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)

JAPANESE ORTHOGRAPHY

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

JAPANESE ORTHOGRAPHY

とける

Hira

- phonographic
- ambiguous (for computer)

溶ける, 解ける

る

- ideographic
- clear sense

Orthographical conversion includes word sense disambiguation

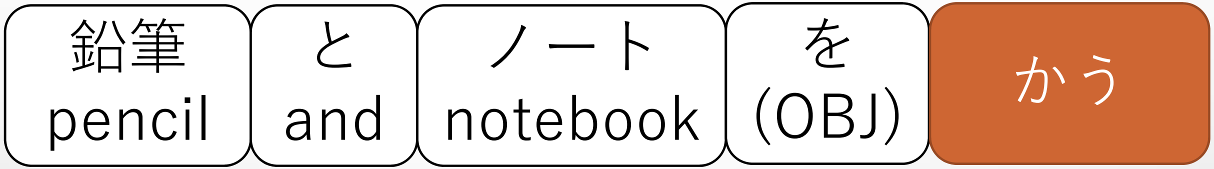
METHOD: OUR POLICY

- 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)	

IDENTIFY TARGET WORD AND CONVERSION CANDIDATES



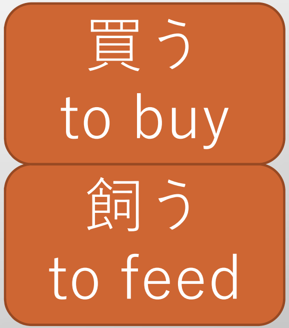
鉛筆
pencil

と
and

ノート
notebook

を
(OBJ)

かう

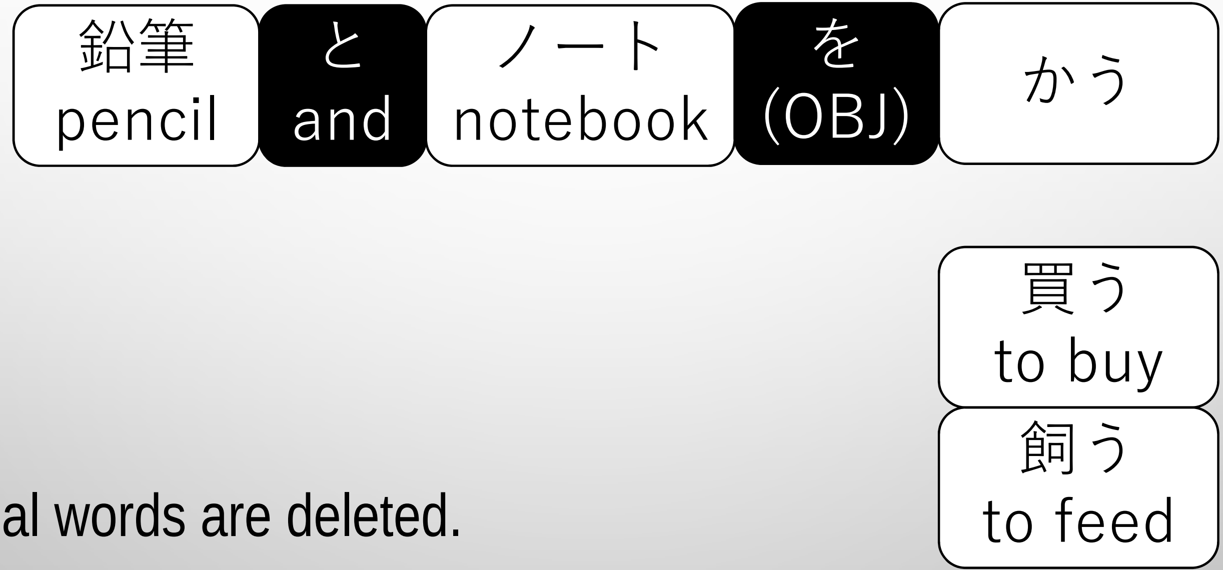


買う
to buy

飼う
to feed

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

SELECT POTENTIAL CLUES



鉛筆
pencil

と
and

ノート
notebook

を
(OBJ)

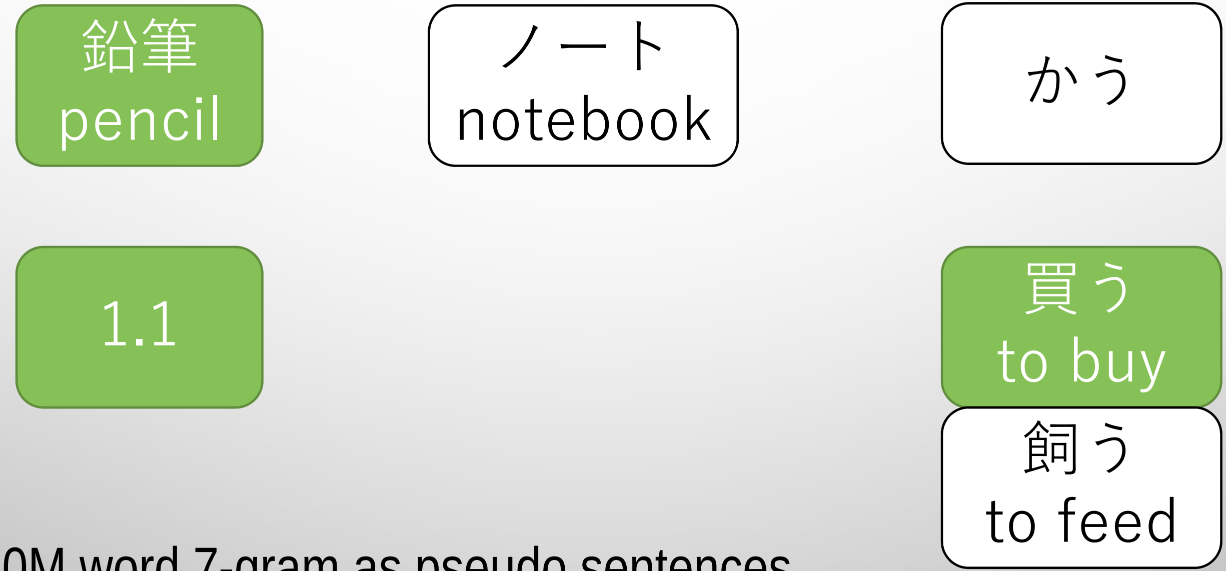
かう

買う
to buy

飼う
to feed

- Functional words are deleted.

COMPUTE PMI SCORE



鉛筆
pencil

ノート
notebook

かう

1.1

買う
to buy

飼う
to feed

- Use of 570M word 7-gram as pseudo sentences

COMPUTE PMI SCORE

鉛筆
pencil

ノート
notebook

かう

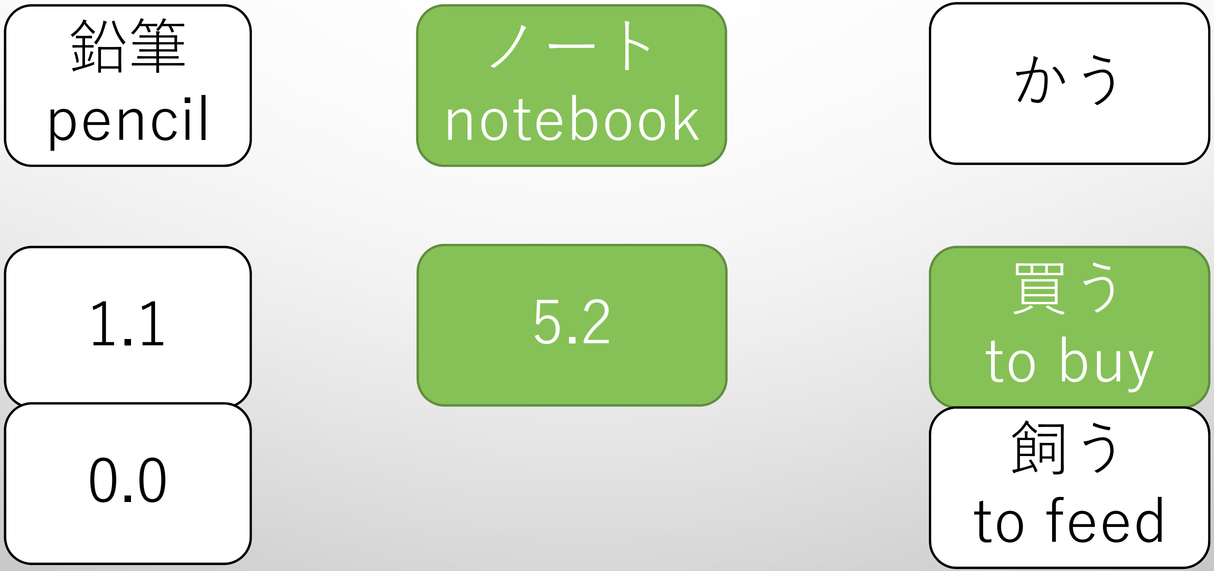
1.1

買う
to buy

0.0

飼う
to feed

COMPUTE PMI SCORE



鉛筆
pencil

ノート
notebook

かう

1.1

0.0

5.2

買う
to buy

飼う
to feed

COMPUTE PMI SCORE

鉛筆 pencil	ノート notebook	かう
1.1	5.2	買う to buy
0.0	0.0	飼う to feed

CONVERT THE WORD

鉛筆 pencil	と and	ノート notebook	を (OBJ)	かう
1.1		5.2		買う to buy
0.0		0.0		飼う to feed

EVALUATION: PRECISION

Candidate	Words	Sentences	Correct	Accuracy
Unambiguous	52	1,040	1,031	99.1%
Ambiguous	71	1,420	1,336	94.1%
Total	123	2,460	2,367	96.2%

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

EVALUATION: COVERAGE

Frequency	< 20	20-10K	> 10K
# of words	16,052	886,649	9,470,468
(# of ambiguous words)	(unknown)	316,190	(very few)
(# of disambiguated words)		80,962	

- 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