Applying the Stop List and Part of Speech Analysis to Processing the IEPG Search Query

Denis Kiselev, Rafal Rzepka, Kenji Araki
Graduate School of Information Science and Technology, Hokkaido University
{dk,kabura,araki}@media.eng.hokudai.ac.jp

Abstract

The proposed approach to searching the Japanese TV program data features processing one program description at a time and matching it with the query where certain morphemes are excluded and certain parts of speech are defined as mandatory or optional matches. The present paper uses TV program guide search examples to illustrate how the proposed method can improve search results.

Key Words- EPG, NLP, Information Retrieval, Query Processing, Morphological Parsing

1. Introduction

The IEPG (Internet Electronic Program Guide) is a searchable WWW-based database providing textual descriptions for television programs. Multiple sites (http://tv.yahoo.co.jp/ and http://www.tvguide.or.jp/ to give a few URLs) make TV program data publicly available in Japan helping viewers choose from a variety of programs.

The data is normally divided into segments, each one for a separate program. The segments natural language text includes such information as the name for the TV channel broadcasting the program, the broadcasting date, time and the program description from a word or two to about a paragraph in length. More details on the IEPG format, including examples, are given in Yamasaki et al. [1].

We have developed a search system that extracts the IEPG data form the website, filters out metadata (i.e. HTML tags, etc.) and retrieves the program description that matches a user's query. Previous research in the English language query processing shows that removing the unnecessary words from the verbose (sentence-like) query and then using the shortened query for the search can improve its results, Huston et al. [2].

Regarding the Japanese language IEPG query, we suggest that removing some particles and endings of some pre-noun adjectivals, defining nouns and adjectives as mandatory matches while other parts of speech as optional, and reversing the query word order can improve search results.

This paper explains the above and related procedures as well as the structure for the search system that we suggest. Differently form n-gram-based statistical approaches to information retrieval, e.g., one proposed by Millar et al. [3], we emphasize taking into account the Japanese language morphology in query segmentation and search pattern matching. Efforts have been made to explain Japanese linguistics issues in such a way that the Japanese language knowledge is unnecessary to understand them.

2. System structure and functions

Figure 1 below outlines functions of the major system components. Arrows represent data flow between them.

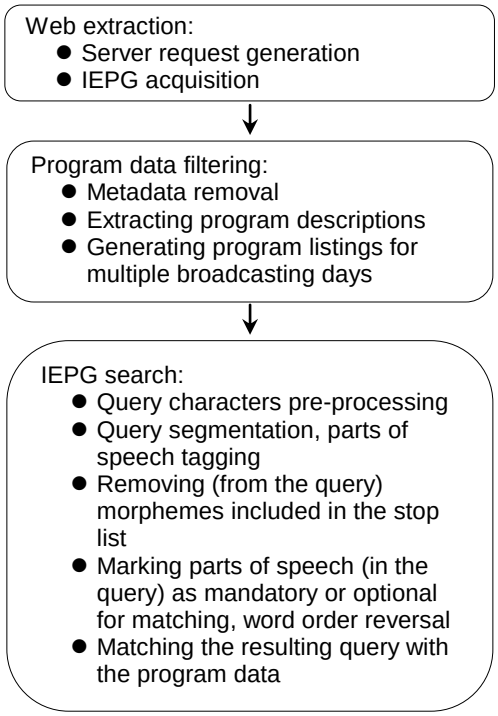

Figure 1. Input-output flowchart

As outlined above, to be extracted from the Web, IEPG data is requested from a server in the HTML format. The request is dynamically generated for the current date and seven days ahead. This eight-day data is as much as one can presently obtain from the major Japanese IEPG sites. At this time the system downloads data broadcast terrestrially in Sapporo by eight TV channels. HTML tags and other metadata are discarded using Perl programming language resources, such as the HTML::Parser module.

The natural language text obtained in this way includes program descriptions along with other irrelevant data, such as selection menus for choosing various broadcasting areas. Program descriptions are extracted and irrelevant data is ignored. To facilitate manual checking of search results, the program descriptions are then grouped, each group for one broadcasting date.

The search is carried out by means of matching the query with one program description at a time. That is, instead of matching the query with all the eight-day data at once the system takes program descriptions one by one to match them with the query. Reasons this approach has been chosen deal with the IEPG data peculiarity and the search precision. The peculiarity consists in the fact that each program description in the guide can most likely be considered semantically independent. In other words, information in one program description is most unlikely to refer to another program description. This peculiarity can affect the search precision. For instance, retrieving two program descriptions one containing "*Tokyo* food", the other "Kyoto *weather*" in response to the query "*Tokyo weather*" is imprecise as the user is looking for *weather* and not for *food*.

The proposed system does not use the n-gram model for segmentation and matching. Instead of statistical segmentation (e.g., applying Microsoft Web N-gram corpus, Wang et al. [4]), linguistic methods are used. The query is segmented by a morphological parser, JUMAN, Kurohashi and Kawahara Laboratory [5]. Before feeding the query to the parser the query character string is pre-processed to ensure it is encoded in uft-8 and all half-width characters are substituted with full-width ones (which is required for using the parser). To divide the string into segments, JUMAN analyzes such Japanese language features as parts of speech combinability and inflections. The segmented query undergoes further processing and is matched with the program description. These stages are explained in the following sections.

3. Query processing

Along with the segmented string, JUMAN output has other information such as tags telling what part of speech each segment (i.e. a character string JUMAN defines as a word) is. Words and their tags are taken out of the output and used for further processing.

3.1. Excluding unnecessary morphemes

In previous research various kinds of stop lists and their applications have been considered. For English, Hiemstra et al. [6] suggest removing words with little conceptual meaning (such as "a", "the" and "it") from the query as well as from the indexed text that is searched. Fukuta et al. [7] describe a system (for processing the Japanese language) that lists words of no potential interest to the user as stop list items.

We suggest using a stop list for multiple reasons. That is, the "unnecessary" in the section heading above refers to parts of words (and sometimes whole ones) that, for structural, semantic and pragmatic reasons, can be omitted or substituted with others with no change to the meaning. The system we propose detects and discards such words and morphemes. Table 1 below lists them and gives examples of the way they may be used in a search query. In Table 1 and onwards, examples written in Japanese are followed by a Romanized transliteration in square brackets and/or an English translation. Transliterations are italicized. In the translation, English articles are sometimes omitted to save the space and preserve the query style. In the "Entry use in a query" column and some other parts of this paper, stop list items appearing as parts of phrases are underlined when that is needed for clarity.

The stop list currently has entries of seven types. The reasons they have been included in the list are explained below. The judgment is based on the human analysis of the search results for multiple queries with the stop list items as parts of them.

It can be said that the particles は[*wa*] and が[*ga*] are interchangeable with no dramatic change to the meaning. In other words, the particles could be roughly compared to English definite and indefinite articles that convey definiteness nuances without changing the lexical meaning of what they modify. Including は[*wa*] or が[*ga*] in the query (like usage example ① above) as a mandatory match, would mean making the search engine look for something not really needed for retrieving the meaning searched for. Moreover, if the engine uses direct matching techniques, as the ones for the IEPG site examples given in Introduction most likely do, for example, 料理<u>が</u>美味しい ([*ryouri <u>ga</u> oishii*] "food is tasty") will not match 料理<u>は</u>美味しい ([*ryouri <u>wa</u> oishii*] "the food is tasty") although the two phrases mean practically the same.

The stop list item の [*no*] is often used as a possessive particle. According to a Japanese dictionary (Goo Dictionary, http://dictionary.goo.ne.jp/) it also can express the idea that "something is a location for something else"

Table 1. Stop list items

No.	Stop list entry	Entry classification	Entry use in a query
①	は[*wa*]	a particle	料理は美味しい [*ryouri wa oishii*] the food is tasty
②	が[*ga*]	a particle	温泉がある地域 [*onsen ga aru chiiki*] area with hot spring (spa)
③	の[*no*]	a particle	札幌の天気 [*sapporo no tenki*] Sapporo weather
④	な[*na*]	a pre-noun adjectival ending that can be substituted with い[*i*] with no change to the word meaning	小さな旅 [*chiisana tabi*] little trip
⑤	い[*i*]	an adjective ending that can be substituted with な[*na*] with no change to the word meaning	小さい町 [*chiisai machi*] small town
⑥	ある[*aru*]	a verb	温泉がある地域 [*onsen ga aru chiiki*] area with hot spring (spa)
⑦	いる[*iru*]	a verb	セレブがいる風景 [*serebu ga iru fuukei*] scene with celebrity

or "that something is the site of a certain action". However, another Japanese dictionary (Sanseido Web Dictionary, http://www.sanseido.net/) suggests that phrases in which の [*no*] is used in a non-possessive meaning be reworded to avoid using it. In many Japanese texts, typically technical, the particle is simply omitted. In fact, in example ③ above, の [*no*] (used in a non-possessive meaning) can also be omitted. Thus searching for it is unnecessary.

Items な[*na*] and い[*i*] can be referred to as variant endings. The same stem can have either of them with no practical change of meaning. It is common knowledge that, for instance, the prenominal adjectival 小さな [*chiisana*] can become 小さい [*chiisai*] and the meaning of both is practically the same, "small". If direct matching is used, a query with the former will not match the text with the latter and vice versa. For search precision reasons the system filter for い [*i*] endings is limited to those adjectives that have な [*na*] pre-noun adjectival counterparts. Counterparts from the indicated Sanseido Web Dictionary are used for the filter.

Items ある [*aru*] and いる [*iru*] are verbs denoting the presence of the inanimate or animate object respectively. As other verbs referring to a certain object often presuppose the presence of that object, removing ある [*aru*] and いる [*iru*] from the query can broaden the search scope. In other words, a Japanese query including ある [*aru*] or いる [*iru*] with an object, can match a text having the same object and other verbs including those presupposing ある [*aru*] or いる [*iru*] meanings. Such broadening of the scope, however, also can result in retrieving verbs with the opposite meaning, "the absence". As it is a common sense matter that a user looking for something or somebody present somewhere also might be interested in the text about the same entity absent from some place, ある [*aru*] and いる [*iru*] are included in the stop list.

3.2. Mandatory and optional matches

As mentioned in the beginning of Section 3, JUMAN divides the query into segments and tags resulting words according to the part of speech. The proposed system analyzes the tags and marks nouns and adjectives as mandatory matches and other words as optional ones. Marking is carried out by means Perl regular expressions.

3.3. Query word order reversal

The query is used for matching in its two variants. One variant with the original word order the other with the reversed, are generated and used for matching with the TV program text. For instance, the system will reverse the word order of the query "Hokkaido spas" to match both "Hokkaido spas" and "spas in Hokkaido" in the TV guide text. The system will also allow the query words to match the same words in the text with zero or more other words between them. This technique, illustrated here by an English example, is used for the Japanese language by the proposed system. It has been observed while testing the system that if one query variant does not match, the other often does.

It is clear that in the above example the query "Hokkaido spas" does not directly match "spas in

Hokkaido". However the described technique makes it possible to match "spas in Hokkaido" in TV guide text. We believe that the proposed technique can be effectively used for languages, such as English and Japanese, that permit such word order flexibility as illustrated by the above example. An example of applying the proposed word order reversal technique to a Japanese query is given in Section 4.

3.4. The resulting query

Having undergone all the above procedures the query is used for matching in two variants schematically described in Figure 2.

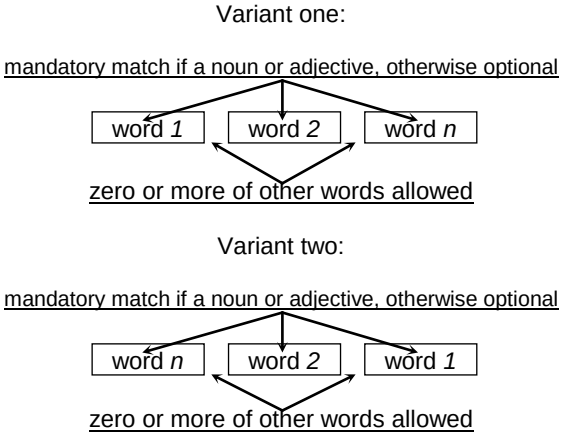

Figure 2. Two query variants

4. The proposed method performance

The purpose of this section is to illustrate how the proposed method (i.e. using the stop list, mandatory and optional matching, and query word order reversal) can be applied to the real-life TV program guide search and how that can improve search results. The guide for programs broadcast terrestrially in Sapporo form May 10 to May 17, 2013 is used for the illustration.

Applying the stop list (Table 1) allows retrieval of relevant TV program text even if it does not exactly match the query. Exact matching, on the other hand, can fail to match such text. For instance, for the query "人気がある焼き肉店 ([*ninki ga aru yakiniku ten*] a popular barbecue restaurant)" among other search results, text including "人気のある焼き肉店 ([*ninki no aru yakiniku ten*] a popular barbecue restaurant)" was also retrieved although the two Japanese phrases do not exactly match. On the other hand, matching the above query exactly would fail to produce this result.

Matching nouns and adjectives mandatorily and other words optionally, allows retrieving relevant text by filtering out unnecessary data. For the query "日本を旅する ([*nihon wo tabi suru*] (literally) do travelling in Japan)" relevant text without "do" was retrieved among other results. The part "する ([*suru*] do)" was filtered out as it does not belong to the noun or adjective part-of-speech category and as it is not necessary to match "する ([*suru*] do)" to retrieve the relevant text.

Word order reversal allows a query to match relevant text where the query words appear in the opposite order. The query "北海道ニュース ([*hokkaidou nyusu*] Hokkaido news)" produced multiple hits for a program "ニュース北海道 ([*nyusu hokkaidou*] News Hokkaido)".

As shown above the suggested method can improve search results by stop-listing unnecessary words, applying mandatory and optional matching and reversing the query word order.

5. Future work

In the future we plan to incorporate the proposed query processing techniques into a system using multiple search methods. We intend to look into applying the proposed techniques along with others allowing searching not only for query words in the opposite order but also for its words in various possible orders. The fact that the current system does not have the capability to do so could be considered its limitation.

Moreover, we intend to look into using the proposed techniques for syntactic-parsing-based as well as semantic-analysis-based information retrieval. We also intend to evaluate the performance of the above search methods and (depending on their effectiveness) to attempt at incorporating them into one IEPG data search system.

References

[1] T. Yamasaki, T. Manabe, T. Kawamura, "Implementation of TV-program Navigation System Using a Topic Extraction Agent", *Computer Software*, Japan Society for Software Science and Technology, Tokyo, 2008, pp. 41-51.

[2] S. Huston, W. B. Croft, "Evaluating verbose query processing techniques", *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010, pp. 291-298.

[3] E. Millar, D. Shen, J. Liu, C. Nicholas, "Performance and scalability of a large-scale n-gram based information retrieval system", *Journal of digital information 1.5*, 2006.

[4] K. Wang, C. Thrasher, E. Viegas, X. Li, B. J. P. Hsu, "An overview of Microsoft Web N-gram corpus and applications", *Proceedings of the NAACL HLT 2010 Demonstration Session*, Association for Computational Linguistics, 2010, pp. 45-48.

[5] Kurohashi and Kawahara Laboratory, Kyoto University

http://nlp.ist.i.kyoto-u.ac.jp/EN/

[6] D. Hiemstra, F. M. G. de Jong, “Statistical Language Models and Information Retrieval: natural language processing really meets retrieval”, *Glot international*, 5 (8), 2001, pp. 288-293.

[7] H. Fukuta, Y. Matsuo, M. Ishizuka, “Browsing support by the keyword extraction from a user's browsing history”, *IEICE Technical Report, NLC, Natural Language Understanding and Models of Communication*, 2002, 101(711), pp. 85-92.