Automatic Extraction of References to Future Events from News Articles Using Semantic and Morphological Information

Yoko Nakajima * † ‡

Advisors:
Fumito Masui †, Hiroshi Yamada †, Michal Ptaszynski †, Hirotoshi Honma ‡

Abstract

In my doctoral dissertation I investigate patterns appearing in sentences referring to the future. Such patterns are useful in predicting future events. I base the study on a multiple newspaper corpora. I firstly perform a preliminary study to find out that the patterns appearing in future-reference sentences often consist of disjointed elements within a sentence. Such patterns are also usually semantically and grammatically consistent, although lexically variant. Therefore, I propose a method for automatic extraction of such patterns, applying both grammatical (morphological) and semantic information to represent sentences in morphosemantic structure, and then extract frequent patterns, including those with disjointed elements. Next, I perform a series of experiments, in which I firstly train fourteen classifier versions and compare them to choose the best one. Next, I compare my method to the state-of-the-art, and verify the final performance of the method on a new dataset. I conclude that the proposed method is capable to automatically classify future-reference sentences, significantly outperforming state-of-the-art, and reaching 76% of F-score.

1 Introduction

In everyday life people use past events and their own knowledge to predict future events. To obtain the necessary data for such everyday predictions, people use widely available sources of information (newspapers, Internet). In my study I focus on sentences that make reference to the future. Below is an example of a future-reference sentence published in a newspaper[1] (translation by the author),

- *Science and Technology Agency, the Ministry of International Trade and Industry, and Agency of Natural Resources and Energy conferred on the necessity of a new system, and decided to set up a new council.*

The sentence claims that the country will construct a new energy system. Interestingly, despite the sentence is written with the use of past tense ("conferred", "decided") the sentence itself refers to future events ("setting up a new council"). Such references to the future contain information (expressions, patterns, causal relations) relating it to the specific event that may happen in the future. The prediction of the event depends on the ability to recognize this information.

A number of studies have been conducted on the prediction of future events with the use of time expressions [Baeza-Yates 2005; Kanazawa et al. 2010], SVM (bag-of-words) [Aramaki et al. 2011], causal reasoning with ontologies [Radinsky et al. 2012], or keyword-based linguistic cues ("will", "is going to", etc.) [Jatowt et al. 2013]. In my research I assumed that the future reference in sentences occurs not only on the level of surface (time expressions, words) or grammar, but consist of a variety of patterns both morphological and semantic.

2 Future Reference Pattern Extraction

The method I propose consists of two stages. Firstly, the sentences are represented in a morphosemantic structure [Levin and Rappaport Hovav 1998] (combination of semantic role labeling with morphological information). Secondly, frequent combinations of such patterns are automatically extracted from training data and used in classification.

Morphosemantic patterns (MoPs) are useful for representing languages rich both morphologically and semantically, such as Japanese (language of datasets used in this research). I generated the morphosemantic model using semantic role labeling (SRL) supported with morphological information. SRL provides labels for words and phrases according to their role in the sentence. 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]`.

To retain words omitted by SRL (particles or function words, not directly influencing the semantic structure, but contributing to the overall meaning) I used morphological

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

[1]Japanese daily newspaper *Hokkaido Shinbun*.

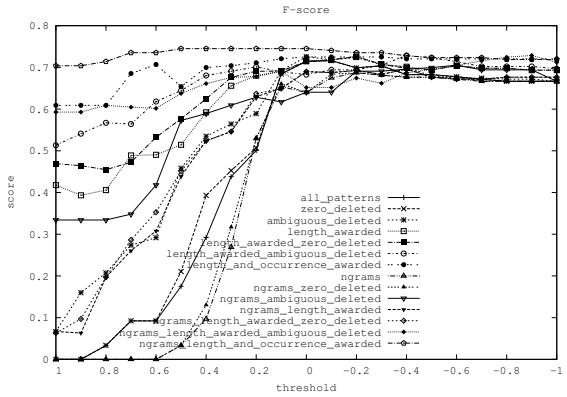

Figure 1: F-score for all tested classifier versions.

analysis to provide information on parts of speech, etc. Below is an example of a sentence generalized on the morphosemantic structure:

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

MoPs: `[Object][Agent][State_change][Action][Noun][State_change][Object][State_change]`

From sentences represented this way I extract frequent MoPs. Firstly, I generate ordered non-repeated combinations from all sentence elements. In every n-element sentence there is k-number of combination groups, such as that $1 \leq k \leq n$. All combinations for all values of k are generated. Additionally, all non-subsequent elements are separated with a wildcard ("*", asterisk). Pattern lists extracted this way from training set are then used in classification of test and validation set.

3 Evaluation

From three newspaper corpora[2] I collected and annotated two datasets containing equal number of (1) sentences referring to future events and (2) other (describing past, or present events).

The datasets were applied in a text classification task on 10-fold cross validation. Each classified test sentence is 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. Moreover, to provide sufficiently objective view on results, I additionally performed threshold optimization to find which modification of the classifier achieved the highest scores. In the evaluation experiment, where I compared 14 different classifier versions, I 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. Experiment results (F-score) for all classifier versions are represented in Figure 1. 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).

[2] *Nihon Keizai Shimbun*, *Asahi Shimbun*, *Hokkaido Shimbun*.

Table 1: Comparison of results 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**
10 pattern with only over 3 elements	**0.42**	0.37	0.40
5 patterns	0.35	0.35	0.35
Optimized model	**0.76**	**0.76**	**0.76**
[Jatowt et al. 2013] (10 phrases)	**0.50**	0.05	0.10

Next, I collected a validation set unrelated to previous data from one year (1996) of *Mainichi Shinbun*. Each sentence was annotated as either future or non-future related by one expert- and two layperson-annotators. The sentences with an agreement between at least one layperson and the expert were left as the validation set.

I compared my method to [Jatowt et al. 2013], who extracted future reference sentences with 10 words unambiguously referring to the future, such as "will" or "is likely to", etc. In comparison, on the new validation set my method obtained much better results even when only 10 most frequent morphosemantic patterns were used (Table 1).

Finally, I verified the performance of the fully optimized model. I re-trained the best model using all sentences from the initial dataset and verified the performance by classifying the new validation set. The final optimized performance is represented in Table 1 ("Optimized model"). The highest reached Precision was .89 (R=.13, F=.22). The highest reached F-score was .78 (P=.65, R=.98). Finally, break-even point (BEP) was at .76, which indicates that the proposed method trained on automatically extracted morphosemantic future reference patterns is sufficiently capable to classify future reference sentences.

4 Conclusions and Future Directions

I proposed a novel method for extracting references to future events from news articles, based on automatically extracted morphosemantic patterns. The evaluation experiment helped me chose the best out of 14 different classifier versions. I validated the optimized method on a new validation set and compared it to the state-of-the-art. The proposed method presented high performance outperforming state-of-the-art.

In the future I plan to increase the size of experimental datasets for more thorough evaluation, and apply the method to estimating probable unfolding of future events in practice.

References

[Aramaki et al. 2011] E. Aramaki, S. Maskawa, M. Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. *EMNLP*, pp. 1568–1576.

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

[Jatowt et al. 2013] A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, K. Kunieda, K. Yamada. Multilingual, Longitudinal Analysis of Future-related Information on the Web. *Cult. and Comp. 2013*.

[Kanazawa et al. 2010] K. Kanazawa, A. Jatowt, S. Oyama, K. Tanaka. 2010. Exracting Explicit and Implicit future-related information from the Web(O) (in Japanese). *DEIM Forum 2010*.

[Levin and Rappaport Hovav 1998] B. Levin, M. Rappaport Hovav. 1998. *Morphology and Lexical Semantics*, pp. 248-271.

[Radinsky et al. 2012] K. Radinsky, S. Davidovich, S. Markovitch. 2012. Learning causality for news events prediction. *WWW 2012*.