	ellipse of FEs	5	0
	annotation error of factuality	3	14
Other (morphological analysis error, etc.)		4	2

not sufficient for factuality analysis. We conducted an error analysis to investigate other factors aside from FEs. Out of 562 errors, 149 events were misclassified as CT+; 413 events were misclassified into other classes. Table 8 shows the error type distribution in 50 events. Other contributing factors included predicates equivalent to FEs, adverbs, and particles. Update rules also remain controversial.

Furthermore, errors caused by the granularity of semantic labels were found.

(7) どうやって色を 判別し てる [んでしょうか]？
(*How does it discriminate between colors?*)
(FE IV: U, Gold factuality: CT+)

For example in (7), んでしょうか is the QUESTION FE; therefore, the factuality of the event 判別し (*discriminate*) is misclassified as U. However, this sentence presupposes that the event 判別し (*discriminate*) is fact, because the author asks how to *discriminate*. There are two methods to resolve the problem: One is to subcategorize semantics labels such as QUESTION into QUESTION-HOW; however, this might lead to a proliferation of labels. Another is to improve the factuality analyzer by considering the scope of FEs or other elements in the sentence.

Annotating events with both FEs and factuality led us to these quantitative investigations for factuality analysis. We showed that our corpus contributes to factuality analysis.

7 Conclusion

In this paper, we designed an annotation scheme for Japanese FEs and constructed a corpus annotated with FE semantic labels based on the scheme. The corpus achieved very high inter-annotator agreement. Our guidelines and the corpus are publicly available. Statistical analysis based on our corpus clarified ambiguous FEs and the distribution of semantic labels. We identified the issues regarding the ambiguity of FE analysis. For factuality analysis, annotating events with both FEs and factuality provided us with some quantitative investigations. We also experienced challenges in applying our corpus to wider areas.

In future work, we will consolidate annotation guidelines by referencing linguistic studies that focus on ambiguous FEs. Futhermore, to obtain better training data, we will redesign the scheme to combine some of the infrequently used labels.

Acknowledgement

This work was supported by MEXT KAKENHI Grant Number 23240018 and by RISTEX, JST.

References

Kathrin Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1402–1407.

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.

Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Linguistic Annotation Workshop*, pages 68–73.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, contrast and contradiction in text processing. In *Proceedings of the 21st national conference on Artificial intelligence*, pages 755–762.

Kenji Imamura, Tomoko Izumi, Genichiro Kikui, and Satoshi Sato. 2011. Jutsubu kinouhyougen-no imiraberu tagaa {Semantic label tagging to functional expressions in predicate phrases}. In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pages 308–311. (in Japanese).

John Lafferty, Andrew K. McCallum, , and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. 2006. Compilation of a Dictionary of Japanese Functional Expressions with Hierarchical Organization. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages (ICCPOL)*, pages 395–402.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus, and source information. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1456–1463.

Yoshiyuki Morita and Masae Matsuki. 1989. *Nihongo Hyougen Bunkei*. ALC Press Inc. (in Japanese).

Kazuya Narita, Junta Mizuno, and Kentaro Inui. 2013. A lexicon-based investigation of research issues in japanese factuality analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 587–595.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

Erik F. Tjong Kim Sang. 2000. Noun phrase recognition by system combination. In *Proceedings of the Language technology Joint Conference ANLP-NAACL2000*, pages 50–55.

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.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 161–170.

Takafumi Suzuki, Yusuke Abe, Itsuki Toyota, Takehito Utsuro, Suguru Matsuyoshi, and Masatoshi Tsuchiya. 2012. Detecting Japanese Compound Functional Expressions using Canonical/Derivational Relation. In *Proceedings of the 8th International Language Resources and Evaluation*.

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

Masatoshi Tsuchiya, Takehito Utsuro, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2005. A Corpus for Classifying Usages of Japanese Compound Functional Expressions. In *Proceedings of the Conference Pacific Association for Computational Linguistics*, pages 345–350.

Masatoshi Tsuchiya, Takao Shime, Toshihiro Takagi, Takehito Utsuro, Kiyotaka Uchimoto, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2006. Chunking Japanese Compound Functional Expressions by Machine Learning. In *Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context*, pages 25–32.

Takehito Utsuro, Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi, and Satoshi Sato. 2007. Chunking and Dependency Analysis of Japanese Compound Functional Expressions by Machine Learning. In *Proceedings of the 7th China Japan Natural Language Processing Joint Research Promotion Conference*.