Paraphrasing Japanese noun phrases using character-based indexing

Tokunaga Takenobu, Tanaka Hozumi and Kimura Kenji
Department of Computer Science
Tokyo Institute of Technology

Paraphrasing

- Definition
  "A process of transforming an expression into another while keeping its meaning intact."
- What is the basis of semantic equivalence?
- What kinds of clues suggest equivalence?

→Application dependent

Applications of paraphrasing

- Machine translation
  - Translation equivalence
  - Parallel corpora
- Information Extraction
  - Denoting the same events
  - Date, Place, Named entities
- Information Retrieval
  - Retrieves the same documents
  - Query expansion by thesauri

Aspects of Paraphrasing

- Approaches
  - Corpus-based
    - Lattice-based matching
  - Rule-based
    - Morpho-syntactic transformation rules
- Target units
  - Words → Thesaurus
  - Phrases
  - Sentences

Our Approach

- Corpus-based
  - Information retrieval
    - Character-based indexing
  - Natural language processing
- Target
  - Japanese noun phrases
- Usable to phrasal index term expansion in information retrieval

Japanese Writing Scripts

- *Kanzi* (Chinese characters): ideograms
  e.g. 学 (study), 通 (commute), 子 (child)
- *Hiragana*: phonograms
  e.g. あ, い, う, え, お
- *Katakana* (imported words): phonograms
  e.g. ア, イ, ウ, エ, オ
- Roman alphabet: phonogram

Paraphrase examples

- 情報/の/検索 (retrieval of information)
  → 情報/検索 (information retrieval)
  - STR: XのY → XY
- 通学/する/子供 (a commuting child)
  学校/に/通う/子供 (a child going to school)
  - STR: ???
- Need to take into account word formation ability of *Kanzi*

Overview of the Proposed Method

- Store passages in the database with character-based indexing
- Given a noun phrase, retrieve passages to give paraphrase candidates
- Filter irrelevant candidates based on syntactic and semantic constraints
- Rank the resulting candidates

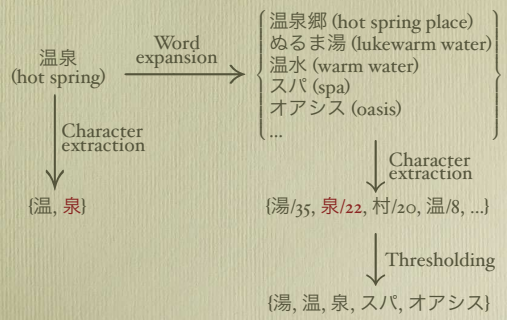
Preparing Database

- Segmentation into passages
- Morphological analysis → content word and unknown words
- Extraction of index: *Kanzi*, *Katakana* words and Numbers (<num>)

Query Expansion

- Replacing an index term in a query with its synonym set
- To solve surface notational variants of index terms
- Referring to a thesaurus which defines equivalence classes of words

Query Expansion: An Example



Term Weighting

$$w(k) = \begin{cases} 100 & \text{if } k \text{ is } Katakana \text{ word or } \langle num \rangle \\ 100 \times \dfrac{\log fr(k, C_t)}{\sum_{k' in E(t)} \log fr(k', C_t)} & \text{if } k \text{ is a } Kanzi \end{cases}$$

e.g. {湯/35, 泉/22, 温/8, スパ, オアシス}

$$w(\text{湯}) = 100 \times \frac{\log 35}{\log 35 + \log 22 + \log 8} = 40.7$$

Retrieving Passages

- Similarity measure

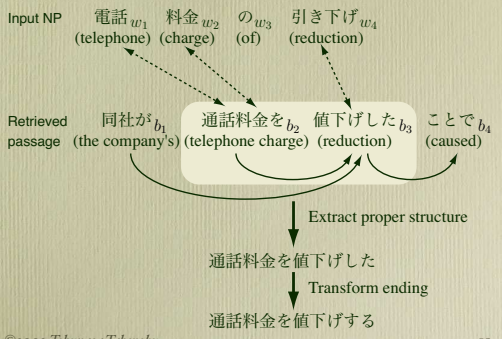
$$sim(I, D) = \sum_{k \in I \wedge k \in D} w(k)$$

I : Input noun phrase
D : passage

Constraints

- Semantic constraints
  Retrieved passages should contain all concepts mentioned in the input noun phrase
- Syntactic constraints
  Retrieved passages should have a syntactically proper structure corresponding to the input noun phrase

Constraints: An Example



Reranking

- Similarity score of passage retrieval
- Distance between words
  "Counterparts of adjacent words in the input should be located closer in the paraphrase."
- Contextual information
  Adopts the idea *one sense per collocation* to disambiguate *Kanzi* meaning.

Contextual Information

- Generate context vectors by extracting content words from:
  - Paraphrase candidate
    the article including the passage
  - Input noun phrase
    all the articles including the noun phrase
- Calculate the context similarity by *tf•idf* term weighting and cosine measure

Experiments

- Queries
  53 queries from BMIR-J2
- Documents
  3 years worth of Newspaper articles (Mainichi Shimbun 1991-1993)
- Tools
  - GETA retrieval engine
  - JUMAN morphological analyzer
  - KNP dependency parser

Qualitative Evaluation

- Correct
  e.g. 冷夏/の/被害 (damage by cool summer)
  → 冷害 (cool summer damage)
- Partially correct
  - Specific
  - General
  - Related
- Incorrect

Partially Correct Examples

- Specific: 農薬 (agricultural chemicals)
  → 殺虫・除草剤 (insecticide and herbicide)
- General: 株価動向 (stock movement)
  → 株価, 為替相場の変動
  (movement of stock and exchange rate)
- Related: 飲料品 (drinks)
  → 国際食品飲料展
  (international exhibition of foods and drinks)

Failure Analysis

- No output for 7 cases.
- No proper paraphrase
  e.g. 液晶 (liquid crystal)
- Limitation of documents collection size
  - Three years worth of newspaper articles is not enough
  - Mismatch of time period of documents and queries

Quantitative Evaluation

- Three independent judges
- 45 paraphrases on average from 46 cases out of 53 queries
- Average accuracy was 10% (Correct and Specific)
- Average precision was 17% (Correct and Specific)
- Agreement measure was very hight in incorrect cases
- Poor results for long inputs

Conclusions and Future Work

- More improvement is necessary for fully automatic paraphrasing
- Usable for suggesting paraphrases to users
  - Novel paraphrases can be extracted
  - Easy to judge incorrect ones
- More precise analysis, such as case analysis
- Integration with syntactic transformation