Modeling Wikipedia Articles to Enhance Encyclopedic Search

Atsushi Fujii

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. Existing search engines provide an enormous volume of information, but retrieved information is not organized. Hand-compiled encyclopedias provide organized information, but the quantity of information is limited. To integrate the advantages of both tools, we have been proposing methods for encyclopedic search targeting information on the Web and patent information. In this paper, we propose a method to categorize multiple expository texts for a single term based on viewpoints. Because viewpoints required for explanation are different depending on the type of a term, such as animals and diseases, it is difficult to manually produce a large scale system. We use Wikipedia to extract a prototype of a viewpoint structure for each term type. We also use articles in Wikipedia for a machine learning method, which categorizes a given text into an appropriate viewpoint. We evaluate the effectiveness of our method experimentally.

1. Introduction

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. Existing search engines provide an enormous volume of information, but retrieved information is not organized. Hand-compiled encyclopedias provide organized information, but the quantity of information is limited.

To solve the quantity and quality problems above, we have been developing a method to produce encyclopedic dictionaries automatically from text information on the Web (Fujii and Ishikawa, 2000; Fujii and Ishikawa, 2001; Fujii and Ishikawa, 2004; Fujii et al., 2002; Fujii et al., 2005). In addition, to collect information related to high-tech terms not found on the Web, we have recently used patent documents (Fujii, 2008). Our method extracts paragraph-style term descriptions from a source text collection and classifies those descriptions into domains. Our method also extracts related terms for each headword and summarizes multiple descriptions into a single text. We have produced an encyclopedic dictionary including approximately 1 900 000 Japanese terms as headwords, which is available at a Web search site called "CYCLONE"[1]. CYCLONE has been used for various research purposes, including looking up a definition for a term and searching for patent documents.

In CYCLONE, a user can submit words and natural language questions to search for paragraph-style term descriptions for a specific term. In addition, a user can also refine a search result by related terms and domains. Figure 1 shows a retrieval result for the term "XML". In the bottom half of Figure 1, three descriptions extracted from Web pages are presented. The following is an English translation for the first description. Because the source sentences in Japanese contain words in English inherently, these English words are in italics.

> *XML* is an extensible markup language (*eXtensible Markup Language*). Unlike *HTML*, which accepts only predefined tags, any original tags are allowed. Thus, annotated with tags related to logical structures, the content of a document can be understandable. *XML* is associated with the following advantages.

Below the input box are domains, such as "computer", and related terms, such as "HTML", which can be used as feedback terms to refine the user's focus.

While we have developed CYCLONE for almost ten years, Wikipedia[2], which is a hand-complied encyclopedia, has recently been growing. In this paper, aiming to enhance our encyclopedic search system, we use Wikipedia to model term descriptions, i.e., how each term is described and structured in Wikipedia. In encyclopedias including Wikipedia, a single term is usually described from one or more viewpoints. In addition, a set of viewpoints required to describe a term is different depending on the type of the term in question. For example, while viewpoints for an animal include "ecology", "species", and "appearance", viewpoints for a disease include "symptom", "diagnosis", and "treatment".

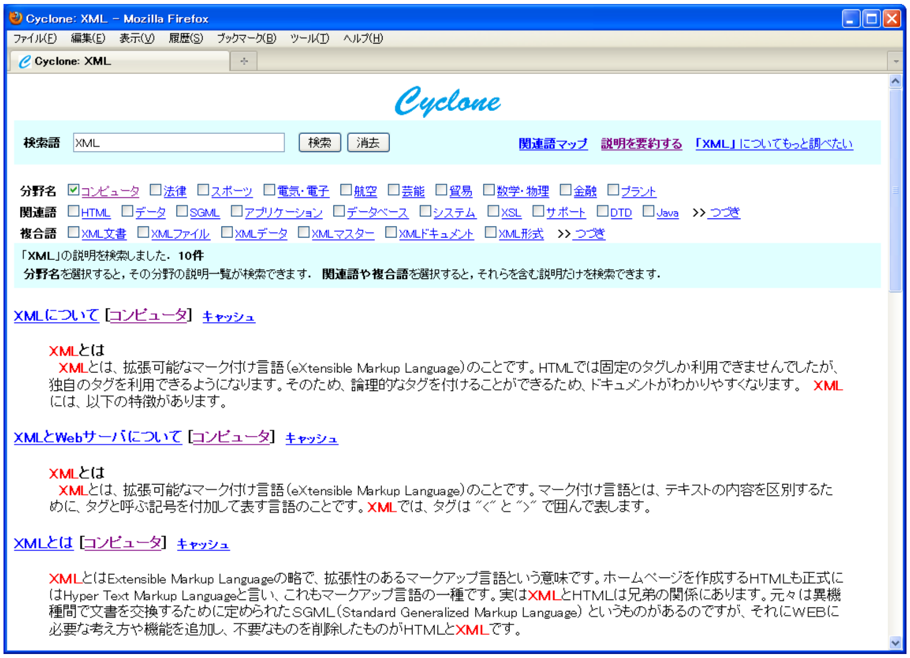
Thus, it is time-consuming to manually model term descriptions for various term types. To address this problem, we automatically extract a template of a viewpoint structure for each term type from Wikipedia. We also use articles in Wikipedia for machine learning purposes, and classify descriptions retrieved by CYCLONE in response to a submitted term, such as "influenza". Although we currently target Japanese, our method to model term descriptions based on Wikipedia articles is language-independent.

2. Related work

In a previous study (Fujii and Ishikawa, 2004), we summarized descriptions retrieved by CYCLONE based on view-

[1] http://cyclone.cl.cs.titech.ac.jp/

[2] http://wikipedia.org/


Cyclone: XML – Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
Cyclone: XML
Cyclone
検索
消去
関連語マップ
説明を要約する
「XML」についてもっと調べたい

検索語 XML

分野名 ☑コンピュータ □法律 □スポーツ □電気・電子 □航空 □芸能 □貿易 □数学・物理 □金融 □プラント

関連語 □HTML □データ □SGML □アプリケーション □データベース □システム □XSL □サポート □DTD □Java >> つづき

複合語 □XML文書 □XMLファイル □XMLデータ □XMLマスター □XMLドキュメント □XML形式 >> つづき

「XML」の説明を検索しました. 10件

分野名を選択すると，その分野の説明一覧が検索できます. 関連語や複合語を選択すると，それらを含む説明だけを検索できます.

XMLについて [コンピュータ] キャッシュ

XMLとは
XMLとは、拡張可能なマーク付け言語(eXtensible Markup Language)のことです。HTMLでは固定のタグしか利用できませんでしたが、独自のタグを利用できるようになります。そのため、論理的なタグを付けることができるため、ドキュメントがわかりやすくなります。 XMLには、以下の特徴があります。

XMLとWebサーバについて [コンピュータ] キャッシュ

XMLとは
XMLとは、拡張可能なマーク付け言語(eXtensible Markup Language)のことです。マーク付け言語とは、テキストの内容を区別するために、タグと呼ぶ記号を付加して表す言語のことです。XMLでは、タグは "<" と ">" で囲んで表します。

XMLとは [コンピュータ] キャッシュ

XMLとはExtensible Markup Languageの略で、拡張性のあるマークアップ言語という意味です。ホームページを作成するHTMLも正式にはHyper Text Markup Languageと言い、これもマークアップ言語の一種です。実はXMLとHTMLは兄弟の関係にあります。元々は異機種間で文書を交換するために定められたSGML(Standard Generalized Markup Language) というものがあるのですが、それにWEBに必要な考え方や機能を追加し、不要なものを削除したものがHTMLとXMLです。

Figure 1: Example descriptions for the term "XML".

points. Figure 2 shows an example summary for "XML", produced from Figure 1. In Figure 2, for each viewpoint, such as "definition", "purpose", and "advantage", a representative sentence is extracted from descriptions related to "XML".

The following is an English translation of the sentences in Figure 2. As in the English translation for Figure 1, the English words in the source sentences are in italics.

- **definition**: *XML* is an extensible markup language (*eXtensible Markup Language*).
- **abbreviation**: an abbreviation for *Extensible Markup Language* (an extensible markup language).
- **purpose**: Because *XML* is a standard specification for data representation, the data defined by *XML* can be reusable, irrespective of the upper application.
- **advantage**: *XML* is advantageous to developers of the file maker *Pro*, which needs to receive data from the client.
- **history**: was advised as a standard by *W3C* (*World Wide Web Consortium*: a group standardizing *WWW* technologies) in 1998,
- **reference**: This book is an introduction for *XML*, which has recently been paid much attention as the next generation Internet standard format, and related technologies.
- **miscellaneous**: In *XML*, the tags are enclosed in "<" and ">".

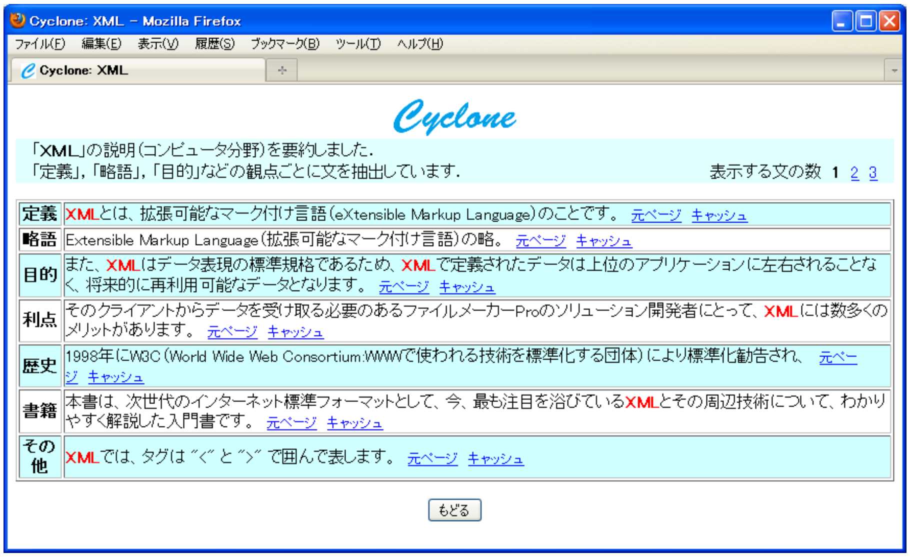
However, because we targeted only the computer domain, a set of viewpoints was always the same. In addition, we used hand-crafted rules to extract sentences for each viewpoint, which is not scalable. In this paper, by modeling Wikipedia articles, we target different viewpoints for various domains.

Biadsy et al. (Biadsy et al., 2008) used Wikipedia to produce a biographic summary for a person. They used articles in Wikipedia to determine whether a sentence is a description for a person or not. However, they targeted only the person domain and did not use viewpoints.

Blair-Goldensohn et al. (Blair-Goldensohn et al., 2008) used aspects, which correspond to viewpoints in this paper, for summarizing customer reviews. For example, reviews for a restaurant are summarized based on aspects, such as "location" and "price". Although aspects are extracted from the texts to be summarized, we use Wikipedia as external knowledge to model viewpoints. In other words, while Blair-Goldensohn et al. performed clustering, we perform categorization, which classifies items into predefined categories. For the same reason, our method is different from Clusy[3], which performs clustering on Web pages retrieved for a query.

The purpose of our research is similar to that of Sauper and Barzilay (Sauper and Barzilay, 2009). In brief, Sauper and Barzilay use Wikipedia articles for a term type, such as "disease", to model descriptions for that term type. Given a specific term, such as "influenza", they use a search engine on the Web to collect texts associated with "influenza" and select representative texts to generate a description for "influenza". However, in their method a type of an input term must be identified manually. In addition, they did not address problems associated with lexical ambiguities (i.e., homonymy and polysemy) for a target term. For example, if a target term is "kiwi (bird/fruit)", their method cannot distinguish different meanings for "kiwi". To resolve this problem, we model term types and viewpoints in a sin-

[3] http://clusty.com/


Cyclone: XML – Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
Cyclone: XML
Cyclone
もどる

「XML」の説明(コンピュータ分野)を要約しました.
「定義」,「略語」,「目的」などの観点ごとに文を抽出しています.

表示する文の数 1 2 3

定義	XMLとは、拡張可能なマーク付け言語(eXtensible Markup Language)のことです。 元ページ キャッシュ
略語	Extensible Markup Language(拡張可能なマーク付け言語)の略。 元ページ キャッシュ
目的	また、XMLはデータ表現の標準規格であるため、XMLで定義されたデータは上位のアプリケーションに左右されることなく、将来的に再利用可能なデータとなります。 元ページ キャッシュ
利点	そのクライアントからデータを受け取る必要のあるファイルメーカーProのソリューション開発者にとって、XMLには数多くのメリットがあります。 元ページ キャッシュ
歴史	1998年にW3C(World Wide Web Consortium:WWWで使われる技術を標準化する団体)により標準化勧告され、 元ページ キャッシュ
書籍	本書は、次世代のインターネット標準フォーマットとして、今、最も注目を浴びているXMLとその周辺技術について、わかりやすく解説した入門書です。 元ページ キャッシュ
その他	XMLでは、タグは“<”と“>”で囲んで表します。 元ページ キャッシュ

Figure 2: Example summary for the term “XML”.

gle framework, which makes our method scalable. While Sauper and Barzilay intended to generate a description for a term, the purpose of CYCLONE is to facilitate searching a large text collection for term descriptions.

3. Methodology

Figure 3 depicts an overview of our method, in which the left and right regions correspond to processes for modeling term descriptions and classifying texts, respectively. The modeling and classification processes are performed offline and online, respectively.

In Figure 3, the numbers of term types and viewpoints for each term type are two and three, respectively, without loss of generality. The numbers of term types and viewpoints for each term type can be arbitrary. We explain the modeling process in terms of Figure 3. To determine target term types, we use categories in Wikipedia, such as “animals” and “diseases”. For each term type, we collect articles describing terms associated with that term type.

However, categories in Wikipedia are not well organized as we expected. For example, although we wish to collect articles for individual animals, an article for a movie associated with animals is also classified into the “animal” category. Because we have not developed an automatic method to discard irrelevant articles, we have to manually select relevant articles from Wikipedia categories. We leave this issue as future work.

Articles in Wikipedia are usually structured based on “sections”. For example, sections for “influenza” include “Etymology”, “History”, “Microbiology”, and “Symptoms and diagnosis”. We use each section as a single viewpoint. However, items in sections for the same term type can vary depending on the author. Thus, we divide the collected articles into sections, and extract high-frequent sections as viewpoints for the term type in question.

We use Support Vector Machines (SVM) to learn two types of classifiers. We use the One-Vs-Rest method to target more than two categories. Given a set of article fragments for each viewpoint, we learn a classifier for viewpoints (“viewpoint classifier”). We use words in article fragments as features. We also learn a classifier for term types (“term classifier”), for which we combine all the article fragments in a term type as a single large text. In Figure 3, we use two large texts for “Animal” and “Disease” to produce a term classifier. We use words in article fragments as features. However, unlike the viewpoint classifier, we also use words in the name of each term as features. For example, both “stomach cancer” and “lung cancer” include the word “cancer” that can be a strong clue associated with diseases.

In summary, our term description model consists of the term classifier and the viewpoint classifier for each term type modeled in the term classifier. This feature contrasts with Sauper and Barzilay (Sauper and Barzilay, 2009), which did not model term types.

The classification process is driven by a set of texts retrieved in response to a keyword, such as “influenza”. For each of the texts, we first use the term classifier to determine the term type of “influenza”. In Figure 3, because “Disease” is selected as the term type for “influenza”, we use the viewpoint classifier for Disease to determine the viewpoint from which “influenza” is described in the input text. Finally, for each viewpoint, we select the top N texts with greater SVM scores, so that a user can obtain information about “influenza” from different viewpoints with a minimal cost. In addition, texts not describing “influenza”, which are usually assigned a small SVM score, can be discarded. Although our primary purpose is to classify descriptions retrieved by CYCLONE, our method can also be used to classify any texts, such as Web pages or snippets retrieved by a general-purpose search engine.

As a result, users can obtain encyclopedic information for terms not included in Wikipedia. Even if the target term is included in Wikipedia, the description is often restricted by author’s viewpoints. However, our method collects various information with respect to that term from the Web and patent archives. In addition, although individual articles in

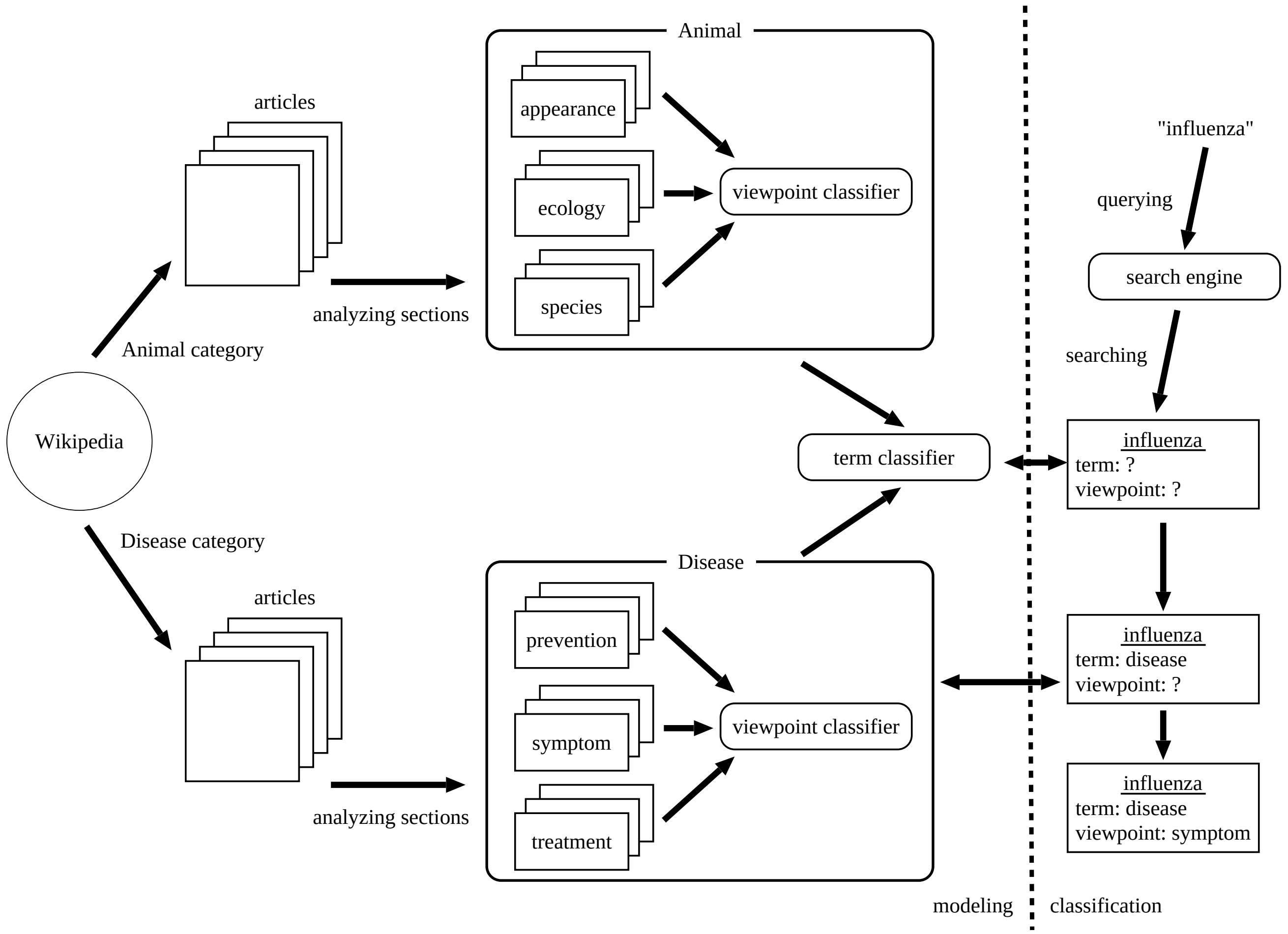


Figure 3: Overview of our method for modeling Wikipedia articles and classifying online texts.

Wikipedia often miss some viewpoints, by collecting high-frequent viewpoints across articles, viewpoints for a term in Wikipedia can be augmented.

We have been testing two different methods to introduce the term description model into CYCLONE. First, we can reduce human supervision in our previous method (Fujii and Ishikawa, 2004). While this previous method used handcrafted rules to classify sentences extracted from descriptions related only to the computer domain, the method in Figure 3 can automatically produce viewpoint classifiers for different domains. Second, we can simply classify retrieved descriptions in Figure 1 as they are. The second method allows a user to restrict retrieved descriptions based on viewpoints and filter out irrelevant descriptions.

4. Experiments

To investigate the validity of the method proposed in Section 3., we performed preliminary experiments. We used 10 general and 10 technical domains as the target term types. In real world usage, target texts for the classification are any texts. However, for an initial stage of research, we used Wikipedia articles divided based on sections as input texts. For each fragment, we used each section name annotated by an author as the correct viewpoint so that we avoided the cost for human judgments. We performed 5-fold cross-validation. Table 1 shows the accuracy of term classification (TC) and viewpoint classification (VC). To evaluate the effectiveness of VC itself, when calculate the VC values, we assumed that TC was performed correctly. Although both TC and VC, on average, were fairly high, the TC values for the last three term types were 0. This is mainly due to that the number of articles is small and that the classification is difficult inherently. For example, the articles for “veterinary” were mistakenly classified into either “animal” or “disease”. In addition, although Wikipedia articles are usually well-written, future work includes evaluating the classification for various texts.

To show the applicability of our method to a general search engine, we classified snippets retrieved by Yahoo! Japan[4] for the term “kiwi”. We used the term description model consisting of the twenty term types in Table 1. As a result, snippets associated with “kiwi (bird)” were classified into “animal” and further classified into viewpoints, such as “ecology”, “distribution”, “relation to human beings”, and “appearance”, correctly. At the same time, snippets associated with “kiwi (fruit)” were classified into “plant” and further classified into viewpoints, such as “usage” and “taxonomy”, correctly. In summary, our method classified texts retrieved by a general search engine into viewpoints depending on the meaning of a target term.

[4] http://www.yahoo.co.jp/

Table 1: Accuracy for classification.

Term type	#Articles	#Viewpoints	TC (%)	VC (%)
animal	1317	7	91.19	91.04
movie	1169	5	98.72	80.24
disease	878	8	97.15	76.77
company	618	3	89.97	97.41
person	500	4	90.80	60.40
plant	276	3	76.81	86.96
mathematics	228	3	86.84	61.40
insect	203	3	68.97	84.24
chemistry	201	3	77.61	84.08
cooking	190	3	72.11	76.84
informatics	103	3	59.22	59.22
fish	96	3	45.83	64.58
sports	86	3	76.74	75.58
law	56	3	71.43	53.57
construction	55	3	30.91	54.55
electricity	49	3	14.29	36.73
astronomy	38	3	55.26	52.63
veterinary	34	3	0	44.12
geology	24	3	0	29.17
physics	23	3	0	21.74
Average	307.2	3.6	86.54	79.66

5. Conclusion

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. In this paper, intended to enhance encyclopedic search, we proposed a method to categorize multiple expository texts for a single term based on viewpoints. Because viewpoints required for explanation are different depending on the type of a term, such as animals and diseases, it is difficult to manually produce a large scale system. We used Wikipedia to extract a prototype of a viewpoint structure for each term type. We also used articles in Wikipedia for a machine learning method, which categorizes a given text into an appropriate viewpoint. We evaluated the effectiveness of our method experimentally. Future work includes developing an automatic method to collect articles related to a specific category and evaluating our methods with a large number of term types.

6. Acknowledgments

This research was supported in part by the Grant-in-Aid Scientific Research on Priority Area of “Cyber Infrastructure for the Information-explosion Era” (Grant No. 19024007 and 21013003) and the Industrial Technology Research Grant Program from the New Energy and Industrial Technology Development Organization (NEDO) of Japan (Grant No. 05A14001a).

7. References

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 807–815.

Sasha Blair-Goldensohn, Ryan McDonald, George Reis, Tyler Neylon, Kerry Hannan, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW2008 Workshop on NLP Challenges in the Information Explosion Era*.

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.

Atsushi Fujii and Tetsuya Ishikawa. 2004. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 645–651.

Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2002. Producing a large-scale encyclopedic corpus over the Web. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1737–1740.

Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2005. Cyclone: An encyclopedic Web search site. In *Special Interest Tracks & Posters of the 14th International World Wide Web Conference*, pages 1184–1185.

Atsushi Fujii. 2008. Producing an encyclopedic dictionary using patent documents. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 208–216.