WORD SENSE IMPLANTATION
BY ORTHOGRAPHICAL CONVERSION

KAZUHIDE YAMAMOTO AND YUKI MIKAMI
NAGAOKA UNIVERSITY OF TECHNOLOGY,
JAPAN
OUTLINE

• We build a tool to determine word sense in Japanese.
  • Although many works, no such tool is available online.
OUTLINE

• We build a tool to determine word sense in Japanese.
  • Although many works, no such tool is available online.

• Word sense implantation rather than word sense disambiguation (WSD).
  • Ambiguous target words are orthographically replaced.
QUESTION:

WHY IS NO WSD TOOL AVAILABLE?
OUR HYPOTHESIS:

IT IS NOT EASY TO USE WORD SENSE AFTERWARD.
PROBLEMS IN WSD

英語の問題がとける
(An English problem is ...)
PROBLEMS IN WSD

英語の問題がとける

Sense dictionary:
#1 to solve
#2 to preach
#3 to comb
#4 to melt
PROBLEMS IN WSD

英語の問題がとける(#1)
PROBLEMS IN WSD

英語の問題がとける(#1)

• Sense dictionary is additionally required.
PROBLEMS IN WSD

英語の問題がとける(#1)

- Sense dictionary is additionally required.
- Not easy to provide additional feature.
PROBLEMS IN WSD

英語の問題がとける(#1)

• Sense dictionary is additionally required.
• Not easy to provide additional feature.
• Not obvious to identify sense into text.
OUR PROPOSAL:

TO IMPLANT WORD SENSE INTO TEXT.
JAPANESE ORTHOGRAPHY

とける

Hiragana word
• phonographic
• ambiguous (for computer)
とける

溶ける, 解ける, 説ける, 梳ける

Hiragana word
- phonographic
- ambiguous (for computer)

Kanji word
- ideographic
- clear sense
Orthographical conversion includes word sense disambiguation
METHOD: OUR POLICY

- Simple and transparent
  - Use of pointwise mutual information (PMI)
- Middle-range frequency words as target
  - Frequently-appeared regarded as natural
  - Hardly-appeared regarded as noise
- High-precision conversion only
  - Target words are selected in advance.
METHOD: EXAMPLE

鉛筆 と ノート を かう
pencil and notebook (OBJ) かう
IDENTIFY TARGET WORD AND CONVERSION CANDIDATES

鉛筆 pencil と and ノート notebook を (OBJ) かう

- Conversion dictionary is built in advance.
- Not only ambiguous but unambiguous words are also converted.
SELECT POTENTIAL CLUES

鉛筆	pencil
to buy
ノート
notebook
飼う
to feed

- Functional words are deleted.
• Use of 570M word 7-gram as pseudo sentences
COMPUTE PMI SCORE

鉛筆 (pencil)  | ノート (notebook)  | かう (to buy)

1.1

0.0

飼う (to feed)
COMPUTE PMI SCORE

<table>
<thead>
<tr>
<th>鉛筆 (pencil)</th>
<th>ノート (notebook)</th>
<th>かう (to buy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>飼う (to feed)</td>
<td>5.2</td>
<td>買う (to buy)</td>
</tr>
<tr>
<td>1.1</td>
<td>0.0</td>
<td>飼う (to feed)</td>
</tr>
</tbody>
</table>
COMPUTE PMI SCORE

鉛筆 (pencil)
ノート (notebook)
かう (to buy)
飼う (to feed)

1.1
5.2
0.0
0.0
**CONVERT THE WORD**

<table>
<thead>
<tr>
<th>鉛筆</th>
<th>と</th>
<th>ノート</th>
<th>を</th>
<th>かう</th>
</tr>
</thead>
<tbody>
<tr>
<td>pencil</td>
<td>and</td>
<td>notebook</td>
<td>(OBJ)</td>
<td></td>
</tr>
</tbody>
</table>

1.1  
0.0  
5.2  
0.0  

買う  
to buy  
飼う  
to feed
### EVALUATION: PRECISION

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Words</th>
<th>Sentences</th>
<th>Correct</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous</td>
<td>52</td>
<td>1,040</td>
<td>1,031</td>
<td>99.1%</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>71</td>
<td>1,420</td>
<td>1,336</td>
<td>94.1%</td>
</tr>
<tr>
<td>Total</td>
<td>123</td>
<td>2,460</td>
<td>2,367</td>
<td>96.2%</td>
</tr>
</tbody>
</table>

- As expected, no problem for precision.
- “Unambiguous” conversion were not 100% unambiguous.
## EVALUATION: COVERAGE

<table>
<thead>
<tr>
<th>Frequency</th>
<th>&lt; 20</th>
<th>20-10K</th>
<th>&gt; 10K</th>
</tr>
</thead>
<tbody>
<tr>
<td># of words</td>
<td>16,052</td>
<td>886,649</td>
<td>9,470,468</td>
</tr>
<tr>
<td>(# of ambiguous words)</td>
<td>(unknown)</td>
<td>316,190</td>
<td>(very few)</td>
</tr>
<tr>
<td>(# of disambiguated words)</td>
<td></td>
<td>80,962</td>
<td></td>
</tr>
</tbody>
</table>

- Coverage: 25.6% (= 80K/316K) by only 71 words.
- Difficult to increase coverage as manual sense-tagging is required.
CONCLUSION

• A tool to determine word sense in Japanese is developed.
  • Word sense implantation as orthographical conversion.
  • Very high precision, although not so high coverage
  • Available on the Web

• Future work: to increase the coverage
SNOWMAN, a Japanese word analyzer
http://snowman.jnlp.org