Methods of Extracting Year References for Chronological-table-generating Text Searching

Yasusi Kanada

Central Research Laboratory, Hitachi Ltd.
Higashi-Koigakubo 1-280, Kokubunji, Tokyo 185, Japan
E-mail: kanada@crl.hitachi.co.jp

Abstract

A method of extracting year references for a textual information retrieval method called the thematic chronological-table search method is explained in this paper. This search method generates an index by extracting and collecting year references from a text collection. The resulting index and a full-text index are used for searching sentences that contain year references and search words. The results are displayed in the form of a chronological table with hyperlinks to the original text.

Seven forms of year or century references are extracted and normalized using string matching patterns. The extraction error rate is reduced by using both local and non-local contexts. If the lower two digits of a Gregorian year, which matches a form, occurs, it is normalized by supplementing the upper digits using the non-local context. This method has been applied to a Japanese encyclopedia. An evaluation shows the precision of extraction to be higher than 99% in most cases.

Keywords

Full-text search, Number extraction, Chronological-table, Encyclopedia, Hypertext, Information extraction, Information retrieval.

1 Introduction

New methods of searching text through which end users can find desired information by using a simple input, and through which they can discover knowledge distributed in a large volume of text will soon be needed as interest in the Internet grows and as more CD-ROM contents are developed. In our attempts to address these needs, we have developed the axis-specified search method [Kan 98] [Kan 98a]. In this method, the user specifies words in the same way as in conventional full-text search. The difference, however, is that the user also selects an axis from a menu. This generates a search-result list ordered along the specified axis. This method allows two or more topics described in a document to be put at the different coordinates on the axis so that the search results on these topics can be extracted separately and sorted. A more fine-grained search is therefore made possible by the axis-specified search.

Thematic chronological-table searching is axis-specified searching with a year axis, and thematic geographical searching [Kan 99] is axis-specified searching with a geographical axis. Hitachi Digital Heibonsha[1] provides these services as a part of the members-only network service called “Encyclopædia on the Net” (“ネットで百科” in Japanese) for users wishing to search the World Encyclopædia [HDH 98]. This service is a first step toward implementing axis-specified searching of encyclopedia text on various axes. In implementing chronological-table searching, however, we have to develop a very precise method of extracting year references. The next section outlines the thematic chronological-table search. The method of extracting year references is explained in Section 3. Examples of extraction errors are shown in Section 4, and the precision is evaluated in Section 5.

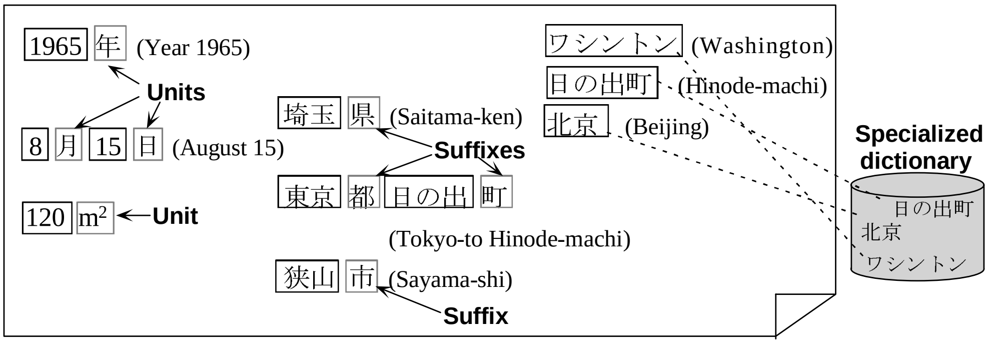
2 Outline of Thematic Chronological-table Search

The thematic chronological-table search method is part of the axis-specified search method. The axis-specified search method is overviewed here, and then the function and its method of implementation are explained.

In an axis-specified search, the user selects an axis, and inputs keywords. The keywords represent the theme of the search, and the axis specifies the general-purpose method of ordering the search result. The search results are ordered along the axis. The result items are distributed in a space that is specified by the axis. In a conventional clustering-based method for organizing search results, a criterion for the organization is selected by the system. However, because the candidates of the axis, i.e., the criterion for organization, is selected by the user in the axis-specified search, the results are ordered just as the user intended. The candidates of the axis are predefined by the search system. In the thematic chrono-

[*] Errata and an updated version of this paper will be available from http://www.kanadas.com/Papers/-Search-papers-j.html#ISDL99.

[1] A Japanese company that publishes CD-ROM and networked versions of encyclopedias (http://www.hdh.co.jp/).



(a) Quantities with specified units

(b) Words with specified word tails or heads

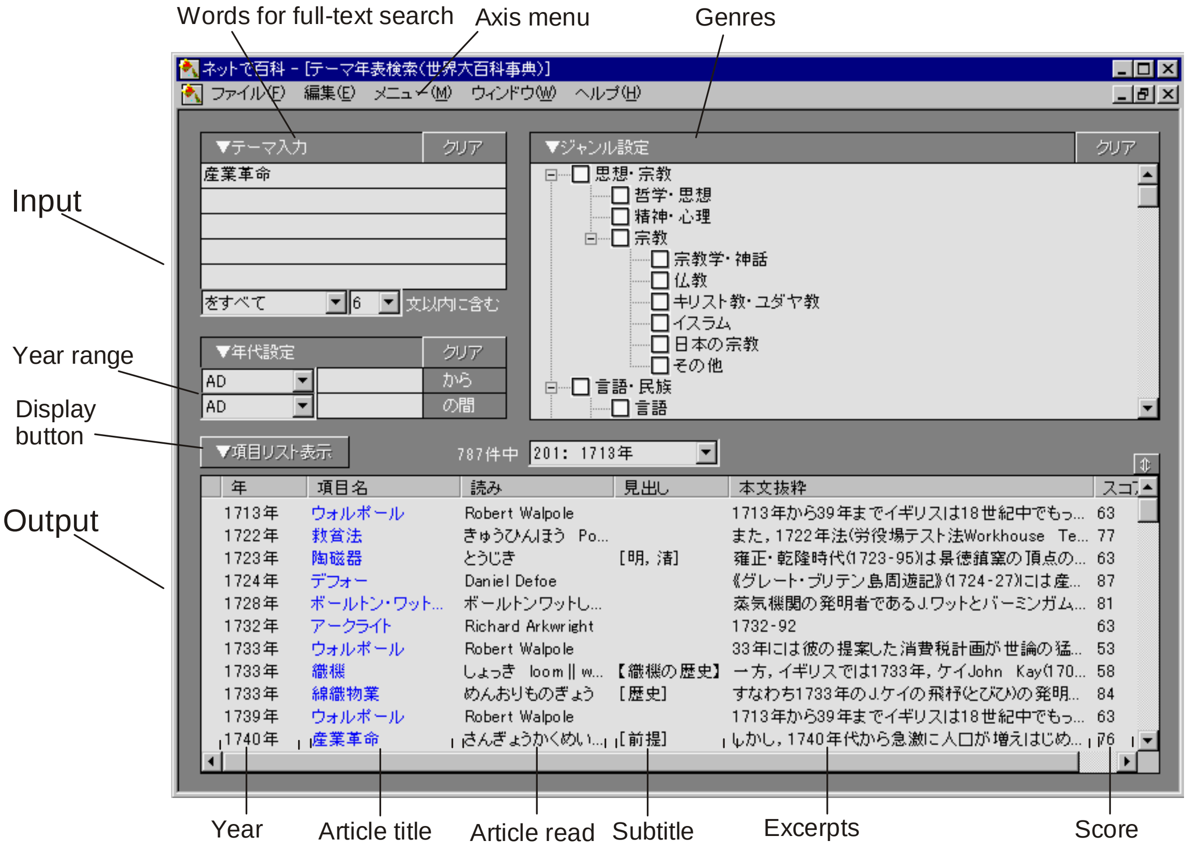
(c) Word category

Figure 1. Three types of axis specification in axis-specified searching

logical-table search, the axis "year" is fixed when the user chooses "thematic chronological-table search" from a menu. The range on the axis can also be specified by the user. When the range is specified, result items out of the range are eliminated.

There are three methods of specifying the axis: specification by the unit of a quantity, specification by the word tail, and specification by the word category (**Figure 1**). Searching in which the axis is specified by the unit of a quantity is called *quantity searching*. Thematic chronological-table searching is a type of quantity searching. Each sentence that is close to both a year reference and a search word is retrieved, the sentences are sorted by year, and the result is displayed in the form of a chronological table. **Figure 2** shows the user interface used for the Internet search service of the World Encyclopædia. This interface was developed by Hitachi Digital Heibonsha and runs on Microsoft Windows and Windows NT. The figure shows an example of search for "industrial revolution". The user can generate a chronological table on a desired theme dynamically.

A query is specified by a combination (conjunction) of search words (with a specification of "and" / "or") and year range (input by Gregorian or Japanese year). If only the search words are specified, sentences close to the search words on all years are collected. And if only the year range is specified, all



産業革命

をすべて 6 文以内に含む

AD から
AD の間

□ 思想・宗教
□ 哲学・思想
□ 精神・心理
□ 宗教
□ 宗教学・神話
□ 仏教
□ キリスト教・ユダヤ教
□ イスラム
□ 日本の宗教
□ その他
□ 言語・民族
□ 言語

787件中 201: 1713年

年	項目名	読み	見出し	本文抜粋	スコ
1713年	ウォルポール	Robert Walpole		1713年から39年までイギリスは18世紀中でもっ...	63
1722年	救貧法	きゅうひんほう Po...		また，1722年法(労役場テスト法Workhouse Te...	77
1723年	陶磁器	とうじき	[明，清]	雍正・乾隆時代(1723-95)は景徳鎮窯の頂点の...	63
1724年	デフォー	Daniel Defoe		《グレート・ブリテン島周遊記》(1724-27)には産...	87
1728年	ボールトン・ワット...	ボールトンワットし...		蒸気機関の発明者であるJ.ワットとバーミンガム...	81
1732年	アークライト	Richard Arkwright		1732-92	63
1733年	ウォルポール	Robert Walpole		33年には彼の提案した消費税計画が世論の猛...	53
1733年	織機	しょっき loom‖w...	【織機の歴史】	一方，イギリスでは1733年，ケイJohn Kay(170...	58
1733年	綿織物業	めんおりものぎょう	[歴史]	すなわち1733年のJ.ケイの飛杼(とびひ)の発明...	84
1739年	ウォルポール	Robert Walpole		1713年から39年までイギリスは18世紀中でもっ...	63
1740年	産業革命	さんぎょうかくめい...	[前提]	しかし，1740年代から急激に人口が増えはじめ...	76

Figure 2. The user interface for the thematic chronological-table search

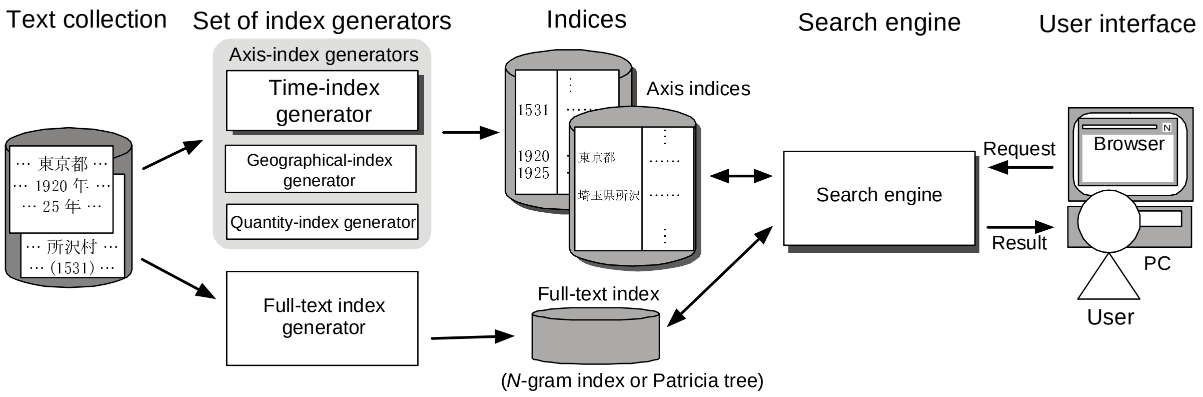


Figure 3. Outline of system structure for axis-specified searches

the sentences in the year range are collected. If these inputs are combined, the number of search results can be effectively limited. The year range is specified by Gregorian year, but it can be inputted by Japanese year, such as the *n*th year of Heisei, and converted by using another window. The result can also be limited by specifying genres in "Encyclopædia on the net".

Each output item contains a year, a sentence extracted from the text, and a hyperlink to the original sentence. By using an optional input, the user can choose whether sentences that contain a year reference or the search word are to be displayed. (The sentences in Figure 2 contain year references because "Year" was chosen as an option.) The user can also choose year, century, or both as the unit of time to be searched. Units such as month, day, or hour are not as important as year or century in an encyclopedia, so they are not collected in the current version of the system.

A hyperlink into the original text is embedded in each row of the chronological table. Thus, the encyclopedia article can be displayed by a Web browser, clicking on a row causes that sentence to be displayed and highlighted at the top of another window that pops up, along with the article in which that sentence appears. If the user scrolls the window, the user can see the whole topic that contains the sentence or whole article.

The implementation of thematic chronological-table searching is now explained. The search server consists of an index generator and a search engine (**Figure 3**). The index generator generates the year-axis index and the full-text index before the user query is made. The year-axis index generator, part of the index generator, extracts character strings that match predefined patterns of year references, normalizes them, and enters them into the year-axis index. The method of information extraction to generate the year-axis index is explained in Section 3. The time required for the search is drastically reduced by using the year-axis index. The full-text index generator, another part of the index generator, generates the full-text index, which is similar to conventional *N*-gram search indices. The unit of the full-text search is a sentence, but long sentences are partitioned at particular commas. The number of resulting "sentences" is approximately 2.7 million.

The search engine invoked by the user searches the year-axis index for sentences that are close to the year references within the specified range. It also searches the full-text index for the specified words in the sentences. It then filters and sorts the results by year, and outputs them. The search results are scored, and those scored too low are dropped.

A score is computed using the distance between the year reference and the nearest word when the search engine is invoked by specifying both an axis and search words.

3 Method of Extracting Year References

3.1 Outline of year references extraction

The index generator inputs all the documents and extracts character strings that match predefined patterns in the axis-specified search. Strings are matched character by character. No natural language processing methods, such as morpheme analysis, are not used. Morpheme analysis is not used for two reasons. First, there are few advantages to morpheme analysis in extracting numbers. Second, a simpler method was preferred in our development because time was limited. In some cases, however, a year reference can be more easily and precisely extracted by using morpheme analysis.

In the axis-specified search, a set of matching patterns is defined for each axis. Extracted strings are normalized and entered into the axis index. In a thematic chronological-table search, some year references are extracted using context-free rules. However, context-sensitive rules are required for extracting certain year references such as abbreviated Gregorian years. Matching patterns and the normalization method for matched strings must be customized to the text type, whether it be encyclopedia-, newspaper-, or Web-based.

The following forms of year references are extracted.

1. One to four digits of Gregorian years followed by “年” (which means “year”). E.g., “1989年.”
2. The lower two digits of Gregorian years followed by “年”, e.g., “89年”, which means the year 1989.
3. One to two digits of Japanese years that are preceded by an era name and followed by “年”. E.g., “平成10年” (10th year of the Heisei era).
4. “…000年前” or “…万年前”, where “年前” means “years ago” and “万年前” means “tens of thousands of years ago.”
5. A parenthesized year. E.g., “ロシア革命 (1917)” (Russian Revolution (1917))
6. Years of birth and/or death. E.g., アインシュタイン Albert Einstein 1879‐1955.
7. “…世紀” (… century A.D.) or “前…世紀” (… century B.C.).

Forms 1 to 4 and form 7 can also be applied to texts other than the World Encyclopædia. Applying forms 5 and 6 to other texts is highly likely to produce garbage. Although most of the above patterns can only be used for Japanese, the patterns in 1, 2, 4, 5, and 7 could be modified and applied to another language, such as English. This will be discussed again later.

Extracted years are normalized to Gregorian years. For example, the first two digits (or one digit) of a year reference in the second form is supplemented by using a preceding four-digit Gregorian year reference. For a year reference before Christ, a negative value is used for the internal representation. Because the year before 1 A.D. is 1 B.C., zero is not used.

The expression “… 時代” (… era), such as “弥生時代” (Yayoi era) or “江戸時代” (Edo era) can be extracted, but it is not extracted in the current version because these expressions usually accompany Gregorian year references in the encyclopedia, and thus their extraction is redundant.

3.2 Extraction of year references that have “年” (year)

The method of extracting numbers followed by “年” (year) but not followed by “年前” (years ago) is explained using some examples.[1]

- References that have a Japanese year name

  If a Japanese year name such as “大化” (Taika) or “平成” (Heisei) precedes the number, the number is interpreted as a Japanese year, and it is converted to a Gregorian year. For example, from the expression “昭和 60 年代まで 4 鉱山があった” (Four mines existed until sometime between the 60th and 69th year of the Showa era), the year 1985 (60th year of the Showa era) is extracted. Some year names other than Japanese ones, e.g., Chinese year names, are also recognized as year names, but currently they are not extracted.[2] For example, the following strings are regarded as year names.

  太始, 嘉永, 建元, …, 延宝, 天武, 隆興.

- References that are preceded by “**後**” (A.D.)

  “後”, which precedes the number, usually means A.D. Thus, this type of year reference is interpreted as a Gregorian year. For example, from the expression “後 70 年にローマ人が第 2 神殿を破壊する” (Romans destroyed the second shrine in 70 A.D.), the year 70 is extracted. However, the numbers coming after “その後” (after that), “この後” (after this), “帰国後” (after returning to home country), “建国後” (after building the country), “感染後” (after the infection), and so on, are not regarded as Gregorian years, and are not extracted.

- References that are preceded by “**前**” (B.C.)

  “前”, which precedes the number, usually means B.C. Thus, the negative number is extracted as the Gregorian year. For example, from the expression “前184年の〈大カトーのバシリカ〉” (the basilica of Cato Major in 184 B.C.), the year –184 is extracted.

- References that are followed by a word that implies a period

  When a word that apparently implies a period of years immediately follows “年”, it is not extracted. For example, no year reference is extracted from “10年間” (10 years). Examples of words that imply periods of years are listed:

  前半 (first half), 後半 (last half), 半 (half), 来 (from), 目 (-th), に及ぶ (as long as), に一度 (once a …), に<*n*>回 (<*n*> times a …), の歴史 (… years of history), の伝統 (… years of tragedy), …, 生 (born in …), 強 (greater than), 弱 (less than).

  However, even if no such word follows “年”, the number may still mean a period of years. Consequently, extraction errors cannot be completely avoided.

- References that are preceded by a word that implies a year relative to another year

  In this case, the year expressions in the text are not regarded as year references, and they are not extracted. For example, nothing is extracted from the expression “独立200 年” (200 years after in-

[1] All examples in this paper are quoted from the World Encyclopædia, 2nd version [HDH 98]. The underline has been added by the author of this paper.

[2] Non-Japanese year names with two-digit year numbers should be recognized for the sake of reducing the error rate of Gregorian years.

dependence). Examples of words that imply relative years are: “齢” (age), “没” (death), “ほぼ” (approximately), “生誕” (birth).

- References that have three to four digits

  When a word that implies range does not follow the digits, a word that implies non-year does not precede them, and the number of digits is three to four, the value is interpreted as a Gregorian year, and it is extracted as is.

- References that have two digits

  When year reference that have two digits occurs after a year reference that have three or four digits, the first two digits of the latter are prefixed to the former. For example, when “64年” (the year 64) appears after “1960年” (the year 1960), the former is regarded as the year 1964. In the World Encyclopædia, more than 99% of two-digit year references are correctly supplemented.

- References that have a century and a two-digit year

  A year is extracted from a reference that contains a century and a two-digit year, such as “16世紀の80年” (80th year in the 16th century), only when the string between them matches predefined patterns.

- References that have an interval

  In expressions that have an interval, such as “1960年から80年” (from 1960 to ’80) or “60〜80年” (from ’60 to ’80), the beginning year (usually the first half) is always to be extracted, but the ending year (the last half) may be discarded.

A pattern of one to four digits followed by “年” can only be used for Japanese. However, Gregorian years in other languages may be extracted in a similar method. For example, Gregorian years in English may be extracted using the pattern of one to four digits preceded by “in”.

3.3 Extraction of year references that have “年前” (years ago)

There are three cases of extracting numbers followed by “年前” (years ago).

- References for years not in units of 10,000

  If the year value is not in units of 10,000 but is more than 10,000, the Gregorian year to be extracted is 2000 minus the value. This is because we are now approximately at 2,000 A.D. For example, the expression “1万 5000年前” ((approximately) fifteen thousand years ago) is interpreted as 13,000 B.C. (–13000).

- Reference for years in units of 10,000

  The Gregorian year to be extracted is a negative year value. 2000 is not added because the fourth digit is not significant. For example, the expression “1万年前” ((approximately) ten thousand years ago) is interpreted as 10,000 B.C. (–10000).[1]

- References that have an interval

  If the year reference contains an interval, such as “約1万〜1万5000年前” (approximately 10,000 to 15,000 years ago), the older year value is extracted. In this case, –13000 is extracted.

If the language is English, this type of year reference may also be extracted using the pattern of a number followed by “years ago”.

3.4 Extraction of year references enclosed by parentheses

If a year reference in a pair of parentheses matches one of the patterns explained in Sections 3.1 to 3.3, the method for that pattern is applied. However, a parenthesized number without “年” (year) is also extracted in the following manner. Numbers between 57 to 2100 that are enclosed by parentheses are extracted from the encyclopedia as years. For example, the year 1917 is extracted from the expression “ロシア革命 (1917)” (Russian Revolution (1917)). A value less than 57 or greater than 2100 is not extracted because such a value usually means a number that does not mean a year in the World Encyclopædia (See Section 4.3). Expressions such as “(1)” or “(2)”, especially, are often used for itemization.[2]

This year reference pattern can probably be used for other languages including English, but whether it is possible depends on the text type.

3.5 Extraction of years of birth/death

Extraction of years of birth and/or death from encyclopedia articles on a specific person is explained here. In the SGML text of the World Encyclopædia, years of birth/death are parenthesized by special tags. The strings between these tags are therefore extracted as years of birth/death when extracting a year reference. The years of birth/death are often fuzzy, and they are expressed using “?”, “ころ” (about), “か” (or), “以前” (before), or “以後” (after). For example, the following expressions can be seen.

1. “行信　ぎょうしん　？– 752？ (天平勝宝 4？)” (Gyoshin ? – 752?) ⇒ The year 752 is extracted.

[1] If “1万年前” ((approximately) ten thousand years ago) means about 8,000 B.C., i.e., two digits are significant, this interpretation is not correct. (See the next item.) In addition, “1万1000年前” ((approximately) eleven thousand years ago) is probably older than “1万年前” ((approximately) ten thousand years ago), but the order is reversed in the extracted information.

[2] The numbers between 57 and 2100 are extracted in our current system except for a few exceptional cases. However, there are non-years even in this range, and they may be wrongly extracted.

2. “ストラボン　Strabon　前64か63 – 後23ころ” (Strabon 64 B.C. to 23 A.D.) ⇒ The years –64 and 23 are extracted, but the year –63 is not extracted.

In our system, it is asserted that there is no abbreviation in the description of a year of birth. However, the year of death is often described by two-digit abbreviated form. So it is supplemented using the year of birth. If “前” (before or B.C.) precedes the number, the number is negated.

3.6 Extraction of year references that have “世紀” (century)

If a number that consists of one or two digits precedes “世紀” (century), e.g., “20世紀” (20th century), the number is regarded as a year reference, and is extracted. If “前” (before) precedes the number, e.g., “前2世紀” (the second century B.C.), the number is negated because the century is B.C. However, if a word that implies an interval follows “世紀” (century), it is not extracted. Zero is not used for representing a century.

4 Examples of Year Extraction Errors

As described in the previous section, detailed rules and exceptions for extracting year references are defined to avoid extraction errors. However, extraction errors cannot always be avoided. Examples of extraction errors are shown in this section. Note that some of the problems below have already been solved.

4.1 Numbers followed by “年” (year)

The following expressions may be wrongly extracted.

- References that have “後” (after) before the number

  In the expression “アメリカで学んだ後73年からブッパータール舞踊団で活躍し” (studied in America, and was active in Wuppertal Dance Group since the year 73), “73年” means the year 1973, but it was extracted as 73 A.D. The possibility that “後73年” means “73 A.D.” cannot be ignored unless the semantics of the sentence are analyzed.

- References that have “前” (before) before the number

  In the expression “変動所得については前2年の変動所得の平均額を超える額であり” (the transitory income exceeds the average transitory income of the preceding 2 years), “前2年” means an interval of two years, but it was extracted as 2 B.C.

- References that have a word implying a period following the number

  Years are not extracted from expressions such as “…年の歴史” (… years of history) or “…年の伝統” (… years of tragedy). However, if there are other words between “年” (year) and a word that implies periods of years (history, life, etc.), an extraction error may occur. For example, in the expression “ほぼ100年のハイネ受容史” (the almost 100-year history of accepting Heine), “史” (history) implies relative years, but two words exist between the year and “史”.

It is more difficult to eliminate errors in two-digit Gregorian year references than in three- or four-digit year references. In the latter, the precision can be improved by sacrificing the recall by suppressing the years. However, this may increase the error rate of two-digit year references that depend on the three- or four-digit year reference, and thus reduce the precision. For example, if “1900年” (the year 1900) occurring after “1890年” (the year 1890) is not extracted for some reason, “20年” (the year 20), which appears after “1900年” and actually means 1920, is normalized to 1820.

4.2 Numbers followed by “年前” (years ago)

Even when “年前” (years ago) follows the number, “前” may be a part of a word. For example, “1866年前橋に移封” (transferred to Maebashi (前橋) in 1866), or “33年前衛美術家・建築家の集団〈ユニット・ワンUnit One〉を結成し” (formed a group of avantgarde (前衛) artists and architects called “Unit One” in the year 33). It is important not to make extraction errors in such expressions. This problem would probably not occur if morpheme analysis was used.

4.3 Years enclosed in parentheses

The following numbers enclosed by parentheses may be wrongly extracted.

1. In the expression “…, (110), (111) の $2^3 = 8$ 通りとなる” (The number of messages, …, (110), and (111), is $2^3 = 8$), 110, 111, and others were wrongly extracted as Gregorian years.
2. In the expression “社民党 (161), 左翼党 (22), 緑党 (18), キリスト教民主社会 (15), 自由党 (26), 中央党 (27), 保守党 (80)” (Social Democratic Party (161), Left Wing Party (22), Green Party (18), Gregorian Democratic Social (15), Liberal Party (26), Central Party (27), Conservative Party (80)), where the parenthesized numbers mean the number of seats, were extracted as Gregorian years.
3. In the expression “長寿を祝う年祝には, 還暦 (61), 古希 (70), 喜寿 (77), 米寿 (88) などがある” (Celebrations of long life includes the sixtieth (61), seventies (70), seventy-seventh (77), and eighty-eighth (88) birthday celebration, and so on), the ages in parentheses were extracted as Gregorian years.

Table 1. Evaluation of year extraction

Search words	Genre	Year range	Year or century	Distance in sentences*	Search-result items	Extraction errors	Precision
アメリカ (America)	Philosophy, Ideology or Religion	All	Year only	4	819	1	0.999
アメリカ (America)	[not specified]	All	Century only	0	632	0	1.000
徳川 (Tokugawa)	[not specified]	All	Year only	0	653	0	1.000
生命 (Life)	[not specified]	All	Year only	5	832	4	0.995
[not specified]	[not specified]	to 5000 B.C.	Year only	-	984	0	1.000
[not specified]	[not specified]	from 2000 A.D.	Year only	-	409	6	0.985
[not specified]	[not specified]	from 1 A.D. to 100 A.D.	Year only	-	276	27	0.902

* This is the maximum number of sentences between the search word and the year references.

4.4 Numbers followed by "世紀" (century)

The following expressions may be wrongly extracted.

1. In the expression "その近代建築が1世紀前後の耐久力しかないとすれば" (If the modern architecture will stay for only around a century), "1世紀" means an interval (a century). However, it is difficult to recognize this.

2. In the expression "アルタシェス (アルタクス) 朝 (前190-前1世紀，〈大アルメニア王国〉とも呼ぶ)" (the Altaces Dynasty (from (the) 190(th) to the first century B.C. and called "the Great Armenian Kingdom"), "前190" means a year, but it is extracted as the 190th century B.C. because the last half of the range contains "世紀" (century).

5 Precision of Year Reference Extraction

Table 1 evaluates the precision of extracting year references. Five thematic chronological-table search problems are used for this evaluation. The first five columns show the search conditions, and the last three show the results. Each result item contains a year reference.[1] The correctness of year references in all the search-result items was judged by human for each search condition. The extraction error rate was less than 1% in most cases in this evaluation; i.e., the precision was better than 0.99. However, the error rate exceeded 1% in certain cases. The precision of the year range from 1 to 100 A.D. (the bottom row in the table) was the worst because this range contains many error cases of two-digit year references.

Recall has not yet been evaluated because precision is more important in the encyclopedia search. Improving recall is the next step.

6 Related Work

There has been much research on information extraction from English text. Examples of research on information extraction from Japanese text are Saito et al. [Sai 98], Sato et al. [Sat 95], Takao et al. [Tak 99], and Hisamitsu et al. [His 97]. Much research remains to be done on extracting year information precisely enough for encyclopedia searches, however.

7 Conclusion

A method of searching text and organizing the result called thematic chronological-table search has been developed. This method generates a year-axis index by extracting year references from the text.

An elaborate method of year information extraction for seven forms of year or century references has been developed, as have techniques for reducing the number of extraction errors. The measured average precision was greater than 99%. Most year references of the form of abbreviated Gregorian years with two digits were also extracted correctly. However, extraction errors in this form causes the precision of references to other years to be lower, especially years between 1 to 100 A.D. Improving the precision is not an easy task because it requires the exhaustive extraction of three- and four-digit year references.

The rules described here were developed for the World Encycloædia among the year extraction rules. However, most of the rules can be applied to other type of texts, such as newspaper text. Most of the rules can probably be modified and applied to texts in other languages, such as English, too.

Future work on extracting year references includes devising more year extraction techniques that increase the precision, extracting year references that are not currently extracted because increasing the precision of the current version was not possible, and applying this method to other types of text, such as text in newsgroups and on the WWW.[2]

[1] A result item contains a sentence, and a sentence may contain two or more year references. However, each search result focuses on only one year reference.

[2] However, the axis-specified search method has already been applied to the Mainichi Newspaper [Kan 98].

Acknowledgments

I am grateful to Yasufumi Fujji and other members of the Hitachi Digital Heibonsha for allowing use of the text of the World Encyclopædia, and for helping me improve the extraction and sorting methods. I also thank Mikio Yamazaki from the Hitachi Tohoku Software Company, who developed the user interface.

References

[HDH 98] *DVD/CD-ROM World Encyclopædia, version 2*, Hitachi Digital Heibonsha, 1998.

[His 97] Hisamitsu, T., and Niwa, Y.: Acquisition of Person Names from Newspaper Articles by Lexical Knowledge and Co-occurrence Analysis, *SIG on Natural Language Processing*, Information Processing Society of Japan, 118-1, pp. 1–6, 1997 (in Japanese).

[Kan 98] Kanada, Y.: Axis-specified Search: A New Full-text Search Method for Gathering and Structuring Excerpts, *3rd Int'l ACM Conf. on Digital Libraries*, pp. 108–117, 1998.

[Kan 98a] Kanada, Y.: Axis-specified Search: A Full-text Search Method for Extracting and Ordering Excerpts from Documents, *Technical Report 98-FI-50-4*, SIGFI, Information Processing Society of Japan, pp. 25-32, 1998

[Kan 99] Kanada, Y.: A Method of Geographical Name Extraction from Japanese Text for Thematic Geographical Search, *18th Int'l Conference on Information and Knowledge Management* (*CIKM 99*), submitted.

[MUC 98] Proceedings of the Seventh Message Understanding Conference (MUC-7), SAIC, 1998.

[Sai 98] Saito, K., Sakota, A., Nakae, T., Iwai, S., Tamura, N., and Nakagawa, H.: Numerical Information Extraction from Newspaper Articles, *SIG on Natural Language Processing*, Information Processing Society of Japan, 125-6, pp. 63–70, 1998 (in Japanese).

[Sat 95] Sato, M., Sato, T., and Shinoda, Y.: Automated Editing for Packaging Netnews Articles, Transaction of the Information Processing Society of Japan, Vol. 36, No. 10, 2371–2379, 1995 (in Japanese).

[Tak 99] Takao, Y., Nagai, H., Nakamura, S., and Nomura, H.: Information Extraction from Newspaper Articles of Multiple Products — classification of expression patterns —, *SIG on Natural Language Processing*, Information Processing Society of Japan, 129-17, pp. 117–124, 1999 (in Japanese).