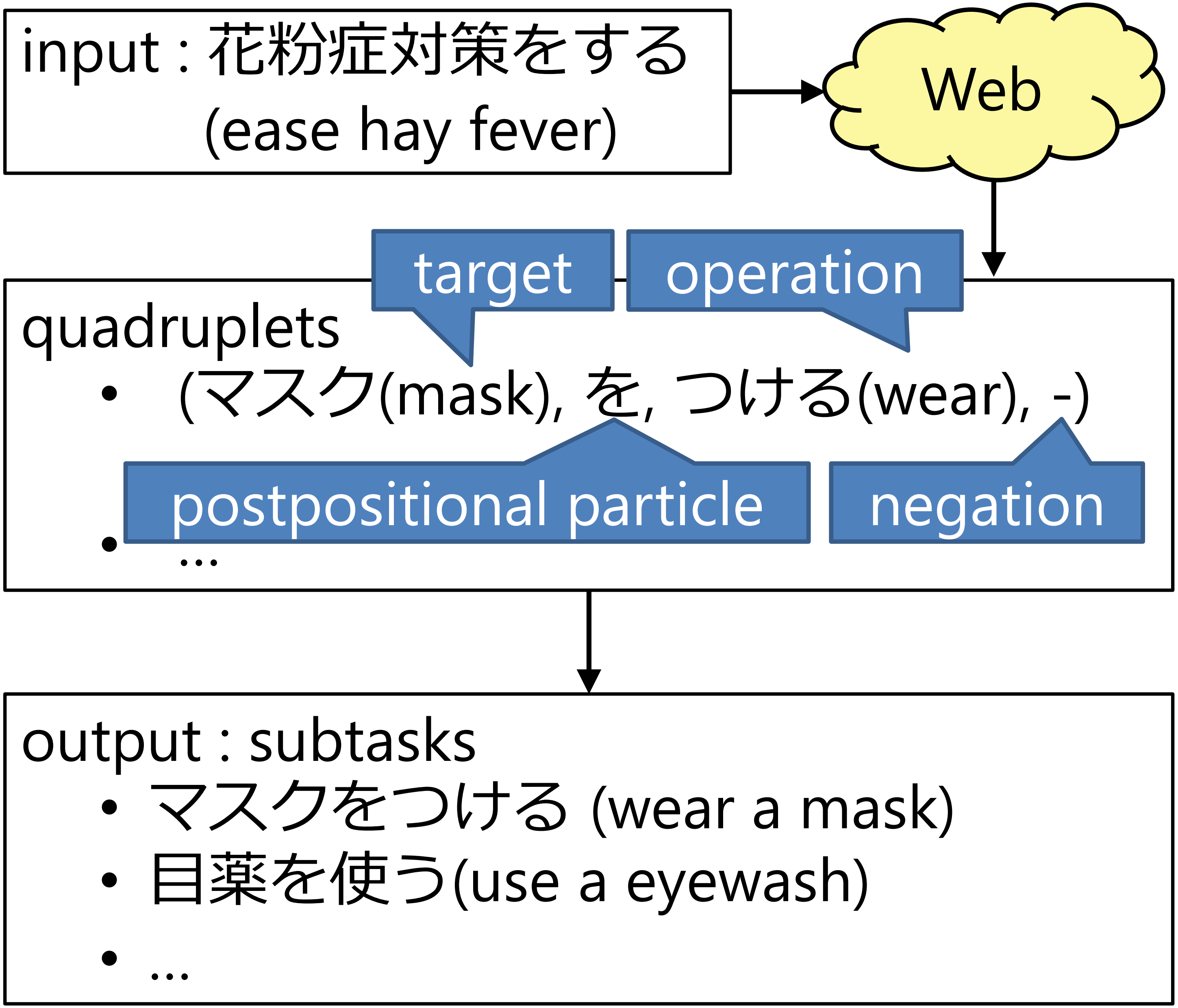
NTCIR-11 Conference (11th Dec., 2014)

University of Hyogo at NTCIR-11 TaskMine by Dependency Parsing

Takayuki Yumoto
University of Hyogo, Japan

Overview



input : 花粉症対策をする (ease hay fever)

Web

quadruplets

- (マスク(mask), を, つける(wear), -)
- …

output : subtasks

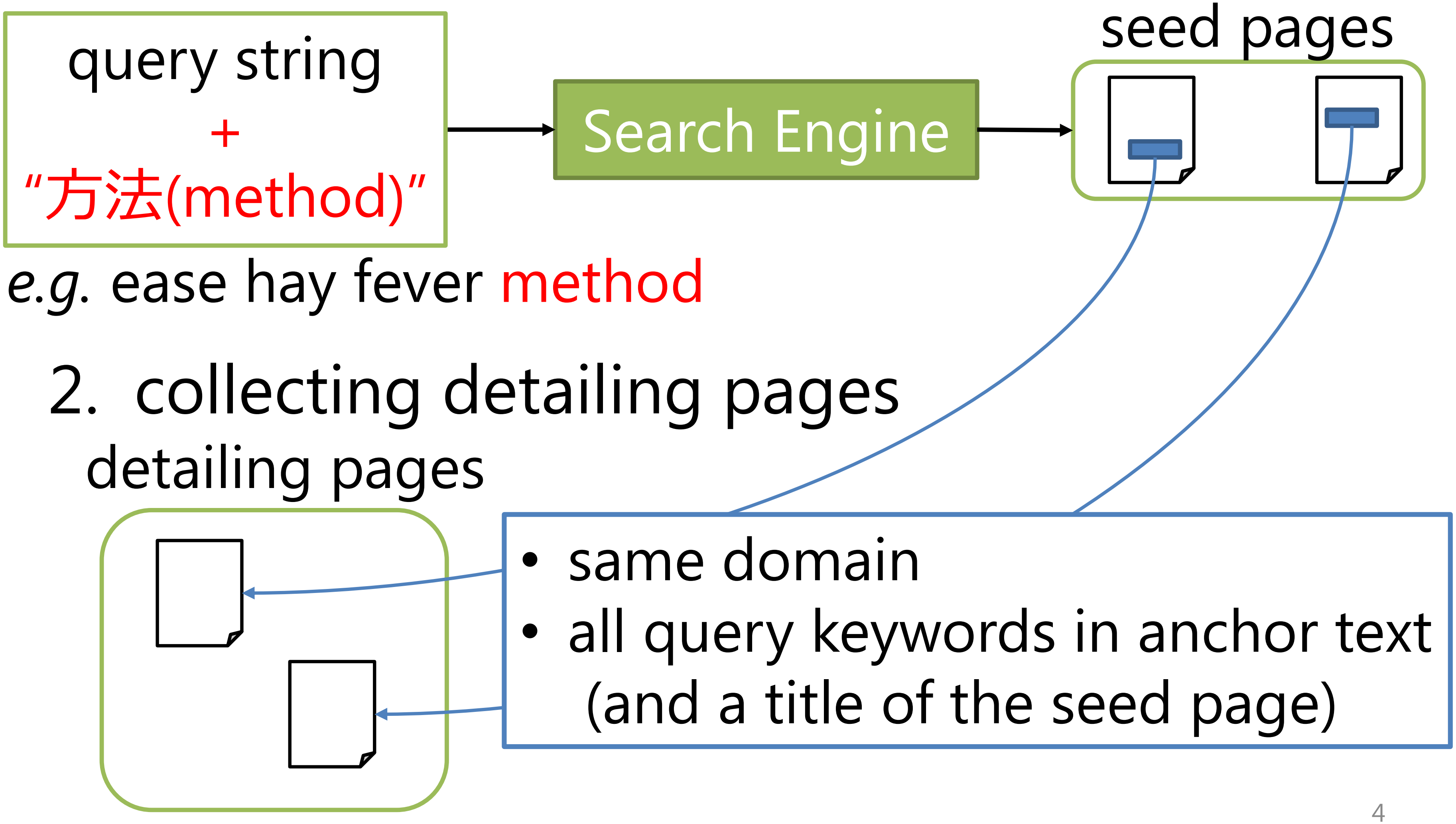
- マスクをつける (wear a mask)
- 目薬を使う(use a eyewash)
- …

Our Approach

1. Collecting pages
   - collecting seed pages by query modification
   - collecting detailed pages by link anchor
2. Extracting quadruplets
   - Finding operation (and negation)
   - Finding target and postpositional particle
3. Ranking quadruplets
   - Synonyms by Wikipedia corpus
   - ranking by site frequency

1. Collecting Pages

1. collecting seed pages by query expansion



query string

+

“方法(method)”

Search Engine

seed pages

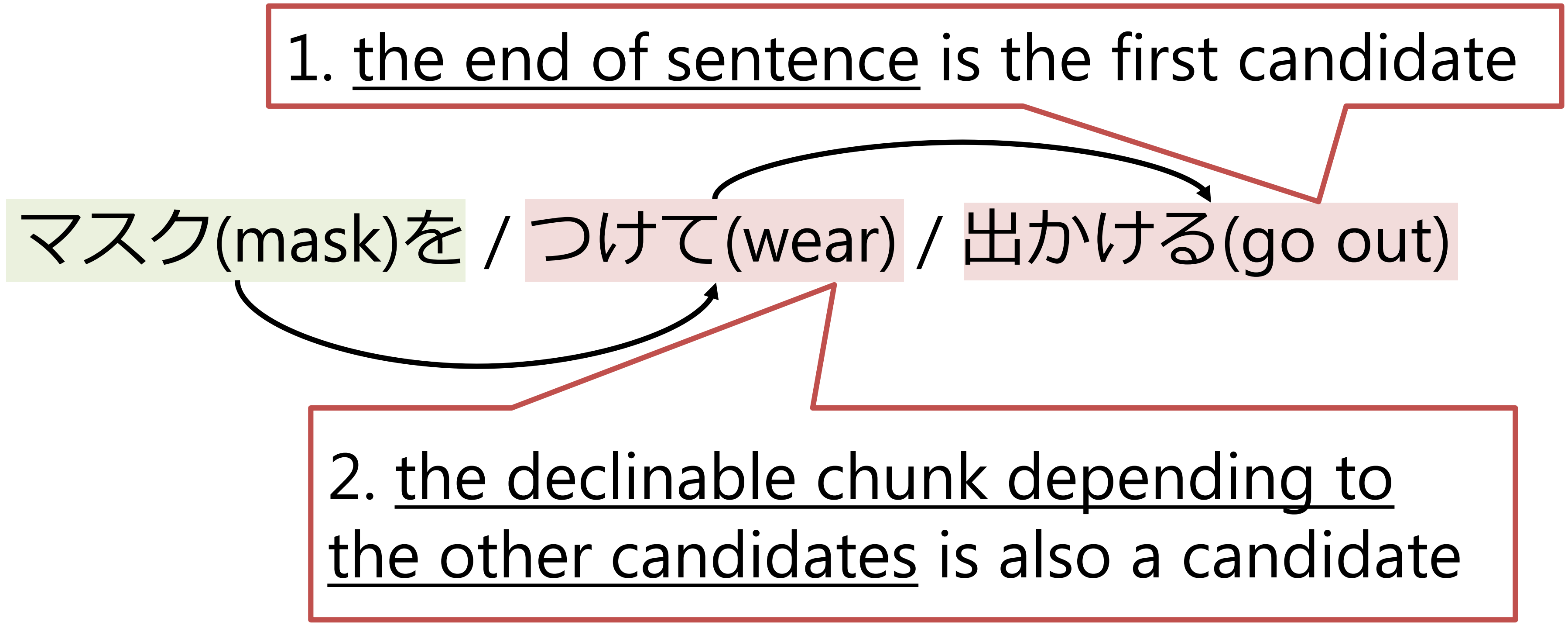
e.g. ease hay fever method

2. collecting detailing pages

detailing pages

- same domain
- all query keywords in anchor text (and a title of the seed page)

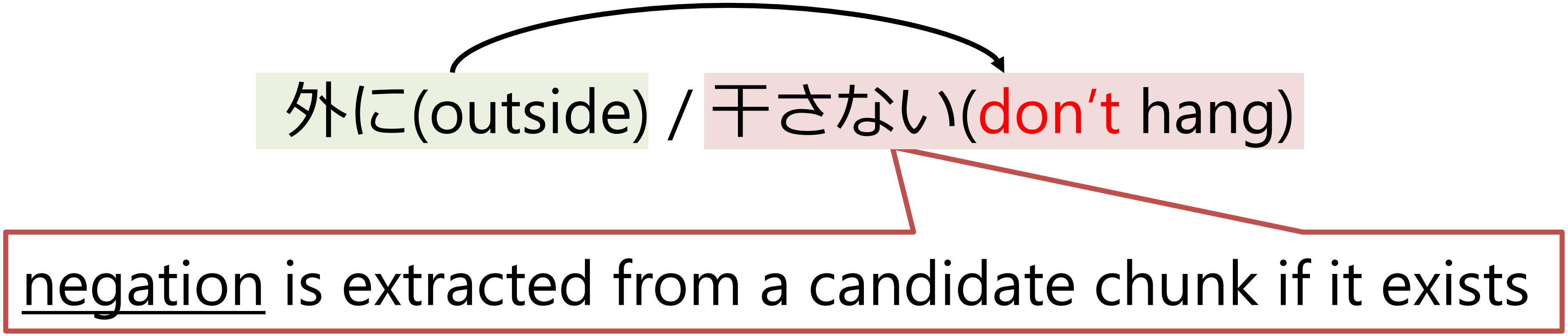
2.1 Finding operation (and negation)



1. the end of sentence is the first candidate

マスク(mask)を / つけて(wear) / 出かける(go out)

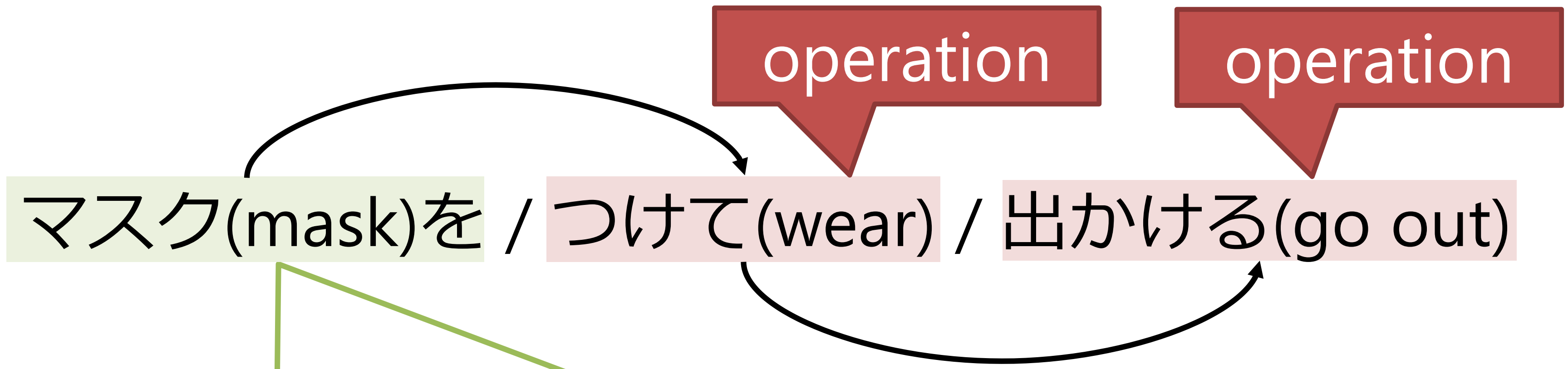
2. the declinable chunk depending to the other candidates is also a candidate



外に(outside) / 干さない(don’t hang)

negation is extracted from a candidate chunk if it exists

2.2 Finding target and postposition



operation

operation

マスク(mask)を / つけて(wear) / 出かける(go out)

the indeclinable chunk depending to operation
→ target and postpositional particle are extracted

extracted quadruplet

(マスク(mask), を, つける(wear), -)

(no) negation

3. Ranking by site frequency

- We use site frequency instead of DF to reduce effects of site template (e.g. copyright statement)

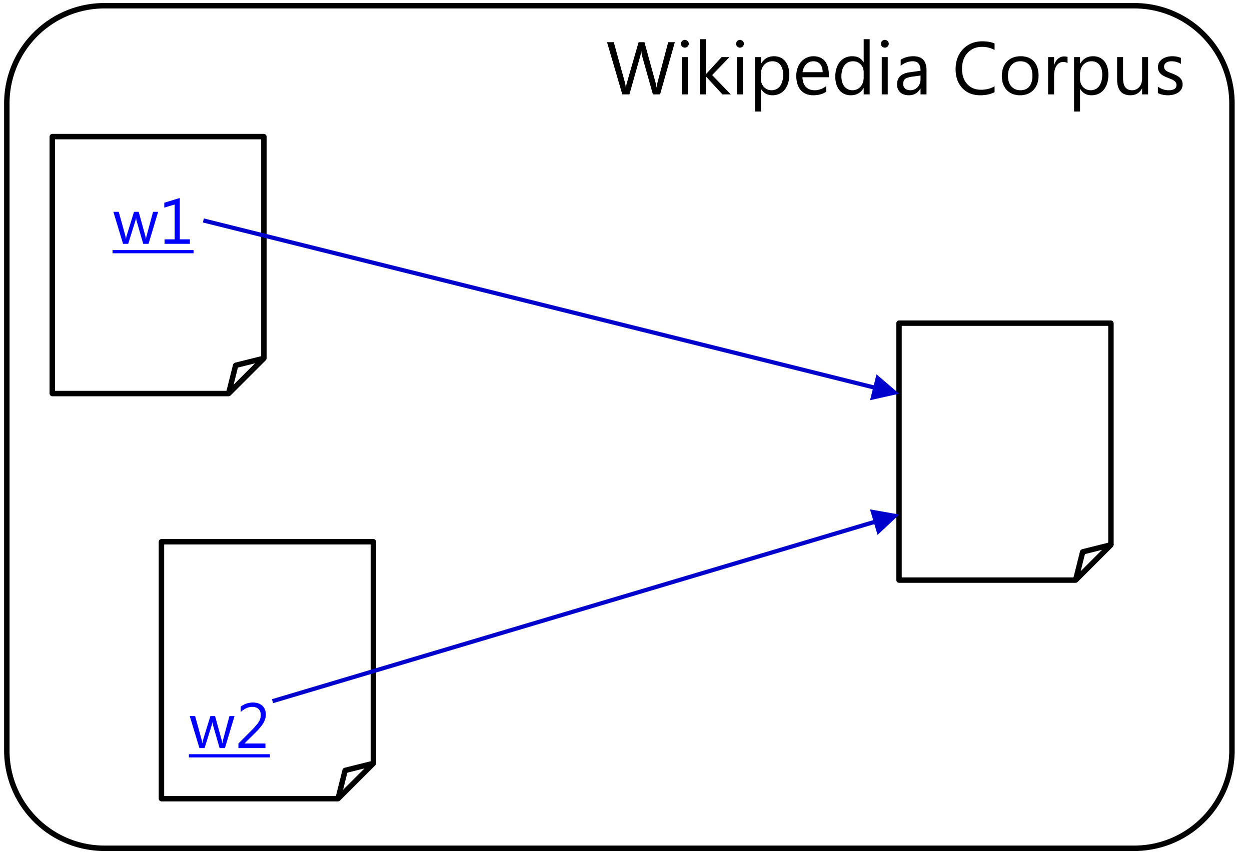
- We propose two ranking methods:

(A) Site frequency of pair
(マスク(mask),を,つける(wear), -)
→ Order by SF(pair of target and operation)

(B) Site frequency of min and max of target and op
consider importance of target and op separately
(マスク(mask),を, つける(wear), -)
→ Order by Min(SF(target), SF(op)),
Max(SF(target), SF(op))

Synonyms by Wikipedia data

- To identify synonyms, we used Wikipedia data
- If two words are used as anchor text liking to the same page, they are regarded as synonyms.



Result

- We tried 50 queries of 4 categories
- We compared nDCG@k

	k=1	k=5	k=10	k=50
(A)	0.109	0.150	0.171	0.191
(B)	0.098	0.119	0.132	0.166
baseline	0.013	0.040	0.053	0.096

Category	k=1	k=5	k=10	k=50
Health	0.140	0.141	0.173	0.191
Education	0.000	0.075	0.107	0.161
Daily life	0.167	0.153	0.150	0.154
Sequential	0.100	0.237	0.253	0.259

Problems

- the part where subtasks are extracted
  e.g. the page describing not only methods to ease hay fever but also mechanism of hay fever

- same subtask in different expressions
  e.g. “wear a mask” = “use a mask”

- limitation of model　: multiple targets are sometimes needed in a single subtask
  e.g. 種に傷をつける (scratch a seed)

target　target

Summary

Our method consists of:

- Collecting pages by query modification and link anchor
- Extracting quadruplets by dependency parsing
  - Finding operation (and negation)
  - Finding target and postpositional particle
- Ranking by site frequency
  - Synonyms by Wikipedia corpus

Our approach is better than the baseline, but it should be improved