Construction of Personalized and Purpose-Oriented Thesaurus

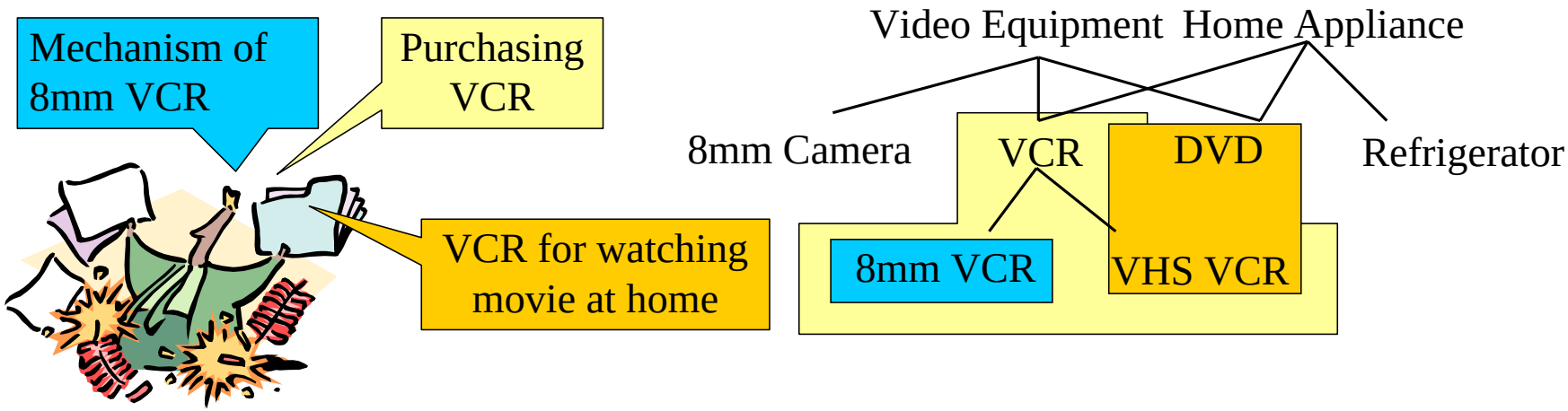
Masaharu YOSHIOKA Hokkaido Univ.
Makoto HARAGUCHI

Background and Objectives

- Background:
  - Concept structure of electronic dictionary and thesauri
    - General purpose thesaurus is too general and they do not have fine granularity abstract concepts to represent users' particular intention
    - Concept category for representing user's particular intention is selective and is not directly corresponds to a single common concept category stored in general thesauri
- Objectives
  - Propose methodology to construct personalized and purpose-oriented thesaurus
    - Modification of concept structure of electronic dictionary for particular user's intention
    - Evaluation of Construct Personalized and Purpose-Oriented Thesaurus by Application to Information Retrieval System

Appropriate Concept Category for Particular Users Intention

- Role of keyword in different situation
  - Use a keyword as a particular concept
    - A keyword corresponds to a concrete concept category
  - Use a keyword as a concept category
    - A keyword corresponds to an abstract concept category
  - Use a keyword as representative concept of intended category
    - A keyword corresponds to an abstract concept category
    - General thesaurus may not have appropriate concept category



Construction of Tentative Concept Category

- Construction of concept category from relevant documents
  - Assumption
    - Tentative concept category should exist more frequently than usual documents
    - Alternative keyword for representing this category exists in relevant documents
  - Method
    - Use mutual information content for selecting an abstract concept category
    - Collect keywords that belongs to selected category to make tentative concept category
  - Evaluation measure for selecting an abstract concept category
    - Appropriate concept category for characterizing relevant documents might have higher mutual information content with relevant documents

$$I(T;W) = I(W;T) = H(W) - H(W \mid T)$$

$$= \sum_{w \in W} \left[- p(w) \log_2 p(w) + \sum_{t \in T} p(t) p(w \mid t) \log_2 p(w \mid t) \right] \cdots (1)$$

$$G(w) = - p(w) \log_2 p(w) + \sum_{t \in T} p(t) p(w \mid t) \log_2 p(w \mid t)$$

$$= p(r) p(w \mid r) \{ \log_2 p(w \mid r) - \log_2 p(w) \}$$

$$+ p(\bar{r}) p(w \mid \bar{r}) \{ \log_2 p(w \mid \bar{r}) - \log_2 p(w) \}$$

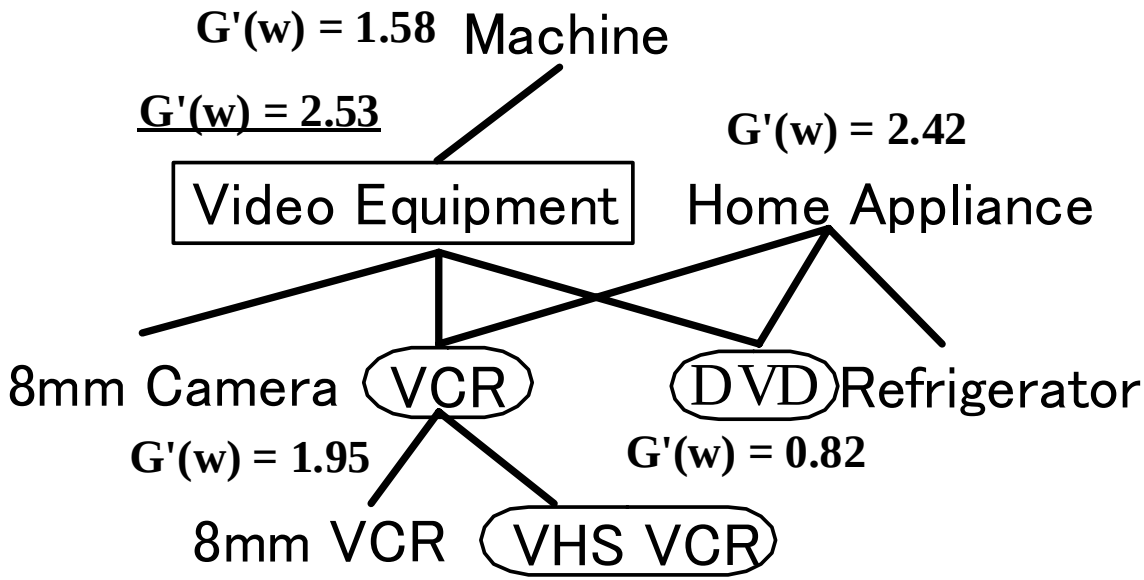
$$\approx p(r) p(w \mid r) \log_2 \frac{p(w \mid r)}{p(w)} \cdots (2)$$

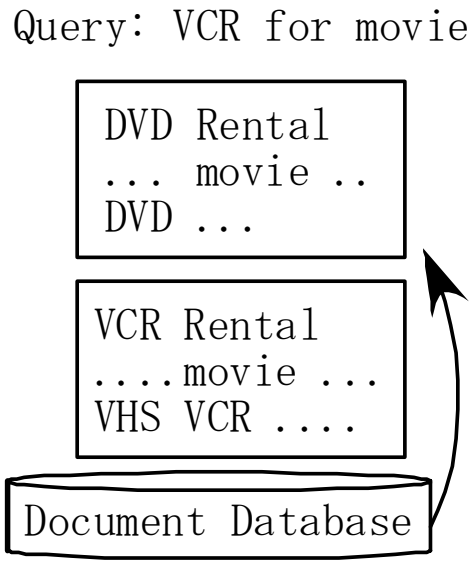
Example of Tentative Concept Category

- Evaluation measure for selecting an appropriate voncept vategory

$$G'(w) = p(w \mid r) \log_2 \frac{p(w \mid r)}{p(w)} \cdots (3)$$

- Construction method
  - Calculate evaluation measure for each keyword and candidate abstract concept category
  - Select concept that has highest evaluation measure for appropriate one



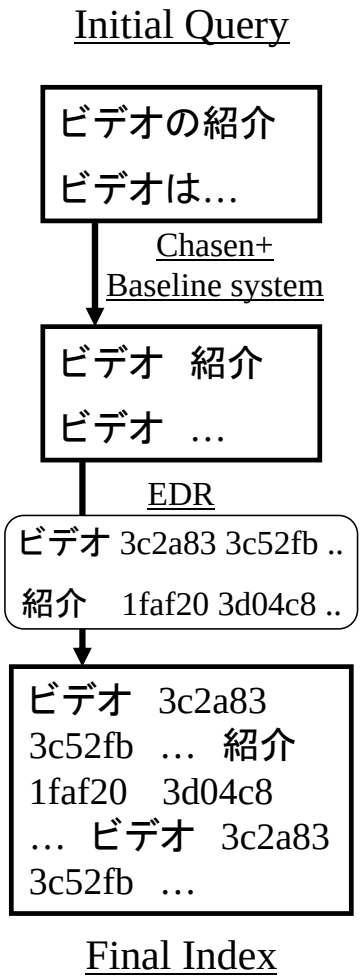


IR System based on Personalized and Purpose-Oriented Thesaurus

- Evaluate effectiveness of tentative concept construction
- IR system that uses personalized and purpose-oriented thesaurus for query expansion
  - Prototype system
    - Implementation based on Information Retrieval Package（Uchiyama, 2001）: Baseline system
      - BM25-base IR system
    - Electronic Dictionary
      - EDR Dictionary
      - Goi-Taikei- A Japanese Lexicon日本語語彙体系
  - Test Collection: NTCIR-1 (Japanese)
    - Document set: Abstract for conference proceedings: 332918 documents
    - Number of topic:53
      - Type of data: Title、Description、Narrative、Concept
      - Relevance judgment：A（Relevant）,B（Partially relevant）,C（Irrelevant）
    - Type of retrieval:
      - long（Title+Description+ Narrative+Concept）
      - short(Description）
    - Use B relevance judgment and calculate mean-average precision

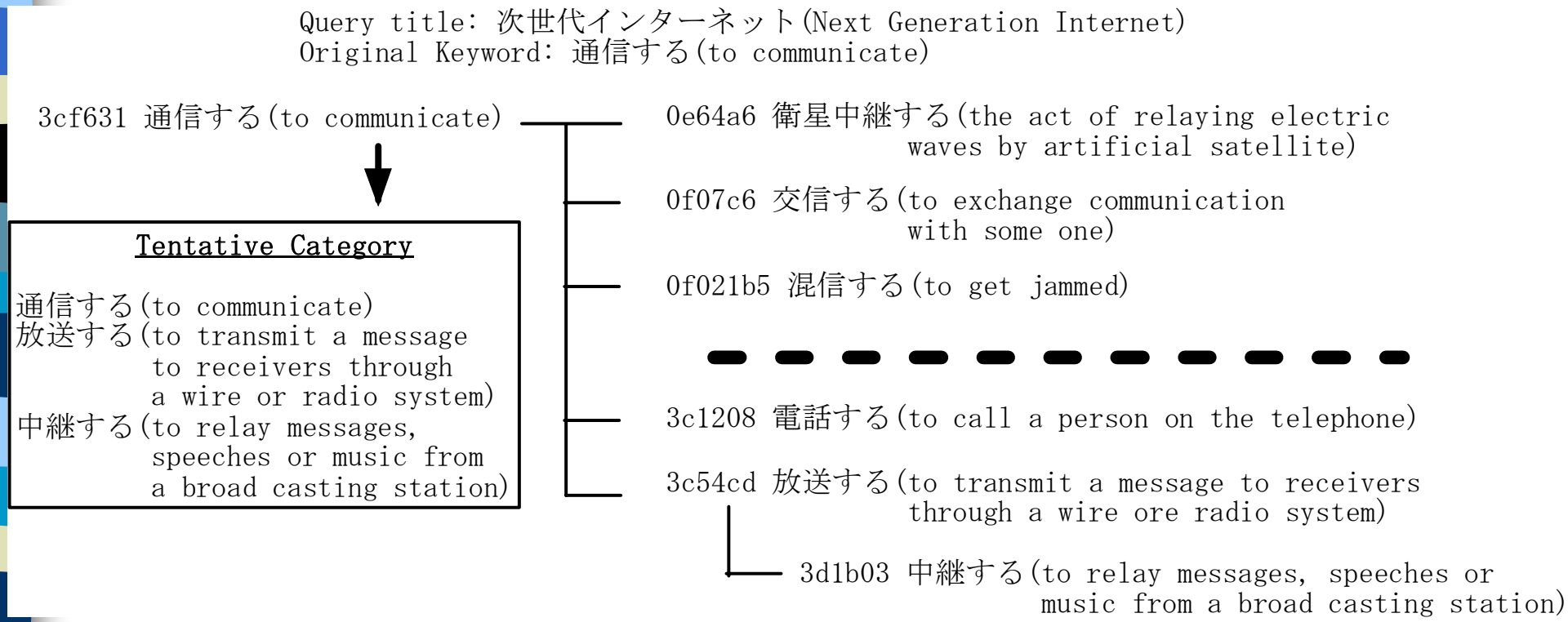
Prototype System

- Construction of retrieval index
  1. Apply morphological analyzer to extract initial index keywords (mostly from noun)
  2. Check EDR (or Goi-Taikei) Dictionary about initial keywords and add concept ID list that corresponds to initial keywords to index
- Information retrieval based on adaptive generalization
  1. Extract initial query keywords by using morphological analyzer
  2. By using relevant document sets (select 5 relevant documents by automatic feedback), the system construct tentative concept category
  3. Add words of constructed category for query expansion



Example of Constructed Concept Category

- Query title: 次世代インターネット(Next Generation Internet)
  - Original Keyword: 通信する(to communicate)
  - Generalized concept: 3cf631 23 subcategories and 56 words
  - Select 3 words from relevant documents



Evaluation Result (EDR)

- PP-oriented thesaurus results has better performance than original thesaurus
- PP-oriented thesaurus has poorer performance than baseline system

Mean Average Precision

System type	Long	Short
PP-oriented Thesaurus	0.4587	0.3587
Original Thesaurus	0.4283	0.3131
Baseline System	0.4880	0.4097

Expanded Query Size

System type	Long	Short
PP-oriented Thesaurus	50.89	8.50
Original Thesaurus	46.22	7.44
Baseline System	209.90	186.47

Evaluation Result (Goi-Taikei)

- Similar to EDR case
- Goi-Taikei has poorer performance than EDR one

Mean Average Precision

System type	Long	Short
PP-oriented Thesaurus	0.4352	0.3106
Original Thesaurus	0.4282	0.3087
Baseline System	0.4880	0.4097

Expanded Query Size

System type	Long	Short
PP-oriented Thesaurus	67.08	9.53
Original Thesaurus	36.28	6.70
Baseline System	209.90	186.47

Discussion

- Coverage of thesauri
  - Since these thesauri don’t have good coverage on technical terms, baseline system that adds all terms in relevant documents works well
  - We need a mechanism to expand queries by using words in relevant documents
- EDR dictionary has better performance than Goi-Taikei
  - Goi-Taikei has approximately 3000 concept categories and EDR has approximately 50000 concept categories.
  - Because of that, Goi-Taikei does not have appropriate abstract concept categories for particular purpose

Conclusion

- Propose methodology to construct personalized and purpose-oriented thesaurus
  - Propose methodology to select relevant term list from general abstract concept category.
  - It improve the performance of IR system compared with original thesaurus.
  - However, it is not sufficient for IR performance compared to automatic relevance feedback query expansion.
- Apply this method to different application domain.
  - Construct purpose-oriented thesaurus for a category in open directory.