Constructing English Reading Courseware

Masao Utiyama (NICT)
Midori Tanimura (Kinki Univ.)
Hitoshi Isahara (NICT)

Contents

- Goal and motivation
- Courseware constructed
- Construction algorithm
- Experiment
- Conclustion

Goal

English reading courseware $\leftarrow$ target corpus + vocabulary

Motivation

- Help students acquire target vocabulary
- Help teachers create courseware

Benefit

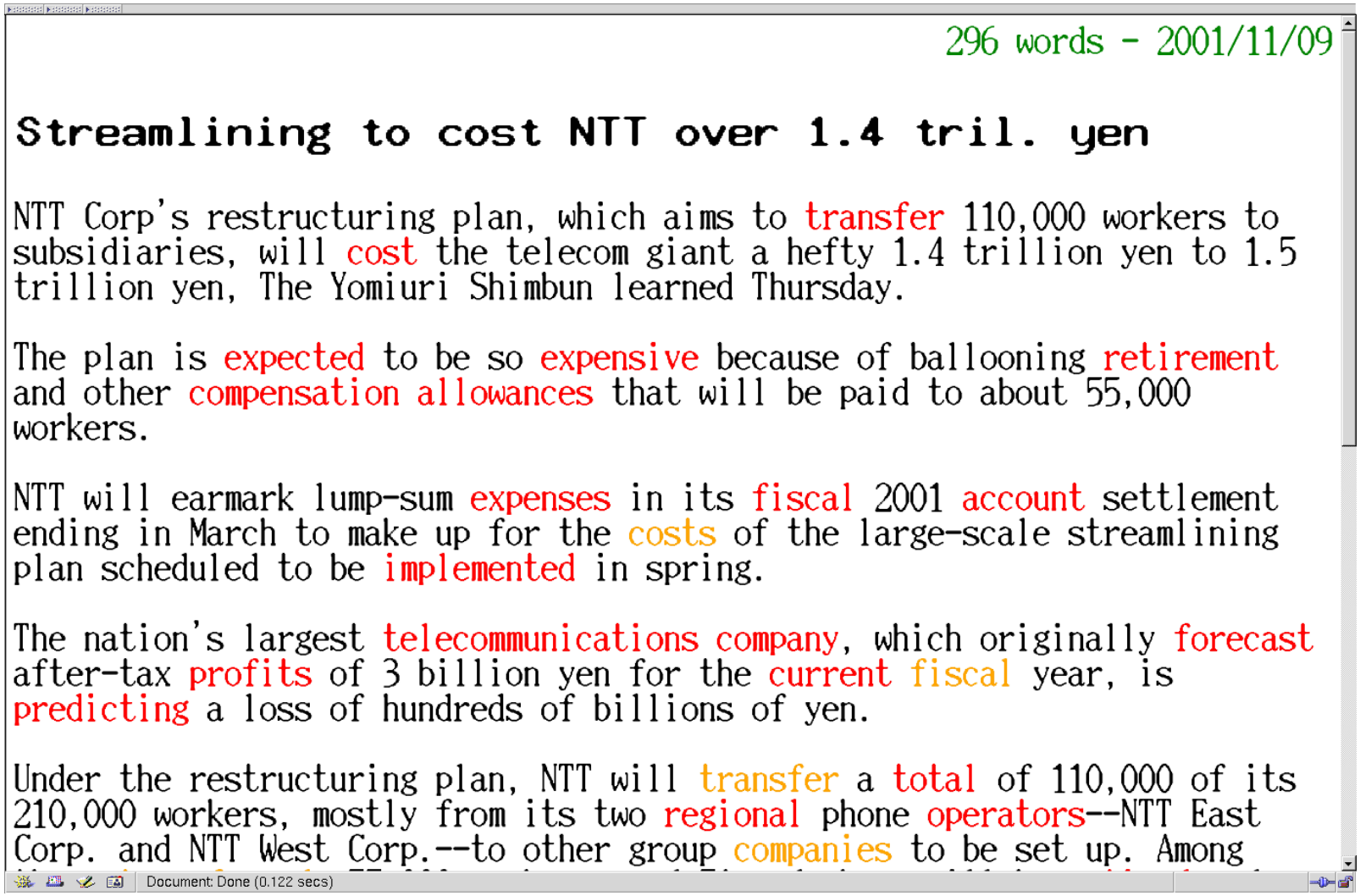
ESP (English for Special Purposes)

Courseware constructed

Vocabulary: TOEIC (Test of English for International Communication) +
Corpus: The Daily Yomiuri newspaper articles
→
Courseware:

- 116 articles
- All of the TOEIC vocabulary
- Distribution of the vocabulary was quite dense

Example article



296 words - 2001/11/09

Streamlining to cost NTT over 1.4 tril. yen

NTT Corp's restructuring plan, which aims to transfer 110,000 workers to subsidiaries, will cost the telecom giant a hefty 1.4 trillion yen to 1.5 trillion yen, The Yomiuri Shimbun learned Thursday.

The plan is expected to be so expensive because of ballooning retirement and other compensation allowances that will be paid to about 55,000 workers.

NTT will earmark lump-sum expenses in its fiscal 2001 account settlement ending in March to make up for the costs of the large-scale streamlining plan scheduled to be implemented in spring.

The nation's largest telecommunications company, which originally forecast after-tax profits of 3 billion yen for the current fiscal year, is predicting a loss of hundreds of billions of yen.

Under the restructuring plan, NTT will transfer a total of 110,000 of its 210,000 workers, mostly from its two regional phone operators--NTT East Corp. and NTT West Corp.--to other group companies to be set up. Among

Document: Done (0.122 secs)

Efficient courseware

Operational definition of efficiency

- As short as possible
- Contains the required vocabulary

Effects

- Exposes students the target vocabulary
- Enable students to learn words in contexts through reading

Optimization: Converting definition into algorithm

$\hat{C} = \arg\min_C \text{Length}(C)$

- C is courseware
- C is a subset of the target corpus
- C contains all of the target vocabulary
- $\hat{C}$ is the minimum length courseware

Greedy method

To construct the minimum length courseware

- Step1: Get a document with the maximum number of new words
- Step2: Put it into the courseware
- Step3: Until the courseware covers all of the target vocabulary

Document score (1/2)

$$\text{Score}(d|\alpha, V_{\text{todo}}, V_{\text{done}}) = \alpha g(d|V_{\text{todo}}) + (1-\alpha) g(d|V_{\text{done}})$$

- Both uncovered (V_{todo}) and covered (V_{done}) vocabulary
- Uncovered vocabulary has priority over covered vocabulary

$$\alpha = \frac{|V_{\text{done}}|}{1 + |V_{\text{done}}|}$$

Document score (2/2)

$$g(d|V) = \frac{k_1 + 1}{k_1((1-b) + b\frac{|W(d)|}{E(|W(\cdot)|)}) + 1}|W(d) \cap V|,$$

- Based on the Okapi BM25 function (information retrieval measure)
- Documents relevant to the target vocabulary
- Large when many words are shared due to $|W(d) \cap V|$
- Large when the document length is short due to $\frac{|W(d)|}{E(|W(\cdot)|)}$

Effects

Short courseware that covers the target vocabulary

Experiment

- TOEIC vocabulary
- The Daily Yomiuri newspaper article corpus
- Statistics of the constructed courseware
- Problems
- Use in the classroom

Vocabulary: TOEIC

- compiled by Chujo 2003 (publicly available)
- 640 entries
- beginner to intermediate level

Corpus: The Daily Yomiuri

- 25,000 articles
- 300 words or less
- Japanese counterparts exist
- lemmatized to match with the vocabulary

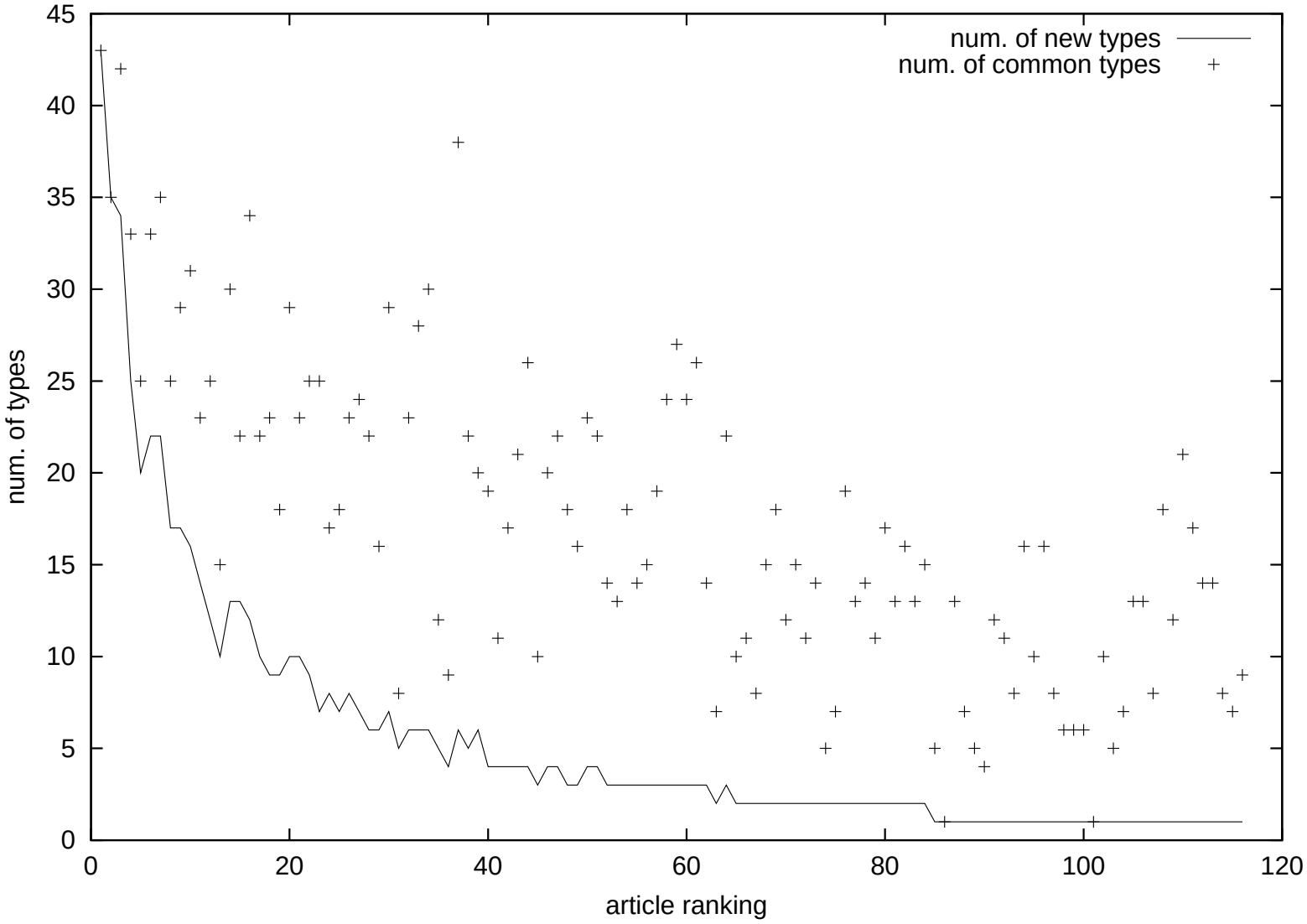
Efficiency comparison with randomly sampled articles

Coursware = 20,900 tokens, 116 articles.

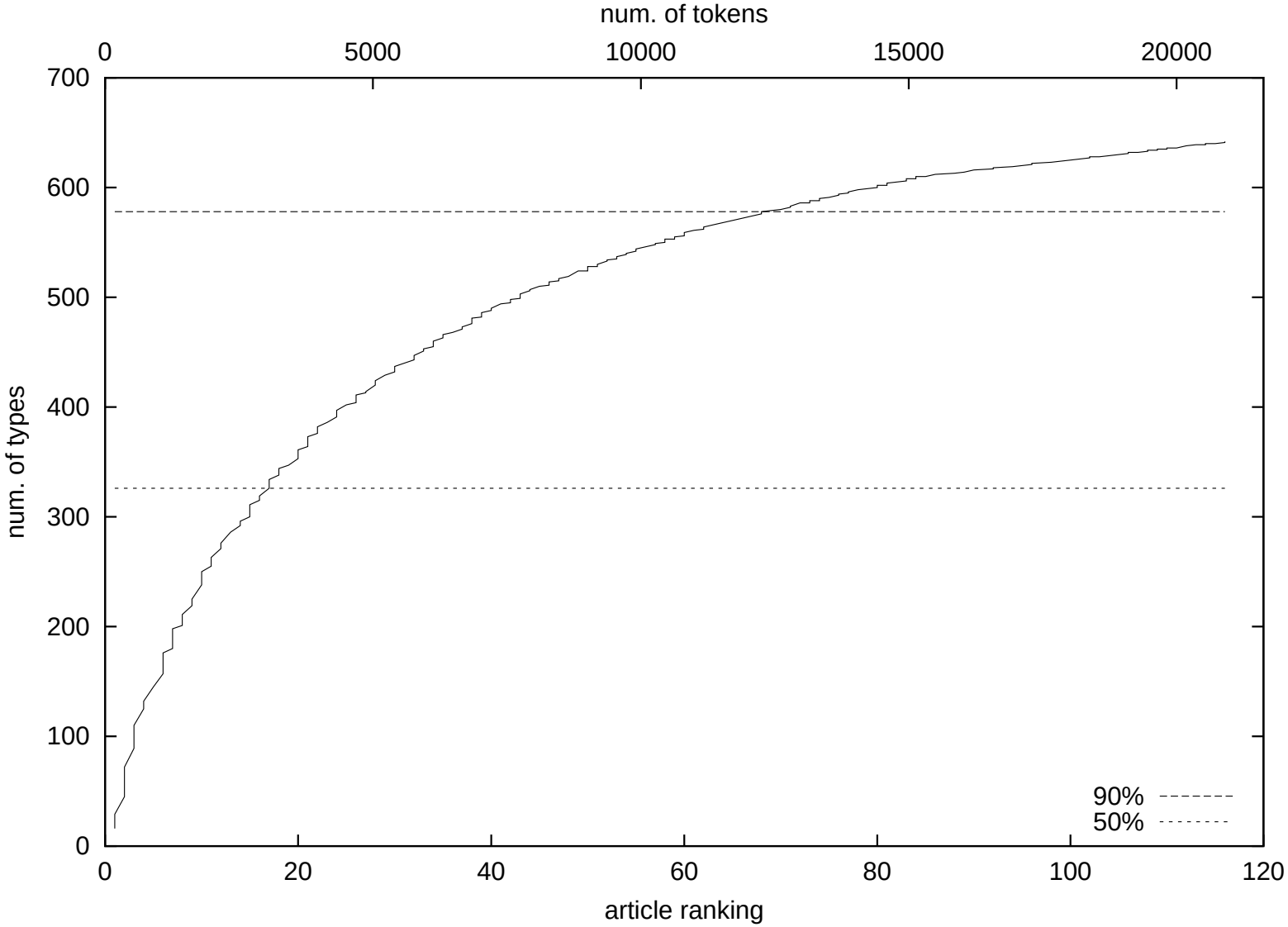
	random		courseware	summary
	average	SD		
avg. num. of common tokens	19.3	1.1	25.3	large
avg. num. of common types	12.8	0.6	17.4	large
coverage	0.616	0.016	1.0	high

Constructed courseware was efficient.

Distribution of the number of types



Increase in the number of covered types



Problems: Usage discrepancies

agency

TOEIC → a business that provides particular services, (an advertising agency)
Yomiuri → an administrative unit of government

appointment

TOEIC → a meeting arranged in advance
Yomiuri → the act of putting a person into a non-elective position

Remedy for the mismatches

- Use a corpus that is similar to the TOEIC vocabulary
- Best is the use of the TOEIC tests.

Use in the classroom (1/2)

- 3 English classes in one university since May 2004
- Beginner to intermediate level
- Supporting material
- Vocabulary quiz

Use in the classroom (2/2)

Suitable to intermediate level students
Motivation is high.

- Vocabulary quiz:

  High scores

- Meaning in contexts:

  Takashi Kitaoka, **president** of Mitsubishi Electric Corp., said...

- Get used to reading:

  The main textbook has become easy to read.

Promising, though detailed evaluation has yet to be done.

Conclusion

- Efficient Courseware ← Corpus + Vocabulary
- Optimization with respect to efficiency
- Promising

Future work

- Detailed evaluation
- Acquisition of phrases