The 9th NTCIR Workshop Meeting, December 7, 2011. National Center of Sciences, Tokyo, Japan.

RMIT and Gunma University at NTCIR-9 GeoTime Task

Michiko Yasukawa* J Shane Culpepper †
Falk Scholer † Matthias Petri †

*Gunma University, Japan
† RMIT University, Australia

Table of Contents

Background

- Inverted indexes
  - A classical solution for search problems.
    - A vocabulary of terms mapped to documents.
  - Terms (words or n-grams) are defined at indexing time, and not changed at query time. ☹
- Self-indexes
  - A new viable alternative to inverted indexes.
    - A data structure for character level pattern matching.
  - Word boundaries are flexibly changed at query time. ☺
    - Search terms are arbitrary patterns of characters.

Ranked Self-Indexing

- Prior work
  - Frequency counting for a single phrase.
  - Search effectiveness has not been evaluated.
- A new search engine, NeWT [Culpepper, et al. 2010]
  - Efficient term frequency counting.
    - two wavelet trees
    - BWT (Burrows-Wheeler Transform)
  - Anything can be a term at query time.
    - Ranked search for multiple phrases, words, morphemes, and/or any character sequences.

Ranking metrics in NeWT

- (1) raw term frequency:

$$\text{RAW} = \sum_{t \in q} f_{t,d}$$

RAW : the aggregate of the term frequency, $f_{t,d}$.
$f_{t,d}$: term frequency counts per document.

- (2) BM25 variant:

$$\text{BM25} = \sum_{t \in q} \log\left(\frac{N - f_t + 0.5}{f_t + 0.5}\right) \cdot \text{TF}_{\text{BM25}}$$

$$\text{TF}_{\text{BM25}} = \frac{f_{t,d} \cdot (k_1 + 1)}{f_{t,d} + k_1 \cdot ((1 - b) + (b \cdot l_d / l_{avg}))}$$

N : the number of documents in the collection.
f_t : the number of distinct documents appearances of t.
l_d : the number of UTF8 symbols in the documents.
l_{avg} : the average of l_d over the collection. $k_1 = 1.2, b = 0.75$

Our Goal for NTCIR-9 GeoTime Task

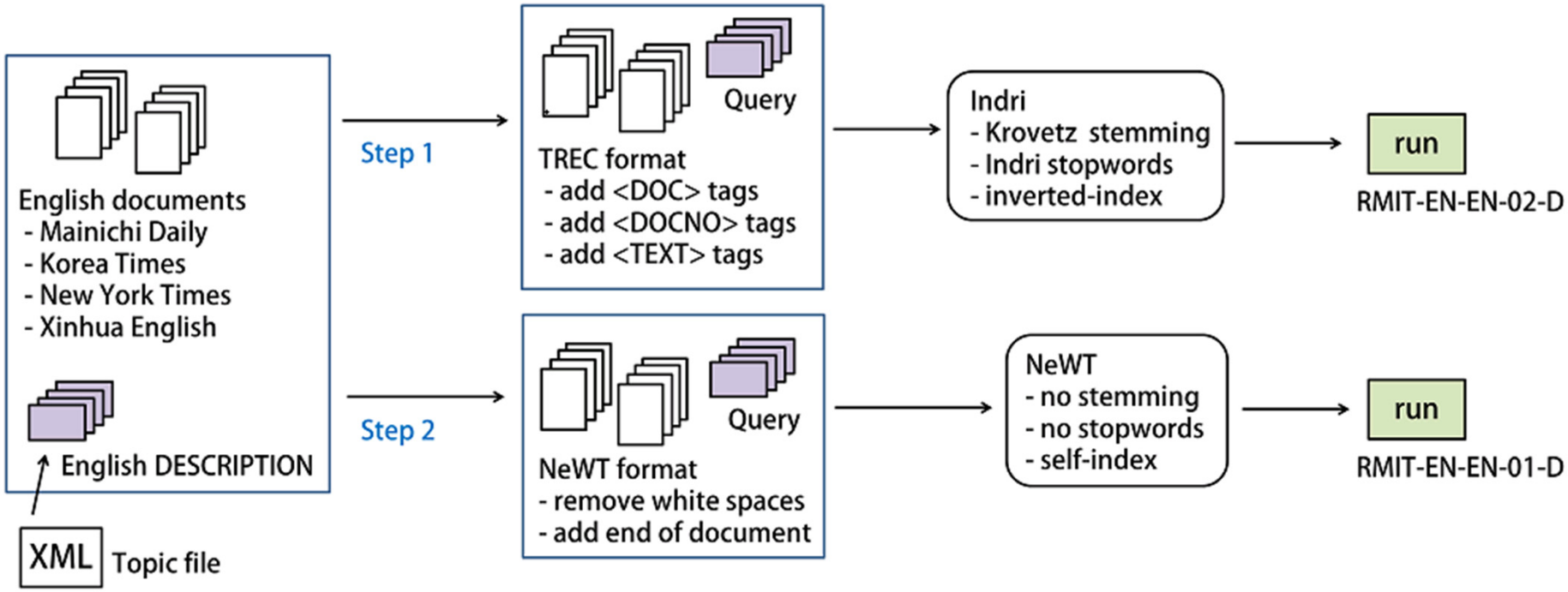
- Compare the search effectiveness:

Indri → *classical*	vs.	NeWT → *innovative*
✓ Inverted index (Terms are static.)		✓ Self-index (Terms are flexible.)
✓ Multilingual support		✓ Language independent

- [Step1] Search in English with Indri.
- [Step2] Experiment in English with NeWT.
- [Step3] Search in Japanese with Indri.
- [Step4] Experiment in Japanese with NeWT.
- [Step5] Query Expansion in Japanese with NeWT.

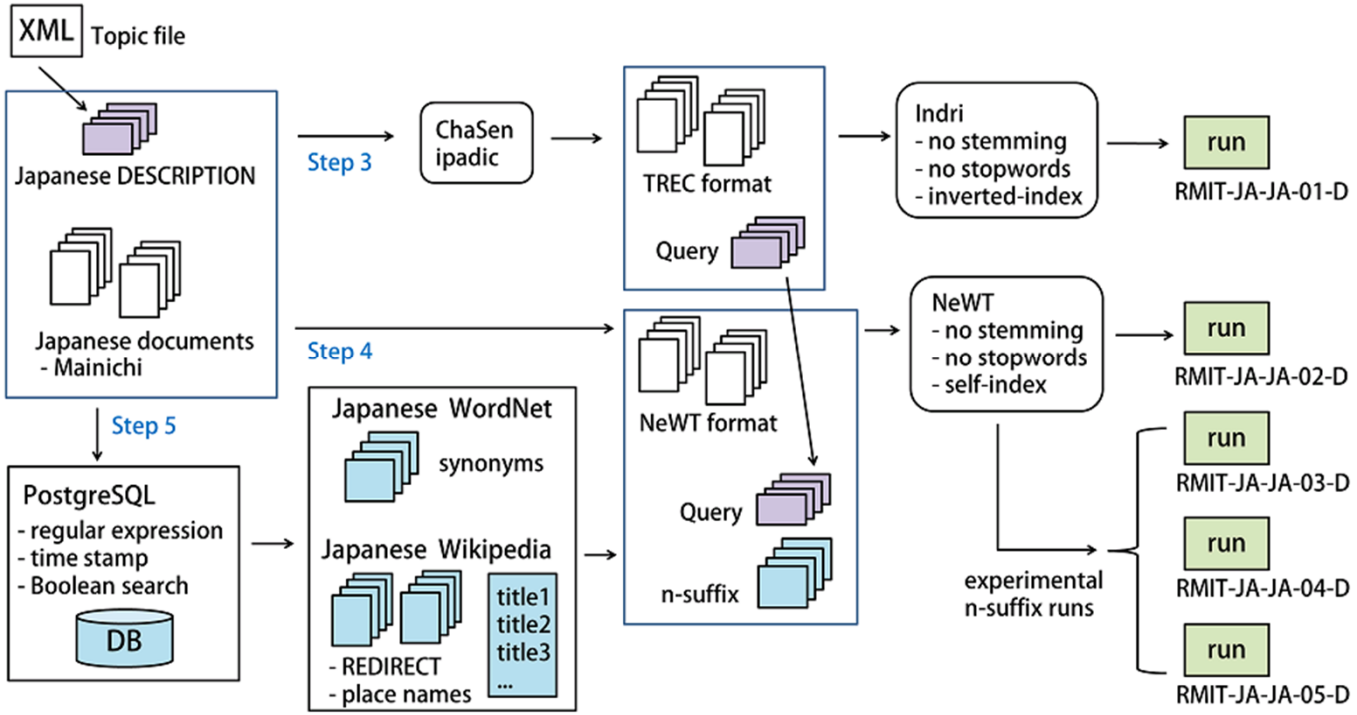
Experimental Framework (for English)

- Step1: English search with Indri
- Step2: English search with NeWT



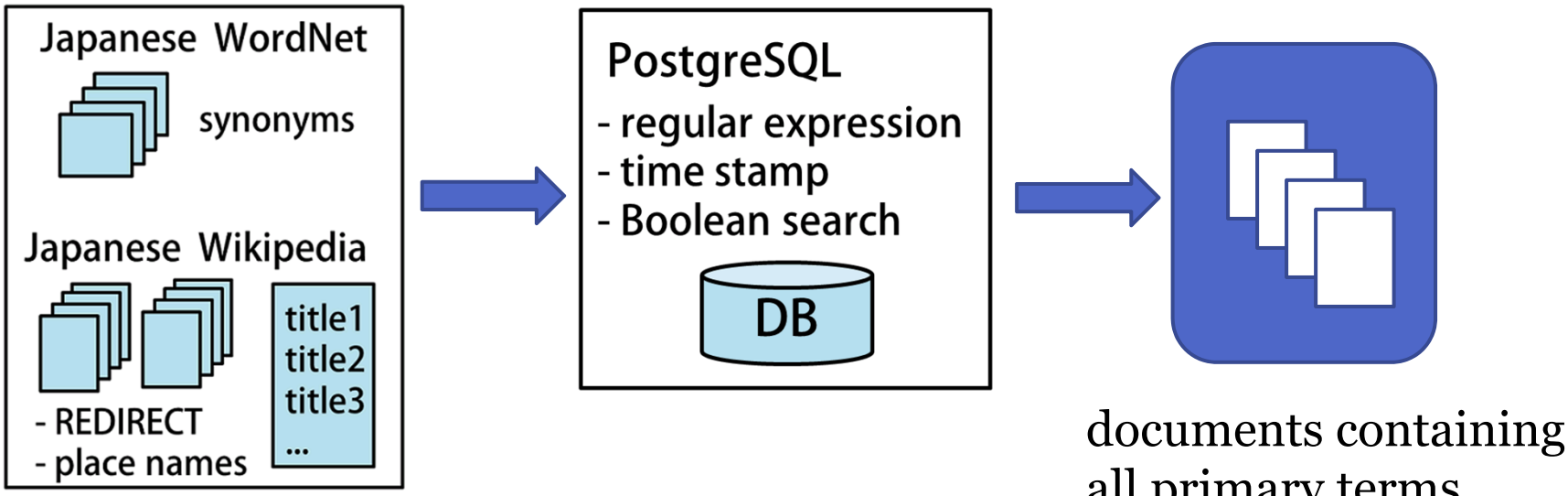
Experimental Framework (for Japanese)

- Step3: Japanese search with Indri
- Step4: Japanese search with NeWT → Substring Mismatch
- Step5: Step4 + *n*-suffix query expansion



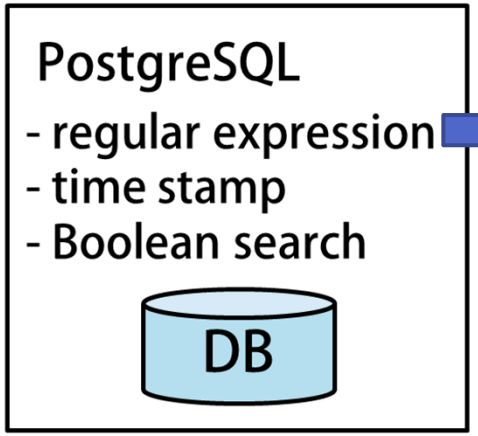
Query Expansion in Japanese

- Boolean search to gather initial documents.
  - All topic terms appear in each document. (AND)
  - Synonyms from Japanese WordNet and Wikipedia. (OR)
- Later documents likely contain “when and where”.
  - Reverse chronological order of time stamp. (ORDER BY)



Query Expansion in Japanese (Cont.)

- Regular Expression in PostgreSQL
  - *n*-suffixes from the gathered documents.
    (*n*-character suffixes at the tail of the query term)
- For the experiment:
  - 100 n-suffixes per topic.
  - *n*-suffixes using *n=2, 3, 4*.



Regular Expression	n-suffixes
('peter piper', 'p.{1}', 'g')	{pe}, {pi}, {pe}
('peter piper', 'p.{2}', 'g')	{pet}, {pip}
('peter piper', 'p.{3}', 'g')	{pete}, {pipe}
('peter piper', 'p.{4}', 'g')	{peter}, {piper}
('peter piper', 'p.{8}', 'g')	{peter pip}

Results in English

- NeWT run EN-01 shows higher performance. (nDCG@10)
- But, more poorly on other effectiveness measures.
- Overall, no statistically significant difference.

Run	System	Ranking	Preprocess	Expansion	MAP	Q	nDCG@10	@100
EN-01	NeWT	BM25	None	None	0.2477	0.2524	0.4282	0.3691
EN-02	Indri	Dirichlet LM	Krovetz	None	0.2830	0.3057	0.3531	0.3763
JA-01	Indri	Dirichlet LM	ChaSen	None	0.3779	0.4119	0.4769	0.5109
JA-02	NeWT	BM25	None	None	0.3084†	0.3239†	0.3510†	0.3936‡
JA-03	NeWT	BM25	None	2-suffixes	0.3282	0.3349	0.4768	0.4653
JA-04	NeWT	BM25	None	3-suffixes	0.3671	0.3714	0.5230	0.5211
JA-05	NeWT	BM25	None	4-suffixes	0.3376	0.3398	0.4988	0.4841

† and ‡ indicate statistical signicance relative to the baseline run at the 0.05 and 0.001 levels respectively, based on a paired t-test.

Results in Japanese

- The NeWT run JA-02 performed worse than the Indri run JA-01.
- The 3- and 4-suffix query expansion runs were effective. (nDCG@10)
- But, the differences were not statistically significant.

Run	System	Ranking	Preprocess	Expansion	MAP	Q	nDCG@10	@100
EN-01	NeWT	BM25	None	None	0.2477	0.2524	0.4282	0.3691
EN-02	Indri	Dirichlet LM	Krovetz	None	0.2830	0.3057	0.3531	0.3763
JA-01	Indri	Dirichlet LM	ChaSen	None	0.3779	0.4119	0.4769	0.5109
JA-02	NeWT	BM25	None	None	0.3084†	0.3239†	0.3510†	0.3936‡
JA-03	NeWT	BM25	None	2-suffixes	0.3282	0.3349	0.4768	0.4653
JA-04	NeWT	BM25	None	3-suffixes	0.3671	0.3714	0.5230	0.5211
JA-05	NeWT	BM25	None	4-suffixes	0.3376	0.3398	0.4988	0.4841

† and ‡ indicate statistical signicance relative to the baseline run at the 0.05 and 0.001 levels respectively, based on a paired t-test.

Conclusions

- A new self-indexing search engine, NeWT
  - Experimented on the multilingual task.
    - Language processing at query time, not at indexing time. ☺
    - Multiple languages can be incorporated into a single index. ☺
  - Search effectiveness was examined.
    - Efficient document ranking with self-indexes. ☺
    - For GeoTime topics, no significant effectiveness . ☹

- Future work:
  - Efficiently determine IDF (Inverse Document Frequency).
  - Explore the substring mismatch problem.

Thank you very much for your kind attention.

Michiko Yasukawa
michi@cs.gunma-u.ac.jp