Investigation of Future Reference Expressions in Trend Information

Yoko Nakajima*† Michal Ptaszynski† Hirotoshi Honma* Fumito Masui†

∗ Department of Information Engineering
Kushiro National College of Technology
2-32-1 Otanoshike, Kushiro, 084-0916, Japan
{*yoko,honma*}@*kushiro-ct.ac.jp*

†Department of Computer Science
Kitami Institute of Technology
165 Koen-cho, Kitami, 090-8507, Japan
{*ptaszynski,f-masui*}@*cs.kitami-it.ac.jp*

Abstract

When we categorize textual data according to time categories the three main types of information that come up are the past, the present, and the future. In this paper we present our study in predicting the future. In particular, we aim at detecting expressions which refer to future events (temporal expressions, etc.) and apply them to support the prediction of probable future outcomes. In order to realize the future prediction support, we firstly need to be able to find out whether a sentence refers to the future or not in general. We propose a method of bi-polar text classification for sentences into either future-related or non-future-related (other). To do this we use a machine learning based sentence pattern extraction system SPEC and report on the accuracy of extracted patterns. We train the classifier using semantic representations of sentences and show that it is possible to extract fully automatically frequent patterns from sentences referring to the future.

Introduction

In recent years, obtaining large-scale data containing Web pages and newspaper articles has become a task requiring less and less effort. Thus a number of research actively developing and discussing the technology to analyze these data has increased dramatically. Large-scale data is the most interesting for the fact that it contains lots of trend information. Trend information is the kind of information from which one can derive hints about possibilities for events which are to unfold. The most common association would be with the prediction of stock trends, but the idea of trend information expands much further, to everyday information as well, and usually does not require any supernatural abilities. For example, by obtaining two hypothetical facts, such as "President of the USA is considering paying a state visit to Egypt" and a later one "A revolution started in Egypt", one could predict that the president would most probably postpone or cancel the visit. This kind of future prediction is a logical inference and we can experience it everyday when we read news articles. As another example, if one reads an article in which it is stated that a country is expected to draw up a law about economic relaxation, one could predict that the situation of the country could change in future in the good direction. Or, if one reads an article about releasing a new product, one could predict that if the product sells well, the finances of companies taking part in producing parts for the product will improve. This way, we believe it is possible to predict a future trend by analyzing articles mentioning events related to the future.

In the following sections we firstly describe our study about expressions mentioning the future in trend documents. Next, we explain our proposed method for classifying sentences which mention future events. Further, we describe the experiments and results of automatic classification of sentences into future-related and non-future-related. Finally, we conclude the paper, point out a number of potential improvements to the method and discuss potential applications.

Previous Research

The validity of practical use of information about the future has been studied by a number of researchers. Baeza-Yates (2005) performed a study of about five hundred thousand sentences containing future events extracted from one day of Google News (http://news.google.com/), and got to the conclusion that scheduled events occur in almost perfect probability and that there is a high correlation between the reliability degree of the occurrence of an event and time proximity of the event. Therefore the information about upcoming events is of a high importance for predicting future outcomes. According to the study of Kanhabua et al., who have investigated newspaper articles, one-third of all articles contains reference to the future (Kanhabua et al. 2011). In another research, Kanazawa et al. extracted implications for future information from the Web using explicit information, such as expressions about future time (Kanazawa et al. 2010). Alonso et al. have indicated that time information included in a document is effective for enhancing information retrieval applications (Omar et al. 2011). Kanazawa et al. 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 searching for a real-world event corresponding to the one predicted automatically (Kanazawa et al. 2011). Jatowt et al. studied relations between future news written in English, Polish and Japanese by using keywords queried on the Web (Jatowt et al. 2013). Popescu et al. have investigated significant changes in the

distribution of terms within the Google Books corpus and their relationships with emotion words (Popescu and Strapparava).

Among the research regarding retrieval of future information, Kanhabua et al. proposed a ranking model for predictions that takes into consideration their relevance (Kanhabua et al. 2011).

When it comes to predicting the probability of an even to occur in the future and its relevance of to the actual event, Jatowt et al. have proposed a model based clustering algorithm for detecting a future phenomenon based on the information extracted from text corpus, and proposed a method of calculating the probability of the event to happen in the future (Jatowt and Au Yeung 2011). In a different research, Jatowt et al. used the rate of incidence of reconstructed news articles over time to forecast recurring events, and proposed a method for supporting human user analysis of occurring future phenomena, by applying a method based on the summarization of future information included in documents (Jatowt et al. 2009). Aramaki et al. used SVM-based classifier on twitter to perform classification of information related to influenza and tried to predict the spread of influenza by using a truth validation method (Aramaki et al. 2011). Kanazawa et al. proposed a method for estimation of validity of the prediction by automatically calculating cosine similarity between predicted relevant news and searching for the events that actually occurred (Kanazawa et al. 2011). Radinsky et al. proposed the Pundit system for prediction of future events in news. They based their method on causal reasoning derived from a calculated similarity measure based on different existing ontologies (Radinsky et al. 2012). However, their approach is based on causality pairs, not on specific future-related expressions. Thus their method might not be able to cope with, e.g., sentences containing causality expressions, but referring to the past.

These findings have lead us to the idea that by using expressions referring to the future included in trend reports (newspaper articles, etc.), it could be possible to support the future prediction as one of the activities of people do everyday. For example, if we take sentences mentioning the future, such as "The technique for applying gas occurring in underground waters to generate power is rare and the company is going to sell it worldwide in Europe, China and other countries." and be able to determine plausibility of these sentences, it could enable us to develop a method supporting prediction of future events. Such a method would be widely applicable in corporate management, trend foresight, and preventive measures, etc. Also, as indicated in previous research, when applied in real time analysis of Social Networking Services (SNS), such as twitter or facebook, it could also become helpful in disaster prevention or handling of disease outbreaks.

Methods using time referring information, such as "year", "hour", or "tomorrow", has been applied in extracting future information and retrieving relevant documents. Moreover, it has been indicated that it is useful to predict future outcomes by using information occurring in present documents. However, although all previously mentioned methods have used future time information, none of them used 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 contribute greatly to the research of future information extraction.

Table 1: Examples of future- and time-related expressions.

Type of expression	Found	Examples; Y=year, M=month (usually appearing as numerical values)
-time-related expressions	70	*Y-Nen M-gatsu kara* (from month M year Y), *kongo Y-nenkan ni* (next in Y years), *Y-gatsu gejun ni mo* (late in year Y), *Y-nen M-gatsu ki made* (till month M of year Y), *kon-getsu chūjun* (this month), *chikai uchi ni* (in the near future), *Nen-matsu made ni* (till the end of the year), etc.
-future expressions	141	*mezasu* ("aim to") (11), *hōshin* ("plan to") (12), *mitooshi* ("be certain to") (9), *kentō* ("consider to") (9), *-suru* ("do") (76), *-iru* ("is/to be") (36), etc.

Our goal is to propose a method for supporting predictions of future events by using not only time referred information or searching for information arranged in chronological order. We propose and evaluate a method for automatic extraction of candidate patterns which refer to the future. By the patterns we mean sentence patterns extracted from future referring sentences generalized using semantic representations and syntactic information.

Investigation of Future Reference Expressions

We performed a study of expressions (words and phrases) which refer to a change in time in general or to the future in particular. The study has been performed by reading through articles from the following newspapers: the Nihon Keizai Shimbun[1], the Asahi Shimbun[2], the Hokkaido Shimbun[3]. All newspapers in their paper and Web version. From these articles we extracted 270 representative sentences which referred to the future. Next, from the sentences we manually extracted future expressions. There were 141 unique future expressions (words, phrases, etc.) and 70 time-related expressions.

Some examples of the expressions are represented in Table 1. There are two kinds of future-related expressions. First consists of concrete expressions which include numerical values, such as "year 2013", or "11 o'clock". Second is derived from grammatical information (verb tense, word order, particles, etc.), such as phrases "will [do something]", "the middle of a month", "in the near future", or particles *-ni* ("in, due, till", point of time), *-made* ("until", implied deadline for continuous action), or *-madeni* ("until", implied deadline for single action). Many of the extracted 270 sentences did not contain typical either time or future expressions. From

[1] http://www.nikkei.com/

[2] http://www.asahi.com/

[3] http://www.hokkaido-np.co.jp/

Table 2: An example of semantic representation of words performed by ASA.

Surface	Semantic (Semantic role, Category, etc.) and grammatical representation
mezasu ("aim to")	No change (activity)-action aiming to solve [a problem]-pursuit; Verb;
hōshin ("plan to")	Other;Noun;
mitooshi ("be certain to")	Action;Noun;
kentō ("consider to")	No change (activity)-action aiming to solve [a problem]-act of thinking;Noun;
-suru ("do")	Change-creation or destruction-creation (physical);Verb;
-iru ("is/to be")	Verb;

all expressions we extracted from the sentences, 55% appeared two or more times, and 45% (nearly half) appeared only once. There could be many variations of future expressions, and many words or phrases could be used as future expressions only in a specific context. However, we can assume that these which appear the most often could be said to have a characteristics of being used as future expressions. Therefore if we consider sentences and their different representations (grammatic, 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. For example, a sentence annotated with semantic roles should provide semantic patterns occurring frequently in future-reference sentences. Below we investigate a method to extract such patterns and verify their practical effectiveness in classification of future-related sentences.

Future Reference Pattern Extraction Method

In this section we describe our method for extraction of word patterns from sentences.

Firstly, 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 a sentence "John killed Mary" the labels for words are as follows: John=`actor`, kill[past]=`action`, Mary=`patient`. Thus the semantic representation of the sentence is `actor-action-patient`.

For semantic role labeling in Japanese we used **ASA**[4], a system, developed by Takeuchi et al., which provides semantic roles of words and generalizes their semantic representation using a thesaurus (Takeuchi et al. 2010). Examples of labels ASA provides for some of the previously mentioned words are represented in Table 2.

Not all words are semantically labeled by ASA. For some words ASA provides only grammatical or morphological information, such as "Proper Noun", or "Verb". We used a heuristic rule to label grammatical information in case when no semantic representation is provided. Moreover, in cases where only morphological information is provided there could be a situation where one compound word is divided.

[4]http://cl.it.okayama-u.ac.jp/study/project/sea.html

Table 3: 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]

For example "Japan health policy" is one semantic concept, but in grammatical representation it takes form of "Noun Noun Noun". To optimize the method we used a set of linguistic rules for specifying compound words. An example of a sentence analyzed this way is represented in Table 3.

All sentences we extracted in the preliminary experiment are labeled this way by ASA. The example sentence when generalized into its semantic representation looks as follows: "[State change] [As] [Place] [Action] [Numeric] [State change] [Place] [Other] [State change]".

Having all sentences analyzed this way we use SPEC, a system for extraction of sentence patterns and text classification developed by Ptaszynski (Ptaszynski et al. 2011). **SPEC**, or **S**entence **P**attern **E**xtraction ar**C**hitecturte is a system that automatically extracts frequent sentence patterns distinguishable for a corpus (a collection of sentences). Firstly, the system generates ordered non-repeated combinations from the elements of a sentence. In every n-element sentence there is k-number of combination groups, such as that $1 \geq k \geq n$, where k represents all k-element combinations being a subset of n. The number of combinations generated for one k-element group of combinations is equal to binomial coefficient, represented in equation 1. In this procedure the system creates all combinations for all values of k from the range of $\{1, ..., n\}$. Therefore the number of all combinations is equal to the sum of all combinations from all k-element groups of combinations, like in the equation 2.

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

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

Next, the system specifies whether the elements appear next to each other or are separated by a distance by placing a wildcard ("*", asterisk) between all non-subsequent elements. SPEC uses all original patterns generated in the previous procedure to extract frequent patterns appearing in a given corpus and calculates their weight. The weight can

be calculated in several ways. Two features are important in weight calculation. A pattern is the more representative for a corpus when, firstly, the longer the pattern is (length k), and the more often it appears in the corpus (occurrence O). Therefore the weight can be calculated by

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

All of those situations are evaluated in the process of classification. Moreover, the list of frequent patterns generated in the process of pattern generation and extraction can be further modified. When two collections of sentences of opposite features (such as "positive vs. negative", or "future-related vs non-future-related") is compared, a generated list of patterns will contain patterns that appear uniquely in only one of the sides (e.g. uniquely positive patterns and uniquely negative patterns) or in both (ambiguous patterns). Therefore the pattern list can be modified by

- using all patterns,
- erasing all ambiguous patterns,
- erasing only those ambiguous patterns which appear in the same number in both sides (zero patterns).

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

- all patterns,
- only n-grams.

Finally, if the initial collection of sentences was biased toward one of the sides (e.g., more positive sentences, or the sentences were longer, etc.), there will be more patterns of a certain sort. Thus agreeing to a rule of thumb in classification (fixed threshold above which a new sentence is classified as either positive or negative) might be harmful for one of the sides. Therefore assessing the threshold is another way of optimizing the classifier. All of the above mentioned modifications are automatically verified in the process of evaluation to choose the best model. The metrics used in evaluation are standard Precision, Recall and balanced F-score.

Experiment

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

Dataset Preparation

We randomly selected 130 sentences out of all collected sentences referring to future events and manually collected another 130 sentences which did not make any reference to the future (describing past, or present events). Out of those sentences we created two experiment sets. The first one containing 100 sentences, with 50 future-referring sentences and 50 sentences not referring to the future (later called "set50"). The second one containing 260 sentences, also with equal distribution of sentences of the two types (later called "set130"). All sentences were preprocessed with ASA.

Table 4: The comparison of the best achieved results (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

Classification Results

We provided both sets (set50 and set130) to the SPEC system to learn and evaluated its classification performance using 10-fold cross validation (90% of sentences for training 10% for test). We compared Precision, Recall and balanced F-score of classification based on patterns and, additionally on n-grams alone.

In the set50, for most situations the F-score reached around 0.67-0.71 for patterns and around 0.67-0.70 for n-grams. In the set130, for most situations the F-score reached around 0.67-0.70 for patterns and around 0.67-0.69 for n-grams. The optimal threshold (from the range 1.0 to -1.0 with 0.0 in the middle) is around 0.0 or slightly biased toward 1.0, which means both sides of the training set in each case were balanced, or slightly biased toward future related sentences. Figure 1 represents the results (F-score) for set50 compared for all patterns and n-grams. Figures 2 and 3 represent Precision and Recall separately for the F-score from Figure 1.

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 way of weight calculation is better, the one using normalized weights, or the one awarding length of patterns. This means we also looked at the cases with length-awarded-weights and zero-patterns deleted, and length-awarded-weights with ambiguous patterns deleted. We will discuss the differences between them in the Discussion section below.

Extraction of Future Reference Patterns

Apart from the automatic classification results, we were also interested in the actual patterns that influenced those results. We extracted the most frequent unique future-reference patterns and non-future-reference patterns from the experiment based on set50. We obtained 1131 patterns for the former and 87 patterns for the latter. Some examples for both kinds of patterns are shown in Table 5.

The patterns are composed of labels described in Section "Future Reference Pattern Extraction Method". The asterisk in some patterns means that the elements are disjoint, for example, in the pattern [Action]∗[State change] there are two elements, [Action] and [State change], which appeared

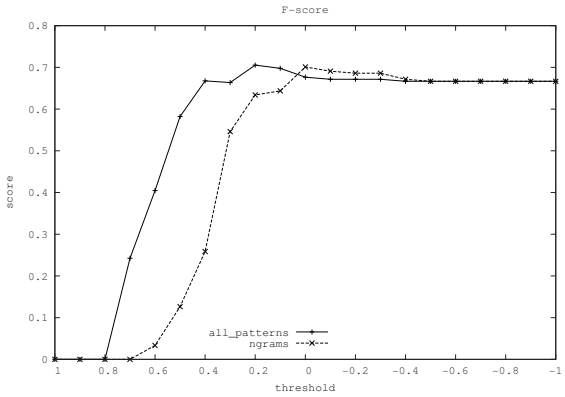

Figure 1: The results (F-score) for set50 compared for all patterns and only n-grams.

in original sentences in exactly this order, and the asterisk indicates that there were other elements between those two.

Discussion

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

Explaining the Classification Results

In general pattern-based approach obtained higher scores than n-grams for most cases, which means that there are meaningful frequent patterns in sentences referring to the future, more sophisticated than simple n-grams. When it comes to the modifications of pattern list and weight calculation, deleting only zero patterns does not influence the results so much. A larger difference can be seen when all ambiguous patterns are deleted and only patterns unique for each side are used. Moreover, awarding pattern length in weight calculation always yields better results. The highest results achieved were F=0.71 with P=0.56 and R=0.98 for the version of the classifier which used pattern list with zero-patterns deleted and length awarded in weight calculation. The greatest improvement of patterns in comparison with n-grams is always in Recall, which means that there are many valuable patterns omitted in the n-gram-only approach. Precision does not change that much and oscillates around 0.55–0.60. For some thresholds n-grams achieve similar or better Precision. This means that the point of around 0.55–0.60 is the optimal maximum there could be achieved with the semantic representation we used in this study. In the future we need to look for a modification to the approach which would generally improve Precision whilst not reducing Recall.

Except comparing patterns with n-grams on the baseline classifier which uses unmodified pattern list and normalized weight calculation, we compare the results also for five other cases (modifying pattern list by deleting zero-patterns, or deleting all ambiguous patterns; and modifying weight calculation by awarding length). When it comes to highest achieved scores 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, due to better Recall usually achieve high F-score even closer to the threshold 1.0, where n-grams usually score lower (compare Figure 1, Figure 2, and Figure 3). To thoroughly verify whether it is always better to use patterns we need to perform more experiments. However, just by using the present data we can conclude that patterns achieve generally better results.

Next we compare the two datasets, **set50** and **set130**. The results of comparison are represented in Table 4. The results do not differ greatly. However, when we look at Figure 4, the F-score for the version of the classifier using pattern list with all ambiguous patterns deleted performs slightly better than the two others (though the differences are not quite statistically significant with $p<0.06$). Comparing these results to the results in Figure 5 indicates that the performance is generally better when the length of patterns is used to modify weight calculation. Especially both modified versions of the classifier (without zero-patterns and without all ambiguous patterns) retain high F-score thorough all threshold span (from 1.0 to -1.0). The same can be said of the results for set130. Comparing Figure 6 and Figure 7 also shows that

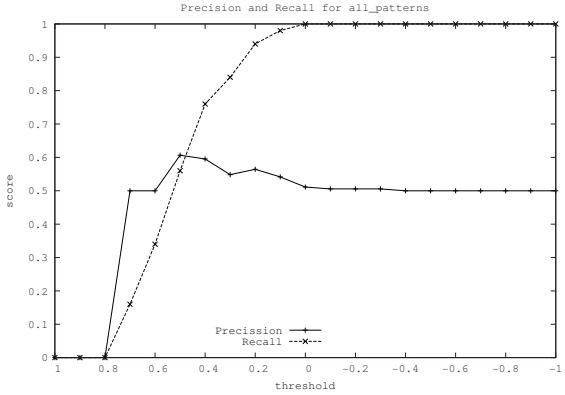

Figure 2: Precision and Recall for patterns.

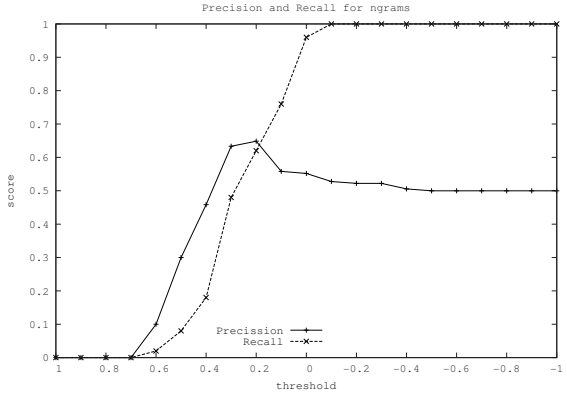

Figure 3: Precision and Recall for n-grams.

Table 5: The 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]

applying pattern length in weight calculation yields better results within the specified threshold. Moreover, it is also advantageous to either exclude zero-patterns or all ambiguous patterns from pattern list. Also, worth mentioning is the fact that the performance for the algorithm as a whole is similar for set50 and set130. Usually the larger the dataset the more ambiguities it contains, thus the results often degrade. With the proposed approach the differences are negligible for most cases (compare Figure 4 and Figure 6) or small (compare Figure 5 and Figure 7).

Inquiry Into Extracted Future Reference Patterns

Using the Language Combinatorics method we could extract frequent patterns distinctive for sentences referring to the future and those not referring to the future. Each time the classifier uses a pattern from the pattern list, the used pattern is extracted and added to a separate list. This extraction is performed for each test in the 10-fold cross validation. By taking the patterns extracted this way from all tests and leaving only the frequent ones (with occurrence frequency higher than 1), we get the refined list of most valuable patterns (those used generally most often). We investigated those patterns and the types of sentences they were used in.

The following example sentences (in order: Romanized Japanese, Translation, Semantic representation) contain the 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 prospected that opening of the tunnel will make it possible to carry 500,000 people a day, which will lead to reducing traffic congestion, according to the government.) [Action]-[Time]-[Object]-[State_change]-[Other]-[Noun]-[Action]-[Organization]-[Action]-[Verb]-[State change]

The following 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 the awareness of the problem, it still is likely to spread the 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 the ones we studied manually in Section "Investigation of Future Reference Expressions". These include time-related expressions ("late May", "from December 1", "from March 1, 1996") and future reference expressions ("is prospected", "is planned to", "is likely to")

Next, we examined sentences containing non-future patterns. The following example sentence contains a 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 a pattern `[Place]*[Place]*[No_state_change(activity)]`

Example 6. *Nagata-ku wa Hanshin Daishinsai de ōkina gai wo uketa chiiki de, koko de wa Betonamu no hito ga kazu ōku hataraite iru.* (Nagata Ward, one of the areas that were

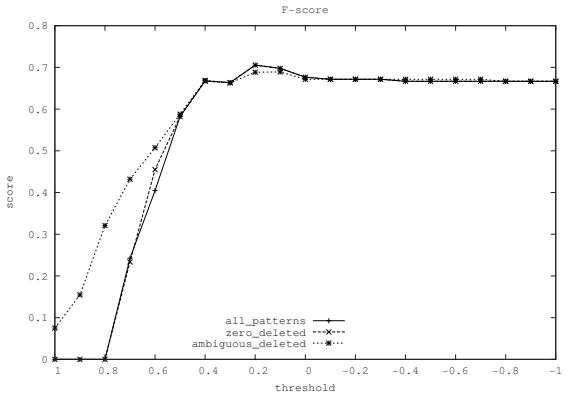

Figure 4: F-scores for the classifier with three different versions of pattern list modification for set50.

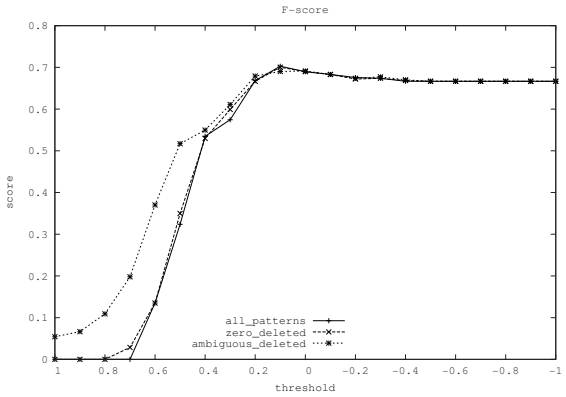

Figure 6: F-scores for the classifier with three different versions of pattern list modification for set130.

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 a 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 a phrase *to wa ienai* ("it cannot be said/considered that"), which is labeled by the semantic role labeling system ASA as [Action], which is a frequent label in future-referring sentences, however, just by this fact the sentence is not yet classified as future-related. Example 6 contains a phrase *–shiteiru* ("to do/is being") which is labeled by ASA as [No state change(action)], however, in future-related sentences this phrase is often labeled as [State change]. As for Example 7, although it contains time-related expressions ("June last year"), the use of sophisticated patterns taking into account wider context, allows correct disambiguation of such cases. Furthermore, since this pattern, although containing time-related expression, is not on the list of future reference patterns, it can be said that the presence of time-related information alone does not influence the classification that much. Instead, other elements of a pattern, such as appropriate tense, etc., together with time-related expressions constitute the pattern as being distinctive for future reference sentences.

There were many future reference patterns with high occurrence frequency (see Table 5), which means the sentences contain many of those patterns. Therefore we can say that "the future" in general has high linguistic expressiveness. For non-future reference patterns, the occurrence frequency of patterns is low, which means that there was in general a large number of patterns, each of them used only once (thus they were not included in the list of general frequent pat-

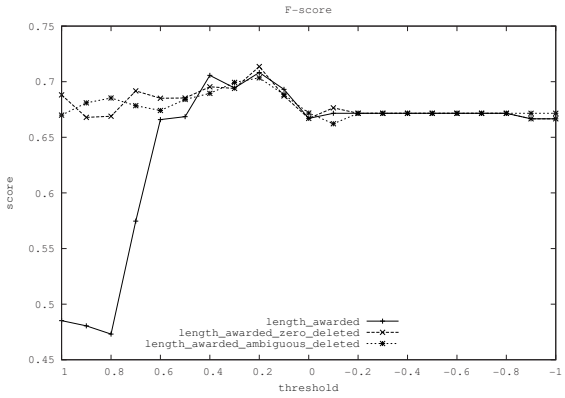

Figure 5: F-scores for length awarding in weight calculation for set50.

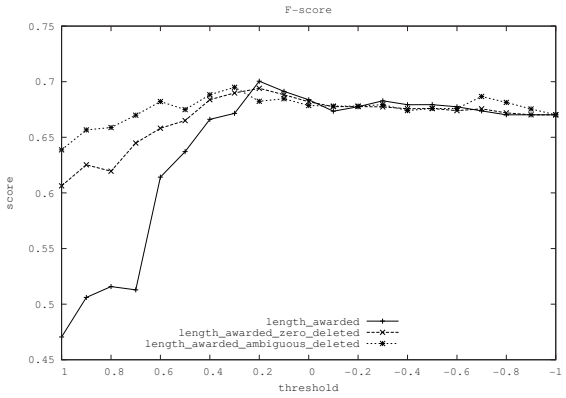

Figure 7: F-scores for length awarding in weight calculation for set130.

terns). Because the variety of patterns is so high, it can be said that there are no particularly distinctive patterns for sentences not referring to the future.

Conclusions and Future Work

We investigated characteristics of expressions referring to the future based on newspapers and classified sentences (future-referring vs. non-future-referring) using a pattern based approach with semantic representations of sentences.

We tested the method on two datasets of different sizes. We found out that the method performs well for both sets (F-score around 0.70–0.71). Although the data we used in the experiments was not large (only 100 sentences (50 future-related and 50 other) for the first set and 260 sentences (130-130) for the second set), we were able to verify that it is possible to determine fully automatically whether a sentence is referring to the future or not. As the results were promising we plan to perform an experiment on even larger dataset and annotate large newspaper corpora according to their time point of reference (future vs past or present).

In the future, except increasing the experiment datasets, we plan to approach the data from different viewpoint to increase Precision of the classifier. We also plan to verify which patterns exactly influence the results positively and which hinder the results. This knowledge should allow determination of a more general model of future-referring sentences. Such a model would be useful in retrieving probable unfolding of events, and would contribute to the task of trend prediction in general.

As our next step, we plan to apply the future reference classification method to real-world tasks by finding new contents with future reference and sorting it in chronological order, which would allow supporting future prediction in everyday life.

References

Eiji Aramaki, Sachiko Maskawa, Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1576.

Sitaram Asur, Bernardo Huberman. 2010. Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 492–499.

R. Baeza-Yates. 2005. Searching the Future. *SIGIR Workshop on MF/IR*.

Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama, Katsumi Tanaka. 2009. Supporting analysis of future-related information in news archives and the Web. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pp. 115–124.

Adam Jatowt, Ching-man Au Yeung. 2011. Extracting collective expectations about the future from large text collections. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1259–1264.

Adom Jatowt, Hideki Kawai, Kensuke Kanazawa, Katsumi Tanaka, Kazuo Kunieda, Keiji Yamada. 2013. Multi–lingual, Longitudinal Analysis of Future-related Information on the Web. *Proceedings of the 4th International conference on Culture and Computing 2013*, IEEE Press.

Kensuke Kanazawa, Adam Jatowt, Satoshi Oyama, Katsumi Tanaka. 2010. Exracting Explicit and Implicit future-related information from the Web(O) (in Japenese). *DEIM Forum 2010*, paper ID: A9-1.

Kensuke Kanazawa, Adam Jatowt, Katsumi Tanaka. 2011. Improving Retrieval of Future-Related Information in Text Collections. *2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 278–283.

Nattiya Kanhabua, Roi Blanco, Michael Matthews. 2011. Ranking related news predictions. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 755-764.

Michael Matthews, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, Hugo Zaragoza. 2010. Searching through time in the New York Times. *Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval*, pp. 41–44.

Omar Alonso, Jannik Strötgen, Ricardo Baeza-Yates, Michael Gertz. 2011. Temporal Information Retrieval: Challenges and Opportunities. *TWAW*, Vol. 11, pp. 1–8.

Octavian Popescu and Carlo Strapparava. 2013. Behind the Times: Detecting Epoch Changes using Large Corpora, *Proc. of IJCNLP 2013*.

Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, Issue 1, pp. 24-36.

Kira Radinsky, Sagie Davidovich and Shaul Markovitch. Learning causality for news events prediction. *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012.

Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, Yuuki Moriyasu. 2010. Construction of Argument Structure Analyzer Toward Searching Same Situations and Actions. *IEICE Technical Report*, Vol. 109, No. 390, pp. 1-6.