PAPER

A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling

Yoko NAKAJIMA[†a)], **Michal PTASZYNSKI**[†], *Nonmembers*, **Hirotoshi HONMA**[††], *and* **Fumito MASUI**[†], *Members*

SUMMARY In everyday life, people use past events and their own knowledge in predicting probable unfolding of events. To obtain the necessary knowledge for such predictions, newspapers and the Internet provide a general source of information. Newspapers contain various expressions describing past events, but also current and future events, and opinions. In our research we focused on automatically obtaining sentences that make reference to the future. Such sentences can contain expressions that not only explicitly refer to future events, but could also refer to past or current events. For example, if people read a news article that states "In the near future, there will be an upward trend in the price of gasoline," they may be likely to buy gasoline now. However, if the article says "The cost of gasoline has just risen 10 yen per liter," people will not rush to buy gasoline, because they accept this as reality and may expect the cost to decrease in the future. In the following study we firstly investigate future reference sentences in newspapers and Web news. Next, we propose a method for automatic extraction of such sentences by using semantic role labels, without typical approaches (temporal expressions, etc.). In a series of experiments, we extract semantic role patterns from future reference sentences and examine the validity of the extracted patterns in classification of future reference sentences.

key words: *natural language processing, future reference expressions, future prediction, information extraction*

1. Introduction

In recent years, obtaining large-scale data from Web pages and newspaper articles has required much less effort. Thus, the number of research actively developing and discussing the technology to analyze such data has increased rapidly. Large-scale data is of high interest for trend prediction, due to containing large amounts of trend information. Trend information is the data from which one can derive hints about the possible unfolding of certain events. The most common association would be with the prediction of stock trends, but the idea of trend information extends also to everyday information, and predicting the outcomes of specific events does not require any special abilities in everyday newspaper readers. For example, if we obtained a hypothetical fact that "the President of the USA is considering paying a state visit to Egypt" and a later one stating that "a revolution has started in Egypt," we could reasonably predict that the President will postpone or cancel the visit. This kind of future prediction is a logical inference which people experience every day when reading news articles. As another example, if one reads an article in which it is stated that a country is expected to draw up a relaxation of economic law, one could predict that the country's economic situation could change for the better in the future. Similarly, if one reads an article about releasing a new product, one could predict that, if the product sells well, the finances of companies involved in producing parts for the product will also improve. In this way, we believe it is possible to predict future trends by analyzing articles mentioning events related to the future, which in practice could be widely applied to corporate management or trend forecasting. In particular we consider future reference sentences to support the prediction of future events.

In the following sections, we first describe our study on expressions mentioning the future in trend documents. Next, we explain our proposed method for classifying sentences that mention future events. Further, we describe a series of experiments, and present results of the automatic classification of sentences into future-related and non-future-related and the extraction of future reference sentences. Finally, we conclude the paper by describing a number of possible further improvements to the method and discussing its potential applications.

2. Previous Research

Linguistically expressed references to the future have been studied by a number of researchers. Baeza-Yates [1] performed a study on about five hundred thousand sentences containing future events extracted from Google News[*] over the course of one day, and concluded that events mentioned in the news as those scheduled to take place, occur with almost perfect probability. A high correlation was also found between the reliability of occurrence and the time proximity of the event. Therefore, information about upcoming events is highly important in predicting future outcomes. Following this discovery, in our research we also chose the news as our data source. This will assure that if we extract the future sentences correctly, the events described in those sentences will have high probability of occurrence in reality.

According to the study of Kanhabua et al. [2], one-

Manuscript received March 31, 2015.
Manuscript revised November 1, 2015.
Manuscript publicized November 18, 2015.

[†]The authors are with the Department of Computer Science, Kitami Institute of Technology, Kitami-shi, 090–8507 Japan.

[††]The author is with the Department of Information Engineering, National Institute of Technology, Kushiro College, Kushiro-shi, 084–0916 Japan.

a) E-mail: yoko@kushiro-ct.ac.jp

DOI: 10.1587/transinf.2015EDP7115

[*]http://news.google.com/

third of all newspaper articles contain some reference to the future. This also supports our choice of the news as our data source. In other research, Kanazawa et al. [3] extracted future implications from the Web using explicitly expressed future reference information. Alonso et al. [4] have indicated that time information included in a document enhances the effectiveness of information retrieval applications. Kanazawa et al. [5] focused on extracting unreferenced future time expressions from a large collection of text, and proposed a method for estimating the validity of the prediction by automatically searching for a real-world event corresponding to the predicted one. Jatowt et al. [6] studied the relation between future news written in English, Polish, and Japanese using keywords queried on the web. Popescu et al. [7] investigated significant changes in the distribution of terms within the Google Books corpus and their relationship with emotion words across a wide time span.

Among the research regarding the retrieval of future information, Kanhabua et al. [2] proposed a ranking model that takes into consideration the relevance of predictions. In terms of predicting the probability of an event occurring in the future and its relevance, Jatowt et al. [8] developed a model-based clustering algorithm for detecting future phenomena based on information extracted from a text corpus, and proposed a method of calculating the probability of the event happening in the future. In a separate research, Jatowt et al. [9] used the incidence rate of reconstructed news articles over time to forecast recurring events. They presented a technique for supporting the human analysis of future phenomena by applying a method based on the summarization of future information included in documents. Aramaki et al. [10] used Support Vector Machines to classify Twitter data related to influenza, and attempted to predict the spread of the virus using a truth validation method. Radinsky et al. [11] proposed the Pundit system for the prediction of future events in news. Their method used causal reasoning derived from a calculated similarity measure based on different existing ontologies. However, as their approach is based on causality pairs, rather than specific future-related expressions, it is not able to cope with certain constructions, e.g., sentences containing causality expressions but referring to the past.

The methods described above often use time-related information, such as "year," "hour," or "tomorrow" to extract future information and retrieve relevant documents. It has also been indicated that using information contained in available present documents is useful for predicting future outcomes. However, although all previously mentioned methods have used future time information, none of them examined more sophisticated expressions, such as sentence patterns referring to the future. Hence, a method using such expressions would approach the problem of future prediction from a new perspective, and could form a significant contribution to research on future information extraction. Below we describe a method for the automatic extraction of such candidate patterns referring to the future.

3. Future Reference Semantic Pattern Extraction Method

In this section, we describe our method for extracting semantic patterns from sentences.

3.1 Theory Basis for Developed Method

In our preliminary study [12] we investigated future reference expressions appearing in newspapers and Web-news corpora. After manually extracting and analyzing 270 future reference sentences, we found out that there were 141 unique future expressions (words, phrases, etc.) and 70 time-related expressions. Furthermore, 55% of future expressions appeared two or more times, and 45% appeared only once. We can assume that these which appear the most often could be said to have a characteristics of being used as future expressions.

Moreover, if we consider sentences and their different representations (grammatical, semantic) as sets of patterns which occur in a corpus (collection of sentences/documents) we should be able to extract from those sentences new patterns referring to the future. As the basic theory we based our idea on the theory of predicate-argument structure [13], which considers both word-formation and semantics. This theory embraces the synergy between the lexical information of a predicate and their semantic and syntactic properties. In practice this can be realized by representing a sentence using semantic role labels. The proposed method takes advantage of such sentence representation and further extracts implicit future reference patterns, not using handcrafted lists of explicit future expressions or temporal expressions, as it was in previous methods.

3.2 Semantic Role Labelling

Firstly, in the method we perform semantic role labeling on sentences. Semantic role labeling provides labels for words according to their role in the present sentence context. For example, in the sentence "John killed Mary," the words are labeled as follows: John=`actor`, kill[past]=`action`, Mary=`patient`. Thus, the semantic representation of the sentence is `actor-action-patient`.

For the semantic role labeling of Japanese text, we used **ASA**[†] system, which provides semantic roles for words and represents them on a semantic structure using a thesaurus [14]. In particular ASA uses 4400 verbs and around 80 labels from Lexeed (basic word-meaning) database [15].

However, there were two basic problems with ASA output. Firstly, due to the limitations of the applied thesaurus, not all words are semantically labeled by ASA. Secondly, as it is shown in Table 1, some labels provided by ASA are too specific. Therefore in order to normalize and simplify the patterns, we specified the priority of label

[†] http://cl.it.okayama-u.ac.jp/study/project/asa

Table 1 An example of a sentence analyzed by ASA.

Example: *Hatsuden no tekichi toshite zenkoku no Megasora keikaku no 4-bun no 1 ga shūchū suru Hokkaidō ni tai suru mikata ga kawari tsutsu aru.* / Opinions about Hokkaido as an appropriate area for power generation, in which 1/4 of the whole country Mega Solar plan is to be realized are slowly changing

No.	Surface	Label
1	*hatsuden-no*	[State change]-[creation or destruction]-[creation (physical)];Verb
2	*tekichi toshite*	[As]
3	*zenkoku-no*	[Place]
4	*Megasora keikaku-no*	[Action]
5	*4 bun no 1 ga*	[Numeric]
6	*shūchū suru*	[State change]-[place change]-[change of place (physical)]-[movement towards a goal]
7	*Hokkaidō ni tai suru*	[Place]
8	*Mikata ga*	[Other]
9	*Kawari tsutsu aru*	[State change]-[change]

groups in the following way.

1. Semantic roles (Agent, Patient, Object, etc.)
2. Semantic meaning (State_change, etc.)
3. Category (Dog → Living animal → Animated object)
4. In case of no output by ASA use parts of speech
5. As post-processing perform compound word clustering

For some words, ASA does not provide semantic information. In such cases we used only grammatical or morphological information, such as “Proper Noun” or “Verb.” Moreover, in cases where only morphological information is provided, there could be a situation in which one compound word is divided by morphological analyzer. For example, “Japan health policy” is one semantic concept, but its grammatical representation has the form “Noun Noun Noun.” To optimize the method, we used a set of linguistic rules to specify compound words. The heuristic rules were hand crafted on the basis of present state of linguistic research regarding compound words in Japanese [16], [17]. An example of a sentence analyzed this way is represented in Table 1.

In the method each sentence is labeled by ASA. An example of a sentence generalized into its semantic representation may be as follows (sentence from Table 1: “`[State change] [As] [Place] [Action] [Numeric] [State change] [Place] [Other] [State change]`.” During the evaluation experiment (see Sect. 4) all sentences from the dataset are first preprocessed this way.

3.3 Pattern Extraction

Once all sentences have been analyzed and assigned semantic roles, we use SPEC (**S**entence **P**attern **E**xtraction ar**C**hitecture), a system for the extraction of sentence patterns developed by Ptaszynski et al. [18]. **SPEC** automatically extracts frequent sentence patterns that are distinguishable for a specific corpus. The patterns are defined in this paper in the following way.

> A sentence pattern is any frequently occurring ordered non-repeated combination generated from elements of the sentence. When the elements are disjoint, the gap is marked by an asterisk (“*”). Sentence elements are defined as parts of the sentence specified by the process of sentence preprocessing and are consistent with available sentence representation selected by the user (e.g., SRL in this paper).

According to this definition, the system generates ordered non-repeated combinations from the elements of a sentence. In every n-element sentence, there are $k \leq n$ combination groups, where k represents all k-element combinations that are a subset of n. The number of combinations generated for one k-element group of combinations is given by the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

In this procedure, the system creates all combinations for all values of k from the range $\{1, \ldots, n\}$. Therefore, the number of combinations is equal to the sum of all combinations from all k-element groups of combinations, given by:

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + \cdots + \frac{n!}{n!(n-n)!} \\ = 2^n - 1 \tag{2}$$

Next, the system specifies whether the elements appear next to each other by placing a wildcard “*” between all non-subsequent elements. SPEC uses all of the initially generated original patterns to extract patterns frequently appearing in a given corpus, and calculates their weight. The weight can be calculated in several ways. Two features are important in weight calculation. Patterns that are more representative of a corpus tend to be long (high values of k) or appear more often (high values of occurrence O). Therefore, the weight can be calculated by considering

- awarding length k,
- awarding length k and occurrence O,
- awarding none (normalized weight).

Moreover, the list of frequent patterns produced during pattern generation and extraction can be further modified. When two collections of sentences with opposite features (such as “positive vs. negative,” or “future-related vs. non-future-related”) are compared, the list of patterns generated will contain some that appear uniquely on one side (i.e., uniquely future patterns and uniquely non-future patterns) and some that appear more than one time on both sides (ambiguous patterns). Therefore, the pattern list can be modified by

- using all patterns,
- erasing all ambiguous patterns,
- erasing only those ambiguous patterns that have the

same number of appearances on both sides (zero patterns).

The list of patterns will contain both sophisticated patterns (with disjoint elements) and more common n-grams. Therefore, the evaluation could be performed on either

- all patterns, or
- only n-grams.

Each of the above mentioned modifications are automatically verified in the process of choosing the best model.

The SPEC system is trained on bipolar training data (e.g., future reference sentences vs. non-future reference sentences), and generates all patterns. Next, it classifies test data using the generated patterns. The performance of the whole system for classification of sentences into either future related or not is tested using a 10-fold cross validation.

4. Evaluation Experiment

In this section, we describe experiments to verify whether the future reference pattern extraction method is effective.

4.1 Experiment Setup

We designed the experiment as a text classification task with the prepared datasets applied into 10-fold cross validation. The classification was performed as follows. Each test sentence was given a score calculated as a sum of weights of patterns extracted from training data and found in the input sentence, as in Eq. (3).

$$score = \sum w_j, (1 \geq w_j \geq -1) \tag{3}$$

The metrics used in evaluation are the standard Precision, Recall, and balanced F-score.

However, if the initial collection of sentences is biased toward one of the sides (e.g., more one kind of sentences, or the sentences are longer, etc.), there will be more patterns of a certain type. Thus, agreeing to a rule of thumb in classification (e.g., fixed threshold above which a new sentence is classified as either future or non-future related) might be harmful to one of the sides and not provide sufficiently objective view on results. Therefore we applied automatic assessing of the threshold as a way of optimizing the classifier.

In the experiment 14 different versions of the classifier are compared under 10-fold cross validation condition. The experiment was performed on two datasets, thus the obtained overall number of experiment runs was 280. There were several evaluation criteria. Firstly, we looked at top scores within the threshold span. Secondly, we checked which version got the highest break-even point (BEP) of Precision and Recall. Finally, we checked the statistical significance of the results using paired *t*-test.

4.2 Dataset Preparation

Firstly, we collected a thousand sentences at random from a corpus containing the following newspapers: Nihon Keizai Shimbun[†], Asahi Shimbun[††], and Hokkaido Shimbun[†††].

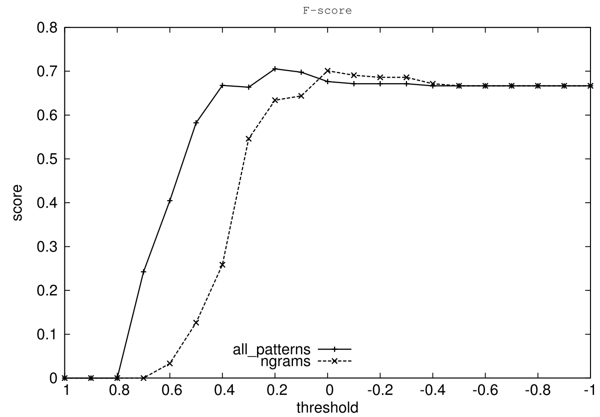

Fig. 1 Comparison of F-scores for set50 for all patterns and n-grams only.

Next, three people manually judged whether these sentences referred to the future or not. The agreement coefficient (multi-rater kappa-value) was 0.456, which indicates somewhat strong agreement between the annotators. We grouped the annotated sentences into three groups: (1) perfect agreement between all three annotators, (2) ambiguous sentences and (3) other sentences (non future referring sentences). From the collected 1000 sentences the group for which all three annotators agreed contained 130 sentences, the ambiguous sentences group contained 330 sentences and the “other” group contained 540 sentences. From the above data, we chose the 130 future referring sentences and additionally selected at random another 130 non-future referring sentences for the experiment. Then, we prepared two sets of data. The first contained 100 sentences (with 50 future-related and 50 non-future-related sentences, later: “set50,”), and the second contained all 260 sentences (with all 130 future-related and addtional 130 non-future-related sentences, later called “set130,”). All sentences were preprocessed with ASA, and semantic role labels were added to each sentence.

4.3 Classification Results

We compared Precision, Recall, and balanced F- score for the classification based on patterns and, additionally, on n-grams alone with semantic role labels.

For set50, the F-score was generally around 0.67–0.71 for patterns, and around 0.67–0.70 for n-grams. The F-score for set130 was around 0.67–0.70 for patterns, and 0.67–0.69 for n-grams. The optimal threshold (from the range 1.0 to −1.0, with 0.0 in the middle) was around 0.0 or slightly biased toward 1.0, which means both sides of the training set were balanced or slightly biased toward future-related sentences. Figure 1 illustrates the F-score results for set50

[†] http://www.nikkei.com/

[††] http://www.asahi.com/

[†††] http://www.hokkaido-np.co.jp/

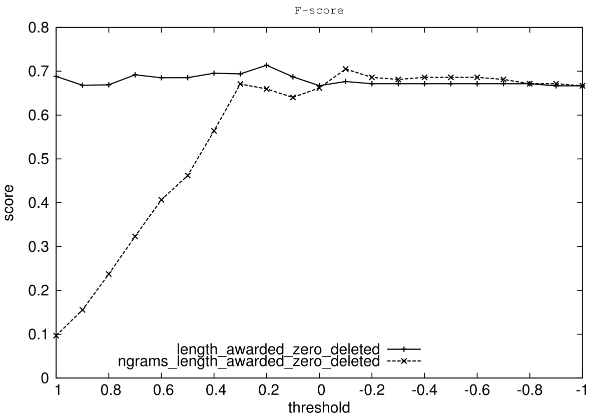

Fig. 2 Comparison of F-scores for set50 for patterns and n-grams for the classifier with length-awarded zero deleted.

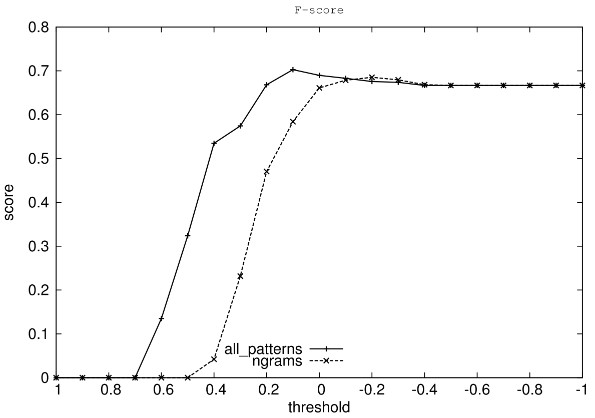

Fig. 3 Comparison of F-scores for set130 for all patterns and n-grams only.

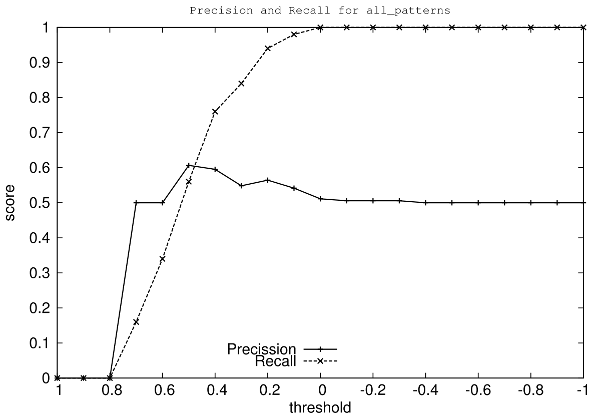

Fig. 4 Precision and Recall for all patterns in set50.

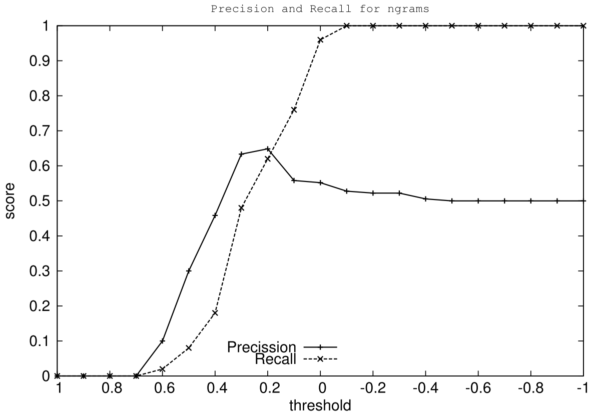

Fig. 5 Precision and Recall for n-grams for set50.

Table 2 Comparison of the best results achieved (Precision, Recall, and F-score) for set50 and set130.

classifier version	**set50**			**set130**		
	Precision	Recall	F-score	Precision	Recall	F-score
unmodified pattern list	0.56	0.94	0.71	0.58	0.90	0.70
zero deleted	0.56	0.94	0.71	0.57	0.90	0.70
ambiguous deleted	0.55	0.92	0.69	0.56	0.91	0.69
length awarded	0.58	0.90	0.71	0.58	0.89	0.70
length awarded zero deleted	0.56	0.98	0.71	0.57	0.87	0.69
length awarded ambiguous deleted	0.55	0.98	0.70	0.56	0.92	0.70

when a list of patterns used in classification contained either all patterns with comparison to n-grams only. Figure 4 and Fig. 5 show the Precision and Recall for patterns and n-grams, respectively, for set50. Figure 2 illustrates the F-scores result for set50 considering patterns and n-grams for the classifier with length-awarded zero deleted. Figure 3 illustrates the F-scores result for set130 considering all patterns and n-grams only.

Furthermore, we compared different versions of the classifier, including those in which the pattern list was modified by deleting either zero patterns or ambiguous patterns. We also verified which method of weight calculation was more effective, the one using normalized weights, or the pattern length-based method. Hence, we also examined the case of length-based weights with zero patterns deleted, and length-based weights with ambiguous patterns deleted. We performed a t-test on the F-scores given by set50 and set130. The p-value was 0.566 for all patterns. This means that the differences between set 50 and set130 were not statistically significant, which is a positive result, since it proves that the performance of our method does not depend on the amount of learning data. The one-sided t-test value was 0.310, which also does not suggest any significant difference. We will discuss the differences in detail in the Discussion section.

4.4 Analysis of Most Useful Future Reference Patterns

Besides the automatic classification results, we were also interested in the actual patterns that influenced the results. We extracted the most frequent unique future reference patterns and non-future reference patterns from set50. We obtained 1131 future patterns and 87 non-future patterns. Ten examples of both pattern types are given in Table 3.

Semantic role label patterns are grouped according to frequency of their appearance on each side, namely, in future reference sentences or in non-future reference sentences. For better comparison the pattern examples presented in Table 3 contain only non-ambiguous patterns

Table 3 Examples of extracted patterns.

Occurrence	Future Reference Patterns	Occurrence	Non-future Reference Patterns
26	[Action]*[State change]	5	[Place]*[Agent]
43	[Action]*[Object]	4	[Numeric]*[Agent]
42	[Action]*[Action]	4	[Verb]*[Artifact]
20	[State change]*[Object]	4	[Person]*[Place]
16	[State change]*[State change]	3	[Numeric]*[Agent]*[Action]
15	[Action]*[Object]*[State change]	3	[Adjective]*[State change]*[State change]
15	[Action]*[State change]*[No state change (activity)]	3	[Place]*[Place]*[No state change (activity)]
14	[Object]*[Action]*[State change]	3	[Place]*[State change]*[Place]
13	[Object]*[Action]*[Object]	3	[Time]*[State change]*[Artifact]
12	[State change]*[Action]*[State change]	2	[Noun]*[Person]*[Noun]*[State change]

which appeared in only in one of the sides.

The asterisk in some patterns means that the elements are disjoint. For example, the pattern [Action]*[State change] contains two elements, [Action] and [State change], which appeared in the original sentences exactly in this order, and the asterisk indicates that there were other elements between these two. Each sentence pattern can appear either within a sentence, or on its edges (beginning, or end of the sentence). The method used for pattern extraction (SPEC) by the definition, does not make this additional distinction. This is due to the fact that sometimes a sentence can a start with a certain pattern, but in another sentence some words could precede this pattern. Making an additional distinction of sentence edges would force treating patterns which are actually the same as different ones only because of their facing the sentence edge or not. For one pattern this would produce four superficial combinations depending on the position of the beginning and the end of the pattern within the sentence (Edge-Inside, Inside-Edge, Edge-Edge, Inside-Inside). Thus although the four types would in fact represent the same one single pattern, its statistics would become dispersed to the four types.

4.5 Discussion

In this section, we present a detailed analysis of the results to facilitate better understanding of the extracted future reference patterns.

In general, the pattern-based approach obtained higher scores than the model trained on n-grams-only. This suggests that there are meaningful frequent patterns, more sophisticated than simple n-grams, in sentences referring to the future. In terms of modifying the pattern list and weight calculation, deleting the zero patterns does not appear to influence the results. A larger difference can be seen when all ambiguous patterns are deleted, and only patterns unique to each side are used. Moreover, the pattern length-based weight calculation always yielded better results. The highest scores of F = 0.71 with P = 0.56 and R = 0.98 were obtained using a pattern list with zero-patterns deleted and a length-based weight calculation. The greatest improvement provided by the use of patterns over n-grams is in Recall, which means that there are many valuable patterns omitted in the n-gram-only approach. Precision does not change significantly, oscillating around 0.55–0.60. For some thresholds, n-grams achieved similar or higher Precision. This means that the range 0.55–0.60 is the optimal maximum that could be achieved with the semantic representation used in this study. In the future, we plan to develop a modification that would improve the Precision without reducing Recall.

As well as comparing patterns with n-grams on the baseline classifier, we compared the results for five other cases (modifying the pattern list by deleting zero-patterns, or deleting all ambiguous patterns and modifying the weight calculation according to pattern length). In general, the highest F-score for patterns was 0.71, while for n-grams it was 0.70. Although the difference is not that large, patterns usually achieve a high F-score because of superior Recall performance, even close to the threshold of 1.0 (compare Fig. 1, Fig. 2, and Fig. 3). In Fig. 2, the highest result of F-score was 0.70 for patterns, and 0.69 for n-grams. In this case the highest achieved F-score is nearly the same between patterns and n-grams. However, patterns achieved better scores for each of the threshold. In case of F-score for set130 (see Fig. 3), the highest result was also 0.7 for patterns and 0.69 for n-grams. However, the results for patterns are higher mostly within the threshold of 1.0 to 0.0, which confirms the results of set50. Since patterns provide better scores for most of the thresholds, we consider patterns as more effective. To thoroughly verify whether it is always better to use patterns, we need to conduct more experiments. However, from the present data, we can conclude that patterns generally produce better results.

Next, we compared the two datasets, **set50** and **set130**. The comparison in Table 2 shows that the results for each dataset did not differ greatly. However, when we look at Fig. 6, the F-score for the classifier using a pattern list with all ambiguous patterns deleted performs slightly better than the other two (although the differences are not quite statistically significant with $p < 0.06$). Comparing these results to those in Fig. 7 indicates that the performance is generally better when the pattern length is used to modify the weight calculation. In particular, both modified versions of the classifier (without zero-patterns and without ambiguous patterns) retain high F-scores across the threshold span (from 1.0 to -1.0). The same can be said of the results for set130. Comparing Fig. 8 and Fig. 9 also shows that the pattern length-based weight calculation yields better results within the specified threshold. Moreover, it is also advantageous to either exclude zero-patterns or all ambiguous pat-

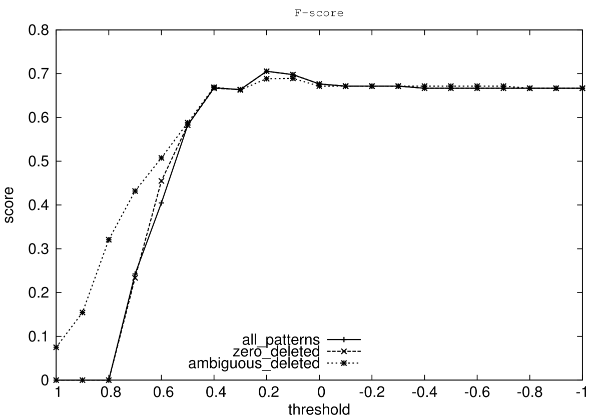

Fig. 6 F-scores for the classifier with three different versions of pattern list modification for set50.

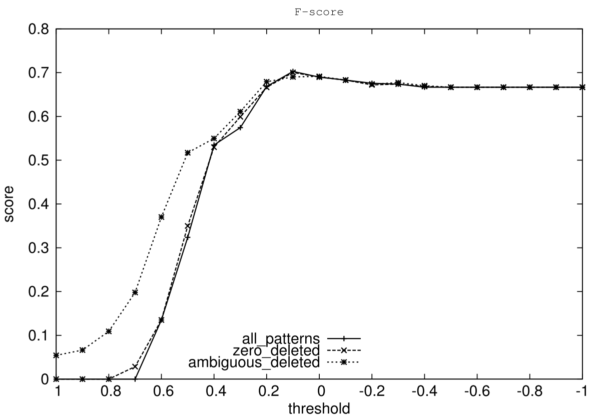

Fig. 8 F-scores for the classifier with three different versions of pattern list modification for set130.

terns from the pattern list. It is also worth mentioning that the performance of the algorithm as a whole is similar for set50 and set130. In general, larger datasets contain more ambiguities, which can decrease the results. With the proposed approach, the differences in results are generally negligible (compare Fig. 6 and Fig. 8) or small (compare Fig. 7 and Fig. 9). Therefore it can be said that the method retains its performance regardless of the amount of data.

4.6 Inquiry into Extracted Future Reference Patterns

Using SPEC we were able to extract frequent patterns from sentences referring to the future and those not referring to the future. Each time a trained pattern was used during the classification, it was also added to a separate list of frequently used patterns. This extraction was performed for each fold in the 10-fold cross-validation. By taking the patterns extracted this way from all tests, and leaving only the frequent ones (used in classification at least two times across all experiment runs), we obtained a refined list of the most valuable patterns (those used most often). We investigated these patterns and the types of sentences in which they were used.

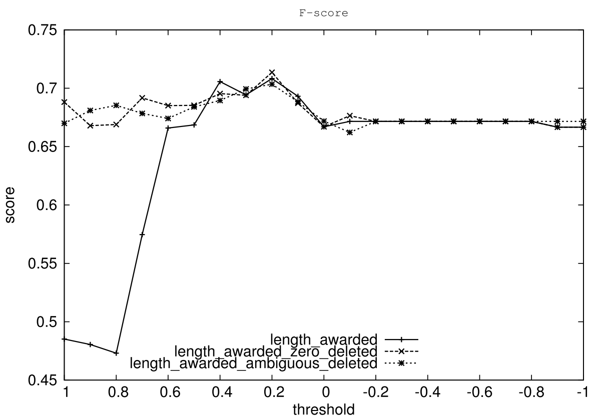

Fig. 7 F-scores for length based weight calculation for set50.

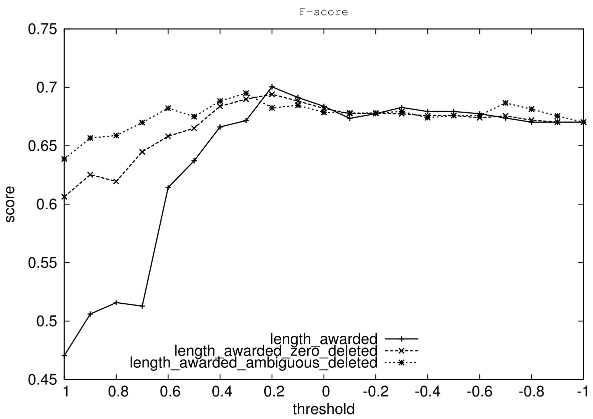

Fig. 9 F-scores for length based weight calculation for set130.

Below we present a number of example sentences used in classification. The information is provided in the following order: Romanized Japanese (transcribed in roman alphabet), English translation, and Semantic representation. The two first examples contain the following pattern: `[Action]*[Object]*[State change]` (pattern in question underlined).

Example 1. *Iryō, bōsai, enerugī nado de IT no katsuyō wo susumeru tame no senryaku-an wo, seifu no IT senryaku honbu ga 5gatsu gejun ni mo matomeru.* (IT Strategy Headquarters of the government will also put together in late May, the draft strategy for advancing the use of IT for health, disaster prevention, or energy.) [Action]-[Other]-[Other]-[No state change(activity)]-[State change]-[Artifact]-[Object]-[Organization]-[Agent]-[Noun]-[Time]-[State change]

Example 2. *Tonneru kaitsū ni yori, 1-nichi 50 man-nin wo hakobu koto ga kanō ni naru mitōshi de, seifu wa jūtai kanwa ni tsunagaru to shite iru.* (It is expected that the opening of the tunnel will make it possible to carry 500,000 people a day, which will lead to a reduction in traffic congestion, according to the government.) [Action]-[Time]-[Object]-[State_change]-[Other]-

[Noun]-[Action]-[Organization]-[Action]-[Verb]-[State change]

The next examples contain a slightly different pattern, namely `[Object]*[Action]*[State change]`.

Example 3. *Nesage jisshi wa shinki kanyū-ryō, kihon ryōkin ga 12gatsu tsuitachi kara, tsūwa ryōkin ga 1996nen 3gatsu tsuitachi kara no yotei.* (The price cut implementation is planned to apply to the new subscription fees, for the basic rate plan from December 1, for call charges from March 1, 1996.) [Object]-[Action]-[Agent]-[Numeric]-[Time]-[Action]-[Time]-[Numeric]-[Time]-[State change]

Example 4. *Kin'yū seisaku wo susumeru ue de no kakuran yōin to shite keishi dekinai, to no mondai ishiki no araware to wa ie, kin'yū-kai ni hamon wo hirogesōda.* (Although they admitted that proceeding with the [new] monetary policy could become a disturbance factor and that it cannot be neglected, which showed an awareness of the problem, it still is likely to spread ripples in the financial world.) [Object]-[State change]-[Reason]-[Action]-[Action]-[Action]-[Agent]-[Place]-[Other]-[State change]

In the above examples, the patterns that were matched comprise those studied in previous research [3], [5], [6]. These include time-related expressions ("late May," "from December 1," "from March 1, 1996") and future reference expressions ("is expected," "is planned to," "is likely to").

Next, we examined sentences containing non-future patterns. The following example sentence contains the pattern `[Numeric]*[Action]*[Action]`.

Example 5. *20man-ji no chōhen shōsetsu kara 2 moji dake wo kopī shite shōbai ni tsukatte mo ihō to wa ienai.* (It cannot be considered illegal to copy only two characters from a two-hundred-thousand-word-novel and use them for commercial purposes.) [Numeric] [Artifact] [Numeric] [State change] [No state change] [No state change] [Action] [Action]

The following example sentence contains the pattern `[Place]*[Place]*[No_state_change(activity)]`.

Example 6. *Nagata-ku wa Hanshin Daishinsai de ōkina higai wo uketa chiiki de, koko de wa Betonamu no hito ga kazu ōku hataraite iru.* (Nagata Ward, one of the areas that were greatly affected by the Great Hanshin Earthquake, is a place where many people from Vietnam are working.) [Place] [Organization] [adjective] [Other] [No state change(state)] [Object] [Place] [Agent] [Adjective] [No state change(action)]

The following example sentence contains the pattern `[Time]*[Noun]*[Role]`.

Example 7. *Sakunen 6gatsu, Kaifu ga Jimintō to tamoto wo wakatte aite jin'ei (gen Shinshintō) ni kumi shita toki mo, rinen to meibun ga hakkiri shinakatta.* (June last year, when Kaifu parted company with the Liberal Democratic Party and joined an opponent camp (now called New Frontier Party), their ideas and causes were unclear.) [Time] [Numeric] [Person] [Organization] [Noun] [State change] [Noun] [Organization] [Verb] [Role] [Place] [No state change(state)]

Example 5 contains the phrase *to wa ienai* ("it cannot be said/considered that"), which is labeled as an [Action] by ASA. This label is frequently used in future referring sentences, but this sentence is not classified as future-related. As for Example 7, although it contains time-related expressions ("June last year"), the use of sophisticated patterns that take the wider context into account allows correct disambiguation in this case. Furthermore, although this pattern contains a time-related expression, it is not listed as a future reference pattern. Thus, the presence of time-related information alone does not influence the classification. Instead, other elements of the pattern, such as the appropriate tense together with time-related expressions, constitute the pattern being distinguished as referring to the future.

Many future reference patterns had a high occurrence frequency (see Table 3), which means the sentences contain many of those patterns. Therefore, we can say that in general, "the future" has high linguistic expressiveness. For non-future reference patterns, the occurrence frequency was low, which suggests a large number of patterns, each used only once (thus, they were not included in the list of frequently used patterns). Because of this variety of patterns, there are no particularly distinctive patterns for sentences that are not referring to the future.

5. Method Validation

In this section, we present an additional experiment to validate the effectiveness of the proposed method in the extraction of future reference sentences.

5.1 Experiment Setup

We collected the following additional validation dataset containing future reference sentences, completely unrelated to any of the training or test sets. From the daily edition of *Mainichi Shinbun* newspaper articles from one year (1996) we extracted 170 sentences from articles appearing on pages 1–3 (which usually covers the most important or featured articles), and articles from the "economics" and "international events" sections. Next, three annotators manually annotated these sentences as either future referring or non-future referring.

In the validation experiment we performed two analyses. Firstly, we compared our method to the state-of-the-art represented by the method of Jatowt et al. [8]. Secondly, we analyzed fluctuation of results when various changes were made to the pattern lists.

Jatowt et al. in their experiment extracted future reference sentences using 10 words such as "will," "may," "be

Table 4 Comparison of results (Precision, Recall, and F-score) for validation set between different pattern groups and the state-of-the-art.

Pattern set	Precision	Recall	F-score
10 patterns	0.39	**0.49**	**0.43**
15 patterns	0.38	**0.49**	**0.43**
5 patterns	0.35	0.35	0.35
10 pattern with only over 3 elements	**0.42**	0.37	0.40
Jatowt et al. 2011 [8]	**0.50**	0.05	0.10

likely to," etc. We used their set of words, translated them into Japanese and used to classify the 170 sentences from the validation set.

Our proposed method generates a large number of patterns even from small amount of training data. Therefore a straightforward comparison of our method to [8] could be considered as unfair. To make the comparison and validation more fair we classified the validation set using only the most frequent patterns generated in the previous experiment. We created four small scale pattern lists, which we used in the comparison and validation. The results are shown in Table 4. In particular, we performed pattern matching on the new sentences with the following pattern sets:

[Pattern sets]
A: 10 patterns (see Table 3)
B: adding 5 patterns of length more than three elements to set A
C: subtracting 5 patterns from set A
D: using only 10 patterns containing more than three elements

5.2 Results and Discussion

The results for the initial set of ten patterns reached F-score of 0.43, which appeared to be a plateau of the performance level, after which increasing the number of patterns made little difference. In future, we will apply a larger validation set to investigate in more detail how this plateau fluctuates according to the size of the data set and the number of patterns used for classification. The performance of pattern set C is poor because only a few patterns are used. The Precision of pattern set D is slightly higher than that of the other sets. This indicates it could be more effective to use frequent future reference patterns containing more than three elements, even when the number of applied patterns is small. It could also be more effective to use patterns consisting of a few (two or three) elements if the focus of the extraction was on Recall, whereas it would be more effective to use patterns consisting of three or more elements if the focus of the extraction is on Precision.

The scores obtained in this experiment were lower than those in the evaluation experiment. However, we were able to extract future reference sentences with approximately 40% Precision using only ten future reference patterns, a score that is not far below the one generated in the evaluation experiment (which used a total of 1131 patterns in Sect. 4.4). The performance could be further improved by training the system on a limited, specific genre of events (e.g., future-reference sentences in economy, or energy areas). Another point worth mentioning is that, as the method uses semantic role labeling, it is not dependent on the grammar or meaning of particular single words.

Finally, we compared our method with the one proposed by Jatowt et al. [8]. Their results were P = 0.50, R = 0.05, F = 0.10. Although the Precision seems higher than our best in these conditions (0.50 comparing to 0.42), one has to notice, that the Recall of the state-of-the-art did not exceed 5%, which in practice means that the number of sentences classified correctly was small (six out of 170 in particular). In comparison our method extracted many future referring sentences also using only 10 patterns. This suggests that the proposed method is valid. Moreover since the proposed method is fully automatic, it is also much more efficient.

6. Conclusions and Future Work

In this paper we investigated the characteristics of expressions referring to the future in newspapers and classified sentences (future referring vs. non-future referring) using a pattern-based approach with semantic representations of sentences. We then extracted future reference sentences from newspaper articles using future reference patterns consisting of the semantic roles automatically extracted with the proposed method.

We tested the proposed method on two datasets of different sizes. We found that the method performs well for both sets (F-score around 0.70–0.71). Although the datasets used in our experiments were not large, we confirmed that our approach can automatically determine whether a sentence is referring to the future or not. Since the results were promising, we plan to increase the experiment datasets and annotate large newspaper corpora according to their temporal references (future vs. past or present).

The experiment results indicated that there could be differences in the performance of future-reference pattern candidates depending on the kind of event. In particular, event- or area-specific future reference patterns could perform more effectively than general future reference patterns. In other words, the accuracy could vary depending on specific time spans / time points (future-related expressions used in newspapers change with time) or events (the kind of expressions used differ across areas).

Moreover, the coverage given by extracting future referring sentences using the proposed method is around ten times that using only temporal-reference words applied in previous research. This indicates that our method can classify a wide range of future referring sentences.

In the near future we plan to increase the size of experimental datasets and approach the data from different viewpoints to increase the precision of the classifier. We also plan to verify which patterns influence the results positively and which hinder the results. This knowledge could determine a more general model of future referring sentences. Such a model could be useful in estimating the probable unfolding of events, and would contribute to the task of trend prediction in general. Furthermore, it is better to use newspaper corpora than other kinds of textual data for training since newspaper articles can be considered as sufficiently reliable source of information. Also, we plan to apply the method in prediction of future events, and compare the performance of the method in classification of sentences referring to the future, past and the present to verify whether the proposed method is applicable only in the classification of future-reference sentences, or time-referring sentences in general. As for our next step, we plan to apply the future reference classification method to real-world tasks by finding new content with future references and sorting it in chronological order, which would allow the support of future predictions in everyday life. Finally, since we verified the validity of patterns, we plan to compare the results using other classifier approaches, such as Support Vector Machines, or Neural Networks, trained on the lists of automatically extracted patterns.

References

[1] R. Baeza-Yates, "The searching the future," SIGIR Workshop on MF/IR, 2005.

[2] N. Kanhabua, R. Blanco, and M. Matthews, "Ranking related news predictions," Proc. 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp.755–764, ACM, 2011.

[3] K. Kanazawa, A. Jatowt, S. Oyama, and K. Tanaka, "Extracting explicit and implicit future-related information from the web (in japanese)," DEIM Forum 2010, pp.A9–1, 2010.

[4] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz, "Temporal information retrieval: Challenges and opportunities," TWAW, vol.11, pp.1–8, 2011.

[5] K. Kanazawa, A. Jatowt, and K. Tanaka, "Improving retrieval of future-related information in text collections," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp.278–283, IEEE, 2011.

[6] A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, K. Kunieda, and K. Yamada, "Multi-lingual longitudinal analysis of future-related information on the web," Proc. 4th International conference on Culture and Computing, IEEE, 2013.

[7] O. Popescu and C. Strapparava, "Behind the times: Detecting epoch changes using large corpora," Proc. IJCNLP, pp.347–355, 2013.

[8] A. Jatowt and C.-M.A. Yeung, "Extracting collective expectations about the future from large text collections," Proc. 20th ACM International Conference on Information and Knowledge Management, pp.1259–1264, ACM, 2011.

[9] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka, "Supporting analysis of future-related information in news archives and the web," Proc. 9th ACM/IEEE-CS joint conference on Digital libraries, pp.115–124, ACM, 2009.

[10] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using twitter," Proc. Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.1568–1576, 2011.

[11] K. Radinsky, S. Davidovich, and S. Markovitch, "Predicting the future with social media," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp.492–499, IEEE, 2010.

[12] Y. Nakajima, M. Ptaszynski, H. Honma, and F. Masui, "Investigation of future reference expressions in trend information," AAAI Spring Symposium Series Big Data Becomes Personal: Knowledge into Meaning, 2014.

[13] J. Bresnan, Lexical-functional syntax. Oxford: Blackwell, 2001.

[14] K. Takeuchi, S. Tsuchiyama, M. Moriya, and Y. Moriyasu, "Construction of argument structure analyzer toward searching same situations and actions," IEICE Technical Report, NLC2009-33, 2010.

[15] K. Takeuchi, "Construction of thesaurus of predicate-argument structure for japanese verbs," The 25th Annual Conference of the Japanese Society for Artificial Intelligence, 2010.

[16] Y. Kobayashi, T. Tokunaga, and H. Tanaka, "Meishi-kan no imiteki kyōki jōhō o mochiita fukugō meishi no kaiseki". (Analysis of compound nouns using semantic co-occurrence information between nouns) [In Japanese]. Shizen Gengo Shori (Natural Language Processing), vol.3, no.1, pp.29–43, 1996.

[17] Y. Matsumoto, "Nihongo no goi-teki fukugō dōshi ni okeru dōshi no kumiawase." (Combinations of verbs in lexical compound verbs in Japanese) [In Japanese]. Gengo Kenkyū (Language Research), vol.114, pp.37–83, 1998.

[18] M. Ptaszynski, R. Rzepka, K. Araki, and Y. Momouchi, "Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion," International Journal of Computational Linguistics (IJCL), vol.2, no.1, pp.24–36, 2011.

Yoko Nakajima graduated from the Electronic Engineering Course, National Institute of Technology, Kushiro College in 1989. She joined the Department of Information Engineering, NIT, Kushiro College in 1989. Currently, she is an Assistant Professor of the Information Engineering Course, NIT, Kushiro College. She is in Ph.D. course at Kitami Institute of Technlology. Her research interests include natural language processing. She is a member of IPSJ.

Michal Ptaszynski was born in Wroclaw, Poland in 1981. He received the MA degree from the University of Adam Mickiewicz, Poznan, Poland, in 2006, and PhD in Information Science and Technology from Hokkaido University, Japan in 2011. In years 2011-2013 he was a JSPS Post-doctoral Research Fellow at the High-Tech Research Center, Hokkai-Gakuen University, Japan. Currently he is an Assistant Professor at the Kitami Institute of Technology. His research interests include natural language processing, dialog processing, affect analysis, sentiment analysis, HCI, and information retrieval. He is a member of the ACL, the AAAI, the IEEE, the HUMAINE, the AAR, and the ANLP.

Hirotoshi Honma was born in 1967. He received the B.E., M.E. and D.E. degrees in Engineering from Toyohashi University of Technology, in 1990, 1992 and 2009, respectively. He joined the Department of Information Engineering, National Institute of Technology, Kushiro College in 1992. He became an Associate Professor in 2001. His research interest includes computational graph theory, parallel algorithms, and natural language processing. He is a member of IEICE and ORSJ.

Fumito Masui received B.S. degree from University of Okayama and Ph.D. degree from Hokkaido University, Japan, in 1990 and 2006, respectively. He was a member of the Kansai Laboratories at Oki Electric Industry from 1990 to 2000 and the Department of Information Engineering at the Graduate School of Mie University as an assistant professor from 2000 to 2009. And he has joined the Department of Computer Science at Kitami Institute of Technology as an associate professor from 2009 to present. He has researched natural language processing, tourism informatics and winter sport informatics. Digital score book iCE, which has developed in his curing informatics research project, has been used by many top curling teams in Japan. He is a member of IEICE, IPSJ, JSAI, ANLP, SOFT, STI, SPEJ, AAAI, ACL, and the Japan Curling Association.