Extracting References to the Future from News Using Morphosemantic Patterns

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

†‡ Kitami Institute of Technology
‡ National Institute of Technology, Kushiro college

IJCAI15 WS
Chance Discovery, Data Synthesis, Curation and Data Market
background

Future References Sentences (FRS)

• Describe the probable future event.
• Contain comments to the future event.
• Include information on past events, background knowledge and professional views etc.

Using FRS people can decide about their action and thinking more effectively.
Research purpose

Extract **future reference sentences** from corpus for support action or thinking of people
Our previous work

1. Investigation of future reference expressions
2. Extract patterns of future reference expressions

Future Reference Expression : FRE
Future Reference Sentence : FRS
Our previous work

1. Investigation of FRE

Corpus: newspapers
data: 270 sentences extracted randomly
  • not depend on morphology or temporal expressions.
  • used variety of words in future references sentences.

Corpus: newspaper
data: 1000 sentences extracted randomly
annotated manually: one expert, two laypeople
(referring to future or not)
  • 13% of newspaper corpus
FRE manually extracted from 270 sentences

<table>
<thead>
<tr>
<th>Type</th>
<th>frequency</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Expression</td>
<td>70</td>
<td>• next year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• tommorrow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• from month M year Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• this month</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• next in ... etc.</td>
</tr>
<tr>
<td>verb</td>
<td>141</td>
<td>• mezasu (aim to)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• hoshin (plan to)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• - suru (do)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• - iru (is/to be) ... etc.</td>
</tr>
</tbody>
</table>
Propose method

Extract patterns of future reference expressions

Morphology + Semantic

Morphosemantic Patterns : MoPs

example

• analysis of Indonesian suffix in Wordnet [*1]
• analysis of Croatian lexis [*2]

Semantic role labelling

a sentence: “John killed Mary.”

predicate argument structure

 semantic role labelling

“actor action patient”

Application examples

- Construction of Japanese Frame Net [*3]
- Collection of Event Ontology [*4]

Morphological analysis

a sentence: “John killed Mary.”

morphological structure

“noun verb(past) noun”

MeCab

Standard tool morphology for Japanese [*5]

Additional Post-processing

• ordering of priority taking semantic roll
  1. Semantic role (Agent, Patient, Object, etc.)
  2. Semantic meaning (State change, etc.)
  3. Category (Dog $\rightarrow$ Living animal $\rightarrow$ Animated object)
  4. Adjunct (Time-Point, Time-Line, Location, etc.)
  5. parts of speech

• compound word clustering

Example:

“International Joint Conference on Artificial Intelligence”
$\rightarrow$ Adjective Adjective Noun Preposition Adjective Noun
$\rightarrow$ Proper Noun
Example of morphosemantic structure (MS)

Japanese: ニホンウナギが絶滅危惧種に指定され、完全養殖によるウナギの量産に期待が高まっている。

Alphabet: nihonunagi ga zetumetu kigushu ni sitei sare, kanzen yo-shoku ni yoru unagi no ryousan 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.

MS: [Object] [Agent] [State change] [Action] [Noun] [State change] [Object][State change]
Extracting morphosemantic patterns

SPEC : Sentence Pattern Extraction arChitecture [*6]

- Generate all combination from all elements of a sentence.
- Calculate occurrence frequency of combinations in a corpus.
- Frequent combinations = patterns

Generating all patterns from a sentence

J : [kinou] [kare ha] [watashi ni] [tegami wo] [okutta]

MS : [Time-Point] [Agent] [Patient] [Object] [State change]

E : [yesterday] [he] [me] [a letter] [sent]

1. [Time-Point] [Agent] [Patient] [Object] [State change]
2. [Time-Point] * [Patient] [tObject] [State change]
3. [Time-Point] [Agent] * [Object] [State change]
4. [Time-Point] [Agent] [Patient] * [State change]
5. [Time-Point] * [Object] [State change]
6. [Time-Point] [Agent] * [State change]
7. [Time-Point] * [State change]

::
Experiment setup

corpus

- Japan Economy Newspaper
- Asahi Newspaper (national)
- Mainichi Newspaper (national)
- Hokkaido Newspaper (regional)
- http://www.nikkei.com/
- http://www.asahi.com/
- http://www.hokkaido-np.co.jp/

- extract 1000 sentences randomly
- annotate FRE or NRE

*FRE : 13% of newspaper corpus.

Training data

set50 and set130
FRS
(50 or 130)
and
other sentences
(50 or 130)

learning by SPEC
Experiment setup

- sophisticated patterns (with disjoint elements)
  - awarding length (LA)
  - awarding length and occurrence (LOA)
  - awarding none (normalized weight, NW)
  - 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, 0P)
  - patterns (PAT)
  - only n-grams (NGR)

- n-fold cross validation
- Results calculated in F-score, Precision, Recall
- Choose the most useful pattern
Experiment 1: Extract MoPs

- Test data: set50, set130
- 10-fold cross validation

Compare to F-scores set130 and set50

<table>
<thead>
<tr>
<th>sofisticated patterns</th>
<th>set50</th>
<th>set130</th>
</tr>
</thead>
<tbody>
<tr>
<td>all_patterns</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>zero_deleted</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>ambiguous_deleted</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>length_awarded</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>length_awarded_zero_deleted</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>length_awarded_ambiguous_deleted</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>
The examples of extracted MoPs

<table>
<thead>
<tr>
<th>occurrence</th>
<th>Future Reference Patterns</th>
<th>occurrence</th>
<th>Other Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>[Action]*[State change]</td>
<td>5</td>
<td>[Place]*[Agent]</td>
</tr>
<tr>
<td>23</td>
<td>[Action]*[Object]</td>
<td>4</td>
<td>[Number]*[Agent]</td>
</tr>
<tr>
<td>22</td>
<td>[Action]*[Action]</td>
<td>4</td>
<td>[Verb]*[Artifact]</td>
</tr>
<tr>
<td>20</td>
<td>[State change]*[Object]</td>
<td>4</td>
<td>[Person]*[Place]</td>
</tr>
<tr>
<td>16</td>
<td>[State change]*[State change]</td>
<td>3</td>
<td>[Number]<em>[Agent]</em>[Action]</td>
</tr>
<tr>
<td>15</td>
<td>[Action]<em>[Object]</em>[State change]</td>
<td>3</td>
<td>[Adjective]<em>[State change]</em>[State change]</td>
</tr>
<tr>
<td>15</td>
<td>[Action]<em>[State change]</em>[No state change(activity)]</td>
<td>3</td>
<td>[Place]<em>[Place]</em>[No state change(activity)]</td>
</tr>
</tbody>
</table>
Experiment 2: Extract FRS with frequent patterns

• corpus
  Mainichi Newspaper (1996)
  topics: economy, international event, energy
  270 sentences

• validation data set
  annotate manually from 270 sentences
  • one expert  • four laypeople

  FRS: 100, other: 170

• frequent patterns
  out of learning set 130 with length awarded
Extract with frequent patterns by pattern matching

A: first 10 patterns
B: adding 5 patterns longer than three elements to set A
C: subtracting 5 patterns from the tail of set A
D: using only first 10 patterns containing more than three elements

<table>
<thead>
<tr>
<th>occurrence</th>
<th>Frequent patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>[action] * [state change]</td>
</tr>
<tr>
<td>23</td>
<td>[Action] * [Object]</td>
</tr>
<tr>
<td>22</td>
<td>[Action] * [Action]</td>
</tr>
<tr>
<td>20</td>
<td>[State change] * [Object]</td>
</tr>
<tr>
<td>16</td>
<td>[State change] * [State change]</td>
</tr>
<tr>
<td>15</td>
<td>[Action] * [Object] * [State change]</td>
</tr>
<tr>
<td>14</td>
<td>[Object] * [Action] * [State change]</td>
</tr>
<tr>
<td>13</td>
<td>[Object] * [Action] * [Object]</td>
</tr>
<tr>
<td>12</td>
<td>[State change] * [Action] * [State change]</td>
</tr>
<tr>
<td>10</td>
<td>[Action] * [Action] * [No state change(Activity)]</td>
</tr>
<tr>
<td>9</td>
<td>[State change] * [Noun] * [Object]</td>
</tr>
<tr>
<td>8</td>
<td>[Action] * [State change] * [No state change(Activity)]</td>
</tr>
<tr>
<td>8</td>
<td>[Object] * [Action] * [Object] * [State change]</td>
</tr>
<tr>
<td>5</td>
<td>[Action] * [State change] * [Action] * [No state change(Activity)]</td>
</tr>
<tr>
<td>5</td>
<td>[Action] * [State change] * [Object] * [No state change(Activity)]</td>
</tr>
</tbody>
</table>
Performance of extracted FRS with most frequent patterns

<table>
<thead>
<tr>
<th>Pattern set</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 10 patterns</td>
<td>0.39</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>B: 15 patterns</td>
<td>0.38</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>C: 5 patterns</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>D: 10 patterns with only over 3 elements</td>
<td>0.42</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>baseline (10 temporal expressions) [*2]</td>
<td>0.50</td>
<td>0.05</td>
<td>0.10</td>
</tr>
</tbody>
</table>

* 2 [Jatowt and Au Yeung 2011]

6/170 sentences
Experiment 3 : Validation for fully optimized model

• Corpus:
  Mainichi Newspaper (1996)
  topics: economy, international event
  270 sentences (FRS:100 Non FRS:170)

• Training data:
  1 cross validation for set130
  result calculated with length-awarded

• Evaluation:
  one expert, four laypeople
Classification result

Precision, Recall, F-score

Precision, Recall, F-score

break-even point

0.76

P=0.89
R=0.13
F=0.22

P=0.65
R=0.98
F=0.78
Example of extracted future referring sentence

1. score=2.27

RJ: *Dosha* wa kore made,*Shigen Enerugi-Cho ni taishi,* do hatsudensho no *heisa,* kaitai ni tsuite *ho shin o setumei shite kita ga,* kaitai ni tsuite no ho teki kisei wanai tame, do chō mo kaitai no kettei o shitam eru koto *ni nari-so da.*

E: So far the *company* has been describing to the *Agency for Natural Resources and Energy* the policy for either *closure* or dismantling of the plant, and since there are no legal regulations found for dismantling, it is most likely that the agency will also lean to the decision of dismantling.

MS: [Agent] [Other] [Organization] [Action] [State-change] [State-change][Object][Role] [State-change] [State-change][Action][Adjective][Thing][Agent][State-change][Other] [Verb]

MoPs: [Agent]*[Verb],
[Agent]*[Organization]*[Verb],
[Agent]*[Action][State-change]*[Verb],
[Agent]*[Organization][State-change][Verb].
Conclusion

• We presented a novel method for extracting references to future events.
• Based the method on automatically extracted morphosemantic patterns.
  - Represent news articles in morphosemantic structure.
  - Extract all possible morphosemantic patterns from the corpus.
• Performed a text classification experiment.
• Compared 14 different classifier versions.
• Compared to the state-of-the-art.
  - The proposed method outperformed the state-of-the-art.
• Validated the method on new dataset.
  - Final score was break even point of precision and recall = 76%.
Future Work

• Increase the size of the experimental datasets.
• Apply in practice
  - Estimating probable unfolding of events.
  - Contribute to trend prediction.
Thank you
Generating all patterns

\[ \text{[ pattern ]} \]

\[ 1 \ 2 \ 3 \ 4 \ \ldots \ \ n \]

number of elements : \( n \)
number of group of combination : \( k \)
in \( k \)-element : \( \quad k \subseteq n \quad 1 \leq k \leq n \)

\[
\left( \begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}
\]  

(1)

\[
\sum_{k=1}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + \ldots + \frac{n!}{n!(n-n)!} = 2^n - 1
\]  

(2)