Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries

Aya Ishino[a],
Hidetsugu Nanba[a]
Toshiyuki Takezawa[a]

[a] Graduate School of Information Sciences, Hiroshima City University,
Hiroshima, Japan

{ishino, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract

For travelers who plan to visit a particular tourist spot, information about it is required. In this paper, we propose a method for extracting and organizing appropriate information from weblogs (blogs). Recently, increased numbers of travelers have been writing of their travel experiences via blogs. We call these travel blog entries, and they contain much useful travel information. For example, some bloggers introduce useful web sites for a tourist spot, while others report on transportation between tourist spots. Here, we extract hyperlinks of web sites for tourist spots from travel blog entries and organize them via automatic classification. We also extract transportation information automatically from travel blog entries. To investigate the effectiveness of our method, we conducted experiments. For the extraction of transportation information, we obtained an 80.3% for Precision. For the classification of hyperlinks, we obtained a high Precision. Finally, we constructed a prototype system, which provides information about (1) transportation between tourist spots and (2) useful web sites for tourist spots.

Keywords: Blog; Information Extraction; Travel Information; Link Classification

1 Introduction

For travelers who plan to visit a particular tourist spot, information about the place is necessary. Travel guidebooks and portal sites provided by tour companies and governmental tourist boards are useful information sources about travel. However, it is costly and time-consuming to compile travel information for all tourist spots and to keep this data up to date manually. Therefore, we have studied the automatic compilation of an online travel portal, which provides useful web sites for travel, and transportation information.

For this compilation, we focused on travel blog entries, which are defined as travel journals written by bloggers in diary form. Travel blog entries are considered a useful information source for obtaining travel information, because many bloggers' travel experiences are written in this form. For example, some bloggers introduce useful web sites for a tourist spot, while others report on transportation between tourist spots.

Nanba *et al*. (2009) identified travel blog entries in a blog database, then extracted pairs comprising a location name and a local product from these entries. In this paper, we propose a method that extracts transportation information from travel blog entries, which are identified automatically by Nanba's method. From these entries, we also extract the hyperlinks by which bloggers describe useful web sites for a tourist spot, and thereby construct collections of hyperlinks for a tourist spot.

The remainder of this paper is organized as follows. Section 2 shows the system behavior in terms of snapshots. Section 3 discusses related work. Section 4 describes our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section 5 reports the experimental results. We present some conclusions in Section 6.

2 System Behavior

In this section, we describe our prototype system, which provides information about (1) transportation between tourist spots and (2) useful web sites for tourist spots.These are the steps in the search procedure.

(Step 1) Input the location name for a tourist spot, such as "Hiroshima", in the search form (shown as ① in Figure 1).

(Step 2) Click the "search" button (shown as ②) to generate a list of transportation options, such as "Hiroshima → Osaka" and "Hiroshima → Tokyo", for the location name.

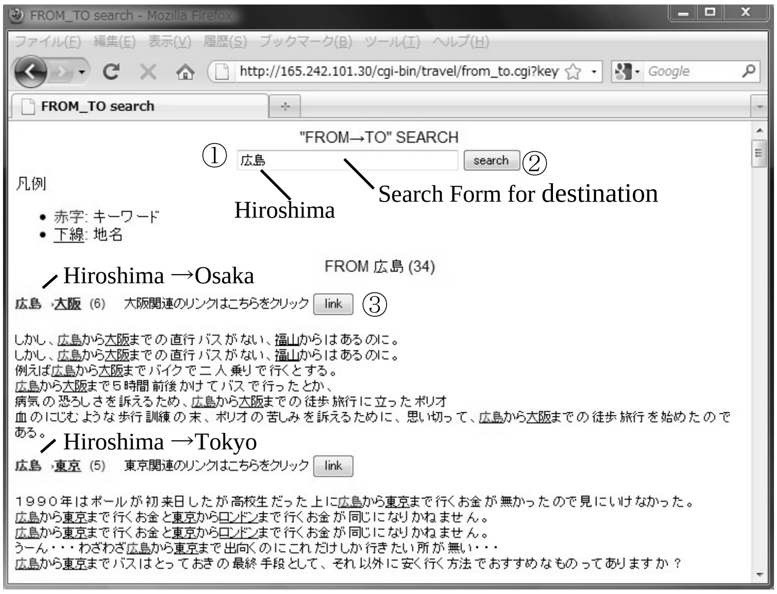
(Step 3) Click the "link" button (shown as ③) to generate a list of URLs for web sites related to the location together with automatically identified link types and the context of citations (we call them "citing areas"), by which the authors of travel blog entries describe the sites. Figure 2 shows a list of links related to "Osaka".

3 Related Work

In this section, we describe some related studies on geographic information retrieval, extraction of transportation information, and link classification.

- **Geographic Information Retrieval**

GeoCLEF (http://ir.shef.ac.uk/geoclef/) is the cross-language geographic retrieval track run as part of the Cross Language Evaluation Forum (CLEF), and it has been operating since 2005 (Gey *et al.*, 2005). The goal of this task is to retrieve news articles relevant to particular aspects of geographic information, such as "wine regions around the rivers in Europe". In our work, we focus on travel blog entries rather than news articles, because bloggers' travel experiences tend to be written as travel blog entries.


FROM_TO search - 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)
http://165.242.101.30/cgi-bin/travel/from_to.cgi?key
Google
FROM_TO search
①
search
②
Hiroshima
Search Form for destination
Hiroshima →Osaka
link
③
Hiroshima →Tokyo
link

"FROM→TO" SEARCH

広島

凡例

- 赤字: キーワード
- 下線: 地名

FROM 広島 (34)

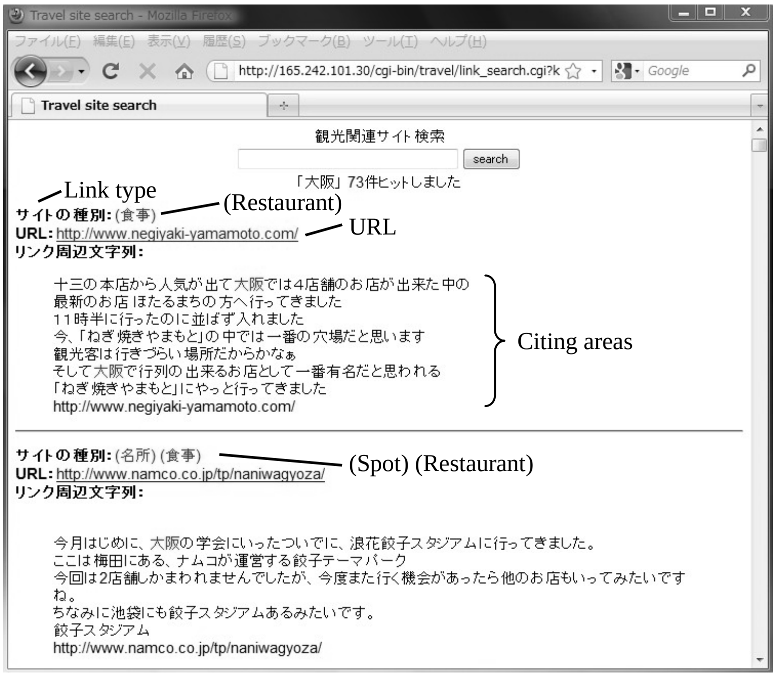
広島 ›大阪 (6) 大阪関連のリンクはこちらをクリック

しかし、広島から大阪までの直行バスがない、福山からはあるのに。
しかし、広島から大阪までの直行バスがない、福山からはあるのに。
例えば広島から大阪までバイクで二人乗りで行くとする。
広島から大阪まで５時間前後かけてバスで行ったとか、
病気の恐ろしさを訴えるため、広島から大阪までの徒歩旅行に立ったポリオ
血のにじむような歩行訓練の末、ポリオの苦しみを訴えるために、思い切って、広島から大阪までの徒歩旅行を始めたのである。

広島 ›東京 (5) 東京関連のリンクはこちらをクリック

１９９０年はポールが初来日したが高校生だった上に広島から東京まで行くお金が無かったので見にいけなかった。
広島から東京まで行くお金と東京からロンドンまで行くお金が同じになりかねません。
広島から東京まで行くお金と東京からロンドンまで行くお金が同じになりかねません。
うーん・・・わざわざ広島から東京まで出向くのにこれだけしか行きたい所が無い・・・
広島から東京までバスはとっておきの最終手段として、それ以外に安く行く方法でおすすめなものってありますか？

Fig. 1. The travel information search system


Travel site search - 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)
http://165.242.101.30/cgi-bin/travel/link_search.cgi?k
Google
Travel site search
search
Link type
(Restaurant)
URL
Citing areas
(Spot) (Restaurant)

観光関連サイト検索

「大阪」73件ヒットしました

サイトの種別: (食事)
URL: http://www.negiyaki-yamamoto.com/
リンク周辺文字列:

十三の本店から人気が出て大阪では4店舗のお店が出来た中の
最新のお店 ほたるまちの方へ行ってきました
11時半に行ったのに並ばず入れました
今、「ねぎ焼きやまもと」の中では一番の穴場だと思います
観光客は行きづらい場所だからかなぁ
そして大阪で行列の出来るお店として一番有名だと思われる
「ねぎ焼きやまもと」にやっと行ってきました
http://www.negiyaki-yamamoto.com/

サイトの種別: (名所) (食事)
URL: http://www.namco.co.jp/tp/naniwagyoza/
リンク周辺文字列:

今月はじめに、大阪の学会にいったついでに、浪花餃子スタジアムに行ってきました。
ここは梅田にある、ナムコが運営する餃子テーマパーク
今回は2店舗しかまわれませんでしたが、今度また行く機会があったら他のお店もいってみたいですね。
ちなみに池袋にも餃子スタジアムあるみたいです。
餃子スタジアム
http://www.namco.co.jp/tp/naniwagyoza/

Fig. 2. A list of web sites for a travel spot.

- **Extraction of Transportation Information**

Davidov (Davidov, 2009) presented an algorithm framework that enables automated acquisition of map-link information from the Web, based on surface patterns such as “from X to Y”. Given a set of locations as initial seeds, they retrieved from the Web an extended set of locations, and produced a map-link network that connects these locations using transport-type edges. In this paper, we propose a method for extraction of transportation information via machine-learning techniques.

- **Link Classification**

There have been several reports on research that automatically classifies links in blog entries (Kale *et al.*, 2007; Martineau & Hurst, 2008). Kale devised a method that classifies links in blog entries as positive or negative, using manually created rules (Kale *et al.*, 2007). Alternatively, Martineau proposed a machine-learning approach for link classification from several viewpoints using words that appear in the context of citations of URLs as features. In our work, we classify links into four categories of travel, which we will describe in Section 4.2.3, using a machine-learning technique.

4 Automatic Compilation of an Online Travel Portal

The task of compiling travel information is divided into three steps: (1) identification of travel blogs, (2) extraction of transportation information, and (3) classification of links in travel blog entries. For Step 1, we use Nanba’s method (Nanba *et al.*, 2009). Steps 2 and 3 are explained in Sections 4.1 and 4.2.

4.1 Extraction of Transportation Information

We use information extraction based on machine learning to extract information, such as “a departure place”, “a destination”, or “a transportation device”, from travel blog entries. First, we define the tags used in our examination.

- **FROM** tag includes a departure place.
- **TO** tag includes a destination.
- **VIA** tag includes a route.
- **METHOD** tag includes a transportation device.
- **TIME** tag includes the time for transportation.

This is a tagged example.

[original]
<FROM>広島</FROM>から<TO>大阪</TO>まで<TIME>5時間</TIME>かけて、<METHOD>バス</METHOD>で行った。
[translation]
It took <TIME>five hours</TIME> from <FROM>Hiroshima</FROM> to<TO>Osaka</TO> by <METHOD>bus</METHOD>.

We formulate the identification of the class of each word in a given sentence and solve it using machine learning. For the machine-learning method, we opted the Conditional Random Fields (CRF) method (Lafferty, McCallum, & Pereira, 2001), whose empirical success has been reported recently in the field of natural language processing. The CRF-based method identifies the class of each entry. Features and tags are used in the CRF method as follows: (1) k tags occur before a target entry, (2) k features occur before a target entry, and (3) k features follow a target entry. We used the value k = 4, which was determined via a pilot study. We use the following 15 features for machine learning. A sequence of nouns (a noun phrase) was treated as a noun. We used MeCab (http://mecab.sourceforge.net/) as a Japanese morphological analysis tool to identify the part of speech.

- A word.
- Its part of speech.
- Whether the word is a quotation mark.
- Whether the word is a cue phrase, detail as follows.

Tag	Cue phase	The number of cues
FROM	Whether the word is a cue that often appears immediately after the “FROM” tag, such as “から” (from) or “を出発” (left).	40
FROM TO	Whether the word is frequently used in the name of a tourist spot, such as “博物館” (museum) or “遊園地” (amusement park).	45
	Whether the word is frequently used in the name of a destination, such as “観光” (sightseeing tour) or “駅” (station).	11
	Whether the word is the name of a tourist spot.	13,779
	Whether the word is the name of a station or airport.	9437
TO	Whether the word is a cue that often appears immediately after the “TO” tag, such as “まで” (to) or “に到着” (arrival).	271
VIA	Whether the word is a cue that often appears immediately after the “via” tag, such as “経由” (via) or “通って” (through).	43
	Whether the word is the name of a highway.	101
METHOD	Whether the word is the name of a transportation device, such as “飛行機” (airplane) or “自動車” (car).	148
	Whether the word is the name of a vehicle.	128
	Whether the word is the name of a train or bus.	2033
TIME	Whether the word is an expression related to time, such as “分” (minute) or “時間” (hour).	77

4.2 Link Classification

The procedure for classifying links in travel blog entries is as follows.

1. Input a travel blog entry.
2. Extract a hyperlink and any surrounding sentences that mention the link (a citing area).
3. Classify the link by taking account of the information in the citing area.

In the following, we will explain Steps 2 and 3.

4.2.1 Extraction of citing areas

We manually created rules for the automatic extraction of citing areas. These rules use cue phrases. When authors of travel blog entries introduce web sites, quotation marks or brackets are often used immediately before and after the title of the site. The authors also use particular words, such as “紹介” (introduction), “公式サイト” (official site), or “の HP” (web page of), or particular marks, such as quotation marks or brackets. Therefore, we manually selected 26 cues and used them for citing area extraction using the following rules.

1. Extract a sentence that includes the link.
2. Extract *X* sentences that appear before or after a web hyperlink, and add them to the candidate. Here, we used the value of $X = 2$, which was determined via a pilot study.
3. Extract keywords from the candidate area in Step2 using the following rules (a) and (b), if the area includes cues.
   (a) Extract character strings within quotation marks or brackets as keywords.
   (b) Extract character strings just before or after particular cues, such as “の HP” (web page of).
4. Extract all sentences including the keywords in the blog entry and the sentences extracted in Step 2 as a citing area.

We explain these rules using the following travel blog entry.

[original]
1 チェックアウト後、いつものようにパパ&ママの寄り道が始まります!!
2 ということで、まずは河津の【バガテル公園】に行ってきました☆
3 四季の蔵から、車で数分圏内にあります。
4 ワンコもお散歩 OK なので、犬連れには嬉しい場所です
5 メッチャ、綺麗でしたよ～□
6 ※バガテル公園の HP は、こちら→
7 http://www.bagatelle.co.jp/index.html
8 ↑いうまでもなく、美しいバラの数々(写真)
9 四季の蔵の朝ごはんがボリューム満点だから、これくらいで充分です!!
10 初めて来たバガテル公園ですが、ワンコ OK だし、
11 季節によってはお花が綺麗なのでいいかも～♪
12 ランチメニューも充実しているし、また今度も来ようっと(ノ▽≦*)キャハッッッ♪

[translation]
1 Dad and Mom started to take a side trip after the checkout!!
2 Firstly, we visited “Bagatelle Park” in Kawazu☆
3 It took a couple of minutes from Shikinokura by car.
4 In this park, we could take a stroll with dogs.
5 Very very beautiful□
6 (*) Following is the web page of the Bagatelle Park
7 http://www.bagatelle.co.jp/index.html
8 ↑ Beautiful roses (pictures).

9 As we had a big breakfast in Shikinokura, the lunch in this park is enough!!
10 We visited the Bagatelle Park for the first time, and we could take a stroll with our dog here.
11 And flowers are beautiful in high season□
12 The lunch menu is abundant. I hope to come again :-) ♪

In Step 1, we extract sentence 7, which includes a hyperlink as an initial candidate area. In Step 2, we also extract the two sentences that appear before and after the hyperlink (5, 6, 8, and 9) and add them to the candidate. In Step 3, we extract "バガテル公園" (Bagatelle Park), which appears just before a cue phrase "の HP" (web page of), as a keywords. In Step 4, we add sentences 2 and 10, both of which include the keyword "バガテル公園" (Bagatelle Park), to the candidate. Finally, we extract the sentences 2, 5, 6, 7, 8, 9, and 10 as a citing area.

4.2.2 Definition of link types

We classify link types into the following four categories.

- **S (Spot)**: Whether the information is about tourist spots.
- **H (Hotel)**: Whether the information is about accommodation.
- **R (Restaurant)**: Whether the information is about restaurants.
- **O (Other)**: Other than types S, H, and R.

It is possible to classify a hyperlink into more than one link type. For example, a hyperlink to "ラーメン博物館" (Chinese noodle museum, http://www.raumen.co.jp/home/) is classified into types S and R, because the visitors to this museum can learn the history of Chinese noodles in addition to eating Chinese noodles.

4.2.3 Method of link type classification

Here, we explain how to classify hyperlinks automatically. We employed a machine-learning technique using the following features. A sequence of nouns (a noun phrase) was treated as a noun.

- A word.
- Whether the word is a cue phrase, detailed as follows, where the numbers in brackets shown for each feature represent the number of cues.

Cues for type S

Cue phrase	The number of cues
A list of tourist spots, collected from Wikipedia.	17,371
Words frequently used in the name of tourist spots, such as "動物園" (zoo) or "博物館" (museum).	138
Words related to sightseeing, such as "見学" (sightseeing) or "散策" (stroll).	172
Other words.	131

Cues for type H

Cue phrase	The number of cues
Words that are frequently used in the name of hotels, such as “ホテル” (hotel) or “旅館” (Japanese inn).	9
Component words for accommodations, such as “フロント” (front desk) or “客室” (guest room).	29
Words that are frequently used when tourists stay in accommodation, such as “泊る” (stay) or “チェックイン” (check in).	14
Other words.	21

Cues for type H

Cue phrase	The number of cues
Dish names such as “omelet”, collected from Wikipedia.	2,779
Cooking styles such as “Italian cuisine”, collected from Wikipedia.	114
Words that are frequently used in the name of restaurants, such as “レストラン” (restaurant) or “食堂” (dining room).	21
Words that are used when taking meals, such as “食べる” (eat) or “おいしい” (delicious).	52
General words that indicate food, such as “ご飯” (rice) or “料理” (cooking).	31
Other words.	31

5 Experiments

In order to investigate the effectiveness of our methods, we conducted two experiments: (1) extraction of transportation information from travel blog entries, and (2) extraction and classification of hyperlinks. We report on these in Sections 5.1 and 5.2, respectively.

5.1 Extraction of Transportation Information

Data Sets and Experimental Settings

We randomly selected 10,000 sentences from 193 travel blog entries, and manually assigned tags to them, as described in Section 4.1. The number of manually assigned tags is shown in Table 1. We used CRF++ (http://www.chasen.org/~taku/software/CRF++) software as the machine-learning package. We used Recall and Precision as evaluation measures.

Table 1. Numbers of manually assigned tags in the extraction of transportation information

	Training	Test
FROM	136	30
TO	384	126
VIA	58	15
METHOD	245	55
TIME	87	27

Results and Discussion

The evaluation results are shown in Table 2. As shown in the table, we obtained a high Precision. Among these results, both the Recall and Precision of “VIA” were low, which is due to the low frequency of this tag in both training and test data (Training: 58, Test: 15).

Table 2. Evaluation results for the extraction of transportation information

	Recall (%)	Precision (%)
FROM	30.0	75.0
TO	45.2	75.0
VIA	33.3	55.6
METHOD	66.0	94.9
TIME	50.0	87.6
Total	46.8	80.3

There were two typical errors causing low Precision: (1) ambiguity of cues (69.6%) and (2) the traveler’s desire (17.4%). We describe these errors as follows:

⑴ Ambiguity of cues (69.6%)

In the following example, the “VIA” tag was assigned to “結局お昼” (noon, after all), because a cue for VIA “過ぎて” (past) appears immediately before it. However, the “VIA” tag should not have been assigned in this case.

VIA	**[original]** **(Correct)** 結局お昼を過ぎても動かなかった。 **(Analysis result)** <VIA>結局お昼</VIA>を過ぎても動かなかった。 **[translation]** **(Correct)** I did not move though it was past noon, after all. **(Analysis result)** I did not move though it was past <VIA>noon, after all</VIA>.

⑵ The traveler’s desire (17.4%)

In the following example, the “METHOD” tag was mistakenly assigned to “Shinkansen bullet train”, which the visitor did not actually use. This was because the “METHOD” cue “で帰る” (return by) appears immediately before it.

METHOD	**[original]** **(Correct)** 新幹線で帰るのがベターだったのですが、ちょっとは旅気分を味わいたいということで、別ルートで。 **(Analysis result)** <METHOD>新幹線</METHOD>で帰るのがベターだったのですが、ちょっとは旅気分を味わいたいということで、別ルートで。 **[translation]** **(Correct)** It was better to return by Shinkansen bullet train, but I chose another route, because I wanted to draw out the journey. **(Analysis result)** It was better to return by <METHOD>Shinkansen bullet train</METHOD>, but I chose another route, because I wanted to draw out the journey.

We now discuss the low Recall of our method. There were two typical errors for low Recall: (1) the lack of contexts (59.1%) and (2) the lack of cues (17.3%). We describe these errors as follows.

(1) The lack of contexts (59.1%)

In the following example, the “TO” tag should be assigned to “寺田屋” (Teradaya), but our method did not assign any tags to this word because there were no cues in this short sentence. To solve this problem, we need to take account of a longer context. For example, the “TO” cue “来ました” (came), which appears in the previous sentence, would be necessary for solving the problem in this case. However, taking account of a longer context might result in lower Precision.

TO	**[original]** **(Correct)** いよいよ来ました！ 「<TO>寺田屋</TO>」 **(Analysis result)**いよいよ来ました！ 「寺田屋」 **[translation]** **(Correct)** Finally I arrived! “<TO>Teradaya</TO>” **(Analysis result)** Finally I arrived!! “Teradaya”

(2) The lack of cues (17.3%)

In the following example, the “FROM” tag was not assigned to “東京駅” (Tokyo station), because “に別れを告げる” (I say good-bye to) was not included in the “FROM” cues. To increase the number of cues, a statistical approach is required.

FROM	**[original]** **(Correct)** 駅弁買い込んで<FROM>東京駅</FROM>に別れを告げる。 **(Analysis result)** 駅弁買い込んで東京駅に別れを告げる。 **[translation]** **(Correct)** I bought a station lunch and said good-bye to <FROM>Tokyo Station</FROM>. **(Analysis result)** I bought a station lunch and said good-bye to Tokyo Station.

5.2 Extraction and Classification of Links

Data Sets and Experimental Settings

Among the 7,412 hyperlinks in 17,266 travel blog entries, we removed 2,987, which link to Wikipedia and news sites. These sites are easily classified into type O by their URLs. We randomly selected 1,000 of the remaining 4,155 links, manually classified them, and used them for our examination. Table 3 shows the number of hyperlinks for each type. We performed a four-fold cross-validation test. We used TinySVM (http://chasen.org/~taku/software/TinySVM/) software as the machine-learning package and used Recall and Precision as evaluation measures.

Table 3. The number of hyperlinks for each type

Link types	S	H	R	O
the number of links	353	98	343	250

Alternatives

To investigate the effectiveness of our method, we classified link types using the following two methods for citing area classification.

- Our method: Extract sentences by manually created rules, as described in Section 4.2.1.
- Baseline method: Extract the *X* sentences before and after the link.

Results and Discussion

We used a value of $X = 2$, which was determined via a pilot study, for the baseline method. The evaluation results are shown in Table 4, where our method generally shows improved Recall and Precision in comparison with the baseline method. In particular, the Recall and Precision for link type S were significantly improved.

Table 4. Evaluation results for link classification

Link Type	Baseline Method		Our Method	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
S	54.5	64.7	62.5	72.7
H	63.3	79.8	64.9	81.3
R	72.3	76.0	71.9	76.7
O	59.2	42.2	71.6	48.6

There were two typical errors in link classification: (1) the lack of cues and (2) ambiguity in cue phrases. We describe these errors as follows.

(1) Lack of cues

For the machine learning, we used manually selected cues, as described in Section 4.2.3. To improve the coverage of cues, a statistical approach, such as applying n-gram statistics to a larger blog corpus, will be required.

(2) Ambiguity in cue phrases

We used “visit” (訪れた) as an S cue. However, “visit” (訪れた) is also frequently used when a visitor eats in a restaurant, as in “食事を取るためレストランを訪れた” (I visited the restaurant for taking a meal). Therefore, our method has misclassified a link type R as a link type S.

6 Conclusion

In this paper, we have proposed two methods: (1) extraction of transportation information from travel blog entries, and (2) extraction and classification of hyperlinks in the travel blog entries. From our experimental results, we have confirmed the effectiveness of our methods. Finally, we have constructed a system that can search for a destination to which the user can travel from the present location and that can provide links about the destination.

In this paper, we focused on travel blog entries written in Japanese. In our future work, we will translate cue phrases from Japanese into other languages, and apply our method into blog entries written in various languages.

References

Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A. & Takezawa, T. (2009). Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper*: 205-208.

Gey, F. C., Larson, R. R., Sanderson, M., Joho, H., Clough, P. & Petras, V. (2005). GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. *Lecture Notes in Computer Science,* LNCS4022: 908-919.

Davidov, D. (2009). Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*: 267-175.

Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T. & Joshi, A. (2007). Modeling Trust and Influence in the Blogosphere Using Link Polarity. *International Conference on Weblogs and Social Media*.

Martineau, J. & Hurst, M. (2008). Blog Link Classification. *Proceedings of International Conference on Weblogs and Social Media*.

Lafferty, J., McCallum, A. & Pereira, F. (2001). Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th Conference on Machine Learning*: 282-289.