Improving Search Query Matching for Electronic TV Program Guide Data Extraction

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

Abstract—**This paper describes a system for searching the Web-based Japanese TV program guide. The system features using morphological parsing and part-of-speech analysis to locate words with nominal and attributive semantic features in the query. Such words are matched mandatorily when searching the TV program guide text, while other words are matched optionally. Moreover, certain words and morphemes are removed from the query as they are considered to have little semantic value. The system checks every query against a stop list of such words and morphemes. Other processing methods, e.g. reversing the search phrase word order and allowing "zero or more words" between the search target words, are also utilized. The present paper uses TV guide search examples to demonstrate how the proposed method can improve Japanese TV program data search results. The paper also contains a few ideas about ways the method could be used for other languages.**

Keywords: NLP, EPG, Information Retrieval, Query Processing, Morphological Parsing, Lexical Semantics

I. INTRODUCTION

Nowadays for the convenience of the spectator, TV program guide text comes in the searchable electronic format. Such data are referred to as EPG (Electronic Program Guide). In Japan the guide can be browsed through built-in features of most television sets, as well as by using a PC interface to access the data in the WWW. For this Internet-based TV guide version, a term "iEPG" (meaning "Internet Electronic Program Guide") has been coined. 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.

Data in the guide are normally grouped, each group describing a single program. The description natural language text includes such information as the name for the TV channel broadcasting the program, the broadcasting date, time and the program contents from a word or two to about a paragraph in length. More details on the iEPG format, including examples, are given in [1].

Using search systems available on major Japanese iEPG websites, such as ones for which URLs are given above, shows 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).

The proposed system is designed to improve the above matching technique by means of analyzing the Japanese morphology as well as such semantic primes as nominal and attributive features. The present paper explains this and related procedures, it also describes the search system that we suggest. Differently form n-gram-based statistical approaches to information retrieval, e.g. one proposed by [2], we emphasize taking into account the Japanese language structure 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.

II. NOMINAL AND ATTRIBUTIVE SEMANTICS ANALYSIS OF THE QUERY

This section explains the essence of the analysis and gives reasons the procedure is implemented. The system input-output flow is shown in the next section.

Existing research demonstrates that morphological features of a word to some extent determine its semantic features. That is, if a word has certain morphemes, it belongs to a certain part-of-speech and basic meaning category. For instance, the suffix *–ist* of the word *guitarist* makes it a "denominal person noun" [3]. Another research states that a word of a language has the "semantic core" also referred to as the "semantic prime" [4]. For example, semantic primes for nouns are classified as "substantives" and those for adjectives as "determiners" (ibid.). Moreover, according to [5] semantic primes are "universal", i.e. present in multiple languages. The method we propose focuses on two types of semantic primes, i.e. *the object* and *the property-of-an-object* meaning features. The former is characteristic of nouns, the latter of adjectives. By the semantic prime for the noun we mean the fact that nouns signify objects, and by the semantic prime for the adjective that adjectives signify properties of objects.

The proposed system bases its analysis of sematic primes in the search phrase on the morphological and part-of-speech analysis. The system attempts to make a judgment on the following three aspects:

978-0-7695-5119-7/19 $26.00 © 5119 IEEE
DOI 10.1109/ICSC.2013.34

1) whether the user is searching for nouns, i.e. words meaning objects;

2) whether the user is searching for adjectives, i.e. properties of objects;

3) whether the user is searching for words different from the above.

When searching TV program data the system uses words it marked as objects and their properties as obligatory matches, while other words are used as optional matches for the following reasons. An existing research shows that nouns “constitute over 70% of query terms”[1] [6]. Moreover, nouns used together with adjectives are “common need information clusters” in English queries for multiple search engines [7]. Another research shows that nouns, such as proper nouns, are numerous in Japanese search queries [8].

To sum it up, the proposed system looks into the universally present core meaning of the query to find object and object property features. If found, words with these meaning features are given priority when matching TV program data because nouns form the majority of query terms and noun-adjective clusters are common in queries for multiple search engines.

We believe that the proposed technique, i.e. matching words meaning objects and their properties obligatorily and other words optionally, could be used not only for Japanese but also for other languages. The fact that object and property semantic features are “universal”, as explained above, justifies this belief.

III. System Input-output Flow

Figure Ⅰ outlines functions of the major system components. Arrows represent data flow between them. As shown in Figure Ⅰ, to be extracted from the Web, iEPG data are requested from a server in the HTML format. The request is dynamically generated for the current date and seven days ahead. At this time the system downloads data broadcast terrestrially in Sapporo by eight TV channels. HTML tags and other metadata are discarded. The natural language text obtained in this way includes program descriptions along with irrelevant data, such as selection menus for choosing various broadcasting areas. TV program descriptions are extracted and irrelevant data are ignored. To facilitate manual checking of search results, the program descriptions are clustered, each cluster has programs for one broadcasting date.

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 [9]), linguistic methods are used. We have preferred such methods because n-gramming a character string does not take into account morphology and semantics, however it is our purpose to take them into account. The segmentation technique described in [9] involves matching a character string with strings repeatedly found in the corpus. Shorter strings that often match are considered to be separate words (ibid.).

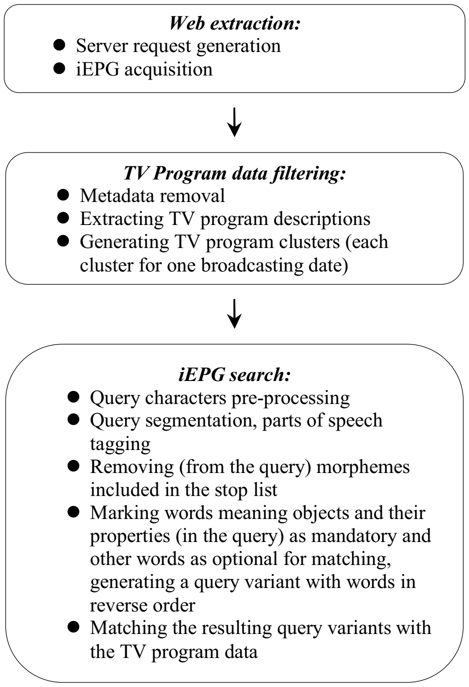
Web extraction:
● Server request generation
● iEPG acquisition
TV Program data filtering:
● Metadata removal
● Extracting TV program descriptions
● Generating TV program clusters (each cluster for one broadcasting date)
iEPG search:
● Query characters pre-processing
● Query segmentation, parts of speech tagging
● Removing (from the query) morphemes included in the stop list
● Marking words meaning objects and their properties (in the query) as mandatory and other words as optional for matching, generating a query variant with words in reverse order
● Matching the resulting query variants with the TV program data

Figure Ⅰ. Input-output flow.

Differently form that technique the proposed system segments the query by means of a morphological parser JUMAN[2], developed by Kurohashi and Kawahara Laboratory [10]. 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. 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 considers a word) is. Words and their tags are taken out of the output and used for further query processing.

Stop-listed words and morphemes are removed from the query and the remaining words are marked as mandatory or optional matches as explained in Section Ⅱ. The query is then 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 allows 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

[1] In the above research this refers to search queries in the English language only.

[2] See http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN for a detailed parser description.

proposed system. We consider that the technique can be used for both English and Japanese as well as for other languages with a flexible word order.

The search is carried out by means of matching the query with one TV 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"[3].

IV. STOP-LISTING THE QUERY

In existing research various kinds of stop lists and their applications have been considered. For English, [11] suggests 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. Another research [12] describes 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, some parts of a word (and sometimes the whole word) 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 I lists them and gives examples of the way they may be used in a search query. In Table I and onwards, examples written in Japanese are followed by a Romanized transliteration in square brackets and/or an English translation. 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 written in Japanese and their transliterations are underlined for clarity, all transliterations are italicized.

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

TABLE I. 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

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 the 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 in the example for item ① in Table I) as a mandatory match, would mean making the search system look for something not really needed for retrieving the meaning searched for. Moreover, if the system uses direct matching techniques, as the ones for the iEPG site examples given in Section I most likely do, for instance, "料理が美味しい ([*ryouri ga oishii*] food is tasty)" will not match "料理は美味しい ([*ryouri wa 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" or "that something is the site of a certain action". 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.

[3] For more clarity let us consider two semantically independent TV program descriptions, one containing the phrase "Tokyo food" and the other the phrase "Kyoto weather". A query "Tokyo weather" matches the word "Tokyo" in the first program description and the word "weather" in the second one, so both descriptions could be retrieved as search results. However this is definitely imprecise because the two program descriptions are semantically independent and the user is looking for "weather" and not for "food".

Items “な [*na*]” and “い [*i*]” can be considered variant endings. The same stem can have either of them with no practical change to the 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 an inanimate or animate object respectively. As other verbs referring to a certain object normally presuppose the presence of that object, removing “ある [*aru*]” and “いる [*iru*]” from the query can broaden the search scope. Thus if “ある [*aru*]” and “いる [*iru*]” are removed from a query including these verbs with their subjects, the query will match a text having the same subjects 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.

V. The Proposed Method Performance

The purpose of this section is to illustrate how the proposed method (i.e. using the stop list, obligatory or 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. For space considerations only several examples are given.

Applying the stop list (shown in Section Ⅳ) 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 words with object and property semantic features obligatorily and other words optionally, allows retrieving relevant text by filtering out unnecessary semantic features. 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 have the object or property semantic feature and as it is not necessary for retrieving 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, filtering out unnecessary semantic features and reversing the query word order.

VI. Future Work

In the future we plan to incorporate the suggested query processing techniques into a system using multiple search methods. Namely, we intend to look into using the proposed techniques along with information retrieval based upon syntactic parsing and a more thorough word-meaning analysis.

References

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

[2] 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, pp. 1-25, 2006.

[3] R. Lieber, Morphology and Lexical Semantics, vol. 104. Cambridge University Press, 2004.

[4] C. Goddard, “The search for the shared semantic core of all languages”, in Cliff Goddard and Anna Wierzbicka (eds). Meaning and Universal Grammar - Theory and Empirical Findings, vol. 1. Amsterdam: John Benjamins, pp. 5-40, 2002.

[5] A. Wierzbicka, Semantics: Primes and Universals, Oxford University Press, 1996.

[6] C. Barr, R. Jones, M. Regelson, “The linguistic structure of English web-search queries,” Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1021-1030, 2008.

[7] R. Baeza-Yates, C. Hurtado, M. Mendoza, G. Dupret, "Modeling user search behavior", Proceedings of the Third Latin American Web Congress (LA-WEB '05), 2005.

[8] I. Arita, H. Kikuchi, K. Shirai, “Word Clustering Using Concurrent Search Queries”, IEICE technical report, NLC, Language Understanding and Models of Communication, 107(158), pp. 115-120, 2007.

[9] 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, pp. 45-48, 2010.

[10] Kurohashi and Kawahara Laboratory, Kyoto University http://nlp.ist.i.kyoto-u.ac.jp/EN/

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

[12] 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, 101(711), pp. 85-92, 2002.