Matching Word-Order Variations and Sorting Results for the iEPG Data Search

Denis Kiselev, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

Rafal Rzepka, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

Kenji Araki, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

ABSTRACT

This paper describes using a finite-state automaton (FSA) to retrieve Japanese TV guide text. The proposed FSA application can be considered novel due to lack of research on the subject. The automaton has been implemented for matching and extracting all possible combinations of search query words in all possible word orders that may be present in the TV guide text. This implementation also sorts the extraction results by analyzing word semantic features (such as “being an object” or “being a property of an object”). The present paper also proposes a search system using the above implementation and compares it with a baseline system that matches query words (of multi-word queries) in exactly the same and exactly the opposite word orders only. Both systems use morphological parsing and apply a stop list to the query. A multi-parameter evaluation has shown advantages of the proposed system over the baseline one.

Keywords: Electronic Program Guide (EPG), Finite-State Automaton (FSA), Information Retrieval, Lexical Semantics, Morphological Parsing, Natural Language Processing (NLP)

MOTIVATION FOR THIS RESEARCH

Japanese is written without spaces between words. That means a search system processing this language needs to “know” what character strings are words, or at least where character strings that could be words start and end. It is even better if a system attempts to find out what those words, or groups of characters, may mean. The same is true for searching the Japanese language iEPG (Internet Electronic Program Guide or, simply, Web pages saying when, what programs are shown on TV).

It can be concluded from the output of search systems available on major Japanese iEPG websites¹ that those systems most likely apply the direct matching technique to the query,
treated by them as a character string. In other words, they most likely match the search phrase without segmenting it into words (i.e. without morphological parsing and inserting spaces at word boundaries).

Kiselev et al. (2013) suggested improvements to that technique and proposed an iEPG search system utilizing morphological parsing and the core meaning analysis for matching the search query with the TV guide text.

The above authors also demonstrated how using that system could improve search results, however matching query words in all possible orders was left for future work.

The system proposed by the above authors can match query words (provided the query has two or more of them) in exactly the same or exactly the opposite orders only (ibid.). For two-word queries “exactly the same” and “exactly the opposite” are all the possible word order options, however there are more options for longer queries. Thus, the system will successfully match text with “観光地は人気で綺麗 ([kankouchi wa ninki de kirei] the sightseeing spot is popular and beautiful)” in response to the query “綺麗で人気な観光地 ([kirei de ninki na kankouchi] a beautiful and popular sightseeing spot)”, but will not match the same text in response to “人気で綺麗な観光地 ([ninki de kirei na kankouchi] a popular and beautiful sightseeing spot)”.

This ability to express (practically) the same meaning using the same words in various orders is described as a characteristic feature of “context-free languages”, i.e. ones allowing more flexible word combinability, by Maruoka (2011). The order flexibility in Japanese word combinations is illustrated in terms of the “context-free grammar” and NLP (Natural Language Processing) by Tanabe, Tomiura and Hitaka (2000).

Implementing a system capable of matching query words in all possible orders characteristic of the Japanese language, has been the primary motivation for the research described in this paper. The system proposed by Kiselev et al (2013) (mentioned earlier in this section) has been used as a baseline.

It should be noted that both the baseline and proposed system implementations are essentially different from large search engines, such as Google. First, large search engines retrieve web documents, such as websites and parts of them, whereas the proposed and the baseline implementations retrieve pieces of text that describe TV programs. To do so, the implementations do not require indexing millions of web documents (the way a Google webpage3 says it does) and do not need any corpora, such as the approximately 24-GB large Google N-gram Corpus described by Lin et al (2010). Because of their size the implementations could be used locally as, say, internal search systems for TV sets. It seems unlikely that, for instance, the Google search engine can be used in the same way. It has been our purpose to develop the search system for the TV program guide by taking into account the above peculiarities of this task.

**INPUT-OUTPUT FLOW OF THE PROPOSED SYSTEM**

This section contains a concise flow description. Sections that follow explain flow stages in more detail.

Figure 1 shows the major flow stages. The initial stage is TV guide data acquisition. Once the data are offloaded from a Web source by the system and a query is input, the guide and the query are preprocessed, i.e. changed into the format suitable for matching. At the preprocessing stage, the query string first undergoes morphological parsing, stop-listing and semantic-feature analysis (the analysis is described in C. Semantic-feature Analysis of Query Words). The query is then matched with the TV guide text by means of a finite-state automaton (FSA). The semantic-feature analysis results are used to sort (according to possible relevance to the query) TV guide data.
JIRL: A C++ Toolkit for JPEG Compressed Domain Image Retrieval
www.igi-global.com/article/jirl/84022?camid=4v1a