Producing a Large-scale Encyclopedic Corpus over the Web

Atsushi Fujii[*], **Katunobu Itou**[†], **Tetsuya Ishikawa**[*]

[*]University of Library and Information Science
1-2 Kasuga, Tsukuba, 305-8550, Japan
{fujii, ishikawa}@ulis.ac.jp

[†]National Institute of Advanced Industrial Science and Technology
1-1-1 Chuuou Daini Umezono, Tsukuba, 305-8568, Japan
itou@ni.aist.go.jp

Abstract

Encyclopedias, which describe general/technical terms, are valuable language resources (LRs). As with other types of LRs relying on human introspection and supervision, constructing encyclopedias is quite expensive. To resolve this problem, we automatically produced a large-scale encyclopedic corpus over the World Wide Web. We first searched the Web for pages containing a term in question. Then we used linguistic patterns and HTML structures to extract text fragments describing the term. Finally, we organized extracted term descriptions based on domains. The resultant corpus contains approximately 100,000 terms. We also evaluated the quality of 2,000 test terms, and found that correct descriptions were obtained for 65% of test terms.

1. Introduction

Encyclopedias, which describe general/technical terms, are valuable language resources (LRs). As with other types of LRs relying on human introspection and supervision, constructing encyclopedias is quite expensive. Additionally, since existing encyclopedias are usually revised every few years, in many cases users find it difficult to obtain descriptions for new terms and new definitions for existing terms.

To cope with the above limitation of existing encyclopedias, it is possible to use a search engine over the World Wide Web as a substitute, expecting that certain Web pages describe submitted keywords. However, since keyword-based search engines often retrieve a large number of extraneous pages, it is time-consuming to identify pages that satisfy the users' information needs.

To resolve these problems, Fujii and Ishikawa (2000; 2001) proposed a method to automatically produce encyclopedias over the Web, in which term descriptions are extracted from Web pages and organized based on domains and word senses. However, their prototype system has been applied to a limited number of terms (approximately one hundred), and needs to be enhanced for real-world applications. Thus, we implemented an operational system, and produced a large-scale Japanese encyclopedic corpus. This paper explains our methodology and reports the result of experiments.

2. System Design

2.1. Overview

Figure 1 depicts the overall design of our system, which generates an encyclopedia for input terms. Our system, which is currently implemented for Japanese, consists of three modules: "retrieval", "extraction" and "organization".

In Figure 1, terms can be submitted either on-line or off-line. A reasonable method is that while the system periodically updates the encyclopedia off-line, terms unindexed in the encyclopedia are dynamically processed in real-time usage. In either case, our system processes input terms one by one. We briefly explain each module in the following three sections, respectively.

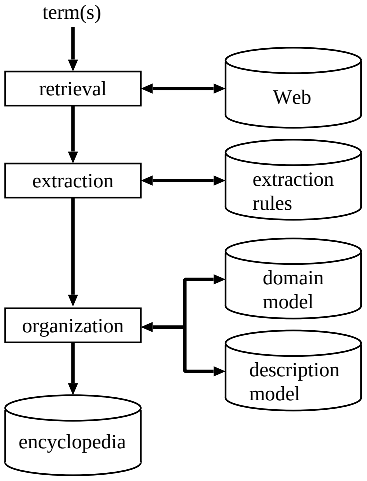


Figure 1: The overall design of our Web-based encyclopedia generation system.

2.2. Retrieval

The retrieval module searches the Web for pages containing an input term, for which we used a software robot to collect pages over the Japanese (.jp) domain. So far, we have collected approximately 20,000,000 pages, and the number of pages is still growing.

To retrieve those pages, we implemented a keyword-based retrieval module. The retrieval module obtains pages containing a input keywords, and sorts them according to the relevance score in descending order. For this purpose,

The first and second authors are also members of CREST, Japan Science and Technology Corporation.

we compute the "PageRank" score for each page based on hyperlink information (Brin and Page, 1998), which is used in "Google"[1].

2.3. Extraction

In the extraction module, given Web pages containing an input term, newline codes, redundant white spaces, JavaScript codes, and HTML tags that are not used in the following processes are discarded so as to standardize the page format. Second, we approximately identify a region describing the term in the page, for which two rules are used.

The first rule is based on Japanese linguistic patterns typically used for term descriptions, such as "X *toha* Y *dearu* (X is Y)". Following the method proposed by Fujii and Ishikawa (2000), we semi-automatically produced 20 patterns based on the Japanese CD-ROM World Encyclopedia (Heibonsha, 1998), which includes approximately 80,000 entries related to various fields. It is expected that a region including the sentence that matched with one of those patterns can be a term description.

The second rule is based on HTML layout. In a typical case, a term in question is highlighted as a heading with tags such as `<DT>`, `<B>` and `<Hx>` ("`x`" denotes a digit), followed by its description. In some cases, terms are marked with the anchor `<A>` tag, providing hyperlinks to pages where they are described.

Finally, based on the region briefly identified by the above method, we extract a page fragment as a term description. Since term descriptions usually consist of a logical segment (such as a paragraph) rather than a single sentence, we extract fragments tagged with a certain HTML tags such as `<P>`, `<LI>`, `<DIV>`, and `<DD>`.

2.4. Organization

Organizing information extracted from the Web is crucial in our framework. For this purpose, we classify extracted term descriptions based on word senses and domains.

Although a number of methods have been proposed to generate word senses (for example, one based on the vector space model (Schütze, 1998)), it is still difficult to accurately identify word senses without explicit dictionaries that define sense candidates.

Since word senses are often associated with domains (Yarowsky, 1995), word senses can be consequently distinguished by way of determining the domain of each description. For example, different senses for "pipeline (processing method/transportation pipe)" are associated with the computer and construction domains (fields), respectively.

To sum up, the organization module classifies term descriptions based on domains, for which we use domain and description models. In Section 3., we elaborate on our organization model.

3. Statistical Organization Model

3.1. Overview

Given one or more (in most cases more than one) descriptions for a single input term, the organization module selects appropriate description(s) for each domain related to the term.

We do not need all the extracted descriptions as final outputs, because they are usually similar to one another, and thus are redundant.

For the moment, we assume that we know *a priori* which domains are related to the input term.

From the viewpoint of probability theory, our task here is to select descriptions with greater probability for given domains. The probability for description d given domain c, $P(d|c)$, is commonly transformed as in Equation (1), through use of the Bayesian theorem.

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \qquad (1)$$

In Equation (1), $P(c|d)$ models a probability that d corresponds to domain c. $P(d)$ models a probability that d can be a description for the term in question, disregarding the domain. We shall call them domain and description models, respectively. We regard $P(c)$ as a constant.

To sum up, in principle we select d's that are strongly associated with a specific domain, and are likely to be descriptions themselves.

In practice, we first use Equation (1) to compute $P(d|c)$ for all the c's predefined in the domain model. Then we discard such c's whose $P(d|c)$ is below a specific threshold. As a result, for the input term, related domains and descriptions are simultaneously selected. Thus, we do not have to know *a priori* which domains are related to each term.

In the following two sections, we explain methods to realize the domain and description models, respectively.

3.2. Domain Model

The domain model quantifies the extent to which description d is associated with domain c, which is fundamentally a categorization task. Among a number of existing categorization methods, we experimentally used one proposed by Iwayama and Tokunaga (1994), which formulates $P(c|d)$ as in Equation (2).

$$P(c|d) = P(c) \cdot \sum_{t} \frac{P(t|c) \cdot P(t|d)}{P(t)} \qquad (2)$$

Here, $P(t|d)$, $P(t|c)$ and $P(t)$ denote probabilities that word t appears in d, c and all the domains, respectively. We regard $P(c)$ as a constant. While $P(t|d)$ is simply a relative frequency of t in d, we need predefined domains to compute $P(t|c)$ and $P(t)$. For this purpose, the use of large-scale corpora annotated with domains is desirable.

However, since those resources are prohibitively expensive, we used the "Nova" dictionary for Japanese/English machine translation systems[2], which includes approximately one million entries related to 19 technical fields as listed below:

[1] http://www.google.com/

[2] Produced by NOVA, Inc.

> aeronautics, biotechnology, business, chemistry, computers, construction, defense, ecology, electricity, energy, finance, law, mathematics, mechanics, medicine, metals, oceanography, plants, trade.

We extracted words from dictionary entries to estimate $P(t|c)$ and $P(t)$, which are relative frequencies of t in c and all the domains, respectively. We used the ChaSen morphological analyzer (Matsumoto et al., 1997) to extract words from Japanese entries. We also used English entries because Japanese descriptions often contain English words.

It may be argued that statistics extracted from dictionaries are unreliable, because word frequencies in real word usage are missing. However, words that are representative for a domain tend to be frequently used in compound word entries associated with the domain, and thus our method is a practical approximation.

3.3. Description Model

The description model quantifies the extent to which a given page fragment is feasible as a description for the input term. In principle, we decompose the description model into language and quality properties, as shown in Equation (3).

$$P(d) = P_L(d) \cdot P_Q(d) \qquad (3)$$

Here, $P_L(d)$ and $P_Q(d)$ denote language and quality models, respectively.

It is expected that the quality model discards incorrect or misleading information contained in Web pages. For this purpose, we use as $P_Q(d)$ the PageRank score for the source page computed in the retrieval module (see Section 2.2.). In other words, we rate the quality of pages based on hyperlink information, and selectively retrieves those with higher quality.

Statistical approaches to language modeling have been used in much NLP research, such as machine translation (Brown et al., 1993) and speech recognition (Bahl et al., 1983). Our model is almost the same as existing models, but is different in two respects.

First, while general language models quantify the extent to which a given word sequence is linguistically acceptable, our model also quantifies the extent to which the input is acceptable as a term description. Thus, we trained the model based on an existing machine readable encyclopedia.

We used the ChaSen morphological analyzer to segment the Japanese CD-ROM World Encyclopedia (Heibonsha, 1998) into words (we replaced headwords with a common symbol), and then used the CMU-Cambridge toolkit (Clarkson and Rosenfeld, 1997) to model a word-based trigram.

Consequently, descriptions in which word sequences are more similar to those in the World Encyclopedia are assigned greater probability scores through our language model.

Second, $P(d)$, which is a product of probabilities for N-grams in d, is quite sensitive to the length of d. In the cases of machine translation and speech recognition, this problem is less crucial because multiple candidates compared based on the language model are almost equivalent in terms of length.

However, since in our case length of descriptions are significantly different, shorter descriptions are more likely to be selected, regardless of the quality. To avoid this problem, we normalize $P(d)$ by the number of words contained in d.

4. Produced Corpus

We collected 169,451 terms in various fields from a number of sources (e.g., machine-readable dictionaries), and submitted them to our encyclopedia generation system. As a result, we produced an encyclopedic corpus including 105,069 terms.

Our corpus is available via a Web browser, in which descriptions for submitted keywords are browsable. Figure 2 shows example descriptions for "*deeta-mainingu* (data mining)". In this figure, the first two paragraphs are associated with the computer domain, and the last paragraph describes "data mining" in the context of the finance domain. Those descriptions were extracted from different pages.

We evaluated the quality of our resultant corpus, for which we selected 2,000 technical terms in the computer domain as a test set. Then, we asked six people to judge each of the resultant descriptions as to whether or not it is a correct description for a term in question.

We analyzed the result on a term-by-term basis, because reading only a couple of descriptions is not crucial. In other words, we evaluated each term (not description), and in the case where at least one of the top ten description was correct for a term in question, we judged it correct. The ratio of correct terms was 64.6%. Since all the test terms were inherently related to the computer domain, we focused solely on descriptions categorized into the computer domain. In this case, the ratio of correct terms was 61.2%.

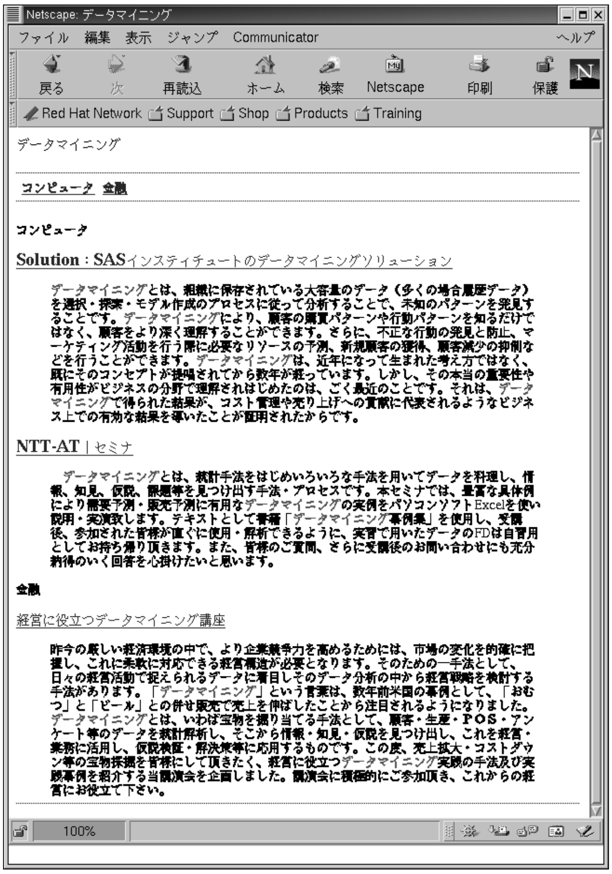
5. Conclusion

The World Wide Web has been an unprecedentedly enormous information source, from which a number of language processing methods have been explored to extract, retrieve and discover various types of information. In this paper, we described a method to produce large-scale encyclopedic corpora consisting of term descriptions from the Web.

Given a term for which encyclopedic knowledge (i.e., descriptions) is to be generated, our method sequentially performs a) retrieval of Web pages containing the term, b) extraction of page fragments describing the term, and c) organizing extracted descriptions based on domains (and consequently word senses). We produced a corpus containing 105,069 terms. We also evaluated the quality of 2,000 test terms, and found that correct descriptions were obtained for 65% of test terms.

6. References

Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.



データマイニング

コンピュータ 金融

コンピュータ

Solution : SASインスティチュートのデータマイニングソリューション

データマイニングとは、組織に保存されている大容量のデータ（多くの場合履歴データ）を選択・探索・モデル作成のプロセスに従って分析することで、未知のパターンを発見することです。データマイニングにより、顧客の購買パターンや行動パターンを知るだけではなく、顧客をより深く理解することができます。さらに、不正な行動の発見と防止、マーケティング活動を行う際に必要なリソースの予測、新規顧客の獲得、顧客減少の抑制などを行うことができます。データマイニングは、近年になって生まれた考え方ではなく、既にそのコンセプトが提唱されてから数年が経っています。しかし、その本当の重要性や有用性がビジネスの分野で理解されはじめたのは、ごく最近のことです。それは、データマイニングで得られた結果が、コスト管理や売り上げへの貢献に代表されるようなビジネス上での有効な結果を導いたことが証明されたからです。

NTT-AT | セミナ

データマイニングとは、統計手法をはじめいろいろな手法を用いてデータを料理し、情報、知見、仮説、課題等を見つけ出す手法・プロセスです。本セミナでは、豊富な具体例により需要予測・販売予測に有用なデータマイニングの実例をパソコンソフトExcelを使い説明・実演致します。テキストとして書籍「データマイニング事例集」を使用し、受講後、参加された皆様が直ぐに使用・解析できるように、実習で用いたデータのFDは自習用としてお持ち帰り頂きます。また、皆様のご質問、さらに受講後のお問い合わせにも充分納得のいく回答を心掛けたいと思います。

金融

経営に役立つデータマイニング講座

昨今の厳しい経済環境の中で、より企業競争力を高めるためには、市場の変化を的確に把握し、これに柔軟に対応できる経営構造が必要となります。そのための一手法として、日々の経営活動で捉えられるデータに着目しそのデータ分析の中から経営戦略を検討する手法があります。「データマイニング」という言葉は、数年前米国の事例として、「おむつ」と「ビール」との併せ販売で売上を伸ばしたことから注目されるようになりました。データマイニングとは、いわば宝物を掘り当てる手法として、顧客・生産・ＰＯＳ・アンケート等のデータを統計解析し、そこから情報・知見・仮説を見つけ出し、これを経営・業務に活用し、仮説検証・解決策等に応用するものです。この度、売上拡大・コストダウン等の宝物採掘を皆様にして頂きたく、経営に役立つデータマイニング実践の手法及び実践事例を紹介する当講演会を企画しました。講演会に積極的にご参加頂き、これからの経営にお役立て下さい。

Figure 2: Example Japanese descriptions for “*deeta-mainingu* (data mining)”.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech’97*, pages 2707–2710.

Atsushi Fujii and Tetsuya Ishikawa. 2000. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 488–495.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 196–203.

Hitachi Digital Heibonsha. 1998. CD-ROM World Encyclopedia. (In Japanese).

Makoto Iwayama and Takenobu Tokunaga. 1994. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 162–167.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Osamu Imaichi, and Tomoaki Imamura. 1997. Japanese morphological analysis system ChaSen manual. Technical Report NAIST-IS-TR97007, NAIST. (In Japanese).

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.