
Doctoral Dissertation

# Study on a Cross-language Search Method for the Japanese TV Guide

（日本語 TV ガイドにおける異言語間検索手法に関する研究）

Denis Kiselev

Graduate School of Information Science and Technology

# CONTENTS

# THESIS ABSTRACT

The present thesis proposes a new method for searching the Japanese TV guide text. The thesis is also to verify the effectiveness of the proposed method. The proposed search techniques have been developed to improve existing ones used for the mentioned kind of search. Below is a summary of the proposed method and work carried out to verify its effectiveness.

An analysis of the Japanese TV guide performed by the thesis author has demonstrated approximately 64% of all the Roman character strings in the guide text to be English words. However, e.g. applying Google Japan customized search to Japanese TV guide website text has shown that the above Google search does not use such words as potential targets. The same is true for other search systems, such as one on a major Japanese TV guide site "Yahoo テレビ　Japan ([*yahoo terebi japan*], Yahoo TV Japan)". The failure to target English words naturally results in the failure to retrieve text that may include relevant information in English. To solve this problem the proposed method utilizes English paraphrases for Japanese search words along with the words themselves. The English WordNet version is used as the source for the paraphrases.

Japanese TV guide search systems, such as that on the Yahoo website mentioned above, seem not to expand the search query by adding other semantically related words to query terms. Similarly, the mentioned Google customized search showed no "hits" for synonyms and other words semantically related to query terms. Using no semantically related words for query expansion makes it impossible to retrieve text that may not have the same words as the query ones but may have the same meaning. The proposed method solves this problem by utilizing the query along with Japanese paraphrases of its words. The paraphrases include synonyms form the Japanese WordNet as well as Kana and Romanized transliterations.

Japanese TV guide search systems, such as one on the website "Yahoo テレビ Japan ([*yahoo terebi japan*], Yahoo TV Japan)", seem to process the search phrase as a character string without segmenting it into words. This may cause the failure to retrieve relevant search results, as query words may appear in the guide text in various orders, with or without other words in between. Search experiments performed on the Yahoo TV Japan iEPG text, demonstrated that for approximately 81% of all the search results that included query words, the order of those words differed from the original search

query word order and/or other words appeared between the original query ones. Thus it can be inferred that, for instance, the “Yahoo TV Japan” search method is most unlikely to be effective in retrieving the above search results. The proposed method solves this problem by applying the finite state automaton (FSA) concept, often used for such tasks as data mining, to the Japanese TV guide search. The implemented FSA can match all query word occurrences and all possible word order variations. To increase the FSA speed, parallel processing is utilized. As a result, an approximately 95% percent decrease in the search time has been demonstrated.

To group search results by relevance, the proposed method not only takes into account word occurrence counts but also analyzes word semantic features. This can be considered an improvement in search result grouping effectiveness. A multi-parameter evaluation has demonstrated an approximately 66% performance improvement for the proposed method as compared with a baseline one. The method can be considered novel due to the fact that no research and implementations combining the described query expansion techniques and semantic feature analysis seem to exist.

The thesis author has analyzed the Japanese online TV guide for the presence and nature of Roman character strings. Results have demonstrated that Roman character strings constitute approximately 10% of all character strings. Moreover, approximately 64% of the Roman character string tokens have been found recognizable as English words in the English WordNet. This clearly shows that along with Japanese words a considerable number of English ones appear in the TV guide text. It is also clear that English words have proven to be more numerous than Romanized variants of Japanese words. These facts demonstrate the need to apply the proposed cross-language search method to the Japanese TV guide, in other words the need for a method making not only Japanese but also English words potential search targets.

Some details of the proposed search method are as follows. Along with the search query itself, the method uses both Japanese and English synonyms from WordNet for the search. The method also uses Hiragana, Katakana and Romanized versions for the original Japanese query. As a result, original Japanese query words along with their English translations, Japanese synonyms, Kana and Romanized versions are matched against the TV guide text. An FSA has been implemented to match all the above search words in all possible orders and combinations. This process can be slow taking into account the matching task and the size of the data. Due to implementing parallel processing of TV guide data portions, the proposed method has demonstrated an approximately 95% decrease in the search time. Thus, the time span between the query

input and the output of search results is very unlikely to cause any bother for the human user.

Before the synonyms and transliterations described above are generated, the original query undergoes stop-listing and morphological analysis. The query is checked against a stop list of words, such as particles, that have no sematic value to them. Such words are not used for the search. Some word endings are also removed.

A morphological analysis has been implemented to segment the search query string into separate words and to determine what semantic features the words possess. The proposed method detects two types of semantic features, i.e. “an object” and “a property of an object”, and later uses the features to sort search results by relevance as explained later. The semantic feature detection is based on the part-of-speech analysis results output by the morphological parser MeCab. Nouns are marked as possessing the “object” sematic feature and adjectives as possessing the “property-of-an-object” one.

Search results are output in groups. Groups with more relevant results are output above those with less relevant ones. Results that include combinations of the above two semantic features are considered more relevant than other search results.

The proposed method has been evaluated using search queries obtained by means of a questionnaire survey involving a group of individuals. The method has shown better performance compared with five baselines. The performance improvement can be illustrated by an approximately 66% average F-measure increase from 0.432 for the baseline implementation that did not use query expansion techniques to 0.718 for that of the proposed method. Thus, it can be concluded that the proposed method for searching the Japanese TV guide has proven effective.

It can be said that the implementation of the proposed cross-language search techniques utilizing paraphrases and transliterations for query expansion, could lead to better performance for Japanese TV guide search systems that are commonly used. This speaks in favor of possible practical applications of the proposed method, for instance on TV guide websites.

# JAPANESE ABSTRACT (日本語の要旨)

本研究の目的は，日本語 TV ガイドにおける検索の既存手法を以下に述べるように改善するための新たな手法を提案し，その有効性を確認することである．

日本語 TV ガイドを分析したところ，TV ガイドに含まれているローマ字文字列の内，約 64%は英単語であり，相当数の英単語が出現することが明らかとなった．一方，Google 日本のカスタマイズサーチを日本語 TV ガイドのウェブサイトにおいて使用してみると，英単語が検索対象にされていないと考えられる．なお,「Yahoo テレビ　Japan」というウェブサイト上の検索システム等を利用した検索でも，英単語が対象外であることがわかる．このため，当然，検索結果として必要となる英語の情報が検出できていない．本問題の解決に向けて，提案手法では本来の検索語とともに，　検索語に対して WordNet から抽出した英語パラフレーズを，検索に利用している．

上述した Google 日本や,「Yahoo テレビ　Japan」の検索システムでは，検索語と，それに意味的に関連がある他の語を検索において使用するクエリ拡張が行われていないと考えられる．実際に Google 日本で,そのカスタマイズサーチ機能を日本語 TV ガイドのウェブサイトに対して使用してみると,同義語等が検出されていないので,クエリ拡張が行われていないと推測される．このようにクエリ拡張を行わなければ，当然，検索語句とは同じ文字ではないが同じ意味を持つ文章の抽出ができない．この問題の解決策として，提案手法では元々の検索語と共にその日本語のパラフレーズが検索に利用されている．パラフレーズの中には日本語 WordNet の同義語，およびカナ，ローマ字で表記された検索語が含まれている．

「Yahoo テレビ　Japan」上の検索システムのような既存のものは，検索語句を単に文字列として扱い，単語単位で分離しないと考えられる．また，文章において検索語が様々な語順や,間に本来の検索語と違う単語が存在することによって,望ましい検索結果が得られない場合がある．例えば,「Yahoo テレビ　Japan」に対して検索実験を行ったところ，検索結果の約 81%が,抽出された検索語の語順が本来の検索語句と異なったり,本来の検索語の間に違う単語が存在していることがわかった．つまり,「Yahoo テレビ　Japan」等の検索方法により，約 81%の検索結果が得られないことが推測される．この解決策として，本提案手法では有限状態オートマトン (a finite-state automaton, FSA) を利用している．有限状態オートマトンはデータマイニング等で多く利用されているが，本提案手法では TV ガイド検索の目的で実装を行った．FSA の実装によって，文章に出現しているすべての検索語がすべての可能な語順で把握できるようになった．さらに長い文章を処理する際の高速化手法として，並列処理を利用している．このことによって検索時間が約 95%減少した．

本提案手法では，検索結果を適切か否か分類する際，語の出現数だけではなく意味素性分析

も行う．この点が検索結果における分類の効率化に繋がると考えられる．なお，提案手法による約 66%の検索効率化がマルチパラメータ評価によって示されている．本提案手法のように，上述のクエリ拡張と意味素性分析を併せ持った手法は，他に類似している研究が見当たらず，このことが本研究の独自性である．

日本語 TV ガイドに含まれる文字列を分析した結果，ローマ字文字列は約 10%であることがわかった．さらにこの文字列において，約 64%のトークンが英語 WordNet に記載されている英単語と一致した．つまり，日本語 TV ガイドに含まれる英単語は，ローマ字表記の日本語より多く，相当数存在するという結論が導き出される．この点は，TV ガイドの日本語だけではなく，英語も対象にできる異言語間検索の提案手法を開発する必要性を示している．

提案手法の概要を以下に述べる．入力された検索語に対して WordNet から日本語と英語の同義語が抽出され，検索語とともに検索に用いられる．また，検索語の平仮名，カタカナ，ローマ字表記も生成されて検索に利用される．結果的に元々のクエリの他に，その英訳，日本語同義語，カナとローマ字表記を用いて TV ガイドの文章が検索される．検索の際に，文章に現れているこれらの語における，全ての可能な語順や組み合わせを把握できるように FSA が実装されている．しかし，データ量が多いためにこの検索過程は遅くなる場合もある．そこで，TV ガイドを分離して，それぞれの部分を並列的に処理するという手法を取ることにより，検索時間は約 95%減少した．このような方法により，ユーザーの待ち時間の負担が軽くなった．

前述した同義語やカナが生成される前に，元々の検索語に対してストップリストを用いたフィルタリングと形態素解析が行われる．フィルタリングの際に検索語が，助詞などのそれ自身では意味を持たない語を含むストップリストと照合され，一致したものは検索で使用されない．一部の語尾もこの時に排除される．

提案手法で使用されている形態素解析の役割は，文字列を語単位で分離すること及び，語において意味素性分析を行うということである．意味素性分析の際に，「オブジェクト」（物体など）と「オブジェクトのプロパティ」（属性）という二種類の意味素性索引をそれぞれ語に付与し，その索引の検索結果が適切か否かの分類を行うことに利用される．索引の付与は，MeCab という形態素解析ツールの品詞情報出力をもとにしている．つまり，名詞には「オブジェクト」の索引が付与されて，形容詞は「オブジェクトのプロパティ」の索引が付与される．

検索結果はグループ化されてから出力される．ディスプレイ上では，より適切な結果のグループが適切性の低いグループよりトップに近くなる．上記の索引が付与されている語の組み合わせを含む結果は，他の結果より適切だと判定される．

提案手法の評価には，アンケート調査によって得られた検索語句が用いられている．この評価において提案手法は，5 つのベースライン手法より良い性能を示した．F 値が，提案手法では 0.718，検索語句拡張を行わないベースラインでは 0.432 となり，約 66 ポイントの検索効率の向

上を示している．このことは日本語 TV ガイド検索における提案手法の有効性を示していると考えられる．

さらに，前述した既存の TV ガイド検索システムにおいても，提案手法と同じような英訳，同義語，カナ表記のクエリ拡張を用いた異言語間検索を使用することにより，性能が向上すると考えられる．つまり，TV ガイドのウェブサイト等でも提案手法の実用性があるものと考えられる．

# LIST OF TABLES

# LIST OF FIGURES

# 1. THE PROPOSED METHOD BACKGROUND

This section first defines the concept of "cross-language search" the present research deals with. Next, the section makes clear the fact that the proposed method can be potentially applied to any text in which both Japanese and English vocabulary is present. It is explained why this research examines the method application to the Japanese TV guide in particular and why the human knowledge base rather than other natural language processing (NLP) resources and techniques is used.

The section also reviews characteristics of the digital TV guide data as used in Japan, providing examples of the user-end data format and ways the TV guide can be accessed in this country. Later in the section the data utilized by the proposed method is described. The section proceeds with a discussion of existing research with regard to cross-language information extraction, query expansion and other techniques the proposed method uses. An analysis demonstrating such facts as a considerable presence of English words spelled in Roman letters in the Japanese TV guide, is described later. At the end of the section reasons for the method implementation, such as the mentioned presence of English words, are given. It is also explained why the proposed search method can be considered effective and novel.

The concept of "cross-language search" the present research deals with implies searching the Japanese text in which English vocabulary is also present. The "cross-language" part of the concept means that for the search query input in Japanese semantically related words in both Japanese and English are generated to expand the query. "Cross-language" also means that the generated words are used, together with the original query, to target vocabulary in both of the two languages present in the text. The method has been developed primarily for processing the query in Japanese because knowledge of this language is presupposed for searching the Japanese text. In other words, the user should understand the text he/she is searching.

Using the described cross-language search concept makes it possible to apply the proposed method to any text in which both Japanese and English vocabulary is present. It should be noted that applying the method to the Japanese TV guide is by no means the only original feature of the present research. Advantages of this research over existing research in cross-language information extraction, query expansion that uses bilingual vocabulary and knowledge-bases are explained in 1.2. Advantages of the proposed

method implementation, such as those over existing search systems are given in 1.4.

The main reason the present research applies the proposed method to the Japanese TV guide it the fact that by analyzing this guide the author has found it to include a considerable number of English words. See 1.3 for more details of the analysis. Another reason is that the TV guide is publicly available online and broadcast to be viewable using home TV sets.

The present research deals with using the human knowledge base for query expansion, the way described in 2.7, for the following reason. It appears hardly possibly to achieve the same level of word semantics coverage if, say, a database generated by means of a machine-learning algorithm based upon, say, n-grams[1] is utilized for query expansion. The present research, instead, concentrates upon utilizing modern NLP methods that cover human language semantics.

## 1.1. Characteristics of the Japanese EPG (Electronic Program Guide)

In the present thesis the phrase "TV guide" refers to textual data that lists descriptions of TV programs and gives the broadcasting schedule for them. 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). The EPG history appears to be not so short, an early EPG version being patented[2] in the USA in 1994. The patent describes a system for rendering TV program information to a television receiver screen.

In contemporary Japan the guide can be browsed through built-in features of most television sets, as well as by using a computer interface to access the data in the Internet. For this Internet-based TV guide version, the term "iEPG" (meaning "Internet Electronic Program Guide") has been coined. Multiple websites[3] make TV program data publicly available in Japan helping viewers choose from a variety of programs.

---

[1] "N-grams" refers to single words or groups of them that are found, usually a by machine, to be used a certain number of times in a certain text.
[2] https://www.google.com/patents/EP0775417B1?cl=en
[3] An example of such sites is located at http://tv.yahoo.co.jp/.

Figure 1 gives examples of the user-end TV guide format. In Figure 1 an example of the “インターネットＴＶガイド ([*intaanetto terebi gaido*], Internet TV Guide)”[4] Japanese iEPG website[5] data format is placed above an example for the “Yahoo テレビ　Japan ([*yahoo terebi japan*], Yahoo TV Japan)” site data format.


Figure 1. iEPG Format Examples

The iEPG shown in Figure 1 has the following format peculiarities.

---

4 Here and throughout the thesis text, examples in Japanese are followed by the romanization and/or translation. Romanizations are italicized and put in square brackets.

5 URL: http://www.tvguide.or.jp. As of Oct. 1, 2014 this site has discontinued displaying iEPG data, according to the information posted on the site in Oct. 2014.

1) Broadcasting areas, e.g. “北海道・札幌 ([*hokkaidou sapporo*] Hokkaido Sapporo)” and broadcasting types, e.g. “地上波デジタル番組表 ([*chijouha dejitaru bangumi hyou*] program listings for terrestrial digital broadcasting)” shown in Figure 1 are choosable by the user.

2) Data in the online TV guide are grouped, each group describing a single program, e.g. “にっぽんの芸能 ([*nippon no geinou*] Japanese performing art)” in Figure 1. The description natural language text includes such information as the broadcasting date and time, and the program content from a word or two to about a paragraph in length.

3) A guide browsing function is provided for navigation by the broadcasting date and time, or by the broadcasting channel, see the slides at the top right corner of Figure 1. Japanese iEPG sites, such one shown in the figure, can usually display TV program listing for the present date and for seven days ahead, i.e. data for a total of 192 hours.

4) A search function for the TV program content natural language text is provided. See “検索する[*kensaku suru*] search” in Figure 1.

Japanese iEPG websites can normally display two types of the guide data. The first type, such as one illustrated by Figure 1, is concise, the second is more detailed. The search method proposed in the present thesis uses the more detailed data. The format and other peculiarities of such data are given below.

Concise TV program descriptions have clickable areas available to the human user (and programming scripts to be used by the machine) for loading detailed data and displaying these data to the user. The clickable area is usually provided for each TV program within the limits of the program title text. Figure 2 gives an example for such clickable area. The example source is “Yahoo Japan テレビ ([*yahoo japan terebi*] Yahoo Japan TV)” , another major Japanese iEPG website[6].

[6] http://tv.yahoo.co.jp/

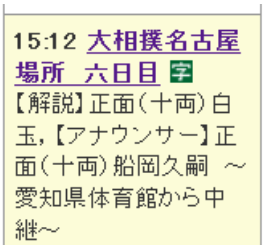


Figure 2. Clickable Area for Loading Detailed iEPG Data

By clicking the program title “大相撲名古屋場所　六日目 ([*oozumou nagoya basho muikame*], Nagoya Grand Sumo Tournament Day Six)” shown in Figure 2, the iEPG user can view more detailed data for this program. Apart from the program content displayed under the title in the same figure, the detailed data includes additional information. A snapshot of that information for the above Sumo tournament, taken at the same website is given in Figure 3. The additional information format has the following peculiarities.

1) Casts or lists of participants may be included. Functions of participants, such as the “解説 ([*kaisetsu*], commentator)” or “アナウンサー ([*anaunsaa*], announcer)” in Figure 3, may be given.

2) Program genre information may be provided. See the “スプーツ – 相撲・格闘技 ([*supootsu sumou kakutougi*], sports, Sumo, martial arts)” in Figure 3.

3) Information regarding video displaying and sound, such as the “字幕 ([*jimaku*], subtitles)” and “ステレオ ([*sutereo*], stereo)” in Figure 3, may be included.

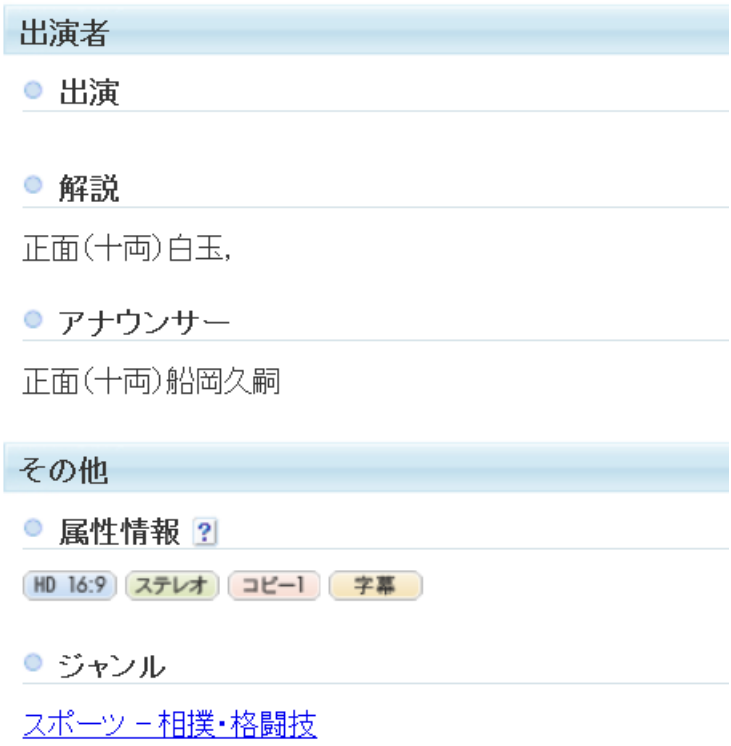


Figure 3. Additional Data form the Detailed iEPG

More information regarding the Japanese iEPG with data examples can be found in Yamasaki et al (2008). Tanaka et al (2000) discuss TV Guide interfaces in detail.

In summary, the present thesis deals with iEPG (meaning “Internet Electronic Program Guide”) as it is used in Japan. The guide lists TV program descriptions and provides broadcasting schedules for the programs. The detailed guide data utilized by the proposed search method includes program content descriptions, i.e. natural language text telling what the program is about, as well as such additional information as lists of program participants, descriptions of the program genre and details on video image displaying and sound.

## 1.2. Related Work in the Cross-language Search, Query Expansion and Other Fields

Following is a discussion of some previous research that can be considered background for the proposed search method. The research work descriptions are given together with explanations of ways the proposed method builds upon the previous research.

A considerable amount of work has been dedicated to analyzing various data sources, such as the WWW, in order to see how frequent certain cross-language counterparts are and how those words can be used for information extraction. Qu, Grefenstette & Evans (2003) propose a method for generating "English-Japanese transliteration pairs" and use them for "Japanese to English retrieval experiments" on bilingual texts. The phrase "English-Japanese transliteration pairs" refers in this work to English words, such as technical terms and proper names, with their *Katakana* counterparts. The research features utilizing bilingual corpora to generate the transliteration pairs. The work also demonstrates improvements in search results due to applying the English-Japanese bilingual lexicon to information extraction. Later, Grefenstette & Evans (2004) use the Web on a large scale to create pairs of English names and their Japanese *Katakana* counterparts for information extraction and other tasks.

As seen from the above, English-Japanese transliteration pairs could be successfully used for the search query expansion, provided that the search target includes both Japanese and English data. However, in the above research the bilingual counterparts appear to be limited to transliterations including English names with their *Katakana* variants. The proposed method does not have that limitation. In fact, it builds upon the described cross-language search techniques by utilizing *Katakana*, *Hiragana* and *Romaji* transliterations for all query nouns, as well as *Katakana* and *Hiragana* transliterations for all query adjectives.

Moreover, the proposed method does not limit the range of search targets to transliterations only. Along with transliteration it uses word paraphrasing techniques

similar to those described by Nanba (2007) whose work deals with applying WordNet[7] to query paraphrasing and expansion. The author describes a system searching not only for query terms as they are, but also for synonyms and hyponyms found in WordNet for the query terms. The author concludes that utilizing synonyms rather than hyponyms to expand the query can considerably improve the performance for searching Japanese and English texts respectively. In other words, the author states that using English synonyms can improve the efficiency of searching English (unilingual) texts and using Japanese synonyms can serve the same purpose for Japanese (unilingual) ones.

The problem is that the target iEPG text of the proposed method is not unilingual. It includes both Japanese and English vocabulary, as demonstrated by the results of the iEPG analysis described in 1. 3. Thus, the search targeting the Japanese vocabulary only, will definitely overlook potentially relevant information in English. The proposed method solves this problem by applying cross-language search to the Japanese TV guide, i.e. by applying the search that uses Japanese query words with English paraphrases to target the data that includes both Japanese and English vocabulary.

Table 1 summarizes the above discussion by listing search techniques used by the mentioned previous research and search techniques used by the proposed method. The "○" indicates that the respective technique is used, "×" that it is not used.

As seen from the table, the proposed method builds upon the previous research by introducing the use of English synonyms as well as *Hiragana* and Romanized transliterations for searching the Japanese iEPG data. Along with the original query words, the proposed method uses their Japanese and English synonyms as well as *Katakana*, *Hiragana* and Romanized transliterations to search the iEPG.

To implement the proposed method it was necessary to integrate the English and Japanese WordNet versions in order to generate bilingual paraphrases for Japanese query words. Robkop et al (2010) propose tools for integrating WordNet versions for Asian languages, including Japanese. The authors' approach maps the Asian WordNet entries onto English ones enabling a user to retrieve synonyms and other related words in multiple languages. For size and ease-of-use considerations, a different WordNet version combining only Japanese and English was chosen to implement the proposed method. The version is called "Japanese WordNnet", its construction and initial sources

[7] WordNet is a database for meaningfully related words, e.g., synonyms. The current English WordNet version is described at http://wordnet.princeton.edu. The current Japanese version is described at http://nlpwww.nict.go.jp/wn-ja/.

*Table 1. Search Techniques Used by the Previous Research and the Proposed Method*

| **Search Technique Description** | **Use by the Previous Research** | **Use by the Proposed Method** |
|---|---|---|
| Bilingual Data Search Utilizing English-*Katakana* Transliterations | ○ | ○ |
| Bilingual Data Search Utilizing Japanese and English Synonyms (Including Transliterations and Translations) | × | ○ |
| Japanese Data Search with Query Expansion Utilizing *Hiragana* and Romanized Transliterations of Japanese Words | × | ○ |
| Japanese Data Search Utilizing Japanese Synonyms | ○ | ○ |

of data are described by Isahara et al (2008). Database improvements and interfaces are described by Bond et al (2009).

The proposed search method has been developed to use the mentioned bilingual WordNet to search the Japanese iEPG. This can be looked upon as another novel feature due to the fact that no other research on this kind of WordNet application has been found by the author. Roughly speaking, the proposed method performs the following three related functions one of which involves using the WordNet.

1) Analyzing query words.

2) Utilizing the analysis results to extract synonyms for those words from WordNet.

3) Utilizing the synonyms and other words to expand the query.

Next, the expanded query is used for the search. See 2 for a detailed search process description.

It is interesting enough that query translation, as opposed to monolingual search, was found "relatively [computationally] inexpensive to implement" by Fujii & Ishikawa (2001) as early as about thirteen years ago. The authors of this work describe using an early version of WordNet to index search-target documents. According to the authors, more information regarding this WordNet version can be found in Fellbaum (1998).

Differently form the above approach the proposed method has more computational costs involved due to the variety of performed operations and the size of the processed data. This problem has been solved by implementing a finite state automaton (FSA) with parallel processing described in detail in 2. 8 and 2. 9. The proposed method does not only use "query translation", i.e. English synonyms for Japanese search words, but also expands the query in multiple other ways described above. The expanded version of some queries may total about twenty words. These words are matched against iEPG data sets for eight broadcasting days, each set including text for over 2,000 TV programs, and each program description including a paragraph of text on the average. This matching process has been found to be computationally expensive. However, parallel processing of iEPG by portion has made possible an approximately 95% reduction of the search time. As a result, the input-output interval, averaging to 0.7 of a second including the server-client transmission time, is most unlikely to cause any bother for the human user. The described implementation of the FSA and parallel processing to search Japanese iEPG data can be considered one more novel feature of the proposed method as no research on this kind of implementation has been found by the author.

In summary, compared with the discussed research, the proposed method introduces the following novel features. First, to target the Japanese iEPG bilingual (Japanese and English) vocabulary, cross-language search has been introduced. Second, the method utilizes a bilingual WordNet version to expand the query for searching the Japanese iEPG. Third, the method applies an FSA with parallel processing capability to searching the Japanese iEPG text.

## 1.3. An Analysis of the Japanese iEPG for the Presence and Nature of Roman Character Strings

This subsection first outlines successive stages of the analysis and explains the reason the analysis was carried out. It proceeds by giving descriptions of the stages.

The stages can be outlined as follows.

1) Segmenting the TV guide text into Japanese words and Roman letter strings. Counting respective numbers of the words and the strings.

2) Tokenizing the Roman letter strings.

3) Checking the Roman strings against the English WordNet to see if they match its entries and so can be considered English words. Counting those words.

Cultural exchange between Japan and English-speaking countries becoming more intense in the mid-nineteenth century resulted in thousands of English word borrowings, "and the 'boom' in adopting foreign culture which became renewed in that period has continued to this day" (Kay, 1995). Moreover, according to Stanlaw (2004) "[the Japanese] population not only tolerates the presence of a foreign language such as English, but indeed, seems to encourage it". The thesis author believes that these tendencies of the Japanese society may have resulted in the fact that the he has observed multiple occurrences of English words in the Japanese iEPG text.

The analysis was motivated by the following assumption. If the iEPG text contains Roman letter strings and those strings are English words, then such words may be potential hits for cross-language search. In other words, if the text contains both Japanese and English words, then searching not only for the former but also for the latter can yield more search results.

For the analysis, a text file containing Japanese TV guide for twenty-four days was used. At present, however, eight-day data is as much as one can obtain from the major Japanese iEPG websites, such as those mentioned in 1. Three eight-day day data sets were joined into the above file without overlapping.

The Roman letters to be found in the strings were limited to the 26 letters of the modern Latin alphabet used for writing in English.

As outlined above, the analysis was carried out in several stages. Frist, the TV guide was segmented into Japanese words and Roman letter strings. The words and the strings were counted.

*Table2. Japanese Words and Roman Letter Strings*

| Total Number of Separate Strings | Number of Japanese Words | Number of Roman Letter Strings |
|---|---|---|
| 117,132 | 133,360 | 11,836 |

Table 2 demonstrates the total number of separate character strings in the analyzed file, and quantities of Japanese words and Roman letter strings respectively. As it is seen from Table 2, the Roman letter strings are not so few, i.e. approximately 10%.

To single out Japanese words, the part-of-speech and morphological analyzer MeCab[8] was used. The tool was chosen as the fastest, according to Ptaszynski et al (2012), and easy-to-use for the above task.

Second, the Roman letter strings were tokenized by assigning one token to a group of repeated strings found in the text. Each token exactly matched the strings in the group it was assigned to. The tokens were utilized to reduce the computation cost. Algorithms matching the same string (hundreds of times, in some cases) were not used. Instead, e.g. the string "NHK"[9] appearing in the analyzed file 3,064 times was matched as one token.

Third, the tokens were matched with the English WordNet and the number of matches was counted. It was found that about 64% of tokens consisting of Roman letters matched English words (and could be considered English words). The matches excluded those of such words as articles and prepositions that, according to WordNet documentation[10], are not listed in WordNet. The remaining tokens could be classified as likely Romanizations of Japanese words, abbreviations and tokens hard to classify.

Table 3 gives a count breakdown for the Roman letter string tokens. The tokens are

[8] MeCab developer site: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.
[9] *Nippon Housou Kyoukai* (NHK) is a broadcasting corporation in Japan.
[10] http://wordnet.princeton.edu/wordnet/man/wninput.5WN.html

separated into those recognized as English words and others.

*Table 3. Tokens Recognized as English Words*

| Total Number of Tokens | Recognized as English Words | Other Tokens |
|---|---|---|
| 442 | 282 | 160 |

As shown in Table 3, 282 out of 442 tokens (i.e. about 64%) matched words in the English WordNet. For matching the package WordNet 3.0 downloadable from the developer website[11] was used. The complete lists of tokens recognized as English words and other tokens are given in Appendix Part1, which also lists frequencies for each token.

Table 4 gives examples of tokens recognized and unrecognized as English words. Frequencies for each token (i.e. numbers of times the tokenized word appears in the analyzed TV guide) are pointed at by arrows.

By listing multiple examples in Table 4, the author intends to demonstrate the variety of human activity fields the recognized English words are likely to be used in. The considerable variety of the fields serves, in the author's opinion, as another proof of the need to apply cross-language search to the Japanese TV guide. As seen form Table 4, the words range from high-flown ("Dramatic") to professional ("Medical") to colloquial ("DJ", "Sexy"). It should be noted here that no guide for "specialized" TV channels (e.g., music or adult entertainment ones) was analyzed. More information regarding the choice of the channels is given in 2. 2.

[11] https://wordnet.princeton.edu/wordnet/download/

*Table 4. Examples of Tokens*

| **Recognized as English Words** | **Others** |
|---|---|
| TV→107 | HBC→734 |
| NEWS→94 | JUVY→18 |
| JAPAN→40 | SMAP→13 |
| EURO→26 | TARAKO→12 |
| SONG→25 | CHIE→9 |
| sports→24 | NARUTO→8 |
| DJ→16 | TAKAHIRO→8 |
| SHOP→16 | BOSAI→7 |
| GiRLS→14 | FFFFF→7 |
| Job→6 | JR→7 |
| Dramatic→4 | AKBINGO→5 |
| Sexy→4 | BUSAIKU→3 |
| Medical→3 | Shoppin→3 |
| LAUNCHER→3 | GALETTe→2 |
| Love→2 | Hemenway→2 |

The Table 4 “Recognized as English Words” entries can be classified as belonging to various fields of human activity in the following way. In the list below brief descriptions for activity fields are followed by Table 4 examples in brackets. The classification is mostly based on the human examination of the TV guide context. Some of the examples can be classified under more than one activity field.

1) Mass Media (“TV”, “NEWS”)
2) General Entertainment (“TV”, “SONG”, “DJ”, “Dramatic”, “Love”, “LAUNCHER”)
3) Economy (“JAPAN”, “EURO”)
4) Politics (“JAPAN”)

5) Sports (“sports”)

6) Commerce (“SHOP”)

7) Gender-specific Activities (“GiRLS”, “Sexy”)

8) Professional Activities (“Job”, “Medical”)

9) Adult Entertainment (“GiRLS”, “Sexy”)

10) Hard-to-classify (“LAUNCHER”)

It should be noted that the examples covered by the above classification are by far not all the English words found in the analyzed TV guide, and by far not all the human activity fields are listed. (For a complete list of the English words see Appendix Part 2.) As stated above, by classifying the words the author intends to demonstrate a considerable variety of human activity fields the words can be used in. The larger the variety is, the more potential search results the data are likely to include.

As seen from Table 4, the tokens unrecognized as English words include such lexical items as an abbreviation (for a TV channel) “HBC”, a Romanized Japanese name “*TAKAHIRO*” and a word “*BUSAIKU*”, a Japanese common noun that can be translated as “an ugly thing”.

In summary, the Japanese TV guide was analyzed for the presence of English words. A considerable number of English words, representing a variety of human activity fields, was found in the text. In view of this fact it has been concluded that applying cross-language search to the Japanese TV guide, i.e. searching not only for Japanese but also for English lexical items, can increase the number of search results.

## 1.4. Regarding the Novelty and Effectiveness of the Proposed Method

This subsection provides reasons to consider the proposed method novel. The reasons emphasize the effectiveness of the search and processing large textual data. Brief accounts on evaluation experiments and other data are used to illustrate explanations. For details on the method evaluation see 4.

The proposed method introduces applying cross-language search to the Japanese TV guide. As explained in 1.3, this type of search has been implemented due to considerable presence (about 64% of all Roman letter strings) of English words belonging to various fields of the human activity. To the best of the author's knowledge, there is no previous research in which the Japanese iEPG is analyzed for the presence of bilingual vocabulary. The same can be said about previous research in cross-language search applied to the Japanese TV guide and using Japanese synonyms to search the guide text.

Experiments the author performed to find out if English-Japanese cross-language search was also used by other search systems demonstrated that no such search was most probably used by them.

In the above experiments Google Japan customized to search iEPG data and three search systems on major Japanese iEPG websites were used. See 4.4 for details.

The experiments demonstrated that those systems retrieved no search results in which the text included English synonyms for the query words. No search results in which the text included Japanese synonyms for the query words were retrieved, either. Thus the proposed method can be considered novel due to using the mentioned cross-language search and query expansion that are not used by other systems.

Implementing the described cross-language search techniques has led to an increase in search effectiveness. The increase has been demonstrated in experiments comparing the performance of the proposed method with that of baselines. The F-measure increase (by approximately 66%) from 0.432 for the baseline implementation that did not use query expansion to 0.718 for that of the proposed method one, illustrates the search effectiveness improvement.

Figures 4 and 5 give examples of search results that include English and Japanese

synonyms for query words. The figures reproduce the output of the proposed method implementation used for the performance comparing experiment described above.

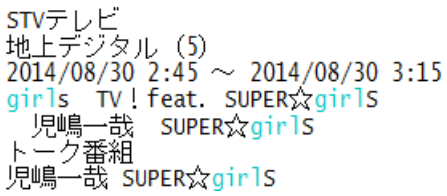


Figure 4. A Search Result with an English Synonym

Figure 4 shows an iEPG listing for the talk show “girls TV!”. The listing was retrieved as a search result for the query “女子が入りやすい居酒屋 ([*joshi ga hairiyasui izakaya*], pubs a female feels comfortable to enter)”. The result can be considered possibly relevant to the query.

The search result in Figure 4 includes the synonym “girl” for the query word “女子 ([*joshi*], a female person)”. The proposed method implementation retrieved the synonym from a bilingual WordNet (the “Japanese WordNet” described in 1. 2) and used the synonym for the search. This, i.e. applying a bilingual WordNet to the iEPG data search, can be looked upon as another novel feature of the proposed method. There is no previous research in this WordNnet application to the best of the author’s knowledge.

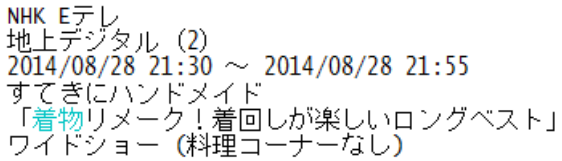


Figure 5. A Search Result with a Japanese Synonym

Figure 5 shows a relevant result that includes the Japanese WordNet synonym “着物 ([*kimono*] , kimono)” for the word “和服 ([*wafuku*], Japanese-style clothing)” used in the query “和服の作り方 ([*wafuku no tsukurikata*], how to make Japanese-style clothes)”. The result in Figure 5 is a listing for the clothes making “wide show” called “すてきにハンドメイド ([*suteki ni handomeido*], Nicely Handmade)”.

Another novel feature of the proposed method is applying the finite state automaton (FSA) concept to searching the iEPG. For more information on the concept and its implementation for the method see 2.8 and 2.9. The purpose for the automaton implementation is to find all occurrences of query words (as well as all query synonyms and transliterations) in all possible orders.

Implementing the FSA has demonstrated an approximately 81% increase in the number of search results for multi-word queries, which is another statement in favor of the proposed method effectiveness. See 4.2 for more details on this effectiveness increase. On the other hand, systems (1) and (2), described earlier in this subsection, have demonstrated no capability of finding occurrences of query words in orders different from that of the original query word order. No capability of finding query words with other words in between was demonstrated by the above two systems, either. System (3) (mentioned earlier in this subsection) seems hardly customizable to retrieve iEPG data by single TV programs the way the proposed method does. The reason is, probably, that Google retrieves what is located at one URL as a single search result and does not look upon parts of the text at the same URL as single search results. On iEPG sites, however, TV program descriptions for the same broadcasting area are usually parts of the text located at the same URL.

Parallel processing of TV guide parts resulted in a large decrease in time spent for the search. The implemented FSA often needs to match text for each of more than 2,000 TV programs with more than 20 words including the original query, its synonyms and transliterations. The parallel processing has been implemented to increase the speed of this process. Experiments have shown an approximately 95% decrease of the search

time resulting from the FSA implementation. See 4.2 for more details on the experiments. The speed decrease is another improvement in the proposed method effectiveness.

To sum it up, the novelty and effectiveness of the proposed method have been experimentally proven and can be outlined as follows.

1) The proposed method is based on the TV guide word stock analysis that has not been performed by other researchers to the best of the author's knowledge.

2) The proposed method introduces a novel cross-language search technique using Japanese and English synonyms together with *Hiragana*, *Katakana* and Romanized transliterations to expand the Japanese query. Semantic feature analysis for grouping search results has also been introduced. No previous research in using the same techniques exists to the best of the author's knowledge.

3) The cross-language search used by the proposed method is, most probably, not used by search systems on major Japanese iEPG sites as well as by Google customized to search a Japanese TV guide site.

4) The query expansion by adding Japanese synonyms to the original query, that is used by the proposed method is, most probably, not used by the above major website systems and the customized Google.

5) The proposed method introduces applying a bilingual WordNet to the Japanese iEPG cross-language search. No research in this area is known to the author.

6) The proposed method demonstrated an approximately 66% F-measure increase.

7) The proposed method introduces applying a finite state automaton to searching the Japanese iEPG data. The automaton can also be applied to other texts in which both Japanese and English words are used. No previous research in this FSA application is known to the author. The automaton implementation has demonstrated an increased word matching capability and an approximately 95% decrease in the search time.

# 2. THE PROPOSED SEARCH METHOD

The present section starts by giving an overall description of the search process that is used by the proposed method. Later in the section each stage of the process is explained in detail.

The first subsection gives an overall process description. The following subsections, each named to concisely define a process stage, provide more details. The section is structured in this way to first demonstrate all the process stages and their order and then provide details on each stage.

## 2.1. The Overall Search Process Description

Figure 6 outlines the stages of the TV guide search process. Arrows represent the iEPG data flow. As shown in Figure 6, 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. The phrase “dynamically generated” refers to automatically computing and formatting year, month and date values for each day the search is performed and seven days ahead. At present the method implementation downloads data broadcast terrestrially in Sapporo, Japan by eight TV channels. A total of over 2,000 TV programs is downloaded.

HTML tags and other metadata are discarded. The natural language text obtained in this way includes program descriptions along with irrelevant textual 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. One cluster has about 250 TV programs.

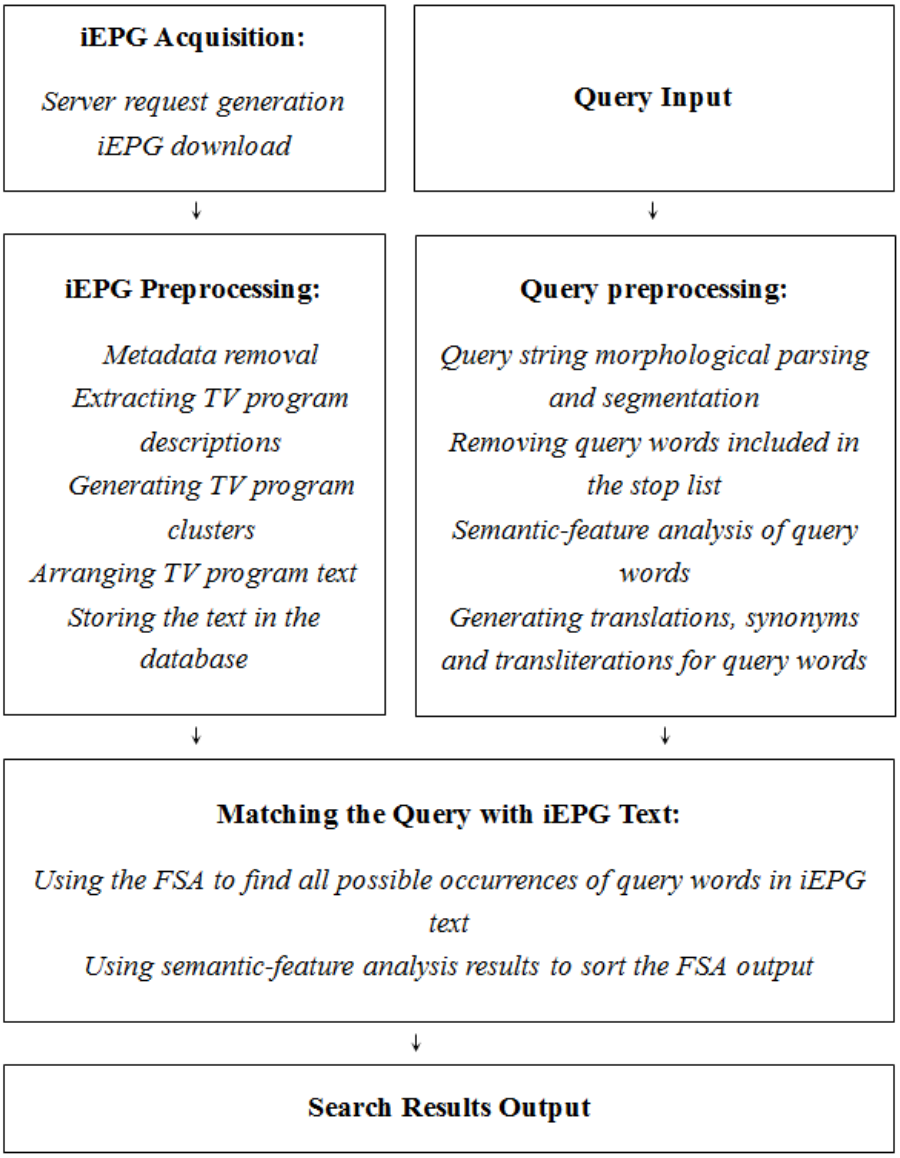


Figure 6. The Overall Search Process

After the search query is input it is segmented (i.e. separated into words) by means of the morphological parser MeCab. Along with the segmented string, MeCab output has other information such as tags telling what part of speech each segment (i.e. a character string MeCab considers a word) is. The part-of-speech tags are taken out of the output and are later used for grouping search results by means of the semantic-feature analysis. Stop-listed words and morphemes are removed from the query.

The iEPG search is carried out by means of matching the query, its bilingual

synonyms and transliterations with one program description at a time. (See 2.7 for details on the query synonyms and transliterations.) That is, instead of matching the query with all the eight-day data at once the proposed method implementation 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.

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”.

The search query words, their bilingual synonyms and transliterations are matched with the iEPG data means of the FSA. The automaton finds all possible combinations of all the above words in all possible word orders, with or without other words in between.

Matching TV programs undergo the sematic feature analysis that is implemented to group search results, i.e. to display more relevant ones closer to the top of the search result list that is output for the user to see. Search results including combinations of nominal and attributive semantic features are looked upon as more relevant than others, as explained in 2.10.

It should be noted that the proposed method implementation is essentially different form large search engines, such as Google. First, large search engines retrieve web documents, such as websites and parts of them, whereas the proposed implementation retrieves pieces of text that describe TV programs. To do so, the implementations do not require indexing millions of web documents (the way a Google webpage[12] says it does) and do not need any corpora, such as the approximately 24 GB large Google N-gram

[12] http://www.google.com/intl/ja/insidesearch/howsearchworks/thestory/

Corpus described by Lin et al (2010). Because of its size the implementation could be used locally as, say, an internal search system for a TV set. It seems unlikely that, for instance, the Google search engine can be used in the same way. It has been the author's purpose to develop the search method for the TV program guide taking into account the above peculiarities of this task.

The overall search process of the proposed method can be summarized as follows. The initial stage is TV guide data acquisition. Once the data are offloaded from a Web source 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 expansion (i.e. adding synonyms and transliterations). Next, the query with its synonyms and transliterations undergoes the semantic-feature analysis. 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 (that match the query). Finally, the results of this procedure are output.

## 2.2. iEPG Acquisition

Procedures implemented at the acquisition stage are listed below.

1) Locating (Java) scripts for loading detailed TV program data on the iEPG Web page.

2) Using the located scripts to generate Ajax requests for the detailed TV guide data.

3) Using the requests to offload the data in JSON and/or HTML format.

As mentioned in 1.1, Japanese iEPG websites normally display two types of the TV Guide. The first type is concise, the second is detailed. Concise TV program descriptions have clickable areas available to the human user (and programming scripts to be used by the machine) for loading detailed data, i.e. displaying these data to the user. The proposed method implementation uses the scripts to offload the detailed data.

After accessing the iEPG Web page and finding the scripts, the implementation makes server requests for the data using the Ajax technology. Eichorn (2006) explains in detail how that technology can be used for a variety of purposes. Figure 7 gives an Ajax request example.

http://www.tvguide.or.jp/TF0010LS.php?type=TVG&packId=78&stationId=62&date=20130517

Figure 7. An Ajax Request Example

The meanings of the parts for the server request in Figure 7 are given in the list below.

- “http://www.tvguide.or.jp/” is the URL of the site from which the data are requested;
- “TF0010LS.php” specifies the programming script used by the site to render the data;
- “type=TVG” specifies the type of the data to be rendered (TV guide data, in this case);

- “packId=78” specifies the part of the TV guide to be rendered;
- “stationId=62” specifies the channel that broadcasts the data to be rendered;
- “date=20130517” specifies the date the rendered data is broadcast (i.e. May 17, 2013, in this case); dates are generated dynamically for the date the data are offloaded and seven days ahead, using the calendar provided by the environment in which the proposed method implementation is used.

The requested data are offloaded for each TV program in the JSON and/or HTML format, i.e. as natural language text with various metadata such as tags indicating the content type for different parts of that text. For example, program title text with a tag indicating that this text is a program title is offloaded for each TV program. This format is useful for data structuring and extraction as it explicitly indicates where, what kind of text can be found. A way the JSON technology can be used for retrieving data as “key-value pairs” (similar to the pair including the program title and the tag mentioned above) is described by Tummarello et al (2010). Figure 8 gives a real-life example of the JSON data as it was loaded by the Internet TV Guide website. (See 1.1 for the site details including the URL). The example contains text for a news program broadcast by the TV channel UHB.

```
{"SystemInfomation":{"login_flg":0,"id":0},
        "MediaInformation":{"mediaId":"1"},
        "StationInformation":{
                "name":"UHB テレビ",
                "additionalDisplayChannel":"地上デジタル（8）",
                "url":"http://uhb.jp/",
                "ProgramInformation":{
                        "programid":"1430",
                        "gcode":"",
                        "objectDate":"2014/10/03 3:24 ～ 2014/10/03 4:30",
                        "title":"ＵＨＢニュース＆ウェザー",
                        "subTile":"",
                        "content":"",
                        "Category":[{"categoryid":"30","name":"ニュース報道"}],
                        "explanation":""
                }
        }
}
```


Figure 8. An Example of TV Program Data in the JSON Format

The program data in Figure 8 includes the program title "UHB ニュース＆ウェザー ([*yu eichi bui niyuusu ando uezaa*], UHB News and Weather)", the program genre "ニュース報道 ([*niyuusu houdou*], news reports)" and other natural language text data that are processed the way described in 2. 3.

At this time the proposed method implementation downloads iEPG for the eight TV channels that are broadcast terrestrially in Sapporo, Japan. The data are downloaded for the current date and seven days ahead and includes descriptions for about 2,000 TV programs.

In summary, the iEPG data including about 2,000 TV program descriptions are acquired by means of locating programming scripts used to render the data on the website, using the scripts to generate Ajax server requests, and offloading the TV program data in the JSON format[13].

[13] This format contains the tag-value pairs explained earlier in 2.2. Figure 8 gives an example of the tag-value pairs.

## 2.3. iEPG Pre-processing

Below is a list of procedures implemented at the iEPG preprocessing stage.

1) Extracting natural language text for TV programs from the JSON data structure.

2) Arranging the text.

3) Storing each TV program text as a separate item in the database.

Natural language text is extracted from the tag-text pairs described in 2. 2 and tags (such as “program title” of “program genre”) indicating contents of text chunks are used to arrange them. The chunks are arranged in the order shown below. Each item of the following list is numbered according to the order and gives the content type of each text chunk.

1) Name of the TV channel broadcasting the program.

2) Broadcasting type (e.g., terrestrial), channel number.

3) Starting date and time, ending date and time for the program.

4) Program title.

5) Program subtitle.

6) Program content summary.

7) Detailed program content explanation.

8) Program genre.

9) Program cast or list of participants.

The above arrangement follows the style iEPG is listed on Japanese websites. If some of the above items (such as “program subtitle” and “program content summary”) are absent from the actual TV guide, the proposed method implementation arranges the available items in the above order without leaving any items blank. Arranged in this way, the text for each TV program is stored as a separate item in the database. The database items are matched one by one with query words at the matching stage.

As mentioned above, sometimes the actual TV guide does not have any text for

some of the above list items (because the TV guide is arranged in this way by the broadcaster, the Web administrator or both.) If an item is missing, i.e. there is blank space instead of it, the implementation ignores the blank space and goes on to the next item for which text is available. For instance, if text for item 5) (“program subtitle” on the above list) is missing from the TV guide, the implementation stores the program text without a subtitle in the database and does not leave any blank space in place of the subtitle. As the proposed method implementation does not ignore any TV program text present in the TV guide and does not process what is absent from the guide, the search accuracy is unaffected.

## 2.4. Search Query Morphological Analysis and Segmentation

Japanese is written without word spacing. After the search query is input by the user, the query is segmented (i.e. separated into words). The words then undergo a part-of-speech analysis that comprises attaching a corresponding part of speech tag to each word. The tags are later used to sort search results by means of the semantic-feature analysis explained in 2.9.

The method implementation does not use the n-gram model for the morphological analysis and segmentation. Instead of e.g., applying Microsoft Web N-gram Corpus to the string segmentation as described by Wang et al (2010), linguistic techniques are used. The thesis author has preferred such techniques because n-gramming a character string does not take into account morphology and semantics, however it is his purpose to take them into account.

The segmentation technique described by Wang et al (2010) involves matching a character string with strings repeatedly found in the corpus. Shorter strings that often match are considered to be separate words.

Differently form that technique the proposed method implementation segments the query by means of the morphological parser MeCab. Before feeding the query to the parser, the query character string is pre-processed to ensure it is encoded in uft-8. To divide the string into segments, MeCab analyzes word combinability by using the statistical data for the given word occurrence in the large corpus.

The segmented query string (i.e. the query separated into words by MeCab) is then checked against the stop-list as explained in 2.5.

## 2.5. Removing Query Words Included in the Stop List

In previous research various kinds of stop lists and their applications have been considered. For English, Hiemstra et al (2001) 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 (2002) describe a system (for processing the Japanese language) that lists words of no potential interest to the user as stop list items.

The author of the present thesis suggests using a stop list for multiple reasons. That is, for structural, semantic and pragmatic reasons some parts of words (and sometimes whole ones) can be omitted or substituted with others with no change to the meaning. The proposed method implementation detects and discards such words and morphemes. Table 5 gives examples of the discarded items and the way they may be used in a search query. In Table 5 (and throughout the thesis text), 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, stop list items appearing as parts of phrases are underlined when that is needed for clarity.

The reasons the examples in Table 5 have been included in the stop list are explained below. The inclusion judgment is based on the human analysis of the search results for multiple queries with the stop list items as parts of those queries.

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, would mean making a search system look for something not really needed for retrieving the meaning searched for. Moreover, if the system uses direct matching techniques, for example, 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[14] it also can express the idea that “something is a location for

[14] Goo Dictionary, http://dictionary.goo.ne.jp/.

*Table 5. Examples of 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 |

something else" or "that something is the site of a certain action". However, another Japanese dictionary[15] 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 Table 5 example (number 3), の [*no*] (used in a non-possessive meaning) can also be omitted. Thus searching for it is unnecessary.

[15] Sanseido Web Dictionary, http://www.sanseido.net/.

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 to the meaning. It is common knowledge that, for instance, the prenominal adjectival 小さな [*chiisa<u>na</u>*] can become 小さい [*chiisa<u>i</u>*] 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 filter for い[*i*] endings is limited to those adjectives that have な[*na*] pre-noun adjectival counterparts. Counterparts with い[*i*] and な[*na*] endings from 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.

## 2.6. Semantic-feature Analysis of Query Words

This subsection gives the essence and theoretical background of what the author refers to as "semantic-feature analysis". It also explains how and why the analysis has been implemented. The analysis is performed on the query after it is segmented and stop list items are removed. The purpose of the analysis is to gather semantic feature information on query words in order to use that information to sort search results the way as described in 2.9.

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" (Lieber, 2004). Another research states that a word of a language has the "semantic core" also referred to as the "semantic prime" (Goddard, 2002). For example, semantic primes for nouns are classified as "substantives" and those for adjectives as "determiners" (ibid.). Moreover, according to Wierzbicka (1996) semantic primes are "universal", i.e. present in multiple languages.

The proposed search method 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.

In other words, by the semantic features that the proposed method implementation analyzes, the author means the semantic primes discussed above. By the semantic prime for the noun the author means the fact that nouns signify objects, and by the semantic prime for the adjective that adjectives signify properties of objects.

The implementation bases its analysis of the sematic features of search phrase words on the morphological analysis and part-of-speech tagging. Using the part-of-speech tags (attached to query words the way described in 2.4) the implementation attempts to make a judgment on the following three aspects:

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 proposed method implementation performs

the semantic-feature analysis for the following reasons. An existing research demonstrates that nouns "constitute over 70% of query terms"[16] (Barr et al, 2008).

Moreover, nouns used together with adjectives are "common need information clusters" in English queries for multiple search engines (Baeza-Yates et al, 2005). Another research demonstrates that nouns, such as proper nouns, are numerous in Japanese search queries (Arita et al, 2007).

To sum it up, the proposed method implementation looks into the universally present core meaning of the query to find object and property-of-an-object features. If found, words with these meaning features become the focus of the analysis because nouns form the majority of query terms and noun-adjective clusters are common in queries for multiple search engines.

The author believes that the proposed analysis technique could be used not only for Japanese but also for other languages. The fact that object and property semantic features are "universal", as stated above, justifies this belief.

[16] In this research the percentage refers to search queries in the English language only.

## 2.7. Generating Translations, Synonyms and Transliterations for Query Words

For each query noun, Japanese and English synonyms are generated using the Japanese WordNet that has bilingual entries. *Hiragana*, *Katakana* and Romanized transliterations for each noun are generated by means of MeCab[17] and the Perl module Lingua-JA-Kana-0.07.

For each query adjective, English and Japanese synonyms as well as *Hiragana* and *Katakana* transliterations are generated. The same generation tools as those for the nouns are used.

As explained by Bond et al (2009), WordNet synonyms for polysemous words are arranged into "synsets" i.e. sense groups. At present, the proposed method implementation uses the first three synonyms from the first synset. The Perl module Lingua-JA-WordNet-0.21 is utilized for the synonym generation task.

Figure 9 reproduces the proposed method implementation output used for testing purposes. The output shows the synonyms and transliterations for a query "冷たいビール ([*tsumetai biiru*] , cold beer)" utilized for testing the implementation performance. See 4 for more information on the evaluation experiments and test queries.

As seen from the Figure 9, a total of fifteen words including synonyms and transliterations have been generated to expand the above query and used for the TV guide search. Following is a legend for Figure 9.

- "___jpn_n_syn" = Japanese synonyms for the query noun;
- "___eng_n_syn" = English synonyms for the query noun;
- "___jpn_adj_syn" = Japanese synonyms for the query adjective;
- "___eng_adj_syn" = English synonyms for the query adjective;
- "___n_hira_kata_roma" = *Hiragana*, *Katakana* and Romanized transliterations for the query noun;
- "___adj_hira_kata" = *Hiragana* and *Katakana* transliterations for the query

[17] Developer site: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

adjective.

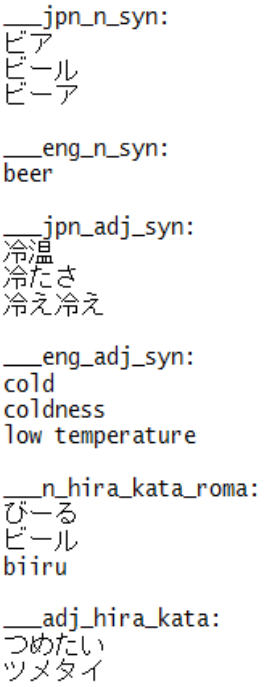


Figure 9. Query Word Synonyms and Transliterations

## 2.8. Utilizing a Finite State Automaton (FSA) and Parallel Processing to Find All Possible Occurrences of Query Terms

The FSA has been implemented for the purpose of solving the following problem. The problem consists in the need to extract TV guide text with all possible combinations of search query words in all possible word orders, with or without other words between the query words.

The FSA concept this thesis deals with can be briefly defined as a program instructing a machine to perform a certain task a finite number of times. Barabás and Kovács (2012) describe using the FSA for detecting and manipulating character strings, such as word stems, in the text. Another research, Patrick and Sabbagh (2011), describes applying the FSA to the extraction of (medical) text chunks. The present thesis introduces applying the FSA to the Japanese iEPG cross-language search. No previous research in this area is known to the thesis author.

The implemented FSA matches all possible combinations of search words in all possible word orders, with or without other words between the search words. TV program descriptions with matching words are extracted and later sorted as described in 2.9.

Below is a mathematical model for the implemented FSA. Figure 10 demonstrates iteration cycles of the automaton. The model and the figure follow the conventional style of the automata theory literature (Aziz et al, 2004; Kumar, 2011).

The following equation shows the way the implemented FSA can be modelled mathematically.

$$M = (Q, I, \partial, S)$$

- M represents the matching automaton implemented.
- Q is the number of states, i.e. the number of times the automaton processes each word in the TV guide.
- I is the set of all words in the TV guide text.
- $\partial$ represents the transition function defined as $\partial(0, i_0) \rightarrow i_n$. In $(0, i_0)$ 0 is the initial transition state at which the FSA attempts to match the very first word

represented by $i_0$. Each word the automaton attempts to match next is represented by $i_n$.

- S is the set of accept states, i.e. the states at which query words match ones in the TV guide text. All the matches are sorted the way described in 2.9.

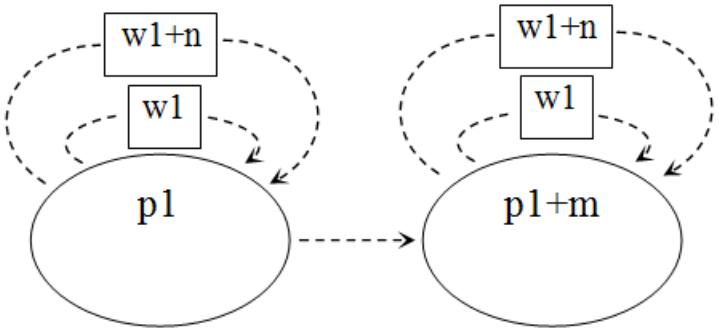


Figure 10. FSA Iterations

The automaton starts matching from the first TV program text, marked p1 in Figure 10. It takes the first query word w1 from the set of all words of the given query, and repeatedly matches it with the p1 text in order to find all occurrences of w1 in p1. The same is repeated for each consecutive word w1+n, where w1+n belongs to the set of all query words of the given query and n is less than or equal to the sum of all words in this set.

The automaton then goes on to each consecutive TV program p1+m and repeats the same matching procedures. Each consecutive TV program p1+m belongs to the set S (i.e. the set of all TV programs in the given TV guide file) and m is less than or equal to the sum of all TV programs in S.

As explained in 2.7, the proposed method implementaion does not only use the "query translation", i.e. English synonyms for Japanese search words, but also expands the query in multiple other ways. Expanded versions of some queries total about twenty words. These words are matched against iEPG data sets for eight broadcasting days, each set including over 2,000 TV program descriptions that in turn include about a paragraph of text on the average. This matching process has been found to be computationally expensive. To reduce the computation cost, parallel processing has been implemented. This has resulted in a large reduction of the search time, as described in 4.2.

The proposed method FSA splits the iEPG text into 12 parts (12 being equal to the number of the used processor cores) and processed each part in parallel. Below is a concise description for the used operating system and hardware.

OS: Linux, CentOS release 5.10;

CPU: 12 physical cores, 2.67GHz;

Memory: appr. 64 GB (including the cashed memory and buffers).

## 2.9. Utilizing Semantic-feature Analysis Results to Sort the FSA Output

According to existing research, finite-state automata are implemented for such tasks as information extraction based on matching words (Smrz and Schmidt, 2009) or matching groups of words (Kwak et al, 2011). The FSA that this thesis proposes not only matches words but also groups the search results. The grouping is based upon semantic features that the matching words possess. (Details on the semantic feature analysis are provided in 2. 6.)

McCandless and Hatcher (2010) state that search results can be grouped according to relevance by means of analyzing how well strings match, or by means of analyzing indexes attached to strings.

The present thesis proposes an approach more similar to the latter analysis, provided part-of-speech tags attached to words by MeCab (as explained in 2. 6) are looked upon as indexes. The implemented automaton "knows" the semantic-feature analysis results based on MeCab tags. In other words it "knows" whether it is matching words meaning objects (e.g., "food"), words meaning properties of objects (e.g., "tasty"), or words without these semantic features.

At the end of each matching cycle (i.e. after the automaton is through with processing each TV program) the above "knowledge" is used to sort matching results.

The implemented FSA filters put the search result (i.e. the chunk of text describing one TV program) into one of five groups. The groups and the order they are output are shown below. The semantic features that search results in each group must possess and the other grouping criteria are described for each item of the following list.

1) TV program descriptions with one or more groups of words meaning objects and properties.

2) TV program descriptions with two or more words that mean objects that are not the same; the text does not include words meaning properties.

3) TV program descriptions with two or more of the same object and no properties.

4) TV program descriptions with one object that is the same; the text does not

include words meaning properties.

5) Other search results.

As explained above, the proposed method implementation not only analyzes sematic features but also attempts to find out if words possessing them are the same.

In accordance with the above grouping criteria, the proposed method implementation does not look upon the number of times the same word appears in the program text as the key factor in judging about the search result relevance. It rather gives priority to groups of words meaning objects and their properties, and groups of ones meaning two or more different objects, provided such words are present in the search query.

# 3. METHOD IMPLEMENTATION ARCHITECTURE

This section describes the elements of the proposed search method implementation and how the elements interact. The description starts with the overall architectural framework. Later in the section the elements and their dependencies (i.e. programs from the software library that they use) are described.

## 3.1. Architectural Framework

The overall structure of the method implementation is shown in Figure 11. Arrows pointed on both sides represent two-direction interaction between elements, arrows pointed on one side represent one-direction interaction. In the figure and onward the element names are capitalized for clarity.

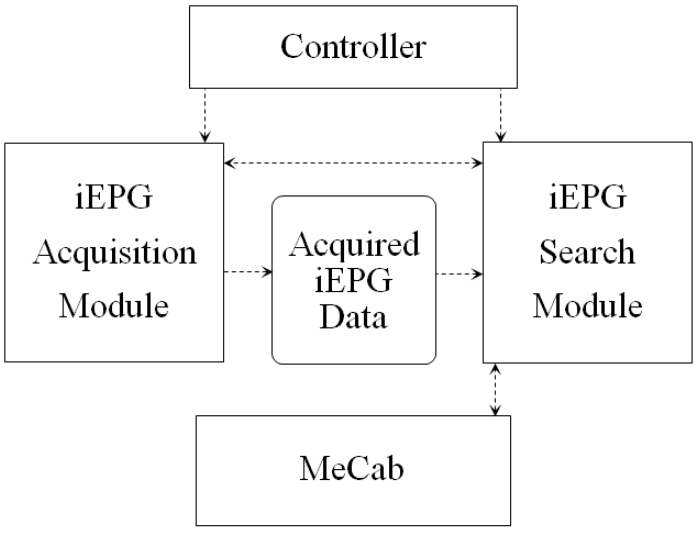


Figure 11. Architectural Framework

The Controller provides a shell interface enabling a user to initiate iEPG data acquisition or search of that data. The Controller receives the user's input and sends commands to the two modules in Figure 11.

The iEPG Acquisition Module offloads TV guide from the Web, arranges the data

and stores it in the Acquired iEPG Data File shown in Figure 11.

The iEPG Search Module performs search on the above file. The module uses MeCab when the search is carried out.

## 3.2. Elements and Their Components

The Controller displays a menu for the user to choose between iEPG acquisition and the iEPG search. If the search is chosen the user is asked to enter a query. The search is initiated on pressing the “Enter” key.

The Controller also provides a query input filter to prevent input errors that may affect the search accuracy. The filter limits the input to the Japanese *Kanji*, *Hiragana*, *Katakana*, English alphabet letters, the semicolon, the hyphen and the space.

The following list describes the components of the iEPG Acquisition Module. Each component in the list is a subroutine, i.e. a function performed by the module.

- Character Converter: changes English alphabet letters, colons, hyphens, digits from full to half width, and changes *Katakana* form half to full width in the TV guide offloaded from the Web. The conversion is carried out to prevent mismatching of the user input and the iEPG data characters. The width of the characters in both the user input and the iEPG is changed as has been specified.
- Offloader for Concise iEPG: offloads the iEPG in the concise format described in 1.1. This component generates the dates for day the data is requested from the server and seven days ahead, uses the dates to offload the concise iEPG, and in that concise data locates Java scripts for loading detailed TV program data the way explained in 2.2.
- Offloader for Detailed iEPG and Program Text Formatter: offloads the iEPG in the detailed format described in 1.1, and changes the program description text into the format suitable for the search as explained in 2.2 and 2.3. This component also stores the TV guide to be used by the iEPG Search Module.

The architecture for the iEPG Search Module is shown in Figure 12.

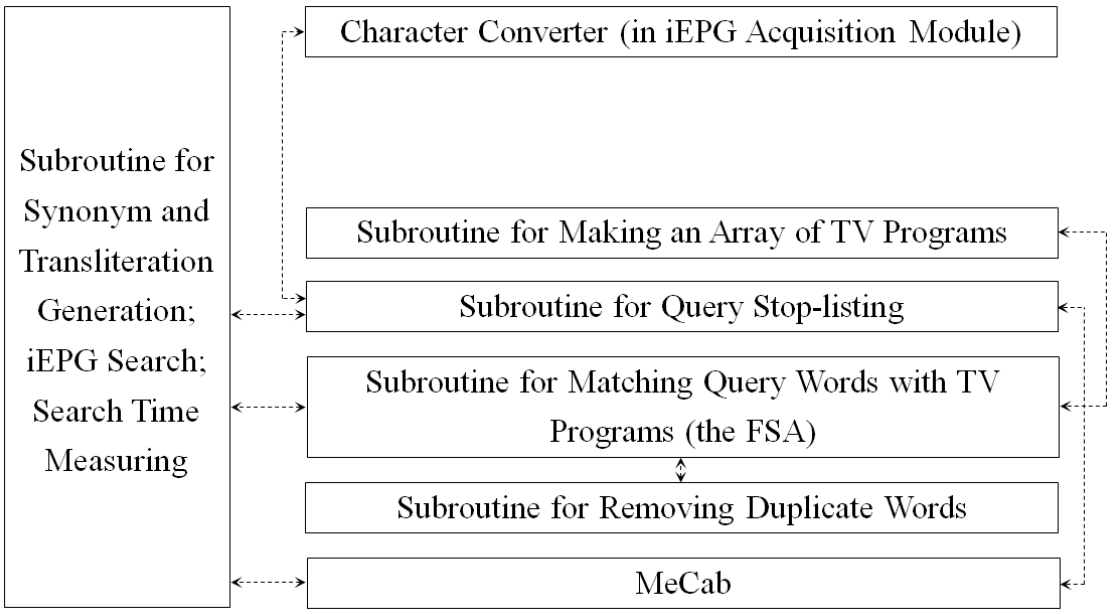


Figure 12. iEPG Search Module Architecture

As seen from Figure 12, some subroutines in this module are used by other subroutines. Arrows pointed on both sides represent two-direction interaction between subroutines. The following list explains the interaction using exactly the same subroutines names as those in figure 12 (and omitting articles before the names).

- Subroutine for Synonym and Transliteration Generation; iEPG Search; Search Time Measuring:
  - calls Subroutine for Query Stop-listing to obtain the search query segmented into words that have part-of-speech tags; the words are used for synonym and transliteration generation as well at as other search stages;
  - calls MeCab to obtain *Hiragana* transliterations of query words; these transliterations are used to generate *Katakana* transliterations;
  - calls Subroutine for Matching Query Words with TV Programs (the FSA) that match query words, their synonyms and transliterations with the text for each TV program and uses the part-of-speech tags for the sematic feature analysis;

- outputs the matching results, time used for the search and testing data such as lists of generated synonyms and part-of-speech tags next to words they have been attached to;

- Subroutine for Query Stop-listing:
  - calls Character Converter that is a part of iEPG acquisition module; for the function of this subroutine see "Character Converter" in the description for iEPG acquisition module above;
  - calls MeCab that segments the query string into words and attaches the part-of-speech tags;
  - removes the words included in the stop list from the query the way explained in 2.5;
- Subroutine for Matching Query Words with TV Programs (the FSA):
  - calls Subroutine for Removing Duplicate Words that makes sure search words do not have repeated ones; this subroutine has been implemented because WordNet 3.0 (used by the implementation) occasionally generates lists of synonyms that have repeated words;
  - calls Subroutine for Making an Array of TV Programs and uses its output to match search words with one TV program at a time the way explained in 2.1;
  - performs the semantic feature analysis explained in 2.6;
  - sorts the search results the way explained in 2.9.

## 3.3. Dependencies for the Elements

The word “dependencies” in the title refers to the Perl library software that the elements of the proposed method implementation use for processing the TV program text and that of the search query. The present subsection gives the name for the software and explains the way the software is utilized.

The iEPG Acquisition Module has the following dependencies.

- LWP::UserAgent 6.03: used to offload the TV guide from the Web;
- Encode 2.49: used to instruct the Perl interpreter to process the offloaded TV guide text as a utf-8 character string;
- File::BOM 0.14: used to mark the offloaded TV guide as a utf-8 string, which helps preventing processing errors when the data is passed among software components;
- Lingua::JA::Regular::Unicode 0.12: used by the Character Converter subroutine to convert characters from full-width to half-width and vice versa as explained in 3. 1.

The iEPG Search Module has the following dependencies.

- Encode 2.49: used to make sure the MeCab output is processed as a utf-8 character string;
- File::BOM 0.14: used to make sure the TV guide text is marked as a utf-8 string when passed among software components;
- Lingua::JA::Kana 0.07: used to generate *Katakana* and Romanized spelling versions for Japanese query words;
- Lingua::JA::WordNet 0.21: used to retrieve Japanese and English synonyms for query words from WordNet 3.0;
- threads 1.86: used to implement the TV guide parallel processing explained in 2. 8.
- Benchmark 1.15: used to measure the time spent for the search.

# 4. EVALUATION EXPERIMENTS

The author carried out three major kinds of evaluation experiments. The purpose of the first kind experiments was to evaluate the FSA implementation performance. Experiments of the second kind evaluated the performance of the proposed method implementation as a whole. Experiments of the third kind were performed to compare, as far as possible, the proposed implementation with similar ones that are commonly used. The following subsections describe the experiments of each kind in detail: 4.1 deals with the data, such as test queries, used for the evaluation; 4.2 explains the way the FSA matching accuracy and speed were evaluated; 4.3 explains the way the proposed method implementation was evaluated as a whole by comparison with baseline implementations; 4.4 compares the proposed method implementation performance with that of existing search systems that are commonly used.

The accuracy of the semantic feature analysis for grouping search results, described in 2.9, was evaluated by an individual asked to mark search results that could be considered irrelevant but appeared above relevant ones in the output of the proposed method implementation. Roughly 12% of all output search results were marked. This evaluation was performed using the test queries described below.

## 4.1. Evaluation Data

For the evaluation experiments test data containing the Japanese iEPG for 24 broadcasting days, and 30 test queries were used.

The test data contained text for 6,634 TV program descriptions from several words to several paragraphs in length. Figure 13 provides an example of that data, the example contains a data snippet with three TV programs. The way the text for each program is arranged is explained in 2.3.

```
_____
NHK総合
地上デジタル(3)
2013/10/03 13:00 ～ 2013/10/03 13:05
ニュース
ニュース報道

_____
NHK総合
地上デジタル(3)
2013/10/03 13:05 ～ 2013/10/03 13:27
連続クイズ ホールドオン！
第266回
クイズ

_____
NHK総合
地上デジタル(3)
2013/10/03 13:27 ～ 2013/10/03 14:00
スタジオパークからこんにちは
必見AKB48高橋大島小嶋の本音！ 高橋みなみ 大島優子 小嶋陽菜 伊藤雄彦 米田弥央
トーク番組
高橋みなみ 大島優子 小嶋陽菜 伊藤雄彦 米田弥央
_____
```

Figure 13. Test TV Guide Data Sample

As mentioned above, the test data sample in Figure 13 includes three TV programs, they are a news program, a quiz and a talk show. The text for each program is separated by horizontal lines for the ease of human viewing and machine parsing.

The test queries were obtained by means of a questionnaire, from five female individuals from 19 to 77 years of age and ten male ones from 19 to 84, i.e. 15 individuals in total. The questioned individuals have not been involved in the present research.

The list below provides 10 samples out of the 30 queries used for the evaluation. An English translation is given for each query. For a list of all the queries used see Appendix Part 3.

1) 今日のニュース (today's news);

2) 子供が好きなアニメ (cartoons that children like);

3) おかしい風景 (weird scenery);

4) 和服の作り方 (how to make Japanese-style clothes);

5) Gackt のいきつけ (places Gackt[18] often visits);

6) 女子が入りやすい居酒屋 (pubs a female feels comfortable to enter);

[18] A Japanese singer-songwriter and actor.

7) 札幌にはない食べ物 (food not found in Sapporo);

8) 人気な温泉 (popular spas);

9) 札幌の人気なレストラン (popular restaurants in Sapporo);

10) 北海道の食べ物 (Hokkaido food).

## 4.2. Evaluating the FSA

The experiments described in this subsection evaluated (1) the FSA capability of matching and retrieving all the input strings in all possible combinations with or without other strings in between, and (2) the FSA speed.

From the 30 test queries 60 separate word strings, including both words written in Japanese and English characters, were used. The 60 words were chosen randomly.

The above 60 words were inserted in random order throughout the above test iEPG text. It was made sure that most of the words had other ones (different from the 60 test words) in between.

The 60 words were fed to the FSA four times, fifteen words at a time. Each time the automaton retrieved all the occurrences of the test words in the iEPG text, i.e. matched and extracted 100.0% of all the test word string occurrences in the test data.

As mentioned in 2. 8, parallel processing was implemented to make the FSA faster. To measure the speed increase due to the parallel processing, the proposed implementation using this technique was compared with a baseline implementation that did not use it. Both implementations used the 24-day test data and the 30 test queries.

The proposed implementation FSA split the iEPG text into 12 parts (12 being equal to the number of the used processor cores) and processed each part in parallel. For the description of the used hardware see 2. 8.

The experiment demonstrated an approximately 95% reduction of the search time. As a result, the input-output interval, averaging to 0.7 of a second, is most unlikely to annoy the human user.

The FSA speed was also evaluated on hardware with lower capacity. The test data for eight broadcasting days and the 30 queries mentioned above were used. This amount of data was utilized because the 8-day data are kept and searched at the Japanese TV guide sites, such as those mentioned in 5. 3. Currently those sites do not use data for more than eight days. Moreover this amount of data is easier to process on the hardware with lower capacity. Below is a concise description for the OS and hardware that were utilized for this evaluation.

OS: Linux Ubuntu 14.04.1 (installed on a laptop computer);

CPU: 4 physical cores, 2.30 GHz;

Memory: appr. 4 GB.

Although no high-performance machinery was used for this experiment, the input-output interval averaging to approximately 0.5 of a second was observed.

In summary, the proposed implementation FSA has demonstrated (1) the capability to retrieve 100.0% of the test string occurrences and (2) an approximately 95% speed increase.

## 4.3. The Proposed Method Multi-parameter Evaluation Using Baselines

The evaluation was carried out by comparing the performance of the proposed implementation and five baseline implementations on the 30 test queries and test iEPG data described in 4. 1.

Table 6 gives evaluation results for the proposed and baseline implementations. The baseline implementations are numbered (2) to (6) in Table 6, the proposed one is numbered (1). Details for each of the baseline implementations are given after the respective numbers below. To generate related words, such as synonyms, the proposed and baseline implementations used the Japanese WordNet.

*Table 6. Results of Evaluation Using Baselines*

| | **Average Recall** | **Average Precision** | **Average F-measure** |
|---|---|---|---|
| **(1) Proposed** | 0.861 | 0.690 | 0.718 |
| **(2) Domain Category (Jp. and Eng.)** | 0.853 | 0.644 | 0.680 |
| **(3) Hypernyms (Jp. and Eng.)** | 0.865 | 0.602 | 0.645 |
| **(4) Synonyms (Jp.)** | 0.644 | 0.528 | 0.530 |
| **(5) Synonyms (Eng.)** | 0.541 | 0.533 | 0.468 |
| **(6) Segmented Input String** | 0.540 | 0.539 | 0.432 |

Baseline (2) generates Japanese and English domain category words for Japanese query nouns. The word "運動競技 [*undou kyougi*] athletics", for instance, is a domain category for "ハイキング [*haikingu*] hiking". (3) generates Japanese and English hypernyms[19] for Japanese query nouns. E.g., "show" is a hypernym for "movie" because the latter is a type of the former. (4) and (5) generate Japanese and English synonyms respectively. These two implementations are based on a previous research (Nanba, 2007) the author of which states that generating the above kinds of synonyms can improve the performance for searching Japanese and English text respectively. (6)

[19] The "hypernym" and other WordNet terms are defined at https://wordnet.princeton.edu/wordnet/man/wn.1WN.html.

uses the words resulting from the query string segmentation described in 2.4 and no related words from the WordNet.

The proposed and the baseline implementations, except baseline (6), generated *Hiragana*, *Katakana* and Romanized transliterations (described in 2.7) for query words and used the transliterations along with the WordNet entries to expand the query.

The baseline implementations marked (2) to (6) were used for the following reasons. The author utilized (2) and (3) to see how the semantic generalization (i.e. expanding the query by adding domain category words and hypernyms respectively) would affect the precision and recall. Baselines (4) and (5) were utilized to verify that using both Japanese and English synonyms for searching the Japanese iEPG is more effective than using only Japanese or English synonyms. Baseline (6) was utilized to see how the proposed query expansion would increase the Japanese iEPG search performance compared with using no query expansion.

As seen from Table 6, the proposed implementation demonstrated a considerably higher F-measure and precision compared with those of all the baselines. The F-measure increase demonstrated by the proposed implementation compared with baseline (6) is approximately 66%. In other words, the performance improvement in comparison with a search method using no query expansion is evident. The proposed implementation has also demonstrated a higher recall compared with the ones for all the baselines except for baseline (3). This and other evaluation results are discussed in 5.

To sum it up, in the evaluation experiments the proposed implementation demonstrated better overall performance in comparison with that of the baseline ones.

The following “traditional” formulas for recall, precision and F-measure were used.

Recall = Number of relevant results retrieved/Number of relevant results in the test data

Precision = Number of relevant results retrieved/Number of all results retrieved

F-measure = 2*Precision*Recall/(Precision+Recall)

The number of relevant results in the test data was obtained as follows. First, the test query was manually segmented into words. Synonyms and transliterations for those words were then generated the way described in 2.7. Next, the implemented FSA (that demonstrated 100.0% accuracy, as explained in 4.2) was used to match the original

query words, the synonyms and the transliterations with the TV program text. The number of relevant results among the matching TV programs was manually counted. The relevance was verified by an individual uninvolved in the present research. The individual was made familiar with the judgment criteria explained below. The same criteria were made clear to another individual, uninvolved in the research, who verified the relevance of the search results output in the evaluation experiments.

The criteria for judging the search result relevance were set as follows. First, the author considered the following guidelines for relevance judgment found in an information retrieval source.

> ... it is generally easier for people to decide between at least three levels of relevance, which are definitely relevant, definitely not relevant, and possibly relevant. These can be converted into binary judgments by assigning the possibly relevant to either one of the other levels ... (Croft et al, 2009)

Speaking the language of this source, “possibly relevant” was assigned to “relevant”, however with a certain amount of caution. When examining search results, the presence of query words in the retrieved text was not considered the only criterion for relevance (or possible relevance). The context was considered along with the word presence. The following example (rendered into English) demonstrates the importance of the context. The program text “this place is not a resort” retrieved in response to the query “popular resort” falls, in the author’s opinion, into the “irrelevant” category.

The technique for grouping search results described in 2.9 was evaluated by an individual asked to mark search results that could be considered irrelevant but appeared above relevant ones in the output of the proposed system. Roughly 12% of all output search results were marked. This evaluation was performed using the same test queries and iEPG as in 4.1.

## 4.4. Comparison of the Proposed Method Performance with That of Some Commonly Used Systems

The following experiments were performed to see if the query expansion by adding Japanese and English synonyms as well as *Katakana*, *Hiragana* and Romanized transliterations was also utilized by some commonly used search systems.

In these experiments the following four systems were used:

(1) Google Japan customized to search the Web-based Japanese TV guide data,

(2) the search system on the website "インターネットＴＶガイド[20] ([*intaanetto terebi gaido*], Internet TV Guide)",

(3) the search system on the website "Yahoo テレビ[21] Japan ([*yahoo terebi japan*], Yahoo TV Japan)",

(4) the search system on the website "G Guide テレビ王国[22] ([*ji gaido terebi oukoku*], G Guide TV Kingdom)".

Fifteen randomly selected queries out of the 30 test queries listed in Appendix Part 3 were submitted to the four systems.

The experiments demonstrated that systems (1), (2), (3) and (4) retrieved no search results in which the text included English synonyms for the query words. No search results in which the text included Japanese synonyms for the query words were observed, either. However, the use of *Hiragana* and *Katakana* transliterations and Romanizations (of query words) by system (1) was observed.

In summary, the query expansion by adding Japanese and English synonyms, utilized by the proposed implementation, was most unlikely utilized by the above four systems.

---

[20] URL: http://www.tvguide.or.jp/. As of Oct. 1, 2014 this site has discontinued displaying iEPG data, according to the information posted on the site in Oct. 2014. Experiments using the system on this site were performed before the mentioned date.
[21] http://tv.yahoo.co.jp/
[22] http://tv.so-net.ne.jp/

# 5. Discussion

The evaluation results demonstrate that the proposed method is on the whole more effective than the baseline ones. The higher recall for baseline (3) (in comparison with the proposed implementation recall) suggests that using hypernyms to expand the query may result in retrieving a larger number of relevant results. However, the "side effect" of using hypernyms is the lower precision and consequently the lower F-measure. This suggests that utilizing hypernyms (i.e. words with more general meanings) may as well result in the search noise. The capability to choose the hypernym with the right degree of generality (or to disregard the hypernym if irrelevant) is desirable to minimize the search noise. A way to solve this problem seems to be the query analysis from the context and common sense viewpoints.

The proposed method performance is considerably better than that of baselines (4), (5) and (6) respectively. This verifies that the query expansion by adding the bilingual synonyms and transliterations, the way the proposed implementation does, is considerably more effective for the Japanese iEPG search than limiting the synonyms to unilingual ones only or than using no query expansion at all.

The approximately 66% performance improvement demonstrated by the proposed method implementation in comparison with baseline (6), is obviously due to utilizing the cross-language search and query expansion techniques. For instance, for the Japanese synonyms "料理店[*ryouriten*]" and "料亭[*ryoutei*]" of a query word "レストラン[*resutoran*]" ("restaurant"), relevant search results were retrieved. Relevant search results were also retrieved for the English synonym "girl" used to expand the query that included the Japanese word "女子[*joshi*] a female".

However, say, transliterating the query word "暖かい[*atatakai*]" ("warm" as in "warm climate") as the *Hiragana* "あたたかい[*atatakai*]" ("warm" in various senses), resulted in retrieving the irrelevant phrase "あたたかい交流[*atatakai kouryuu*]" ("warm exchange"). The precision for the proposed method as well as for baselines (2), (3), (4) and (5) could have been higher if the right sense had been chosen for the paraphrase when generating paraphrases for ambiguous query words.

The difficulty to choose the right sense for the query paraphrase has been pointed out by Andrade et al (2013). The authors of this work propose using bilingual corpora for choosing the correct sense of the synonym. Another way to solve the problem

appears to be utilizing a database with the meaning domain annotation such as the Japanese FrameNet, developing which is not an easy task, according to Ohara (2013). The difficulty, as stated by the author, consists in the difference between Japanese and English languages and the need to map the original English FrameNet version onto the Japanese one.

The precision for both the proposed and baseline implementations could have been higher had it been possible to avoid word-segmentation errors of the morphological parser used. For instance the query word “入りやすい [*hairiyasui*]” describing the feeling of comfort when entering e.g. a building, was segmented into “入り[*hairi*] entering” and “やすい[*yasui*]”. As the word “やすい[*yasui*]” used separately can have such meanings as “easy”, “cheap” etc., irrelevant search results were retrieved.

A way to solve this problem seems to be accurate incorporation of morphological rules, usage statistics as well as, possibly, the semantic analysis into the word segmentation procedure.

# 6. Conclusions and Future Work

In view of the fact that the Japanese iEPG analysis, conducted by the author, has demonstrated considerable presence of Roman letter strings including English words, the author has developed and implemented a new cross-language search method for the Japanese TV guide. Along with query words as they are, the implementation uses transliterations as well as Japanese and English synonyms for the TV guide search. The proposed method implementation demonstrated improved performance.

By examining the evaluation experiment results as well as existing research, the author has arrived at the following conclusions regarding the novelty and effectiveness of the proposed method.

1) The proposed method introduces a novel cross-language search technique using Japanese and English synonyms together with Hiragana, Katakana and Romanized transliterations to expand the Japanese query. Semantic feature analysis for grouping search results has also been introduced. Utilizing these techniques has resulted in a considerable search effectiveness improvement. However, the search precision can be further improved by means of more accurate query paraphrase selection and word segmentation.

2) The proposed method is evidently more effective for searching text with vocabulary in both Japanese and English than the baseline using unilingual synonyms only or than the baseline using no query expansion at all.

3) The proposed method is more precise and on the whole more effective than the baselines using hypernyms or semantic domain words to expand the search query.

4) The proposed method is based on the analysis of the Japanese TV guide for the presence of English words and other Roman letter strings as potential search targets. Such analysis can be definitely considered an effective means of improving the search effectiveness, however no similar analyses have been performed by other researchers to the best of the author’s knowledge.

5) The cross-language search used by the proposed method is, most probably, not used by search systems on major Japanese iEPG.

6) The query expansion by adding Japanese synonyms to the original query, that is

used by the proposed method is, most probably, not used for the Japanese iEPG search by Google.

7) The proposed method introduces applying a bilingual WordNet to the Japanese iEPG cross-language search. This technique can potentially be applied to any text in which Japanese as well as English words are used. No research in this area is known to the author.

8) The proposed method introduces applying a finite state automaton to searching the Japanese iEPG data. No previous research in this area is known to the author. The automaton implementation has demonstrated an increased word matching capability and an approximately 95% decrease in the search time.

In the future, the author plans to improve the proposed search method by integrating new techniques for choosing query paraphrases. Analyzing the phrase structure and the word semantic combinability by utilizing such sources as FrameNet appears to be effective for choosing the right paraphrase.

Moreover, the author intends to conduct experiments in order to find how various semantic relation data that WordNet provides can be more effectively utilized to disambiguate ambiguous query words and generate correct paraphrases for such words.

The author also plans to look into ways of improving word segmentation techniques in order to avoid search noise resulting from wrong segmentation. A way to improve the techniques seems to be not only improving the dictionary used by the morphological analyzer, but also implementing the sematic analysis as a part of the morphological analysis.

# BIBLIOGRAPHY


Andrade, D., Tsuchida, M., Onishi, T., & Ishikawa, K. (2013). Synonym acquisition using bilingual comparable corpora. *International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 1077–1081.

Arita, I., Kikuchi, H., Shirai, K. (2007). Word clustering using concurrent search queries. *IEICE technical report, NLC, Language Understanding and Models of Communication*, 107(158), pp. 115-120.

Aziz, A. D., Cackler J., & Yung, R. (2004). Automata Theory. Eric Roberts' Sophomore College, Stanford University, Retrieved September 29, 2014, from http://www-cs-faculty.stanford.edu/~eroberts/courses/soco/projects/2004-05/automata-theory/basics.html

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

Barabás, P., & Kovács, L. (2012). Efficient encoding of inflection rules in NLP systems. *Scientific Bulletin*, 9, pp. 11-16.

Barr, C., Jones, R., & Regelson, M. (2008). The linguistic structure of English web-search queries. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1021-1030.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009, August). Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 1-8. Association for Computational Linguistics.

Croft, B., Metzler, D., & Strohman, T. (2009). Search Engines: Information Retrieval in Practice, Addison-Wesley.

Eichorn, J. (2006). Understanding AJAX: Using JavaScript to Create Rich Internet Applications. Prentice Hall PTR.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Fujii, A., & Ishikawa, T. (2001). Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities, 35*(4), pp. 389-420.

Fukuta, H., Matsuo, Y., Ishizuka, M. (2002). 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.

Goddard, C. (2002). 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.

Grefenstette, G., Qu, Y., & Evans, D. A. (2004). Mining the web to create a language model for mapping between English names and phrases and Japanese. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference*, pp. 110-116. IEEE.

Hiemstra, D., de Jong, F. M. G. (2001). Statistical language models and information retrieval: Natural language processing really meets retrieval. *Glot international*, 5 (8), pp. 288-293.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pp.2420-2423.

Kay, G. (1995). English loanwords in Japanese. *World Englishes, 14*(1), 67-76.

Kumar, R. (2010). Theory of Automata, Languages and Computation, Tata McGraw-Hill Education.

Kwak, M., Leroy, G., & Martinez, J. D. (2011). A pilot study of a predicate-based vector space model for a biomedical search engine. *Bioinformatics and Biomedicine Workshops (BIBMW)*, 2011 IEEE International Conference on (pp. 1001-1003). IEEE.

Lieber, R. (2004). Morphology and Lexical Semantics, vol. 104. Cambridge University Press.

Lin, D., Church, K. W., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., ... & Narsale, S. (2010). New tools for Web-scale N-grams. In *LREC*.

McCandless, M., & Hatcher, E. (2010). Lucene in Action, Second Edition. Manning Publications.

Nanba, H. (2007). Query expansion using an automatically constructed thesaurus. In *Proceedings of the 6th TCIR Workshop Meeting*, pp. 414-419.

Ohara, K. (2013). Toward constructicon building for Japanese in Japanese FrameNet. *Veredas On-line, Frame Semantics and Its Technological Applications*, 1/2013, pp. 11-27.

Patrick, J., & Sabbagh, M. (2011). An active learning process for extraction and standardisation of medical measurements by a trainable FSA. In *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, pp. 151-162.

Ptaszynski, M., Rzepka, R., Araki, K., & Momouchi, Y. (2012). Annotating syntactic information on 5.5 billion word corpus of Japanese blogs. In *Proceedings of the 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 385-388.

Qu, Y., Grefenstette, G., & Evans, D. A. (2003). Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 353-360. ACM.

Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2010). WNMS: Connecting the distributed Wordnet in the case of Asian WordNet. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010), India*. Narosa Publishing.

Smrz, P., & Schmidt, M. (2009). Information extraction in semantic Wikis. *SemWiki*, 464.

Stanlaw, J. (2004). Japanese English: Language and Culture Contact (Vol. 1). Hong Kong University Press.

Tanaka, K., Nadamoto, A., Kusahara, M., Hattori, T., Kondo, H., & Sumiya, K. (2000). Back to the TV: Information visualization interfaces based on TV-program metaphors. In *Multimedia and Expo*, 2000, ICME 2000, 2000 IEEE International Conference on, Vol. 3, pp. 1229-1232. IEEE.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., & Decker, S. (2010). Sig. ma: Live views on the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4), pp. 355-364.

Wang, K., Thrasher, C.., Viegas, E., Li, X., & Hsu, B. J. P. (2010). 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.

Wierzbicka, A. (1996). Semantics: Primes and Universals. Oxford University Press.

Yamasaki, T., Manabe, T., & Kawamura, T. (2008). Implementation of TV-program navigation system using a topic extraction agent. *Computer Software*, Japan Society for Software Science and Technology, Tokyo, pp. 41-51.

# APPENDIX

## Part 1: Roman Letter Strings Found in the TV Guide

Below is a list of the Roman letter strings found in the Japanese TV guide as a result of the analysis described in 1. 3. A total of 11,836 strings, or 442 string types (tokens) were found. In the list each token is numbered, the token number is followed by a colon. For each token, the number of times it appears in the TV guide text is pointed at by an arrow (--->). The tokens are arranged by the frequency or appearance in the decreasing order.

1: NHK--->3064
2: E--->1707
3: TVH--->992
4: UHB--->855
5: HTB--->746
6: HBC--->734
7: STV--->724
8: TVh--->150
9: SP--->111
10: TV--->107
11: NEWS--->94
12: YOU--->64
13: VS--->56
14: A--->42
15: JNN--->42
16: JAPAN--->40
17: L--->40
18: N--->38
19: News--->36
20: AKB--->33
21: R--->32
22: S--->31
23: min--->31
24: Do--->28
25: EURO--->26
26: TOKIO--->26
27: SONG--->25
28: ANN--->24
29: ZERO--->24
30: neo--->24
31: sports--->24
32: LIVE--->21
33: Q--->21
34: SHELLY--->21
35: U--->21
36: vs--->21
37: B--->20
38: DAIGO--->20
39: FNN--->20
40: Sports--->20
41: V--->20
42: M--->19
43: JUVY--->18
44: Season--->18
45: My--->17
46: SUPER--->17
47: Z--->17
48: DJ--->16
49: LiLiCo--->16
50: PM--->16
51: SHOP--->16
52: m--->16
53: DO--->15
54: GO--->15
55: NMB--->15
56: ZIP--->15
57: D--->14
58: GiRLS--->14
59: IKKO--->14
60: MUSIC--->14
61: THE--->14
62: EXILE--->13
63: Ft--->13
64: J--->13
65: Kis--->13
66: SMAP--->13

67: de--->13
68: Episode--->12
69: KENCHI--->12
70: NNN--->12
71: No--->12
72: TARAKO--->12
73: WEB--->12
74: air--->12
75: man--->12
76: MATSU--->11
77: AIR--->10
78: ALICE--->10
79: C--->10
80: DALE--->10
81: F--->10
82: Jr--->10
83: NACS--->10
84: TEAM--->10
85: BIG--->9
86: CHIE--->9
87: COUNTDOWN--->9
88: ETV--->9
89: SHOW--->9
90: TKO--->9
91: TOYOTA--->9
92: TXN--->9
93: in--->9
94: DASH--->8
95: HUNTER--->8
96: Hit--->8
97: IMALU--->8
98: ISOPP--->8
99: K--->8
100: MAKIDAI--->8
101: MAKO--->8
102: NARUTO--->8
103: NO--->8
104: TAKAHIRO--->8
105: TETSUYA--->8
106: TO--->8
107: USA--->8
108: and--->8
109: com--->8
110: mummy--->8
111: BOSAI--->7
112: Ep--->7
113: FFFFF--->7
114: JR--->7
115: W--->7
116: s--->7
117: BOSS--->6
118: Berryz--->6
119: CHI--->6
120: DMAT--->6
121: DX--->6
122: Dragon--->6
123: FUJIWARA--->6
124: Going--->6
125: Good--->6
126: Job--->6
127: Kei--->6
128: MAKI--->6
129: MEGUMI--->6
130: MEY--->6
131: SL--->6
132: SONGS--->6
133: T--->6
134: WAO--->6
135: YU--->6
136: ZEXAL--->6
137: saku--->6
138: AD--->5
139: AKBINGO--->5
140: BOARD--->5
141: Before--->5
142: Gables--->5
143: Green--->5
144: HKT--->5
145: JUJU--->5
146: MATTER--->5
147: MAX--->5
148: MC--->5
149: Night--->5
150: OK--->5
151: Perfume--->5
152: X--->5
153: or--->5
154: season--->5
155: the--->5
156: AHN--->4
157: Actors--->4
158: BANG--->4
159: BS--->4
160: Break--->4
161: Dramatic--->4
162: File--->4
163: G--->4
164: Girls--->4
165: H--->4
166: I--->4
167: JUN--->4
168: Japan--->4
169: KENN--->4
170: Kids--->4
171: KinKi--->4
172: Little--->4
173: MIKA--->4
174: MISATO--->4

175: May--->4
176: Mr--->4
177: NEO--->4
178: NG--->4
179: Out--->4
180: P--->4
181: PART--->4
182: Sexy--->4
183: Studio--->4
184: TOKIE--->4
185: TOKYO--->4
186: VTR--->4
187: Witch--->4
188: Zone--->4
189: every--->4
190: feat--->4
191: nd--->4
192: news--->4
193: yukarie--->4
194: yukky--->4
195: AAA--->3
196: ABChanZOO--->3
197: Ash--->3
198: BLACK--->3
199: BRAIN--->3
200: BUSAIKU--->3
201: Bee--->3
202: Bogey--->3
203: COUNT--->3
204: Chan--->3
205: Crossroad--->3
206: DIRECTION--->3
207: DOWN--->3
208: Dr--->3
209: ER--->3
210: FAIR--->3
211: FC--->3
212: FNS--->3
213: FOOT--->3
214: Future--->3
215: GT--->3
216: HOMELAND--->3
217: IMF--->3
218: KAT--->3
219: KEIBA--->3
220: LAUNCHER--->3
221: LINE--->3
222: MAP--->3
223: MILLION--->3
224: MONSTER--->3
225: Medical--->3
226: Motion--->3
227: ONE--->3
228: On--->3
229: ROCK--->3
230: Regulus--->3
231: Room--->3
232: SD--->3
233: SEASON--->3
234: SUNDAY--->3
235: SWITCH--->3
236: Sarah--->3
237: Shoppin--->3
238: Standard--->3
239: TBA--->3
240: TM--->3
241: TOWN--->3
242: TRY--->3
243: TUN--->3
244: TVSP--->3
245: Team--->3
246: VI--->3
247: VOICE--->3
248: XY--->3
249: days--->3
250: fun--->3
251: music--->3
252: plus--->3
253: AKIRA--->2
254: ANIMAL--->2
255: Answer--->2
256: At--->2
257: BBR--->2
258: BiS--->2
259: CRUDE--->2
260: CS--->2
261: Circus--->2
262: Classic--->2
263: Cyntina--->2
264: DEEN--->2
265: DVD--->2
266: Dorothy--->2
267: EYELAND--->2
268: Every--->2
269: FINAL--->2
270: FLOWER--->2
271: Fairies--->2
272: Fighters--->2
273: GALETTe--->2
274: GARDEN--->2
275: Get--->2
276: Happy--->2
277: Hemenway--->2
278: ICONIQ--->2
279: IGNITION--->2
280: JOY--->2
281: JPOP--->2
282: JoJo--->2

283: KABA--->2
284: KIZUNA--->2
285: KOBE--->2
286: KiLa--->2
287: LGYankees--->2
288: LIFE--->2
289: LIST--->2
290: Love--->2
291: MALTA--->2
292: MIT--->2
293: MOCO--->2
294: MelodiX--->2
295: Miss--->2
296: Mummy--->2
297: NANA--->2
298: NON--->2
299: OKAMOTO--->2
300: PGA--->2
301: PLAY--->2
302: RED--->2
303: RIKACO--->2
304: ROAD--->2
305: ROLLY--->2
306: RYO--->2
307: Revolution--->2
308: Round--->2
309: SCANDAL--->2
310: SILVA--->2
311: SNOW--->2
312: SQUARE--->2
313: STYLE--->2
314: SWEET--->2
315: Saki--->2
316: Silent--->2
317: Silver--->2
318: Siren--->2
319: SmaSTATION--->2
320: Sound--->2
321: Spoon--->2
322: Step--->2
323: TAKUYA--->2
324: TGC--->2
325: TOMOMI--->2
326: Thing--->2
327: UFO--->2
328: UNISON--->2
329: VIP--->2
330: WAGYU--->2
331: WILL--->2
332: glee--->2
333: hitomi--->2
334: km--->2
335: miwa--->2
336: mummyD--->2
337: ute--->2
338: AC--->1
339: ATM--->1
340: BBQ--->1
341: BET--->1
342: BIGBANG--->1
343: Best--->1
344: Blizzard--->1
345: CDTV--->1
346: CG--->1
347: CL--->1
348: CM--->1
349: COPD--->1
350: CT--->1
351: Communicate--->1
352: DNA--->1
353: DXDX--->1
354: DYMP--->1
355: DeNA--->1
356: EK--->1
357: ELT--->1
358: EXP--->1
359: Egirls--->1
360: FBS--->1
361: FOR--->1
362: Final--->1
363: For--->1
364: Friend--->1
365: GCG--->1
366: GOLD--->1
367: GOODSUN--->1
368: GP--->1
369: GTO--->1
370: HD--->1
371: HOUSE--->1
372: HRK--->1
373: HSP--->1
374: IMAT--->1
375: IN--->1
376: IQ--->1
377: Invisible--->1
378: JSB--->1
379: KARAOKE--->1
380: KO--->1
381: KOBAN--->1
382: LG--->1
383: LOVE--->1
384: LR--->1
385: Lesson--->1
386: Let--->1
387: MBS--->1
388: MIB--->1
389: Made--->1
390: NPO--->1

391: NPS--->1

392: NW--->1

393: NY--->1

394: OG--->1

395: OHA--->1

396: PARK--->1

397: PC--->1

398: PHASE--->1

399: POP--->1

400: PRECIOUS--->1

401: PRINCESS--->1

402: Part--->1

403: Qtube--->1

404: REIJI--->1

405: Rainbow--->1

406: Rose--->1

407: SF--->1

408: SKE--->1

409: SNS--->1

410: SPORTS--->1

411: STUDY--->1

412: SexyZone--->1

413: Sexyzone--->1

414: Someone--->1

415: TBS--->1

416: TMR--->1

417: TORE--->1

418: TPD--->1

419: TPP--->1

420: TRICK--->1

421: Tube--->1

422: UP--->1

423: Uta--->1

424: VBA--->1

425: Vol--->1

426: WBC--->1

427: Wish--->1

428: YMO--->1

429: because--->1

430: by--->1

431: face--->1

432: ing--->1

433: mol--->1

434: my--->1

435: natane--->1

436: of--->1

437: piece--->1

438: presents--->1

439: rd--->1

440: shigure--->1

441: to--->1

442: vol--->1

TTL count of strings: 11836

TTL count of string types: 442

# Part 2: Roman Letter String Types Matching English Words in WordNet

Below respective counts for the string types (tokens) matching English words in WordNet 3.0 (the English version), and for those not matching the English words are given. The total of 442 tokens found in the TV guide was matched. Within that number, 160 tokens did not match the English words and 282 tokens did. Each token listed below is numbered, the number appearing in brackets before the token.

Number of items processed: 442

Number of string types not found in English Wordnet 3.0: 160

String types not found: (1)NHK (2)TVH (3)UHB (4)HTB (5)HBC (6)STV (7)TVh (8)SP (9)YOU (10)VS (11)JNN (12)AKB (13)ANN (14)SHELLY (15)vs (16)DAIGO (17)FNN (18)JUVY (19)My (20)LiLiCo (21)NMB (22)IKKO (23)THE (24)SMAP (25)KENCHI (26)NNN (27)TARAKO (28)MATSU (29)ALICE (30)CHIE (31)ETV (32)TXN (33)IMALU (34)ISOPP (35)MAKIDAI (36)NARUTO (37)TAKAHIRO (38)TETSUYA (39)TO (40)and (41)com (42)BOSAI (43)Ep (44)FFFFF (45)Berryz (46)DMAT (47)DX (48)FUJIWARA (49)Kei (50)MAKI (51)MEGUMI (52)MEY (53)WAO (54)YU (55)ZEXAL (56)saku (57)AKBINGO (58)HKT (59)the (60)AHN (61)JUN (62)KENN (63)KinKi (64)MIKA (65)MISATO (66)TOKIE (67)VTR (68)yukarie (69)yukky (70)ABChanZOO (71)BUSAIKU (72)Chan (73)Dr (74)FC (75)FNS (76)GT (77)KEIBA (78)Shoppin (79)TBA (80)TVSP (81)AKIRA (82)BBR (83)Cyntina (84)DEEN (85)Dorothy (86)EYELAND (87)GALETTe (88)Hemenway (89)ICONIQ (90)JPOP (91)JoJo (92)KABA (93)KIZUNA (94)KiLa (95)LGYankees (96)MOCO (97)MelodiX (98)NANA (99)OKAMOTO (100)PGA (101)RIKACO (102)ROLLY (103)RYO (104)SmaSTATION (105)TAKUYA (106)TGC (107)TOMOMI (108)WAGYU (109)hitomi (110)miwa (111)mummyD (112)BBQ (113)BIGBANG (114)CDTV (115)CG (116)COPD (117)DXDX (118)DYMP (119)DeNA (120)EK (121)ELT (122)EXP (123)Egirls (124)FBS (125)FOR (126)For (127)GCG (128)GOODSUN (129)GTO (130)HD (131)HRK (132)HSP (133)IMAT (134)JSB (135)KOBAN (136)LG (137)NPO (138)OG (139)OHA (140)Qtube (141)REIJI (142)SF (143)SKE (144)SexyZone (145)Sexyzone (146)TMR (147)TPD (148)TPP (149)VBA (150)Vol (151)YMO (152)because (153)ing (154)my (155)natane (156)of (157)rd (158)shigure (159)to (160)vol

Number of word types found in English Wordnet 3.0: 282

Word types found: (1)E (2)TV (3)NEWS (4)A (5)JAPAN (6)L (7)N (8)News (9)R (10)S (11)min (12)Do (13)EURO (14)TOKIO (15)SONG (16)ZERO (17)neo (18)sports (19)LIVE (20)Q (21)U (22)B (23)Sports (24)V (25)M (26)Season (27)SUPER (28)Z (29)DJ (30)PM (31)SHOP (32)m (33)DO (34)GO (35)ZIP (36)D (37)GiRLS (38)MUSIC (39)EXILE (40)Ft (41)J (42)Kis (43)de (44)Episode (45)No (46)WEB (47)air (48)man (49)AIR (50)C (51)DALE (52)F (53)Jr (54)NACS (55)TEAM (56)BIG (57)COUNTDOWN (58)SHOW (59)TKO (60)TOYOTA (61)in (62)DASH (63)HUNTER (64)Hit (65)K (66)MAKO (67)NO (68)USA (69)mummy (70)JR (71)W (72)s (73)BOSS (74)CHI (75)Dragon (76)Going (77)Good (78)Job (79)SL (80)SONGS (81)T (82)AD (83)BOARD (84)Before (85)Gables (86)Green (87)JUJU (88)MATTER (89)MAX (90)MC (91)Night (92)OK (93)Perfume (94)X (95)or (96)season (97)Actors (98)BANG (99)BS (100)Break (101)Dramatic (102)File (103)G (104)Girls (105)H (106)I (107)Japan (108)Kids (109)Little (110)May (111)Mr (112)NEO (113)NG (114)Out (115)P (116)PART (117)Sexy (118)Studio (119)TOKYO (120)Witch (121)Zone (122)every (123)feat (124)nd (125)news (126)AAA (127)Ash (128)BLACK (129)BRAIN (130)Bee (131)Bogey (132)COUNT (133)Crossroad (134)DIRECTION (135)DOWN (136)ER (137)FAIR (138)FOOT (139)Future (140)HOMELAND (141)IMF (142)KAT (143)LAUNCHER (144)LINE (145)MAP (146)MILLION (147)MONSTER (148)Medical (149)Motion (150)ONE (151)On (152)ROCK (153)Regulus (154)Room (155)SD (156)SEASON (157)SUNDAY (158)SWITCH (159)Sarah (160)Standard (161)TM (162)TOWN (163)TRY (164)TUN (165)Team (166)VI (167)VOICE (168)XY (169)days (170)fun (171)music (172)plus (173)ANIMAL (174)Answer (175)At (176)BiS (177)CRUDE (178)CS (179)Circus (180)Classic (181)DVD (182)Every (183)FINAL (184)FLOWER (185)Fairies (186)Fighters (187)GARDEN (188)Get (189)Happy (190)IGNITION (191)JOY (192)KOBE (193)LIFE (194)LIST (195)Love (196)MALTA (197)MIT (198)Miss (199)Mummy (200)NON (201)PLAY (202)RED (203)ROAD (204)Revolution (205)Round (206)SCANDAL (207)SILVA (208)SNOW (209)SQUARE (210)STYLE (211)SWEET (212)Saki (213)Silent (214)Silver (215)Siren (216)Sound (217)Spoon (218)Step (219)Thing (220)UFO (221)UNISON (222)VIP (223)WILL (224)glee (225)km (226)ute (227)AC (228)ATM (229)BET (230)Best (231)Blizzard (232)CL (233)CM (234)CT (235)Communicate (236)DNA (237)Final (238)Friend (239)GOLD (240)GP (241)HOUSE (242)IN (243)IQ (244)Invisible (245)KARAOKE (246)KO (247)LOVE (248)LR (249)Lesson (250)Let (251)MBS (252)MIB (253)Made (254)NPS (255)NW (256)NY (257)PARK (258)PC (259)PHASE (260)POP (261)PRECIOUS (262)PRINCESS (263)Part (264)Rainbow (265)Rose (266)SNS (267)SPORTS (268)STUDY (269)Someone (270)TBS (271)TORE (272)TRICK (273)Tube (274)UP (275)Uta (276)WBC (277)Wish (278)by (279)face (280)mol (281)piece (282)presents

# Part 3: Test Queries

Listed below are the 30 search queries used for testing the proposed method implementation. Each query is followed by an English translation in brackets. The queries were obtained by means of a questionnaire, from five female individuals from 19 to 77 years of age and ten male ones from 19 to 84, i.e. 15 individuals in total. The questioned individuals have not been involved in the present research.

1) 今日のニュース (today's news);
2) 子供が好きなアニメ (cartoons that children like);
3) おかしい風景 (weird scenery);
4) 和服の作り方 (how to make Japanese-style clothes);
5) Gacktのいきつけ (places Gackt often visits);
6) 女子が入りやすい居酒屋 (pubs a female feels comfortable to enter);
7) 札幌にはない食べ物 (food not found in Sapporo);
8) 人気な温泉 (popular spas);
9) 札幌の人気なレストラン (popular restaurants in Sapporo);
10) 北海道の食べ物 (Hokkaido food);
11) AKB48のコンサート (AKB48 concerts);
12) 北海道のニュース (Hokkaido news);
13) ハイキングコースがある地域 (areas with hiking trails);
14) 山登りができる場所 (places you can do mountain climbing);
15) 冷たいビール (cold beer);
16) 日帰り温泉 (one-day trip spas);
17) 磯田彩実が出ている番組 (programs with Ayami Isoda );
18) 水曜はどうでしょう(an English translation found in the Internet: "How do you like Wednesday?" );
19) 暖かい地域の旅 (travelling in a warm place);
20) 良いニュース (good news);
21) 大きな経済の動き (big economy changes);
22) コンサドーレの勝敗 (Consadole game results);
23) SMAPは旅をする (SMAP goes on a trip);
24) 札幌の天気 (Sapporo weather);
25) ダンスのイベント (dancing events);

26) ドラマ (drama);
27) 映画 (movie);
28) スポーツ (sports);
29) 野球 (baseball);
30) サッカー (soccer).

# ACKNOWLEDGEMENTS

I would like to express deep appreciation to the associate examiners Prof. Miki Haseyama (Laboratory of Media Dynamics, Graduate School of Science and Technology, Hokkaido University) and Prof. Tsuyoshi Yamamoto (Laboratory of Information Media Environment, Graduate School of Science and Technology, Hokkaido University).

I would like to express great appreciation to my doctoral course academic advisers Prof. Kenji Araki and Asst. Prof. Rafal Rzepka (Language Media Laboratory, Graduate School of Science and Technology, Hokkaido University) for their guidance and support in the field of natural language processing.

I would like to express deep gratitude to my master's course academic adviser Prof. Yoshihiko Ono (Graduate School of Letters, Hokkaido University) for his guidance and support in the field computational linguistics.

I wish to acknowledge the help provided by Assoc. Prof. Yuzu Uchida (Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai-Gakuen University).

Advice given by Assoc. Prof. Yasutomo Kimura (Department of Information and Management Science, Otaru University of Commerce) has been a great help.

I would also like to thank all my fellow students who helped me in various ways.

# RESEARCH ACHIEVEMENTS

**Academic Journal Publication:**

- Denis Kiselev, Rafal Rzepka and Kenji Araki: “Matching Word-Order Variations and Sorting Results for the iEPG Data Search”, International Journal of Multimedia Data Engineering and Management, 5(1), pp. 52-64, January-March 2014.
  DOI: 10.4018/ijmdem.2014010104

**International Conference Proceedings Publications:**

- Denis Kiselev, Rafal Rzepka and Kenji Araki. “Applying the Stop List and Part of Speech Analysis to Processing the IEPG Search Query”. Proceedings of the 14th Conference of the Pacific Association For Computational Linguistics 2013, Tokyo, Japan (Paper Sep3-8)

- Denis Kiselev, Rafal Rzepka and Kenji Araki: “Improving Search Query Matching for Electronic TV Program Guide Data Extraction”, in the Proceedings of 2013 IEEE Seventh International Conference on Semantic Computing, pp. 146-149, Irvine, USA, 2013.

**Award:**

- Excellent Presentation Award at Hokkaido University Symposium for Young Researchers, 2013