Automatic Extraction of Future References from News Using Morphosemantic Patterns with Application to Future Trend Prediction

Yoko Nakajima (National Institute of Technology, Kushiro College; yoko@kushiro-ct.ac.jp)
Michal Ptaszynski (Kitami Institute of Technology; ptaszynski@cs.kitami-it.ac.jp)
Hirotoshi Honma (National Institute ofTechnology,Kushiro College; honma@kushiro-ct.ac.jp)
Fumito Masui (Kitami Institute of Technology; f-masui@mail.kitami-it.ac.jp)

DOI: 10.1145/3008665.3008671

Introduction

In everyday life people use past events and their own knowledge to predict future events. In such everyday predictions people use widely available resources (newspapers, Internet). This study focused on sentences referring to the future, such as the one below, as one of such resource.

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 energy system, and decided to set up a new council. (Japanese daily newspaper *Hokkaido Shinbun*, translation by the author.)

The sentence claims that the country will construct a new energy system. However, although the sentence is set in the past ("conferred", "decided") the sentence itself refers to future events ("setting up a new council"). Such references to the future contain information (expressions, 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", "shall", etc.) (Jatowt et al. 2013). In this research I assumed that future references in sentences occur not only on the level of surface (time expressions, words) or grammar, but consist of a variety of patterns both morphological and semantic.

Future Reference Pattern Extraction

The proposed method consists of two stages: (1) sentences are represented in a morphosemantic structure (Levin and Rappaport Hovav 1998) (combination of semantic role labeling with morphological information), and (2) frequent morphosemantic patterns (MoPs) are automatically extracted from training data and used in classification. MoPs are useful for representing languages rich both morphologically and semantically, such as Japanese (language of datasets used in this research). Morphosemantic model was generated using semantic role labeling (SRL) supported with morphological information. SRL provides labels for words and phrases according to their role in the sentence. To retain information omitted by SRL (particles, function words, not directly influencing the semantic structure, but contributing to the overall meaning) morphological analysis provided information on parts of speech, of omitted words. Below is an example of a sentence generalized on the morphosemantic structure:

Japanese: *AI gijutsu ha mujinhikōki nado no seihin ya sābisu ni katsuyō ga kitaisarete iru.* **English:** AI technology is expected to be used in products and the services such as a pilotless planes. **MoPs:** `[Object][Noun][Thing]-[Agent][Object][No_State_change]`

From sentences represented this way frequent MoPs are extracted as follows. Firstly, ordered non-repeated combinations from all sentence elements are generated. 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, with non-subsequent elements separated by an asterisk. Frequent pattern lists extracted this way from training set are used in classification of test and validation set.

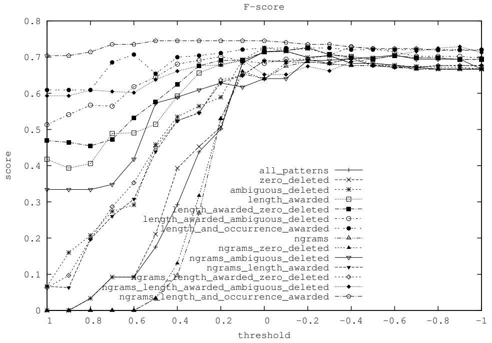

Figure 1: F-score for all tested classifier versions.

Table 1: Comparison of results for different pattern groups, state-of-the-art, and fully optimized model.

Pattern set	Precision	Recall	F-score
10 patterns	0.39	**0.49**	**0.43**
10 pattern (3 elements or longer)	**0.42**	0.37	0.40
5 patterns	0.35	0.35	0.35
Optimized (see Fig. 2)	**0.76**	**0.76**	**0.76**
(Jatowt et al. 2013) (10 phrases)	**0.50**	0.05	0.10

Evaluation

From three newspaper corpora (*Nihon Keizai Shimbun*, *Asahi Shimbun*, *Hokkaido Shimbun*.) two datasets were collected and manually annotated to contain 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. Each classified test sentence was given a score calculated as a sum of weights of patterns extracted from training data and matched with the input sentence. The results were calculated with Precision (P), Recall (R) and F-measure (F). Fourteen classifier versions were compared (see Figure 1) for performance based on the highest statistically significant F within the threshold, and the highest break-even point (BEP) of P and R. The highest overall performance was obtained by the version using pattern list containing all patterns (including ambiguous patterns and n-grams).

In comparison with (Jatowt et al. 2013), who extracted future reference sentences (FRS) with 10 words unambiguously referring to the future, such as “will” or “is likely to”, etc. the proposed method obtained much higher results even when only 10 most frequent MoPs were used (Table 1). Moreover, the performance of a fully optimized model, retrained on all training data with the best settings reached break-even point (BEP) at 76% (Figure 2).

Future Prediction Support Experiment

We performed an experiment to verify that our method is useful for the support of predicting future trends. Thirty laypeople answered questions from Future Prediction Competence Test (*http://homepage3.nifty.com/genseki/kentei.html#kentei*) using only FRS provided by our method. The FRS for each question were gathered from Mainichi Newspaper using: (1) topic keywords related to questions and (2) MoPs generated using the fully optimized model.

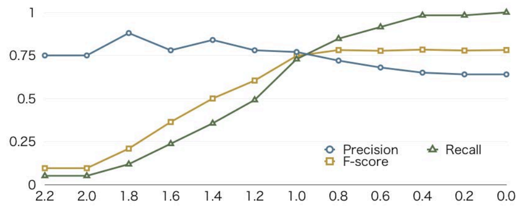

Figure 2: Final results of fully optimized model.

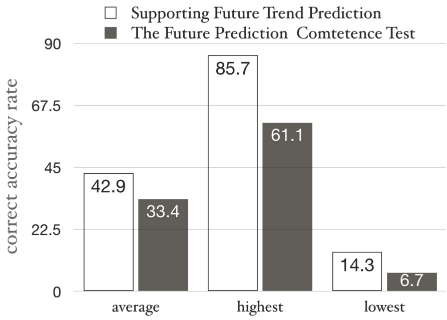

Figure 3: Correct accuracy rate in future trend prediction experiment compared to the original Future Prediction Competence Test.

The correct accuracy rate (Figure 3) of the proposed method was higher than for the original test participants both in average, and for the highest and lowest score achieved. Only 9% higher but, FRS is clearly usefulness supporting to future trend prediction.

Conclusion and Future Directions

We proposed a novel method for extracting references to future events from news articles, based on automatically extracted morphosemantic patterns. From 14 different classifier version compared an optimized model was selected and validated on a new data set. The model achieved high performance outperforming state-of-the-art. Moreover, we performed a future trend prediction experiment and found out that the method is capable to automatically extract sentences providing support for future event prediction. As for further work, we consider applying the method in statistical data interpretation, and key sentence extraction from the Web for supporting business-related judgments.

References

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

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

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*.

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

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

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

Yoko Nakajima received her PhD from Kitami Institute of Technology in 2016. Currently works as an Assistant Professor at the Information Engineering Course, of NIT, Kushiro College. Her research interests include natural language processing and information extraction. She is a member of IPSJ.

Michal Ptaszynski received PhD in Information Science and Technology from Hokkaido University in 2010. JSPS PD Research Fellow at Hokkai-Gakuen University (2010-2012). Since 2013 an Assistant Professor at Kitami Institute of Technology. Research interests: NLP, affect analysis. Member of: ACL, AAAI, IEEE, IPSJ, ANLP.

Hirotoshi Honma received B.E., M.E. and D.E. degrees in Engineering from Toyohashi University of Technology, in 1990, 1992 and 2009, respectively. Joined the National Institute of Technology, Kushiro College in 1992, Associate Professor since 2001. Research interest: computational graph theory, parallel algorithms, and NLP. Member of: IEICE and ORSJ.

Fumito Masui graduated Okayama University Faculty of Science in 1990. Since 2009 an Associate Professor at Faculty of Engineering, Kitami Institute of Technology. Ph.D. in Engineering. Research interests: NLP, tourism informatics, and curling informatics. Member of: JSAI, IEICE, IPSJ, SOFT, Hokkaido Regional Tourism Society, AAAI, ACL, and Japan Curling Association.