Automatic Discovery of Attribute Words from Web Documents

Kosuke Tokunaga, Jun'ichi Kazama, and Kentaro Torisawa

Japan Advanced Institute of Science and Technology (JAIST),
Asahidai 1-1, Nomi, Ishikawa, 923-1292 Japan
{kosuke-t, kazama, torisawa}@jaist.ac.jp

Abstract. We propose a method of acquiring attribute words for a wide range of objects from Japanese Web documents. The method is a simple unsupervised method that utilizes the statistics of words, lexico-syntactic patterns, and HTML tags. To evaluate the attribute words, we also establish criteria and a procedure based on question-answerability about the candidate word.

1 Introduction

Knowledge about how we recognize objects is of great practical importance for many NLP tasks. Knowledge about *attributes*, which tells us from what viewpoints objects are usually understood or described, is one of such type of knowledge. For example, the attributes of *car* objects will be *weight*, *engine*, *steering wheel*, *driving feel*, and *manufacturer*. In other words, attributes are items whose values we want to know when we want to know about the object. More analytically, we tend to regard A as an attribute for objects of class C when A works as if function $v = A(o), o \in C$ where v is necessary to us to identify o (especially to distinguish o from $o'(\neq o) \in C$). Therefore, obvious applications of attributes are ones such as summarization [1,2] and question-answering [3]. Moreover, they can be useful as features in word clustering [4] or machine learning. Although the knowledge base for attributes can be prepared manually (e.g., WordNet [5]), problems are cost and coverage. To overcome these, we propose a method that automatically acquires attribute knowledge from the Web.

To acquire the attributes for a given class, C (e.g., *car*), the proposed method first downloads documents that contain class label C (e.g., "car") from the Web.[1] We extract the candidates of attribute words from these documents and score them according to the statistics of words, lexico-syntactic patterns, and HTML tags. Highly scored words are output as attributes for the class. Lexico-syntactic patterns and other statistics have been used in other lexical knowledge acquisition systems [3,4,6,7,8]. We specifically used lexico-syntactic patterns involving the Japanese postposition "no" as used in [8] such as "C no A" where A is an attribute word, which is almost equivalent to pattern "A of C" used in [7] to

[1] We use C to denote both the class and its class label (the word representing the class). We also use A to denote both the attribute and the word representing it.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 106–118, 2005.

find part-whole relations. Novel features of our method are its use of Web search engines to focus on documents highly relevant to the class and its use of statistics concerning attribute words and surrounding HTML tags.

One of the difficulties in studying attribute knowledge is that there are no standard definitions of attributes, or criteria for evaluating obtained attributes. In this paper, we propose a simple but effective definition of attributes that matches our motivation and applications, i.e., whether we can ask a question about the attribute and whether there is an answer to that question (*question answerability*). For example, one can ask as "Who is the manufacturer of this car?", and someone might answer "Honda", because we want to know the *manufacturer* when we concerned about cars. We designed a procedure for evaluating attributes based on this idea. As the literature points out [9,10], attributes can include many types of relations such as property (e.g., *weight*), part-of (e.g., *engine*), telic (e.g., *driving feel*), and agentive (e.g., *manufacturer*). However, we ignored type distinctions in this study. First, because attributes are useful even if the type is not known, and second, because defining attributes as one of these types and evaluating them only complicates the evaluation process, making the results unstable. The use of linguistic tests to define attributes is not that new. Woods [11] devised a test on whether we can say "The A of o is v." Although we followed this procedure, we focused more on attributes that are important for our understanding of an object by using *question-answerability* as our criterion.

2 Acquisition Method

2.1 Basic Observations on Attributes

Our method is based on the following three observations.

1. Attributes tend to occur in documents that contain the class label and not in other documents.
2. Attributes tend to be emphasized by the use of certain HTML tags or occur as items in HTML itemizations or tables in Web documents.
3. Attributes tend to co-occur with the class label in specific lexico-syntactic patterns involving the postposition "no."

2.2 Extraction of Candidate Words

To acquire the attributes of class C, we first download documents that contain class label C using a Web search engine, according to the first observation. We refer to this set of documents as a *local document set* ($LD(C)$). All the nouns appearing in the local document set are regarded as candidates of attribute words. Here, the nouns are words tagged as "proper nouns", "sahen nouns" (nouns that can become a verb with the suffix "suru"), "location", or "unknown" (e.g., words written in katakana) by a Japanese morphological analyzer, JUMAN [12]. Note that we restricted ourselves to single word attributes in this study. The obtained candidate words are scored in the next step.

Table 1. Lexico-syntactic patterns for attribute acquisition. (We added possible English translations for the patterns in parenthesis).

C no A ha (A of C [*verb*])	C no A de (by A of C)	C no A e (to A of C)
C no A ga (A of C [*verb*])	C no A made (even/until A of C)	C no AA(A of C,)
C no A wo ([*verb*] A of C)	C no A kara (from A of C)	
C no A ni (at/in A of C)	C no A yori (from/than A of C)	

2.3 Ranking of Candidate Words

We rank the candidate words according to a score that reflects the observations described in Sect. 2.1. The overall score takes the following form.

$$V(C, A) = n(C, A) \cdot f(C, A) \cdot t(C, A) \cdot dfidf(C, A), \tag{1}$$

where A is the candidate word to be scored and C is the class. $n(C, A)$ and $f(C, A)$ are scores concerning lexico-syntactic patterns. $t(C, A)$ is a score concerning the statistics of HTML tags to reflect the second observation. Finally, $dfidf(C, A)$ is the score related to word statistics. This reflects the first observation. By multiplying these sub-scores, we expect that they will complement each other. We will explain the details on these sub-scores in the following.

As previously mentioned, we use lexico-syntactic patterns including the Japanese postposition "no" as clues. The patterns take the form "C no A $POST$" where $POST$ is a Japanese postposition or a punctuation mark.[2] The actual patterns used are listed in Table 1. Score $n(C, A)$ is the number of times C and A co-occur in these patterns in the local document set $LD(C)$.

Score $f(C, A)$ requires more explanation. Roughly, $f(C, A)$ is the number of times C and A co-occur in the patterns without the last postposition (i.e., pattern "C no A") collected from 33 years of parsed newspaper articles.[3] Note that pattern matching was done against the parsed dependency structures.[4] The reason this score was used in addition to $n(C, A)$ was to obtain more reliable scores by increasing the number of documents to be matched. This may sound contradictory to the fact that the Web is the largest corpus in the world. However, we found that we could not obtain all the documents that contained the class label because existing commercial Web search engines return URLs for a very small fraction of matched documents (usually up to about 1,000 documents). Although we could use hit counts for the patterns, we did not do this to avoid overloading the search engine (each class has about 20,000 candidate words).

Score $t(C, A)$ is the number of times A appears in $LD(C)$ surrounded by HTML tags. More precisely, we count the number of times A appears in the form: "<*tag1*>A<*tag2*>" where the number of characters between HTML tags

[2] Note that there are actually no spaces between words in Japanese. The spaces are for easier understanding.

[3] Yomiuri newspaper 1987–2001, Mainichi newspaper 1991–1999, and Nikkei newspaper 1983–1990; 3.01 GB in total. We used a Japanese dependency parser [13].

[4] The differences from $n(C, A)$ were introduced to reuse the existing parsed corpus.

<B>タイ風・カレー</B>
材料
鶏肉 400g，なす 2 個，バイマックルー 2 枚，ナンプラー大さじ 1.5
赤唐辛子 1.5 本，砂糖小さじ 1，ココナッツミルク，バジル<P>スパイス
コリアンダー，クミン<P>作り方
<OL><LI>材料をペースト状にして，カレー

Fig. 1. Example HTML document

(i.e., the length of A) is 20 at maximum. The tags (<*tag1*> and <*tag2*>) can be either a start tag (e.g., <A>) or an end tag (e.g., </A>). This score is intended to give high values for words that are emphasized or occur in itemizations or tables. For example, in the HTML document in Fig. 1, the words "タイ風・カレー (Thai-curry)", "材料 (ingredient)", "スパイス (spice)", "コリアンダー，クミン (coriander, cumin)", and "作り方 (recipe)" are counted.

Finally, $dfidf(C, A)$, which reflects the first observation, is calculated as:

$$dfidf(C, A) = df(A, LD(C)) \cdot idf(A), \quad idf(A) = \log \frac{|G|}{df(A, G)}.$$

$df(A, X)$ denotes the number of documents where A appears in documents X. G is a large set of randomly collected Web documents, which we call the *global document set*. We derived this score from a similar score, which was used in [14] to measure the association between a hypernym and hyponyms.

3 Evaluation Criteria

This section presents the evaluation criteria based on *question-answerability* (QA tests). Based on the criteria, we designed an evaluation procedure where the evaluators were asked to answer either by yes or no to four tests at maximum, i.e., a hyponymy test (Sect. 3.4), a QA test (Sect. 3.1) and a suffix augmented QA test (Sect. 3.2) followed by a generality test (Sect. 3.3).

3.1 Question-Answerability Test

By definitions we used, attributes are what we want to know about the object. Therefore, if A is an attribute of objects of class C, we can arrange questions (consisting of A and C) that require the values for A as the answer. Then someone should be able to answer the questions. For example, we can ask "Who is the director of this movie?" because *director* is an attribute of *movie*. The answer might be someone such as "Stanley Kubrick." We designed the QA test shown in Fig. 2 to assess the correctness of attribute A for class C based on this criterion. Several points should be noted. First, since the value for the attribute is actually defined for the object instance (i.e., $v = A(o), o \in C$), we should qualify class label C using "kono (this)" to refer to an object instance of class C.

Second, since we cannot know what question is possible for A beforehand, we generate all the question types listed in Fig. 2 and ask whether any of them are acceptable.

Are any of the following questions grammatically correct, natural, and answerable?
1. この C の A は何? (kono C no A ha nani?/What is the A of this C?)
2. この C の A は誰? (kono C no A ha dare?/Who is the A of this C?)
3. この C の A はいつ? (kono C no A ha itu?/When is the A of this C?)
4. この C の A はどこ? (kono C no A ha doko?/Where is the A of this C?)
5. この C の A はどれ? (kono C no A ha dore?/Which is the A of this C?)
6. この C の A はいくつ? (kono C no A ha ikutu?/How many is the A of this C?)
7. この C の A はどう? (kono C no A ha dou?/How much is the A of this C?)

Fig. 2. Question-answerability Test

Third, the question should be *natural* as well as grammatically correct. Naturalness was explained to the evaluators as positively determining whether the question can be their first choice in usual conversations. In our point of view, attributes should be important items for people in describing objects. We assumed that attributes that conformed to the naturalness criterion would be such important attributes. For example, *stapler* is not an attribute of *company* in our sense, although almost all companies own *stapler*s. Our naturalness criterion can reflect this observation since the question "What is the stapler of this company?" is unnatural as a first question when talking about a company, and therefore we can successfully conclude that *stapler* is not an attribute. Note that Woods' linguistic test [11] (i.e., whether "the attribute of an object is a value" can be stated or not) cannot reject *stapler* since it does not have the naturalness requirement (e.g., we can say "the stapler of [used by] SONY is Stapler-X").[5] In addition, note that such importances can be assessed more easily in the QA test, since questioners basically ask what they think is important at least at the time of utterance. However, we cannot expect such an implication even though the declarative sentence is acceptable.

Finally, the answer to the question does not necessarily need to be written in language. For example, values for attributes such as *map*, *picture*, and *blueprint* cannot be written as language expressions but can be represented by other media. Such attributes are not rare since we obtain attributes from the Web.

3.2 Suffix Augmented QA Test

Some attributes that are obtained can fail the QA test even if they are correct, especially when the surface form is different from the one they actually mean. This often occurs since Japanese is very elliptic and our method is restricted to single word attributes. For example, the word seito (students) can be used to represent the attribute seito suu (number of students) as in the sentence below.

kono	gakko	no	seito	ha	500	nin
this	school	of	students	is	500	NUM

(The number of students of this school is 500.)

[5] *Stapler* might be an important attribute of companies for stationery sellers. However, we focus on attributes that are important for *most people* in *most situations*.

- 数 (number of)　方法 (method for) 名 (name of)　者 (-er)
- 時間　([amount of] time of) 時刻 (time of)　時期　(period of) 場所 (location of)
- 金額 (amount of money for) 程度 (degree of)　具合 (state of)
- の〜さ (nominalized adjectives e.g., "height of" "prettiness of")

Fig. 3. Allowed augmentation

These attributes whose parts are elided (e.g., seito representing seito suu) are also useful since they are actually used in sentences as in the above example. Therefore, they should be assessed as correct attributes in some way. Although the most appropriate question for seito representing seito suu is (6) in Fig. 2, it is unfortunately ungrammatical since ikutu cannot be used for the number of persons. Therefore, seito representing seito suu will fail the QA test.[6]

In Japanese, most of the elided parts can be restored by adding appropriate suffixes (as "suu" (number of) in the previous example) or by adding "no" + nominalized adjectives. Thus, when the attribute word failed the first QA test, we asked the evaluators to re-do the QA test by choosing an appropriate suffix or a nominalized adjective from the list of allowed augmentations and adding it to the end of the evaluated word. Figure 3 lists the allowed augmentations.[7,8]

3.3 Generality Test

Although our primal aim was to acquire the attributes for a given class, i.e., , to find attributes that are common to all the instances of the class, we found, in preliminary experiments, that some uncommon (but interesting) attributes were assessed as correct according to the QA test depending on the evaluator. An example is *subtitle* for the class *movie*. Strictly speaking, *subtitle* is not an attribute of all movies, since all movies do not necessarily have subtitles. For example, only foreign films have subtitles in Japan. However, we think this attribute is also useful in practice for people who have a keen interest in foreign films. Thus, the evaluators were asked whether the attribute was common for most instances of the class when the attribute was judged to be correct in the QA test. We call attributes that passed this generality test *general attributes*, and those that failed but passed the QA test *relaxed attributes* (note that general attributes is a subset of relaxed attributes). We compare the accuracies for the relaxed and general attributes in the experiments.

[6] Seito (representing *students*) might pass the QA test with question type (2) in Fig. 2. However, this is not always the case since some evaluators will judge the question to be unnatural.

[7] Postposition "no (of)" before the suffix is also allowed to be added if it makes the question more natural.

[8] The problem here might not occur if we used many more question types in the first QA test. However, we did not do this to keep the first QA test simple. With the same motivation, we kept the list of allowed suffixes short (only general and important suffixes). The uncovered cases were treated by adding nominalized adjectives.

3.4 Hyponymy Test

Finally, we should note that we designed the evaluation procedure so that the evaluators could be asked whether candidate A is a hyponym of C before the QA tests. If A is a hyponym of C, we can skip all subsequent tests since A cannot be an attribute of C. We added this test because the output of the system often contains hyponyms and these tend to cause confusion in the QA tests since expression "C no A" is natural even when A is a hyponym of C (e.g., "anime no Dragon Ball (Dragon Ball [of/the] anime)").

4 Experiments

4.1 Experimental Setting

We first selected 32 word classes from 1,589 classes acquired from the Web with an automatic hypernym-hyponym acquisition method [14]. Here, we regarded the hypernym as the class label. Since our purpose was just to evaluate our method for classes from the Web, we selected classes that were obtained successfully. We randomly chose the 22 classes listed in Table 2 for human evaluation from these 32 classes.[9] The hyponyms were used to help the evaluators to disambiguate the meaning of class labels (if ambiguity existed).

To collect $LD(C)$, we used the Web search engine goo (http://www.goo.ne.jp). The size of $LD(C)$ was 857 documents (URLs) on class average. There were about $20,000$ candidate words on class average. As global document set G required for the calculation of $dfidf(C, A)$, we used 1.0×10^6 randomly downloaded Web documents.

Table 2. Classes used in evaluation

都市 (city), 博物館 (museum), 祝日 (national holiday), 警察 (police), 施設 (facility), 大学 (university), 新聞 (newspaper), ごみ (garbage), 神社 (shrine), 鳥 (bird), 病院 (hospital), 植物 (plant), 川 (river), 小学校 (elementary school), 曲 (music tune), 図書館 (library), 支店 (branch office), サイト (web site), 町 (town), センサー (sensor), 研修 (training), 自動車 (car)

We output the top 50 attributes for each class ranked with our proposed method and with alternative methods that were used for comparison. We gathered outputs for all the methods, removing duplication (i.e., taking the set union) to achieve efficient evaluation, and re-sorted them randomly to ensure that the assessment was unbiased. Four human evaluators assessed these gathered attributes class-by-class in four days using a GUI tool implementing the evaluation procedure described in Sect. 3. There were a total of $3,678$ evaluated attributes. Using the evaluation results, we re-constructed the evaluations for the top 50 for each method. The kappa value [15], which indicates inter-evaluator agreement, was 0.533 for the general attribute case and 0.593 for the relaxed attribute case. According to [15], these kappa values indicate "moderate" agreement.

[9] This selection was due to time/cost limitations.

4.2 Accuracy of Proposed Method

Figure 4 has accuracy graphs for the proposed method for relaxed attributes. The graph on the left shows per-evaluator precision when the top n (represented by x axis) attributes were output. The precision is the average over all classes. Although we cannot calculate the actual recall, the x axis corresponds to approximate recall. We can see that ranking with the proposed method has a positive correlation with human evaluation, although the assessments varied greatly depending on the evaluator. The graph on the right shows curves for average (with standard deviation), 3-consensus, and 4-consensus precision. 3-consensus (4-consensus) is precision where the attribute is considered correct by at least three (four) evaluators. Figure 5 has graphs for the general attribute case the same as for the relaxed case. Although there is a positive correlation between ranking with the proposed method and human evaluators, the precision was, not surprisingly, lower than that for the relaxed case. In addition, the lower kappa value (0.533 compared to 0.593 for the relaxed case) indicated that the generality test was harder than the QA tests.

The accuracy of the proposed method was encouraging. Although we cannot easily determine which indicator is appropriate, if we use the majority rule (3-

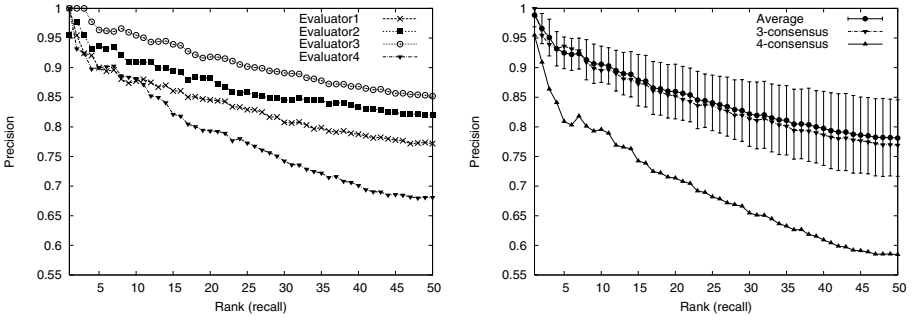

Fig. 4. Accuracy of relaxed attributes

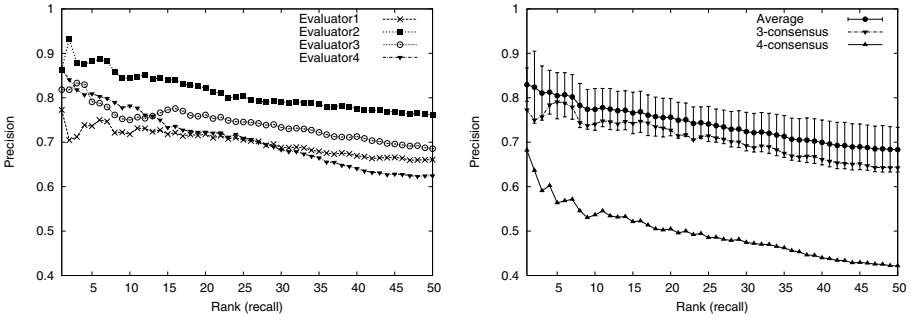

Fig. 5. Accuracy of general attributes

Table 3. Top 20 attributes of several classes obtained by proposed method

Classes	Attributes
鳥 (bird)	写真 (picture)[4/4] 名前 (name)[4/2] 種類 (sort)[4/4] イラスト (illustration)[3/3] 特徴 (characteristics)[4/4] 病気 (disease)[4/2] 生活 (life)[4/4] 話題 (topic)[3/2] 関係 (relation)[0/0] イメージ (image)[4/4] 巣 (nest)[4/4] 鳴き声 (song)[4/4] 姿 (shape)[4/4] 情報 (info.)[4/4] 世界 (world)[0/0] 声 (song)[4/4] 動物 (animal)[0/0] ページ (page)[3/2] 生態 (ecology)[4/4] 羽 (wing)[4/4]
病院 (hospital)	ホームページ (home page)[4/1] 施設 (facility)[3/3] 情報 (info.)[4/4] 紹介 (intro.)[4/4] 窓口 (info. desk)[4/4] 認定 (authorization)[3/3] 名称 (name)[4/2] 医師 (doctor)[4/4] 精神科 (psychiatry)[4/2] 評判 (reputation)[4/4] 対応 (handling)[4/4] 電話 (phone)[2/2] 診療 (medical care)[4/4] 治療 (treatment)[4/4] 医療 (medical service)[3/3] 機能 (function)[3/3] 院長 (director)[4/4] 評価 (valuation)[4/4] 診察 (medical examination)[4/4] ページ (page)[2/2] 管理 (admin.)[4/3] 一部 (part)[1/1]
植物 (plant)	名前 (name)[4/2] 種類 (species)[4/4] 写真 (picture)[4/4] 種子 (seed)[4/4] 栽培 (cultivation)[4/3] 観察 (observation)[4/3] 特徴 (characteristics)[4/4] 説明 (explanation)[4/4] 画像 (image)[4/4] 調査 (surveillance)[4/3] データ (data)[4/4] 進化 (evolution)[3/3] 解説 (description)[4/4] リスト (list)[2/2] 葉 (leaf)[4/3] 保存 (preservation)[2/2] デザイン (design)[1/1] 生育 (growth)[4/4]
川 (river)	水位 (water level)[4/4] 上流 (upstream)[4/4] 名前 (name)[4/2] 環境 (environment)[4/4] 水質 (water quality)[4/4] 歴史 (history)[4/4] 源流 (head stream)[4/4] 写真 (picture)[4/4] 水 (water)[4/4] 水面 (surface)[4/4] 場所 (location)[4/4] 流れ (current)[4/4] 水辺 (waterside)[4/4] 水源 (river head)[4/4] 四季 (four seasons)[3/3] 特徴 (characteristics)[4/4] 中 (inside)[1/1] ほとり (streamside)[4/4] 自然 (nature)[4/4] せせらぎ (babbling)[4/4]
小学校 (elementary school)	活動 (activity)[4/4] 取り組み (efforts)[4/3] 運動会 (athletic meeting)[4/4] 子ども (child)[4/4] ホームページ (home page)[4/0] 校長 (head teacher)[4/4] 教室 (classroom)[4/4] 校歌 (school song)[4/4] 児童 (student)[4/4] 校舎 (school building)[4/4] 行事 (event)[4/4] 学習 (learning)[3/3] 給食 (feeding service)[4/3] ページ (page)[2/2] 体育館 (gym)[4/4] 学級 (class)[3/3] メール (mail)[0/0] 学年 (grade)[1/1] 始業式 (opening ceremony)[4/4] 音楽 (music)[2/2]
曲 (music tune)	歌詞 (lyrics)[4/1] タイトル (title)[4/2] 演奏 (performance)[4/4] リスト (list)[0/0] イメージ (image)[4/4] 作詞 (lyrics writing)[4/1] 楽譜 (musical score)[4/4] 名前 (name)[4/2] 内容 (content)[3/3] ジャンル (genre)[4/4] 情報 (info.)[4/4] ポイント (point)[4/4] 世界 (world)[1/1] メロディー (melody)[4/4] 最後 (end)[3/2] 題名 (title)[4/2] 中 (inside)[0/0] 作曲 (composition)[4/4] テーマ (theme)[4/4] データ (data)[4/2]
図書館 (library)	資料 (source material)[4/4] ホームページ (home page)[4/2] ページ (page)[3/1] 歴史 (history)[4/4] 設置 (establishment)[4/4] システム (system)[4/4] 蔵書 (book stock)[4/4] コピー (copy)[2/2] 本 (book)[4/4] 場所 (location)[4/4] 利用 (use)[4/4] サービス (service)[4/4] データベース (database)[4/3] 図書 (book)[4/4] 新聞 (newspaper)[4/4] 休館 (close)[4/4] 目録 (catalog)[3/3] 展示 (display)[4/2] 施設 (facility)[2/2] 情報 (info.)[4/4]
町 (town)	人口 (population)[4/4] 歴史 (history)[4/4] ホームページ (home page)[4/0] 観光 (sightseeing)[4/4] 情報 (info.)[3/3] 財政 (finance)[4/4] 施設 (facility)[4/4] 文化財 (heritage)[4/2] 環境 (environment)[4/4] 温泉 (hot spring)[3/1] 話題 (topic)[3/2] 四季 (four seasons)[3/3] イベント (event)[4/3] 図書館 (library)[4/3] 文化 (culture)[4/4] 風景 (landscape)[4/4] シンボル (symbol)[4/3] 産業 (industry)[4/3] 農業 (agriculture)[4/2] 議会 (town council)[3/3]
センサー (sensor)	情報 (info.)[4/4] 感度 (sensitivity)[4/3] 種類 (sort)[4/3] 位置 (position)[4/4] 取り付け (install)[4/4] 開発 (development)[4/4] 精度 (accuracy)[4/4] サイズ (size)[4/4] 仕様 (specification)[4/4] 温度 (temperature)[2/1] データ (data)[4/4] セット (set)[4/4] 設置 (install)[4/4] 機能 (function)[4/4] 技術 (technology)[4/4] 特長 (feature)[4/4] ページ (page)[3/3] 高さ (height)[3/2] 採用 (adoption)[3/3] 応用 (application)[4/4]
研修 (training)	内容 (content)[4/4] 目的 (purpose)[4/4] 実施 (practice)[4/4] テーマ (theme)[4/3] プログラム (program)[4/4] 講師 (lecturer)[4/4] 予定 (plan)[4/4] 名称 (name)[4/2] メニュー (menu)[4/4] 報告 (report)[4/4] 対象 (target)[4/4] 成果 (outcome)[4/4] 充実 (satisfaction)[2/2] 場 (place/atmosphere)[3/3] あり方 (state of existence)[2/2] 詳細 (detail)[4/4] 機会 (opportunity)[1/1] 定員 (capacity)[4/4] 受講 (participation)[4/4] ほか (other)[0/0]

consensus in our case) employed in [7], the proposed method obtained relaxed attributes with 0.852 precision and general attributes with 0.727 precision for the top 20 outputs. Table 3 lists the top 20 attributes obtained with the proposed method for several classes. The numeral before (after) "/" is the number of evaluators who judged the attribute as correct as a relaxed (general) attribute. We can see that many interesting attributes were obtained.

4.3 Effect of Scores

In this analysis, we assessed the effect that sub-scores in Eq. (1) had on the acquisition accuracy by observing the decrease in precision when we removed each score from Eq. (1). First, we could observe a positive effect for most scores in terms of the precision averaged over evaluators. Moreover, interestingly, the tendency of the effect was very similar for all evaluators, even though the assessments varied greatly depending on the evaluator as the previous experiment showed. Due to space limitations, we will only present the latter analysis here.

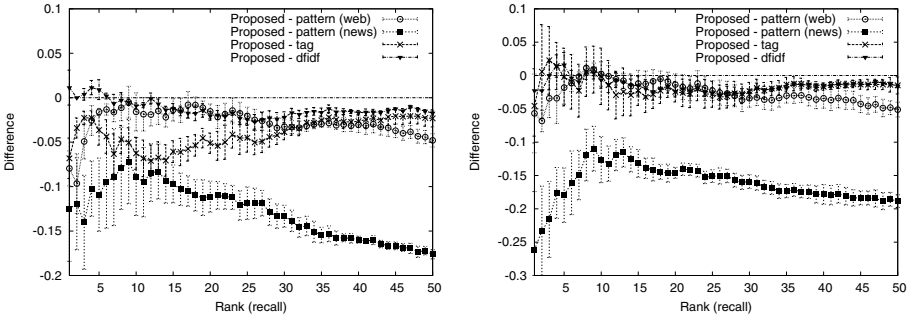

Fig. 6. Effect of scores. Left: relaxed attribute. Right: general attribute.

We calculated the change in precision "per evaluator", and then calculated the averaged change, i.e., the change averaged over evaluators. Figure 6 plots the averaged change and standard deviations. The effect of $n(C, A)$ is represented by "Proposed - pattern (web)", that of $f(C, A)$ by "Proposed - pattern (news)", that of $t(C, A)$ by "Proposed - tag", and that of $dfidf(C, A)$ by "Proposed - dfidf". In the relaxed attribute case, we can see that most of the scores were effective at almost all ranks regardless of the evaluator (negative difference means positive effect). The effect of $f(C, A)$ and $t(C, A)$ was especially remarkable. Although $n(C, A)$ has a similar curve to $f(C, A)$, the effect is weaker. This may be caused by the difference in the number of documents available (As we previously described, we currently cannot obtain a large number of documents from the Web). The effect $dfidf(C, A)$ had was two-fold. This contributed positively at lower ranks but it contributed negatively at higher ranks (around the top 1-5). In the general attribute case, the positive effect became harder to observe although the tendency was similar to the relaxed case. However, we can see that $f(C, A)$ still contributed greatly even in this case. The effect of $t(C, A)$, on the other hand, seems to have weakened greatly.

4.4 Effect of Hypernym

If we have a hypernym-hyponym knowledge base, we can also collect the local document set by using the hyponyms in the class as the keywords for the search engine instead of using the class label (hypernym). In this experiment, we compared the proposed method with this alternative. We collected about the same number of documents for the alternative method as for the proposed method to focus on the quality of collected documents. We used hyponyms with the alternative method instead of class label C in patterns for $n(C, A)$ (thus $n(Hs, A)$ to be precise). $f(C, A)$ was unchanged. Figure 7 plots the results in the same way as for the previous analysis (i.e., difference from the proposed method). We can see that the class label is better than hyponyms for collecting local documents at least in the current setting.

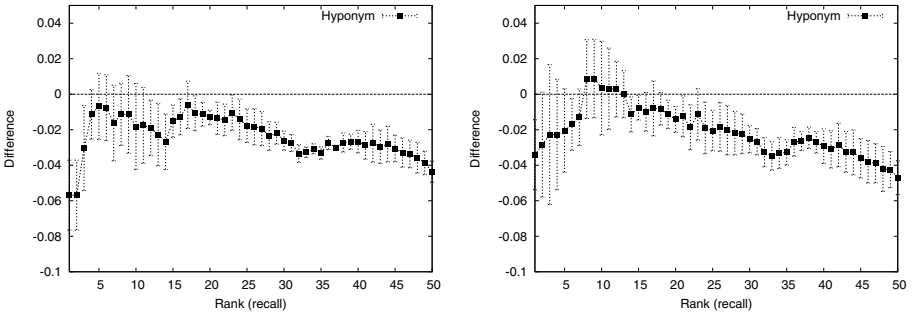

Fig. 7. Effect of hypernyms. Left: relaxed case. Right: general case.

5 Discussion

5.1 Related Work

Several studies have attempted to acquire attributes or attribute-value pairs [1,3,7,8,16]. Yoshida [1] proposed a method of integrating tables on the Web. Although his method consequently acquired attributes, he did not evaluate the accuracy of attributes. Yoshida et al. [16] proposed a method of identifying attribute-value pairs in Web documents. However, since this method only identified the attributes obtained with the method in [1], the coverage might be bounded by the coverage of tables for attributes. Moreover, these methods did not utilize the statistics for words or lexico-syntactic patterns as ours did. Takahashi et al. [8] extracted triples $(object, attribute, value)$ from newspaper articles using lexico-syntactic patterns and statistical scores. However, they focused only on proper nouns and selected the attribute candidates manually. Freishmann et al. [3] extracted attribute-value pairs with a high degree of precision by filtering the candidates extracted with lexico-syntactic patterns by using a model learned with supervised learning. Although this approach is promising, their method was limited to person names and we must prepare training data to apply the method to other types of objects.

5.2 Future Directions

Clues based on QA tests. The current ranking, Eq. (1), does not exploit the observation behind the criteria in Sect. 3. Only the lexico-syntactic patterns "C no A" slightly reflect the criteria. Higher accuracy might be achieved by using patterns that directly reflect the QA tests, e.g., statistics from FAQ lists. The hyponym tests in Sect. 3.4 can also be reflected if we use a hyponymy database. In addition, it is not surprising that the proposed method was not efficient at acquiring general attributes since the score was not meant for that (although the use of class labels might be a contributing factor, ambiguous class labels

cause problems at the same time). The hyponym database might be exploited to measure the generality of attributes.

Full use of the Web. The current method cannot use all Web documents due to limitations with search engines. The more Web documents we have, the more useful the score $n(C, A)$. We are currently planning to prepare our own non-restricted Web repository. Using this, we would also like to elaborate on the comparison described in Sect. 4.4 between the use of hypernyms (class labels) and hyponyms (instance words) in collecting the local document set.

Assessment of Coverage. Currently, the actual recall with the proposed method is unknown. It will be important to estimate how many attributes are needed for practical applications, e.g., by manually analyzing the use of pattern "C no A" exhaustively for a certain class, C. In addition, since we selected classes that were successfully obtained with a hyponymy acquisition method, we cannot deny the possibility that the proposed method has been evaluated for the classes for which reliable statistics can easily be obtained. Thus, the evaluation of more difficult (e.g., more infrequent) classes will be an important future work.

Type Acquisition. What types of questions and what types of suffix augmentations are possible for a given attribute (i.e., the type of attribute value) might also be useful, e.g., in value extraction and in determining type of the attribute (in the sense of "property or part-of"). This was left for the evaluators to chose arbitrarily in this study. We would like to extract such knowledge from the Web using similar techniques such as word statistics and lexico-syntactic patterns.

6 Conclusion

We presented a method of acquiring attributes that utilizes statistics on words, lexico-syntactic patterns, and HTML tags. We also proposed criteria and an evaluation procedure based on question-answerability. Using the procedure, we conducted experiments with four human evaluators. The results revealed that our method could obtain attributes with a high degree of precision.

References

1. Yoshida, M.: Extracting attributes and their values from web pages. In: Proc. of the ACL 2002 Student Research Workshop. (2002) 72–77
2. Yoshida, M., Torisawa, K., Tsujii, J.: Integrating tables on the world wide web. Transactions of the Japanese Society for Artificial Intelligence **19** (2004) 548–560
3. Fleischman, M., Hovy, E., Echihabi, A.: Offline strategies for online question answering: Answering questions before they are asked. In: Proc. of ACL 2003. (2003) 1–7
4. Almuhareb, A., Poesio, M.: Attribute-based and value-based clustering: An evaluation. In: Proc. of EMNLP 2004. (2004) 158–165
5. Fellbaum, C., ed.: WordNet: An electronic lexical database. The MIT Press (1998)

6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. of COLING '92. (1992) 539–545
7. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proc. of ACL '99. (1999)
8. Takahashi, T., Inui, K., Matsumoto, Y.: Automatic extraction of attribute relations from text (in Japanese). IPSJ, SIG-NLP. NL-164 (2004) 19–24
9. Guarino, N.: Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge base. Data and Knowledge Engineering (1992) 249–261
10. Pustejovsky, J.: The Generative Lexicon. The MIT Press (1995)
11. Woods, W.A.: What's in a Link: Foundations for Semantic Networks. In: Representation and Understanding: Studies in Cognitive Science. Academic Press (1975)
12. Kurohashi, S., Nagao, M.: Japanese morphological analysis system JUMAN version 3.61 manual (1999)
13. Kanayama, H., Torisawa, K., Mitsuishi, Y., Tsujii, J.: A hybrid Japanese parser with hand-crafted grammar and statistics. In: Proc. of COLING 2000. (2000) 411–417
14. Shinzato, K., Torisawa, K.: Acquiring hyponymy relations from web documents. In: Proc. of HLT-NAACL04. (2004) 73–80
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorial data. Biometrics **33** (1977) 159–174
16. Yoshida, M., Torisawa, K., Tsujii, J.: Chapter 10 (Extracting Attributes and Their Values from Web Pages). In: Web Document Analysis. World Scientific (2003)