# A Prototype Method for Future Event Prediction
# Based on Future Reference Sentence Extraction

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

†Department of Creative Engineering                    ‡Department of Computer Science
National Institute of Technology, Kushiro College      Kitami Institute of Technology
2-32-1 Otanoshike, Kushiro, 084-0916, Japan            165 Koen-cho, Kitami, 090-8507, Japan
{yoko,honma}@kushiro-ct.ac.jp                          {ptaszynski,f-masui}@cs.kitami-it.ac.jp

## Abstract

This paper presents our study on future prediction support method based on state-of-the-art natural language processing and pattern extraction techniques. We propose two practical applications of the method. The first is supporting future prediction in human users. The second is automatic future prediction based on provided data. We conduct experiments on the developed prototype method and evaluate its effectiveness. In the development of the method we assumed that sentences from official sources such as newspapers that refer to future events could be useful for future prediction. By using sophisticated patterns combining morphemes and semantic roles, we successfully extracted future reference sentences and effectively used them in future prediction performed both by human users as well as by the fully automatic prototype method. In the experiments we tested the method on a number of future prediction questions from official Future Prediction Competence Test, performed yearly in Japan. Both, the results of the support application as well as the automatic method were higher then the results of original test participants. Moreover, the prototype fully automatic method greatly outperformed all human users. This suggests that the method can be expected to not only reduce the response time and the amount of information needed for prediction in humans, but also perform the prediction automatically on a level comparable and exceeding an average human.

## 1   Introduction

One of the main tasks in the field of Artificial Intelligence (AI) is providing means for understanding of the reality around us and predicting possible outcomes of certain decisions. The understanding is often restricted to analysis of specific general meaning of, e.g., a particular instance of language behavior (sentences, documents, etc.). For example, in subfields of AI such as Natural Language Processing (NLP), or Sentiment Analysis, speaker attitudes and emotions expressed in a sentence are in the focus of analysis. A different kind of task from the field of NLP, which we focus on in this research, is predicting trends of future events on the basis of provided limited information.

In everyday life people often apply their knowledge and experience about past events as well as their own general knowledge to predict future events. In such everyday predictions people often use widely available resources (newspapers, Internet). Especially people who bare significant social responsibility, such as politicians, managers, strategists, planing specialists, or policy developers in large companies are in need of tools that would support them in their predictions and decision making, since the company's results and profits depend on how accurate their competence in future trend prediction is.

The goal of this study is to provide a tool, that would help in such predictions. In particular, we aim at creating a system that would perform such predictions automatically. To achieve this we focus on sentences referring to the future as potentially useful in predicting future unfolding of events.

When humans consider future events, they usually process the information from various domains to join it in one reasoning. For example, when a company finds out that a decision they need to take depends on two events/factors, 'X' and 'Y' (although in real life the number of factors is much larger), they can prepare four strategic decisions A, B, C or D for their company management, depending on the predictions on what would happen in the future when they select:

1. Strategic decision 'A' when both events 'X' and 'Y' take place.

2. Strategic decision 'B' when the event 'X' takes place but the event 'Y' does not.

3. Strategic decision 'C' when the event 'Y' takes place but the event 'X' does not.

4. Strategic decision 'D' when both events 'X' and 'Y' do not take place.

When people select their actions from a range of possible options, usually they consider and combine a wide spectrum of information, including one's and also other people's experiences and expertise regarding the events. Obtaining such information for future predictions is a challenging task requiring much time and labor with a lot of information to process and deep foresight ability for making the decision.

When predicting future events such as X and Y, for example, "Will consumers do shopping with augmented reality (AR) in two years time?" we can think of the following possibilities: (1) More than half of consumers will shop with AR; (2) Several percent of consumers will shop with AR; (3) The way of shopping will not change comparing to current situation. This way we can reduce the problem of future prediction to predicting which of the limited number of potential answers (two or more) has higher probability of occurring.

The predicting thus can be formulated as selecting the correct answer, even if the answer is not yet specified at the time of prediction.

Previous studies have shown that data mining using simple statistics can support such predictions regarding future outcomes of events. However, to achieve that, one needs to process numerous numeric data, which requires professional skills, and the expertise to explain the numbers in a comprehensible manner.

There have also been studies on predicting future outcomes of events with the use of NLP techniques. Some of them have proposed applying causality information and past events [Radinsky et al. 2012], which assume that when the event A happens, the event B will usually follow. However, such methods usually are limited to general events from the range of widely perceived common sense (e.g., "what will happened when an apple falls on ones head"). Others applied methods based on keyword extraction with their occurrence frequencies in a timeline using past events, temporal expressions and event-related keywords [Kanazawa et al. 2011].

As for a different research, [Nakajima et al. 2016] have proposed a method for automatic extraction of future reference sentences using combined morphological and semantic (morphosemantic) information and suggested that future reference sentences could be applied in supporting predictions about future events, since they usually contain various related background information, which is also used as the source of knowledge for prediction.

In this research we propose a prototype method which applies such future reference sentences in the process of future prediction. In the experiments we compare the performance of laypeople and the proposed automated approach and discuss its effectiveness.

The outline of this paper is as follows. In Section 2 we describe previous research related to the prediction of future events. Section 3 describes the proposed method applying automatic extraction of references to future events and the experiments evaluating the method. Section 4 describes the experiment to verify the effectiveness of future reference sentences applied to real-world future prediction events. Section 5 describes the prototype method for automated prediction of event unfolding. Finally, section 6 contains conclusions and our plans for improvement of the prototype method and possible applications of the proposed method in future tasks.

## 2 Previous Research

There has been a number of studies in linguistically expressed future reference detection.

For example, [Baeza-Yates 2005] investigated half a million of sentences containing future events extracted from one day of Google News (http://news.google.com/), and found out that scheduled events occur with high probability and with correlation between the occurrence of an event and its time proximity. Therefore the information about upcoming events is of a high importance for predicting future outcomes.

[Jatowt et al. 2009] also focused on news articles, and used a rate of incidence of reconstructed news articles over time to forecast recurring events, which they used for supporting human user analysis of occurring future phenomena.

[Kanhabua et al. 2011] in their investigation of newspaper articles, found out that one-third of all sentences contains some reference to the future.

[Kanazawa et al. 2010] extracted implications for future information from the Web using explicit information, such as time expressions.

When it comes to predicting the probability of an event to occur in the future, [Jatowt and Au Yeung 2011] have proposed a clustering algorithm for detecting future phenomena based on the information extracted from text corpus, and proposed a method of calculating the probability of an event to happen in the future.

Also, [Kanazawa et al. 2011] extracted 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.

[Aramaki et al. 2011] used SVM-based classifier on Twitter to perform classification of information related to influenza and tried to predict the spread of the disease by using a truth validation method.

[Kanazawa et al. 2011] 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.

[Radinsky et al. 2012] proposed the Pundit system for prediction of future events in news based on causal reasoning derived from a similarity measure calculated using different ontologies.

[Jatowt et al. 2013] studied relations between future news in English, Polish and Japanese by using keywords queried on the Web.

Recently, [Zhang et al. 2016] performed a variation analysis of the evolution of technology for techniques to learn causality relations from past events for the extraction of features that cause future changes.

The above findings have lead us to the idea that by using expressions referring to the future included in newspaper articles it could be possible to support human users in the process of future prediction as one of the activities people perform everyday. Moreover, by applying appropriate reasoning algorithm, it could be possible to automatize the process and create a system capable of automatic future prediction. A method like that could have a number of applications in various fields, such as in corporate management, trend foresight, and preventive measures, etc. Also, as indicated by previous research, when applied in real time analysis of Social Networking Services (SNS), such as Twitter or Facebook, it could also be helpful in disaster prevention or handling of disease outbreaks.

As for practical applications used in future trend prediction, Stanford Temporal Tagger[1] converts natural language input such as "next Wednesday at 3pm" into particular calendar based schedule such as "2016-02-17 T 15:00" depending on the assumed current reference time. Similar is possible

---

[1]https://nlp.stanford.edu/software/sutime.html

with HidelTime[2] [Strötgen and Gertz] is as well.

Methods like above, using time referring information, such as "year", "hour", or more general "tomorrow", etc., has been applied before in extracting future information and retrieving relevant documents. It has also been indicated that it is useful to predict future outcomes by using information occurring in present documents. A main difference with our research is the fact that we focused not only on the explicit simple and obvious patterns, such as time expressions, but on more sophisticated expressions, combining both morphological and semantic information, and automatically extracted such ***morphosemantic*** sentence patterns.

# 3 Automatic Extraction of Future Reference Sentences

In this section, we describe the method for extraction of future reference sentences from news corpora.

Future reference sentences include both explicit as well as implicit expressions referring to the future. Explicit expressions include e.g., future temporal expressions, or words and phrases referring to the future (e.g. will∼, is expected to∼, plan to∼, etc.).

However, many important sentences do not contain such explicit expressions, but the information regarding future outcomes is implicit. See the example below regarding the future of America's army troops dispatch to Afghanistan.

*"He rejoiced to hear that President Obama had reemphasized the need to focus on the War on Terror in Afghanistan, increasing the likelihood of an early withdrawal of U.S. troops from Iraq."*

The sentence does not contain any future referring expressions. Moreover, the sentence is in past tense ("rejoiced", "had reemphasized"), and therefore it is not possible to specify that the sentence refers to the future by using standard methods. Yet, the sentence clearly presents potential future outcomes ("withdrawal of U.S. troops from Iraq") with the use of implicit information.

The method proposed here deals with both explicit as well as implicit information, such as above. It consists of two stages. Firstly, the sentences are represented in a morphosemantic structure [Levin and Rappaport Hovav 1998] (combination of semantic role labeling and morphological information). Secondly, frequent combinations of such patterns are automatically extracted from training data and used in classification.

## 3.1 Morphosemantic Patterns

In the first stage of the method, all sentences are represented in **morphosemantic structure** (MS) for further extraction of **morphosemantic patterns** (MoPs).

The idea of MS has been described widely in linguistics and structural linguistics. [Levin and Rappaport Hovav 1998] distinguish morphosemantics as one of the basic type of morphological operations on words, modifying the Lexical Conceptual Structure (LCS) of a word.

MoPs have been applied in analysis of an Indonesian suffix *–kan* [Kroeger 2007], improving links between WordNet

Table 1: An example of a sentence analyzed by ASA.

**Example I: Romanized Japanese (RJ):** *Ashita kare wa kanojo ni tegami o okuru darō.* **/ Glosses:** Tomorrow he TOP her DIR letter OBJ send will (TOP: topic particle, DIR: directional particle, OBJ: object particle.) **/ English translation (E):** He will [most probably] send her a letter tomorrow.

| No. | Surface | Label |
|---|---|---|
| 1 | *ashita* | [Time-Point] |
| 2 | *kare ha* | [Agent] |
| 3 | *kanojo ni* | [Patient] |
| 4 | *tegami o* | [Object] |
| 5 | *okuru darou* | [State_change]-[Place_change]-[Change_of_place(physical)] |

synsets [Fellbaum et al. 2009], or analysis of a Croatian lexicon [Raffaelli 2013].

Below we describe the process of morphosemantic representation of sentences we applied in this research.

At first, the sentences from the datasets (Japanese newspaper corpora) are analyzed using **semantic role labeling** (SRL), which provides labels for words and phrases according to their role in sentence context.

For SRL in Japanese we used ASA, a system which provides semantic roles for words and generalizes their semantic representation using an originally developed thesaurus [Takeuchi et al. 2010]. An example of SRL provided by ASA is represented in Table 1.

However, not all words are semantically labeled by ASA. The omitted words include, e.g., grammatical particles, or function words not having a direct influence on the semantic structure of the sentence, but in practice contributing to the overall meaning. For those remaining words we used a morphological analyzer MeCab[3] in combination with ASA to provide morphological information, such as "Proper Noun", or "Verb". Moreover, as a post-processing procedure we added a set of linguistic rules for specifying compound words in cases where only morphological information was provided.

Finally, for cases where the labels provided by ASA were too specific (see Table 1), we normalized and simplified the labels according to the following label priorities.

1. Semantic role (Agent, Patient, Object, etc.)
2. Semantic meaning (State_change, etc.)
3. Category (Dog → Living animal → Animated object)
4. In case ASA does not provide any of the above labels, perform compound word clustering for parts of speech (e.g., "International Joint Conference on Artificial Intelligence" → `Adjective Adjective Noun Preposition Adjective Noun` → `Proper_Noun`)

4.1 If a compound word can be specified, output the part-of-speech cluster.

4.2 If it is not a compound word, output part-of-speech for each word.

Below is an example of a sentence represented in the above morphosemantic structure.

**Romanized Japanese:** *Nihon unagi ga zetsumetsu kigushu ni shitei sare, kanzen yōshoku ni yoru unagi no ryōsan ni*

---

Table 2: Examples of extracted morphosemantic patterns (MoPs).

| Occ. | Future Reference Patterns | Occ. | Non-future Reference Patterns |
|---|---|---|---|
| 43 | [Action]*[Object] | 4 | [Numeric]*[Agent] |
| 42 | [Action]*[Action] | 4 | [Verb]*[Artifact] |
| 26 | [Action]*[State change] | 5 | [Place]*[Agent] |
| 20 | [State change]*[Object] | 4 | [Person]*[Place] |
| 16 | [State change]*[State change] | 3 | [Numeric]*[Agent]*[Action] |
| ⋮ | | ⋮ | |

*kitai ga takamatte iru.*

**English:** As Japanese eel has been specified as an endangered species, the expectations grow towards mass production of eel in full aquaculture.

**SRL:** `[Object][Agent][State_change][Action]-` `[Noun][State_change][Object][State_change]`

## 3.2 Future Reference Pattern Extraction

From sentences represented in morphosemantic structure we extract frequent MoPs, by firstly, generating ordered non-repeated combinations from all sentence elements. In every $n$-element sentence there is $k$-number of combination groups, such as that $1 \leq k \leq n$. Next, all non-subsequent elements are separated with a wildcard ("*", asterisk). Pattern lists extracted this way from training sets are then used in classification of test and validation sets.

For all patterns generated this way their occurrences $O$ are calculated, and frequent ($O \geq 2$) patterns are retained. Next, the occurrences are used to calculate pattern weight. Two features are important in weight calculation: pattern length $k$ (number of elements it contains) and its occurrence $O$ (how many times it occurs in the dataset) Thus in the experiments we modified the weight by

- awarding length (`LA`),
- awarding length and occurrence (`LOA`),
- awarding none (normalized weight, `NW`).

The generated list of frequent patterns can be also further modified. When two collections of sentences of opposite features (such as "future-related vs. non-future-related") are compared, the list 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). Thus we also modified pattern lists by

- using all patterns (`ALL`),
- erasing all ambiguous patterns (`AMB`),
- erasing only those ambiguous patterns which appear in the same number in both sides (zero patterns `OP`, since their normalized weight is equal zero).

Moreover, a list of patterns will contain both the sophisticated patterns (with disjointed elements) as well as more common n-grams. Therefore the system can be trained on a model using

- patterns (`PAT`), or
- only n-grams (`NGR`).

All combinations of the above modifications are tested in the experiments.

Examples of extracted MoPs of FRS and non-FRS with their occurrences were shown in Table 2.
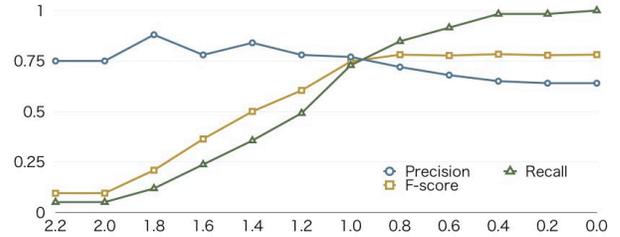


Figure 1: Final overall results of fully optimized model.

## 3.3 Future Reference Sentence Extraction with Morphosemantic Patterns

From three newspaper corpora[4] we collected and annotated a dataset containing equal number of (1) sentences referring to future events and (2) other (describing past, or present events). We conducted an evaluation experiment with training dataset containing 130 sentences each, furthermore as the test data we used randomly extracted additional 170 sentences from the news corpora.

The test datasets were applied in a text classification task with 10-fold cross validation. Each classified test sentence was given a score calculated as a sum of weights of patterns extracted from training data and found in the input sentence. The results were calculated with Precision, Recall and balanced F-score. We compared fourteen classifier versions. The results indicated that the highest overall performance was obtained by the version using pattern list containing all patterns (including ambiguous patterns and n-grams). We looked at top scores within the threshold, checked which version got the highest break-even point (BEP) of Precision and Recall, and calculated statistical significance of the results.

Finally, we compared the proposed method to [Jatowt et al. 2013], who extracted future reference sentences with 10 words explicitly referring to the future, such as "will" or "is likely to", etc. In comparison, the proposed method obtained better results even when only 10 most frequent MoPs were used (see Table 3 for details).

Moreover, we verified the performance of the fully optimized model. We retrained the best model using all sentences from the initial training dataset and verified the performance by classifying the new validation set. The final overall performance was represented in Figure 1. Finally, the obtained BEP was 0.76.

Table 3: Comparison of results for validation set between different pattern groups and the state-of-the-art.

| Pattern set | Precision | Recall | F-score |
|---|---|---|---|
| top 10 patterns | 0.39 | **0.49** | **0.43** |
| top 10 patterns with over 3 elements | **0.42** | 0.37 | 0.40 |
| top 5 patterns | 0.35 | 0.35 | 0.35 |
| Optimized (see Fig. 1) | **0.76** | **0.76** | **0.76** |
| [Jatowt et al. 2013] (10 phrases) | **0.50** | 0.05 | 0.10 |

---

[4] *Nihon Keizai Newspaper*, *Asahi Newspaper*, *Hokkaido Newspaper*.

## 4 Future Prediction Support Experiment

The validity of the method described in previous section needs to be tested twofold. Firstly, by verifying the capability of the method to provide a support for human users performing a task of prediction of how a future event will unfold. Secondly, by testing a fully automated process of prediction.

In this section we present a validation experiment for the effectiveness of using Future Reference Sentences in the task of supporting human users in predictions regarding future events.

### 4.1 Experiment Setup

In the experiment for supporting future trend prediction we used the fully optimized model of FRS trained on MoPs described in section 3.3. The model was applied to extract new FRS concerning a specific topic, from the available newspaper data. Such sentences were further called **future prediction support sentences** (FPSS). Future prediction task was performed by a group of thirty laypeople[5], who were told to read the FPSS and reply to questions asking them to predict the future in 1–2 years from now, or from the starting point of prediction.

The questions were taken from the Future Prediction Competence Test (FPCT, japanese: *Senken-ryoku Kentei*), released by the Language Responsibility Assurance Association(LRAA, japanese: *Genron Sekinin Hoshō Kyōkai*)[6], a nonprofit organization focused on supporting people of increased public responsibility (managers, politicians) and people responsible of making decisions influencing civic life. In particular, the organization helps preparing public speeches and responsibility bound presentations, by training individuals in predicting possible outcomes of future events. A part of this training consists of taking part in the FPCT.

The FPCT is an examination that measures prediction abilities in humans regarding specific events that are to happen in 1–2 years in the future. It has been initiated in 2006 and from that time it has been performed six times. The test consists of various questions, including multiple choice questions (e.g., "Will US Army contingent in Afghanistan increase or decrease during next year?"), essay questions (e.g., "Describe economic situation of a country after next two years"), and questions that must be answered using numbers (e.g., "What will be the exchange rate of Japanese Yen to US Dollar after two years"), and they are scored after those particular events have come to light.

The questions for the experiment were selected from the 4th of the past six FPCTs, as it had the largest total number of questions, and respondents, which would assure the highest possible objectivity of the evaluation. Implemented in 2009, the 4th FPCT contained questions regarding predictions for 2010 and 2011, and the scoring was performed in 2011. Respondents were to choose to answer at least 15 questions from a total of 25 questions in six areas, namely, politics, economics, international events, science and technology, society,

---

[5]25 males and 5 females, age groups from university students studying computer science (28 user samples) to their fifties (2 user samples).

[6]http://genseki.a.la9.jp/kentei.html

---

```
Question 3:  Predict the stationing status
of US troops in Afghanistan at the end of
June 2011.

(A) The U.S. troops will be still present and
    further reinforced comparing to October
    2009.
(B) The U.S. troops will be still present on
    similar level comparing to October 2009.
(C) The U.S. troops will be still present but
    in decreased number comparing to October
    2009.
(D) The U.S. troops will be completely
    withdrawn.

Answer:
[ 1st candidate:        / 2nd candidate:
/ 3rd candidate    ]

Specify which sentence (number ID) from
the prepared Future Prediction Support
Sentences was most useful in making the
above decision:

1st candidate: [    ]    2nd candidate: [
]    3rd candidate: [    ]
```

Figure 2: An example of one multiple choice question from the 4th Future Prediction Competence Test with an additional question about which of the automatically extracted sentences presented to the user was the most useful.

and leisure. The test contained a large number of multiple choice questions and several questions requiring predicting specific numbers. There was also a small number of questions requiring a written explanation of the reasoning for the prediction. When participating in the test, respondents can browse any and all available materials, and are free to seek opinions of others in answering the question, but the submission deadline was fixed and set at December 31st, 2009 (end of the year). The scoring is set at 90 total points on prediction questions and 30 total points for descriptive questions, with a total of 120 points.

In the future prediction support experiment the developed method extracts FPSS related to a given question and provides assistance for human users on which answer to choose during the test. Therefore for its evaluation we limited the questions to multiple-choice questions. Questions with two or more (multiple) possible answers were selected from the 4th FPCT and applied as questions for the experiment. One of the questions was represented in Figure 2.

### 4.2 Data Preparation

In this section we describe the process of data preparation for the experiment. Firstly, a total of 7 multiple-choice questions were selected from the 4th FPCT. Next, for each question we extracted a number of FPSSs from news corpus to be read by participants. Differently to the original settings of the Future Prediction Competence Test, where the participants could refer to any information and had the whole year to prepare their answers, participants of our experiment were to only use the provided FPSSs answer the questions at the time in the experiment.

The FPSSs for each question is collected in the following

Table 4: Keywords for collecting future reference sentences (FRS) for each of the questions.

| No. | Keywords |
|---|---|
| Q1-1 | Participation in regional government by foreigners with permanent resident status \| Participation in regional government permanent resident alien |
| Q1-2 | Husband and wife retaining separate family names |
| Q2 | Midterm elections \| (Republic Party \| Democratic Party) & (United States \| America) |
| Q3 | Afghanistan |
| Q4 | Analog broadcasting \| Digital broadcasting |
| Q5 | Child allowance |
| Q6 | (Democratic Party \| Liberal Party \| Ruling Party) & election |

steps. At first we extracted from the *Mainichi Newspaper*'s entire 2009 year all sentences related to the questions on the basis of topic keywords (Table 4), selected as nouns that appeared in the original questions or answering options. We also manually expanded search query by adding semantically related keywords.

The sentences were represented in morphosemantic structure, analyzed using our fully optimized model, and sorted in a descending order of resemblance to FRS. Next, we retained only those FPSS which scores were over 0.0 and presented the highest 30 of them to the subjects in chronological order (date the sentence appeared in a newspaper), so the subjects had a better image of how the events unfolded, which would make the prediction more natural.

The limit of thirty sentences was set so the subjects did not become tired of the task. However, we kept the rest of the sentences in case the subjects insisted on further reading. In situations where the list of initial sentences extracted with topic keywords was less than thirty, we presented to the subjects all sentences which had a probability of being FRS.

The questions were answered directly after reading only the FPSS. Additionally, the laypeople were asked to report the ID number of the FPSS they referred to in their answer (or the FPSS considered as the most informative or useful).

We evaluated their answers based on the original scoring schema. In particular, each of the questions 1, 2, and 7 were allocated up to 3 points. Moreover, in questions 2–5 the laypeople were allowed to make up to three candidate choice answers: primary, secondary and tertiary candidate, assigned 3, 2 points and 1 point, respectively.

### 4.3 Experiment Results and Discussion

The results of experiment were summarized in Table 5.

In the performed experiment, the average score of the experiment participants was 38.10% (see Table 5). In comparison, in the original FPCT, the average score percentage of the test participants was 33.44%. Therefore the results of our experiment participants were slightly higher.

One issue with the performed experiment was that the questions for the prediction task were in fact past events for the laypeople. Therefore, although the questions were very specific, meaning it was not likely that the participants knew or remembered the events, there was always a chance

Table 5: Performance of subjects in the future prediction task supported with the proposed method with comparison to original results of the 4th Future Prediction Competence Test.

| | correct answer accuracy | | |
|---|---|---|---|
| | average | highest | lowest |
| Original FPCT | 0.3344 | 0.6111 | 0.0666 |
| Our experiment | 0.3810 | 0.6190 | 0.1428 |

that some of the experiment participants might have already known the unfolding of the events in question and use this knowledge in their advantage. Therefore just in case, we also warned the experiment participants that in answering, they should use only the knowledge provided in the FPSS.

Although participants of our experiment obtained higher scores than the participants of original test, the improvement was not large. However, as the major contribution of our method for future prediction support the following can be considered. Even if we acknowledge that the improvement was not sufficient, and that our subjects performed similarly to original participants, our subjects made their decisions based only on about thirty specifically extracted FPSSs and were given only a short time for decision, whereas the original participants had over one year term for preparing the answer, unlimited access to all available data and receiving to help from any other people including experts.

It indicates that accurately extracted future reference sentences are useful in prediction of future event unfolding for people who have no knowledge regarding those future events. Hence, supporting future prediction with FPSS for specific topics can be considered at least as efficient as collecting available information by oneself for one year time.

## 5 Prototype Method for Automatic Future Prediction

In this section, we describe and validate a prototype method for prediction of future event unfolding. In the validation experiment we aim to perform the previous task – described in Section 4 – fully automatically.

### 5.1 Method Description

We developed the prototype method for automatic future prediction to analyze the questions from the Future Prediction Competence Test used in previous experiment. Although the method can be adapted to analyze any content, in this research we limited its functionality to the existing data to make the evaluation possible and as objective as possible.

The method consists of following steps.

1. Building an optimized model for Future Reference Sentence (FRS) extraction (Section 3.3),
2. Extracting topic keywords from questions about future unfolding of events (Section 4.2),
3. Applying the FRS extraction model and the topic keywords to extract FRS related to question from a limited corpus data,
4. Train a new event-topic-specific model on the extracted topic-related FRS, using the method for Automatic Extraction of Future Reference Sentences (Section 3),
5. Analyze answers to the each Question and choose the one with highest score as the correct answer.

## 5.2 Experiment Setup and Data Preparation

We evaluated the performance of the prototype method for automatic a future prediction. Originally, the evaluation task was that laypeople were to read the automatically extracted Future Prediction Support Sentences (FPSS) related to Questions from the Future Prediction Competence Test (FPCT) and to select those answers to the questions they considered as correct using only the provided FPSS.

The method for automatic prediction takes the human out of the loop in the prediction task. Therefore in practice the method would account for automatically reading through the limited corpus and provide automatic inference regarding answers to the FPCT questions based only on the automatically learned information.

In the evaluation, as the reference corpus for the method for learning we applied the same newspaper corpus as in Section 3.3, but limited to one year, namely 2009, which presumably contained news articles related to the questions.

For each of the questions we used the extracted topic keywords (Table 4) with the fully optimized model (Section 3.3) to extract FRS related to each question. Next, the newly obtained FRS were used as training data to train a new model for each question. Finally, the newly created topic-oriented FRS-based model was used to analyze the answers for each question (see example in Figure 2) and the answer with the highest score was selected as the correct one.

Moreover, in order to analyze the influence of FRS on the accuracy rate of correct answers, we developed two versions of the prototype method.

Ver. 1: Using for training thirsty or less FRS (condition similar to the one under which experiment participants performed the future prediction task in the future prediction support experiment, explained in Section 4),

Ver. 2: Using for training all FRSs which scored over 0.98 (condition experimentally selected as optimal for FRS, explained in Section 3.3).

## 5.3 Experiment Results

To put the developed prototype method in the same standpoint as human participants, in evaluation of the prototype method we adopted the same weighted scoring schema as in the future prediction support experiment (Section 4.2). Namely, for questions 1, 2, and 7 if the prototype method answered correctly, it obtained 3 points for each question. Furthermore, for questions 2–5 if the correct answer was selected by the prototype method as either first, second or third candidate, it was assigned 3, 2 points or 1 point, respectively.

The results of the prototype method for each question were shown in Table 6. An example of scoring of answers for two questions (Q1-1 and Q3), for both versions of the prototype method were represented in Table 7. For each question the answer with the highest score is selected as the correct one.

The version of the method using thirty FRS, obtained the accuracy rate of correct answers 57.14%. It is an improvement of over 20 percentage-points over the results obtained by human participants.

Additionally, although the the scores assigned by the prototype method to each answer were different for both versions

of the method, there was no difference in final ratio of correct answers between the version using thirty FRS or less, and all FRS with over 0.98 of FRS-resemblance score (see Table 6). Considering that the number of FRS used in training did not influence the results, it could be more efficient to use the version of the method using fewer number of sentences.

## 5.4 Discussion

In this experiment, we automatized the task of reading future reference sentences and responding to future prediction questions. The experiment results showed an improvement of over 20 percentage points of the developed prototype method over human participants who took part in the prediction support experiment. Moreover, the result was 23.7 percentage-point higher than for average results of participants of the original 4th FPCT. In fact, with Accuracy on the level of 57.14% it was very close to the highest results obtained by participants of original test (61.11%) and our future prediction support experiment (61.9%). Therefore we can clearly say that the prototype method was nearly as good in predicting the unfolding of future events as the best humans, and it was almost twice as good as an average human, both using all available resources and preparing their answer for a year, and using our support method and making the prediction at the time of the experiment. The final results were compared in Table 8. In addition, when the correct answer was allowed till the third candidate, the 5 out of 7 questions could be considered correct, which gives a 71.43% of Accuracy, being over twice as high as an average and over 10 percentage points higher the best scoring human. Furthermore, if the tendencies of correct answer rates for each question is compared between the prototype method, and the future prediction support experiment, the tendencies of correct and incorrect answers were very similar, meaning, the inference resembles, and exceeds human performance.

The result was more than satisfactory, although we acknowledge that there were many limitations imposed by the controlled character of the experiment. Therefore we need to convey additional experiments on other real world events to obtain a clearer image of the capabilities of our method, most desirably on events that in reality will unfold in future from the time of the prediction.

## 6 Conclusions and Future Works

In this paper we conducted two experiments to determine whether Future Reference Sentences (FRS) are effective in supporting future trend prediction by (1) laypeople and (2) by a prototype fully automatic method. We applied questions from the official Future Prediction Competence Test (FPCT) and, using topic keywords from those questions, gathered newspaper articles from the entire applicable year. Then we extracted topic-related FRS from those articles, to be used in predicting the unfolding of the events.

In the results we obtained a small improvement over the original FPCT. The original test allowed preparing answers for over a year and using any available information. On the other hand, participants of our experiment answered the questions immediately after reading the provided support material

Table 6: Results of classification of answer options by both versions of the prototype method for Question 1-1 and Question 3.

| correct (*) | answer option | score ≥ 0.00 | score ≥ 0.98 |
|---|---|---|---|
| Q1-1 | | (16) | (10) |
| | (a) The right to participate in regional politics will be established for foreign permanent residents by the end of June, 2010. | 1.82 | 1.82 |
| * | (b) The right to participate in regional politics will not be established for foreign permanent residents by the end of June, 2010. | 2.00 | 2.00 |
| Q3 | | (30) | (172) |
| * | (a) The U.S. troops will be still present and further reinforced comparing to October 2009. | 1.71 | 2.17 |
| | (b) The U.S. troops will be still present on similar level comparing to October 2009. | 1.82 | 2.22 |
| | (c) The U.S. troops will be still present but in decreased number comparing to October 2009. | 2.04 | 2.41 |
| | (d) The U.S. troops will be completely withdrawn. | 1.50 | 1.91 |

Table 7: Results of prototype method in comparison to future prediction support experiment for each of the applied questions.

| type of experiment | score average | | | | | | | total (max. 21 pt) | Accuracy (% of correct) |
|---|---|---|---|---|---|---|---|---|---|
| | Q1-1 | Q1-2 | Q-2 | Q-3 | Q-4 | Q-5 | Q-6 | | |
| Prototype method | 3.00 | 3.00 | 3.00 | 1.00 | 0.00 | 2.00 | 0.00 | 12.00 | 57.14% |
| Experiment with laypeople | 1.91 | 0.55 | 1.36 | 2.36 | 0.27 | 1.27 | 0.27 | 8.00 | 38.10% |

Table 8: Comparison of the correct answer accuracies between the prototype method, the future prediction support experiment and original Future Prediction Competence Test.

| correct answer accuracy rate | | |
|---|---|---|
| Prototype method | Our experiment | Original FPCT |
| 0.5714 | 0.3810 | 0.3344 |

which consisted of only thirty (or less) automatically selected sentences. The time spent and the amount of information to be processed for answering the future related questions was greatly reduced with our support method. This indicates that FRS are useful in supporting the prediction of future events.

However, the most interesting discovery was made in the experiments with the proposed prototype fully automatic method for prediction of future unfolding of event, which results showed that our prototype method exceeded even the best humans in the prediction task and the average human over two times. This provides a strong suggestion that full automation of future prediction is possible.

In the future, we plan to use this method with other corpora to conduct experiments on real-world problems, including various lengths of the term of prediction to specify to what extent the method is applicable in future prediction. Also, carrying out a chronological analysis of FRS and the addition of sentiment analysis could lead to the discovery of additional new knowledge. We also plan to take part in the next FPCT.

# References

[Aramaki et al. 2011] 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.

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

[Fellbaum et al. 2009] Christiane Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting semantics into WordNet's "morphosemantic" links. *Human Language Technology. Challenges of the Information Society*, Springer, pp. 350-358.

[Jatowt et al. 2009] Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama, Katsumi Tanaka. 2009. Supporting analysis of future-related information in news archives and the Web. In *9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 115-124.

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

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

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

[Kanazawa et al. 2011] 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.

[Kanhabua et al. 2011] 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.

[Kroeger 2007] Paul Kroeger. 2007. Morphosyntactic vs. morphosemantic functions of Indonesian *-kan*. In *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, edited by A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Maling and C. Manning, pp. 229-251. Stanford, CA: CSLI Publ.

[Levin and Rappaport Hovav 1998] Beth Levin and Malka Rappaport Hovav. 1998. *Morphology and Lexical Semantics*. In Spencer and Zwicky, eds., pp. 248-271.

[Nakajima et al. 2016] Yoko Nakajima, Michal Ptaszynski, Fumito Masui, Hirotoshi Honma. 2016. A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling, *IEICE Trans. on Inf. and Sys.*, Vol.E99-D, No.2, pp.514-524.

[Radinsky et al. 2012] Kira Radinsky, Sagie Davidovich and Shaul Markovitch. 2012. Learning causality for news events prediction. *The 21st International Conference on World Wide Web*.

[Raffaelli 2013] Ida Raffaelli. 2013. The model of morphosemantic patterns in the description of lexical architecture. *Lingue e linguaggio*, Vol. 12, No. 1 (2013), pp. 47-72.

[Strötgen and Gertz 2010] Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp.321–324.

[Takeuchi et al. 2010] 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.

[Zhang et al. 2016] Yating Zhang, Adam Jatowt, Katsumi Tanaka. 2016. Detecting Evolution of Concepts based on Cause-Effect Relationships in Online Reviews. *Proceedings of the 25th International World Wide Web Conference (WWW 2016), ACM Press*, pp. 649-660.