Ranking of Coordinate Terms and Hypernyms Using a Hypernym-Hyponym Dictionary

Kosetsu Tsukuda, Hiroaki Ohshima, Katsumi Tanaka
Department of Social Informatics
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{tsukuda, ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

***Abstract*—In this paper, methods for ranking coordinate terms and hypernyms of a given query according to their appropriateness are proposed. Although previous studies have proposed methods for discovering coordinate terms or hypernyms of a query, they focused on only discovering such terms and evaluating discovered terms based on a binary evaluation: appropriate or inappropriate. Unlike these studies, we rank coordinate terms and hypernyms of a query and evaluate the terms by considering their appropriateness. In the proposed method, a bipartite graph is created based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. Subsequently, we apply a HITS-based algorithm to the bipartite graph and rank coordinate terms and hypernyms based on their appropriateness. The experimental results obtained using 50 queries demonstrate that our method could rank appropriate coordinate terms and hypernyms higher than other comparable methods.**

Keywords*—*Coordinate term, Hypernym, Wikipedia

I. Introduction

Given a term t, there are various types of relationships between t and other terms. For example, hypernyms and hyponyms are defined as terms that are more general and specific than t, respectively. A synonym is a term with the same meaning for t as another term, and a coordinate term is a term that has one or more common hypernyms with t. There are also other relationships such as antonyms and related terms. This study focuses on hypernyms and coordinate terms to identify appropriate hypernyms and coordinate terms for a given query.

Discovering coordinate terms for a given query is useful in various situations. For instance, suppose a user inputs a query to a Web search engine, and is not familiar with Web search or does not have sufficient knowledge about the search domain. In such a case, displaying coordinate terms of the query would support his Web search. For example, if a user needs information about digital cameras but knows only "LUMIX," then displaying appropriate coordinate terms, such as "EXLIM," "FinePix," and "Cyber-Shot" for comparison may be useful to him. Similarly, discovering hypernyms of terms is also useful in some situations such as connecting diverse concepts to form a semantic taxonomy [1].

Some studies have proposed methods for discovering coordinate terms or hypernyms of a term [1]–[8]. The aim of these studies is only discovering these terms from unstructured data such as Web pages and query logs of a commercial search engine. The studies evaluate discovered hypernyms or coordinate terms based on a binary evaluation: appropriate or inappropriate. In this paper, we use a hypernym-hyponym dictionary (described in Section III-A), which enables us to easily obtain hypernyms and coordinate terms of a given term. However, from the dictionary we obtain a large number of coordinate terms and hypernyms. For example, for the query "Lionel Messi," we obtain 16 hypernyms and 112,489 coordinate terms from the dictionary. However, as will be described in Section III-B, there are appropriate and inappropriate hypernyms as well as coordinate terms among these results. Thus, although both "Cristiano Ronaldo" and "Stevie Wonder" are coordinate terms of "Lionel Messi," "Cristiano Ronaldo" is more appropriate than "Stevie Wonder." Similarly, for "Lionel Messi," "football player" is a more appropriate hypernym than "human beings."

In this research, we propose methods for ranking coordinate terms and hypernyms of a term based on their appropriateness. Our method first creates a bipartite graph based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. We apply a HITS-based algorithm to the graph and rank coordinate terms and hypernyms based on their appropriateness. Although we use a Japanese hypernym-hyponym dictionary, our methods are language-independent.

The experimental results obtained using 50 queries demonstrate that our method could rank appropriate coordinate terms and hypernyms higher than other comparable methods.

The contributions of this study are twofold:

- We propose methods for ranking coordinate terms and hypernyms by considering their appropriateness. Most of the previous studies have focused on only discovering coordinate terms and hypernyms for a given query, whereas our objective is the ranking of coordinate terms and hypernyms according to their appropriateness.
- We evaluate coordinate terms and hypernyms based on their appropriateness. Most previous studies have evaluated discovered coordinate terms and hypernyms based on a binary evaluation, whereas we evaluate coordinate terms and hypernyms by considering their appropriateness.

The remainder of the paper is organized as follows. In Section II, we discuss previous related studies. In Section III, we describe the hypernym-hyponym dictionary used in this study and our proposed method. In Section IV, we report our evaluation experiments. In Section V, we discuss the results obtained. Finally, in Section VI, we provide our conclusion and present possible suggestions for future studies.

II. Related Work

A. Coordinate Term Mining

Various methods have been proposed for collecting coordinate terms for a given query. Ohshima *et al.* [4] used a Web search engine to perform two searches where queries were generated by connecting the user's query with the conjunction "OR." Yamaguchi *et al.* [8] utilized query logs of a Web search engine with the basic assumption that terms representing common topics tend to co-occur with coordinate terms in query logs of search engines. Wang and Cohen [7] and Kawai *et al.* [3] proposed methods for mining coordinate terms from seed terms based on the bootstrapping technique using table tags and list tags in HTML documents.

Although these studies aimed to discover coordinate terms from the Web, we can easily get many coordinate terms from a hypernym-hyponym dictionary. Our study differs from previous studies, because our objective is the ranking of such coordinate terms according to their appropriateness.

B. Hypernym Mining

Using lexical patterns is a popular approach to automatically discover relations between hypernyms and hyponyms from given text. Hearst [2] proposed some effective lexical patterns, for example, "such as" and "including." Snow *et al.* [6] learned syntactic dependency paths automatically by using hypernym/hyponym word pairs from WordNet [9], [10]. Ritter *et al.* [5] proposed methods that used machine learning techniques and lexical patterns. Although Shinzato and Torisawa [1] did not use lexical patterns, they used clues such as itemization or listing in HTML documents as well as statistical measures such as document frequencies or verb-noun co-occurrences.

As is the case with coordinate terms, the aims of these studies are only the discovery of hypernyms and hyponyms, but our aim is to rank hypernyms obtained from a dictionary according to their appropriateness.

C. Related Term Mining

Many methods that discover semantically related terms for a given query have also been proposed. Chen *et al.* [11] proposed an approach to automatically construct a domain-specific thesaurus based on hyperlink structure analysis. Ito *et al.* [12] proposed a method for constructing an association thesaurus from Wikipedia [1] based on link co-occurrences. For example, given a query "Google," these studies can discover coordinate terms such as "Yahoo!" and "Oracle Corporation." However, terms that are not coordinate terms such as "Sergey Brin" and "PageRank" are also included.

[1]http://ja.wikipedia.org/

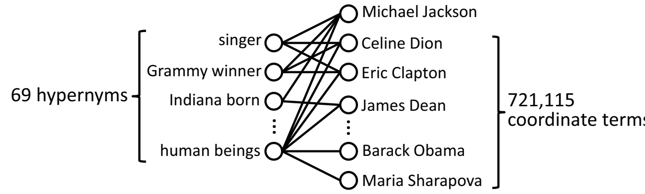

Fig. 1. Examples of Michael Jackson's hypernyms and coordinate terms.

III. Method

In this section, we describe the hypernym-hyponym dictionary used in this study, discuss the characteristics of appropriate coordinate terms and hypernyms, and present methods to rank these coordinate terms and hypernyms.

A. Hypernym-Hyponym Dictionary

In this research, we use an open source "hypernym/hyponym extraction tool [2]." This tool contains approximately 200,000 hypernyms and approximately 2.45 million hyponyms. These hierarchized terms are category names and nouns that occur in the titles of articles in Japanese Wikipedia. Using this data, we can easily extract hypernyms of a term and coordinate terms that have hypernyms in common with the term. For instance, "Michael Jackson" has 69 hypernyms such as "singer" and "Guinness world record holder." Thus, if a term has at least one common hypernym with "Michael Jackson," then the term is a coordinate term of "Michael Jackson," and "Michael Jackson" has 721,115 coordinate terms (Figure 1).

B. Characteristics of Appropriate Coordinate Terms and Appropriate Hypernyms

In this research, we define a coordinate term of a term q as "a term that has one or more common hypernyms with q," as defined by Ohshima *et al.* [4]. Similarly, a hypernym of a term is defined as "a term that is more general than q." However, among coordinate terms and hypernyms obtained using the aforementioned dictionary, there are gaps in the degrees of appropriateness of coordinate terms and hypernyms.

First, we studied the characteristics of appropriate coordinate terms of q and found the following characteristics:

(1-A) An appropriate coordinate term shares many hypernyms with q.

(1-B) An appropriate coordinate term shares hypernyms that have fewer hyponyms with q.

We explain these characteristics using "Lionel Messi" as an example. Thus, given two terms, "Cristiano Ronaldo" and "Stevie Wonder," "Cristiano Ronaldo" is a more appropriate coordinate term of "Lionel Messi," which can be explained by considering (1-A). In this case, "Stevie Wonder" shares only one hypernym, "human beings," with "Lionel Messi" in the dictionary, whereas "Cristiano Ronaldo" shares both "human beings" and "football player." Similarly, given two additional terms, "Wayne Rooney" and "Jorge Luis Borges," "Wayne Rooney" is more appropriate as a coordinate term

[2]http://nlpwww.nict.go.jp/hyponymy/index.html

of "Lionel Messi." However, when we consider only (1-A), the appropriateness of these two terms is equivalent, because "Wayne Rooney" shares two hypernyms, "human beings" and "football player," with "Lionel Messi" and "Jorge Luis Borges" shares two hypernyms, "human beings" and "from Argentina." Hence, in this case, the difference can be explained by considering (1-B); the number of "football player" is fewer than "from Argentina." Therefore "Wayne Rooney" is more appropriate as a coordinate term.

Second, we studied the characteristics of appropriate hypernyms of q and determined the following characteristics:

(2-A) An appropriate hypernym has only appropriate coordinate terms of q as its hyponyms.

(2-B) An appropriate hypernym has many appropriate coordinate terms of q as its hyponyms.

We will also explain these characteristics using "Lionel Messi" as an example. Thus, given two hypernyms, "football player" and "human beings," "football player" is a more appropriate hypernym of "Lionel Messi," which can be explained by considering (2-A). In this case, "human beings" has appropriate coordinate terms of "Lionel Messi" such as "Cristiano Ronaldo" and the inappropriate coordinate terms such as "Stevie Wonder" and "Barack Obama," but "football player" only has appropriate coordinate terms such as "Cristiano Ronaldo" and "Wayne Rooney." Similarly, given two additional terms, "football player" and "winner of UEFA Best Player in Europe Award," we think "football player" is more appropriate as a hypernym of "Lionel Messi" because "winner of UEFA Best Player in Europe Award" is too narrow as a hypernym. However, when we consider only (2-A), the appropriateness of these two terms is equivalent, because both hypernyms have only appropriate coordinate terms of "Lionel Messi" as their hyponyms. Hence, in this case, the difference can be explained by considering (2-B); "football player" has more coordinate terms of "Lionel Messi" as its hyponyms than "winner of UEFA Best Player in Europe Award." Therefore, "football player" is more appropriate as a hypernym of "Lionel Messi."

C. Ranking of Coordinate Terms

First we will define some symbols. Let q denote a query and $hyper(t)$ and $hypo(t)$ denote the set of hypernyms and set of hyponyms of a term t, respectively. H_q and C_q are defined as follows.

- $H_q = \{x | x \in hyper(q)\}$,
- $C_q = \{x | x \in hypo(y), y \in H_q, x \neq q\}$.

That is, H_q and C_q are the set of hypernyms and the set of coordinate terms of q, respectively.

We consider a bipartite graph $G = (\{q\} \cup C_q \cup H_q, E)$, where E is a set of edges between H_q and $\{q\} \cup C_q$. An edge exists between $h_i \in H_q$ and $c_j \in \{q\} \cup C_q$ when h_i is a hypernym of c_j. When q is "Michael Jackson," Figure 1 represents the bipartite graph.

To calculate the appropriateness of each coordinate term in C_q, we propose a method that reflects characteristics (1-A) and (1-B) based on the HITS [13] algorithm. Originally the HITS algorithm was used to evaluate Web pages based on link structure. In the HITS algorithm, a Web page that provides important information is called an *authority*, and a Web page that links to important authorities is called a *hub*. A good hub is a page that points to many good authorities, and a good authority is a page that is pointed to by many good hubs. In our bipartite graph, a hypernym and a hyponym correspond to a hub and an authority, respectively. We denote the hub score of $h_i \in H_q$ and the authority score of $c_j \in \{q\} \cup C_q$ as $hub(h_i)$ and $authority(c_j)$, respectively, and calculate these scores as follows:

$$hub(h_i) = \sum_{c_j \in \{q\} \cup C_q} w_{ji}^{ch} \cdot authority(c_j), \tag{1}$$

$$authority(c_j) = \sum_{h_i \in H_q} w_{ij}^{hc} \cdot hub(h_i), \tag{2}$$

where w_{ji}^{ch} and w_{ij}^{hc} represent the weight of edges, and w_{ji}^{ch} represents the weight from c_j to h_i. In the HITS algorithm, the weight of an edge is equal to 1 if there is an edge between two vertices, otherwise the weight of an edge is equal to 0. If we apply the HITS algorithm to the bipartite graph G then vertices that have a very large number of hyponyms, such as "human beings" and "from Argentina," have high scores. Thus, each hyponym of "human beings" or "from Argentina" has a misleading high score, and terms sharing hypernyms that have many hyponyms become appropriate coordinate terms of q. To solve this problem, we change the weight of edges from hypernyms to hyponyms by considering the number of hyponyms of each hypernym as mentioned in (1-B). Lempel and Moran [14] proposed the SALSA algorithm, considering the weight of edges in the HITS algorithm. In the SALSA algorithm, the more edges a vertex has, the smaller the weights of the edges become. Specifically the weight of the edge from h_i to c_j is represented by $w_{ij}^{hc} = \frac{1}{|hypo(h_i)|}$.

We set the initial value of q as 1 and the initial values of the remaining vertices as 0, because the objective of our method is to calculate the degree of coordination to q. Let $f_{coordinate}(q, c_j)$ and $f_{multitude}(q, h_i)$ denote the convergent scores of $c_j \in C_q$ and $h_i \in H_q$, respectively. When we rank coordinate terms of q based on their appropriateness, we sort $c_j \in C_q$ in descending order of $f_{coordinate}(q, c_j)$.

D. Ranking of Hypernyms

The score $f_{multitude}(q, h_i)$ reflects only characteristic (2-B) from Section III-B. Therefore, hypernyms such as "human beings" have high scores.

To reflect characteristic (2-A), "an appropriate hypernym has *only* appropriate coordinate terms of q as its hyponyms," we calculate the score of $h_i \in H_q$ as follows:

$$f_{purity}(q, h_i) = \frac{1}{|hypo(h_i)|} \sum_{t_j \in hypo(h_i)} f_{coordinate}(q, t_j) \ . \tag{3}$$

That is, $f_{purity}(q, h_i)$ is the average score of the degree of coordination for all of h_i's hyponyms. Finally the appropriateness score of h_i as a hypernym of q is given by:

$$f_{hypernym}(q, h_i) = f_{purity}(q, h_i)^{\beta} \cdot f_{multitude}(q, h_i)^{(1-\beta)} \ , \tag{4}$$

where β is a parameter that ranges from 0 to 1.

TABLE I. EXAMPLES OF QUERIES (ENGLISH TRANSLATION).

category	queries
person	Paul McCartney, Tom Cruise, Ichiro Suzuki, Ludwig van Beethoven, Nobunaga Oda
place	United Kingdom, Paris, Tokyo, the Pacific Ocean, Brazil
product	digital camera, Nintendo DS, refrigerator, frying pan, organ
facility	department store, the University of Tokyo, Universal Studios Japan, Narita International Airport, Osaka Castle
company	Microsoft Corporation, Panasonic, McDonald's Corporation, Adidas, Toyota Motor Corporation

IV. EXPERIMENTS

This section reports on the evaluation of the proposed methods.

A. Query Set

We created a query set comprising 50 queries in five categories: names of people, places, products, facilities, and companies. Each category contains ten queries. These queries are Wikipedia pages, where the title of the page is the query. If a query is unpopular, evaluating is difficult for assessors. Therefore, we have selected popular queries as follows. First, we compute PageRank [15] scores for all Wikipedia articles based on their link structures. Queries with high PageRank scores are considered popular, and we then select the top 100 queries for each category. Finally, we randomly select ten popular queries for each category. Examples from the query set are presented in Table I.

B. Comparative Methods

1) Coordinate term: In this experiment, two comparative methods were used to compute the degree of coordination. The first method, denoted the CommonHypernym method, hypothesizes that the more hypernyms a term $c_j \in C_q$ shares with a query q, the higher the degree of coordination of c_j becomes. That is, the score of $c_j \in C_q$ is calculated as follows:

$$f_{common\_hypernym}(q, c_j) = |hyper(q) \cap hyper(c_j)| \quad (5)$$

The second method, denoted the SALSA method, sets $w^{ch}_{ji} = \frac{1}{|hyper(c_j)|}$ and $w^{hc}_{ij} = \frac{1}{|hypo(h_i)|}$ in Equation 2. The SALSA method hypothesizes that the fewer hypernyms a term $c_j \in C_q$ shares with q and the fewer hyponyms each of the hypernyms have, the more appropriate coordinate term c_j is. More intuitively, a term that shares only rare hypernyms with q is an appropriate coordinate term of q.

2) Hypernym: Two comparative methods were used to compute the hypernym score. The first method, denoted the ManyHyponyms method, hypothesizes that the more hyponyms a hypernym $h_i \in H_q$ has, the more appropriate hypernym h_i is: i.e., the appropriateness score of h_i is calculated by $|hypo(h_i)|$.

In contrast, the second method, denoted the FewHyponyms method, hypothesizes that the fewer hyponyms a hypernym $h_i \in H_q$ has, the more appropriate hypernym h_i is: i.e., the appropriateness score of h_i is calculated by $\frac{1}{|hypo(h_i)|}$.

C. Evaluation Procedure

1) Evaluation of Coordinate Terms: For a given a query, the proposed method and two comparative methods can calculate the degree of coordination for all coordinate terms of the query. However, the average number of coordinate terms for queries used in this experiment was extremely high (263,143.98 terms per query). Manually evaluating the degree of coordination of all terms is difficult; thus, for a given query, we pooled the top 50 coordinate terms from each method to solve this problem. The pooled terms were then randomly sorted and evaluated.

Assessors were recruited through Lancers [3], which is a popular crowd sourcing marketplace in Japan. First, we presented a query and asked the assessors to label each of the query's coordinate terms from 0 to 2, where 0 indicates that the term is not appropriate as the coordinate term, 1 indicates that the term is reasonably appropriate, and 2 indicates that the term is absolutely appropriate. If the assessors were not able to attribute a score for a coordinate term because they did not understand the term, we asked them to label it "unknown" rather than attributing a score. Each coordinate term was labeled by 11 assessors.

2) Evaluation of Hypernyms: For hypernyms, the average number of hypernyms of queries used in this experiment was reasonable (46.4 hypernyms per query). Thus, we used all hypernyms of the queries in this experiment. Again, we used Lancers to recruit assessors. Initially, we displayed a query and asked the assessors to label each of its hypernyms on a scale from 0 to 2. For a given hypernym, 0 indicates that the term is not appropriate, 1 indicates that the term is reasonably appropriate, and 2 indicates that the term is absolutely appropriate. If the assessors were not able to label the score for a hypernym because they did not understand the term, we asked them to label it "unknown" rather than attributing a score. Each hypernym was labeled by 11 assessors.

D. Evaluation Metrics

We used Normalized Discounted Cumulated Gain (nDCG) [16] and Mean Average Precision (MAP) as evaluation metrics. To compute both metrics for coordinate terms, we first listed coordinate terms that more than seven assessors had labeled "unknown." Hereafter, we denote such terms "unknown terms." As mentioned previously, each of the three methods has a term list of the top 50 ranked coordinate terms. Unknown terms were discarded from the list, and the remaining coordinate terms were re-ranked according to their degrees of coordination. Then, we computed the average assessor scores for each remaining coordinate term and regarded this score as the answer score. To compute both metrics for hypernyms, we followed a similar procedure and computed the answer score for each hypernym.

To compute the MAP for coordinate terms, the coordinate terms must be divided into two groups: appropriate and inappropriate coordinate terms. In this experiment, we considered

[3] http://www.lancers.jp/

TABLE II. MAP OF COORDINATE TERMS. THE HIGHEST SCORES IN EACH CATEGORY ARE INDICATED IN BOLD. PAIRED t-TESTS WITH BONFERRONI CORRECTIONS WERE USED FOR SIGNIFICANCE TESTING. SIGNIFICANT DIFFERENCES BETWEEN THE PROPOSED METHOD AND COMMONHYPERNYM ARE INDICATED BY $*(\alpha = 0.05)$, AND SIGNIFICANT DIFFERENCE BETWEEN THE PROPOSED METHOD AND SALSA IS INDICATED BY $\dagger\dagger$ $(\alpha = 0.01)$.

	CommonHypernym	SALSA	Proposed
person	0.535	0.557	**0.578**
place	0.505	0.535	**0.549**
product	0.425	0.468	**0.548***
facility	0.701	0.646	**0.714**
company	0.637	0.601	**0.651**
all categories	0.561	0.560	**0.608***††

TABLE III. COMPARISON OF NDCG AMONG ALL METHODS. THE HIGHEST SCORES AT EACH RANK ARE SHOWN IN BOLD. PAIRED t-TESTS WITH BONFERRONI CORRECTIONS WERE USED FOR SIGNIFICANCE TESTING. SIGNIFICANT DIFFERENCES BETWEEN THE PROPOSED METHOD AND COMMONHYPERNYM ARE INDICATED BY $*(\alpha = 0.05)$.

	CommonHypernym	SALSA	Proposed
@5	0.709	0.709	**0.743**
@10	0.713	0.715	**0.747***
@20	0.732	0.739	**0.762***
@30	0.769	0.774	**0.793**

TABLE IV. NDCG FOR EACH CATEGORY COMPUTED BY THE PROPOSED METHOD.

	person	place	product	facility	company
@5	0.734	0.705	0.689	0.773	0.814
@10	0.766	0.709	0.694	0.766	0.799
@20	0.798	0.745	0.700	0.779	0.787
@30	0.823	0.789	0.731	0.818	0.803

coordinate terms with answer scores ≥ 1 as appropriate, while coordinate terms with answer scores < 1 were treated as inappropriate terms. Hypernyms were treated in the same manner and were also divided into two groups: appropriate and inappropriate.

E. Results

1) Results of Coordinate Terms: Table II presents the MAP for each category. Paired t-tests with Bonferroni corrections were used for significance testing. The proposed method significantly outperformed both the CommonHypernym and SALSA methods for the average of 50 queries. Moreover, the proposed method outperformed other comparable methods in all five categories.

Table III presents a comparison of nDCG for all methods. Although nDCG at rank 40 or 50 cannot be calculated for some queries because unknown terms were discarded (See Section IV-D), the nDCG at rank ≤ 30 can be calculated for all queries. Thus, the nDCG at rank 5, 10, 20 and 30 are presented in Table III. Paired t-tests with Bonferroni corrections were used for significance testing. The results obtained indicate that the proposed method achieved the highest nDCG at any rank (from nDCG@5 to nDCG@30), and this method significantly outperformed the CommonHypernym method at rank 10 and 20.

Table IV presents the nDCG for queries from each category computed by the proposed method. From the results in Table IV, we can say that the proposed method was effective in the person, facility, and company categories, and less effective in product category.

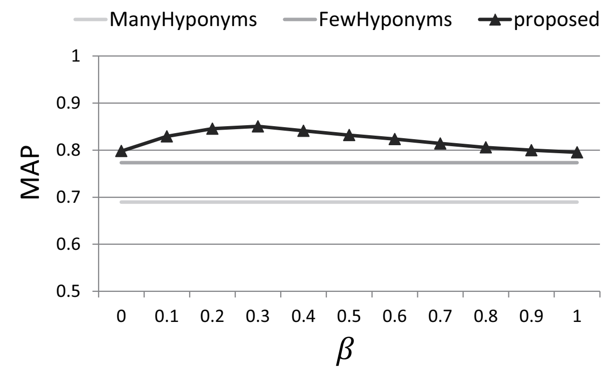

Fig. 2. MAP for the average of 50 queries in each method (β ranges from 0 to 1 in increments of 0.1).

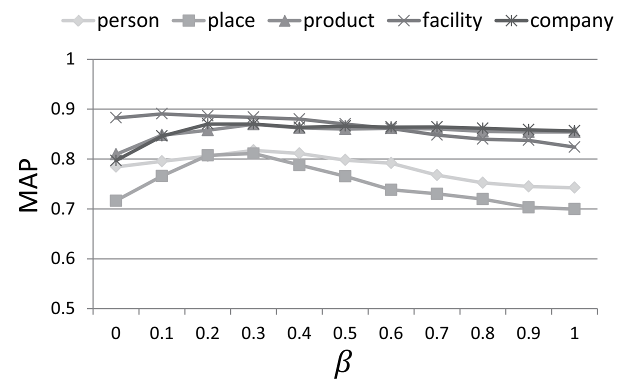

Fig. 3. MAP in each category (β ranges from 0 to 1 in increments of 0.1).

2) Results of Hypernyms: Figure 2 presents MAP result comparisons for the average of 50 queries for all methods. The proposed method has a parameter β, which ranges from 0 to 1 in increments of 0.1. Two comparative methods have no parameter, and have scored constant MAP values regardless of β. Figure 2 determines that the proposed method outperformed two comparative methods for any value of β. The proposed method achieved the highest value (0.850) when β was 0.3, indicating the effectiveness of considering the characteristics of both (2-A) and (2-B) from Section III-B.

Figure 3 illustrates the MAP for each category when β ranged from 0 to 1 in increments of 0.1. The MAP achieved the highest value when β was 0.1 in the facility category and 0.3 in other categories.

Figure 4 presents the average nDCG for the average of 50 queries when β ranged from 0 to 1 in increments of 0.1. At any rank, the nDCG achieved the highest value when β was 0.3 (@5, @20, and @30) or 0.4 (@10). These results indicates the effectiveness of combining the characteristics of both (2-A) and (2-B) from Section III-B.

V. DISCUSSION

In this section, we discuss the results with some specific examples.

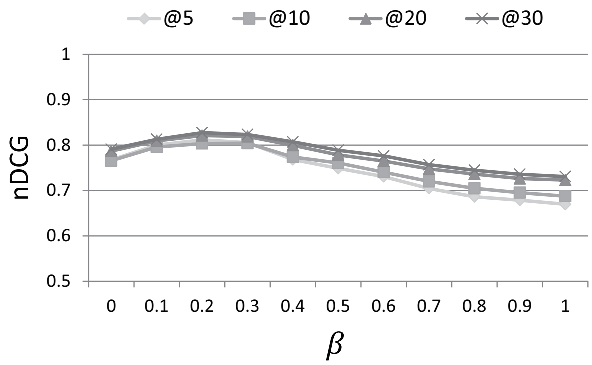

Fig. 4. nDCG of all queries. (β ranges from 0 to 1 in increments of 0.1).

A. Coordinate Terms

Table V presents the results for an example in which the proposed method determined appropriate coordinate terms with high accuracy. Each column of the table displays the top 20 coordinate terms of the CommonHypernym method, SALSA method, and the proposed method, as well as the top 20 terms in terms of answer scores.

Table V shows the results for the query "Paul McCartney." In the answer data, famous western singers were regarded as appropriate coordinate terms of "Paul McCartney," and the proposed method ranked such terms higher. In the CommonHypernym method, terms that do share many hypernyms with the query were ranked higher. However, the method does not consider the importance of each hypernym. Terms such as "Keisuke Kuwata" and "Tsuyoshi Nagabuchi," which are names of famous Japanese singers, were labeled as inappropriate coordinate terms by many assessors and were ranked higher in the CommonHypernym method. They share unimportant hypernyms such as "a singer who plays different instruments when playing different forms of music" with "Paul McCartney." The SALSA method also placed high priority on such hypernyms; thus, the nDCG was lower than that of the proposed method.

According to our observations, there are two principal cases when our methods did not work efficiently. The first case is for a query that has multiple meanings. For example, "Japan Sea" has totally 31 hypernyms. Among the hypernyms of the query, 13 hypernyms are related to a train's name, six hypernyms are related to a sea's name, and eight hypernyms are related to a song's name. In the proposed method, only the names of trains, such as "Twilight Express" and "Hatsukari," were included in top 50 coordinate terms because our method was profoundly affected by hypernyms that were related to a train's name. Average people will think that the names of seas, such as "the Pacific Ocean" and "Okhotsk Sea," are appropriate coordinate terms of "Japan Sea," and they do not know that "Japan Sea" could be related to the name of a train or a song. Thus, the appropriateness of names of trains and songs are low, and the proposed method does not achieve satisfying results. One approach to solve this problem is to cluster hypernyms based on n-gram similarities between hypernyms and the degree of duplication of their hyponyms, and to discover appropriate coordinate terms in each cluster using the cluster's hypernyms.

Another case is for a query that has few hypernyms. For example, the query "vending machine" had only two hypernyms, "sales method" and "business operator/distributor." In the proposed method, 49 terms had the same degree of coordination and were ranked first. This result defies our objective, which is to rank coordinate terms according to appropriateness. One approach to solve this problem is to combine the hypernym-hyponym dictionary used in this research with other dictionaries, such as WordNet [9], [10]. This would enable us to obtain more hypernyms and hyponyms, and to construct a larger bipartite graph. Another approach is to collect hypernyms and hyponyms that are not included in dictionaries using methods proposed by related work discussed in Section II.

B. Hypernyms

Table VI presents results for an example for which the proposed method determined appropriate hypernyms with high accuracy. The table presents the top 10 hypernyms from the proposed method and the top 10 terms in terms of answer scores.

Table VI presents the results of a query "Nintendo DS." When β was 0, hypernyms with many hyponyms, such as "work" and "product," were ranked higher. When β was 1, hypernyms labeled inappropriate because of the meaning being too narrow, such as "brain training game", were ranked higher. When β was 0.3, the results were well balanced and achieved the best nDCG value of 0.837.

VI. Conclusion

In this paper we have proposed methods for ranking coordinate terms and hypernyms of a query according to their appropriateness. The proposed method first creates a bipartite graph based on hypernyms of a query and hyponyms of each hypernym using a hypernym-hyponym dictionary. Subsequently, we applied a HITS-based algorithm to the graph and ranked coordinate terms and hypernyms. The experimental results using 50 queries indicate that the proposed method can rank appropriate coordinate terms and hypernyms higher than other comparable methods.

In the future, we will conduct more detailed experiments. Although we discarded terms that assessors did not understand, we plan to allow assessors to search the meanings of unknown terms and label their appropriateness, which will enable us to evaluate methods more accurately and will facilitate more in depth discussions.

In this paper, we only targeted queries that occur in the titles of articles in the Japanese Wikipedia because we use a hypernym/hyponym extraction tool. Thus, in order to solve this problem, applying the proposed method to other data, such as WordNet [9], [10], would also be work of future interest.

Acknowledgements

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013, 24680008, and 12J03993) from MEXT of Japan.

TABLE V. RANKING RESULTS FOR COORDINATE TERMS FROM THE PROPOSED METHOD AND COMPARISON METHODS FOR THE QUERY "PAUL MCCARTNEY" (NUMBERS IN THE PARENTHESES INDICATE ANSWER SCORE).

rank	CommonHypernym	SALSA	Proposed	answer data
1	Elton John	Elton John	Eric Clapton	John Lennon (1.82)
2	Eric Clapton	Sting	Ringo Starr	BEATLES (1.82)
3	Sting	Eric Clapton	Elton John	Ringo Starr (1.70)
4	John Lennon	John Lennon	John Lennon	Michael Jackson (1.50)
5	Keisuke Kuwata	Ringo Starr	David Bowie	George Harrison (1.45)
6	Mariah Carey	Keisuke Kuwata	BEATLES	Linda McCartney (1.43)
7	Stevie Wonder	Mariah Carey	Sting	Elton John (1.29)
8	Mick Jagger	Stevie Wonder	Celine Dion	Wings (1.25)
9	Paul Simon	George Harrison	Mariah Carey	Stevie Wonder (1.20)
10	Tsuyoshi Nagabuchi	Mick Jagger	George Harrison	Prince (1.14)
11	Keith Richards	Aerosmith	U2	Paul Simon (1.11)
12	Aerosmith	Michael Jackson	Bon Jovi	Eric Clapton (1.10)
13	Michael Jackson	Prince	Jeff Beck	Janet Jackson (1.0)
14	Prince	Tsuyoshi Nagabuchi	Prince	Rod Stewart (1.0)
15	U2	Bob Dylan	Mick Jagger	Bob Dylan (1.0)
16	Neil Young	Paul Simon	George Michael	George Michael (1.0)
17	Bryan Adams	Masaharu Fukuyama	Aerosmith	Mariah Carey (1.0)
18	Rod Stewart	Keith Richards	Stevie Wonder	Tina Turner (1.0)
19	Tomoyasu Hotei	U2	Wings	Bjork (1.0)
20	KinKi	Bryan Adams	Paul Simon	Richard (1.0)
	nDCG@20 = 0.808	nDCG@20 = 0.817	nDCG@20 = 0.879	nDCG@20 = 1.0

TABLE VI. RANKING RESULTS OF HYPERNYMS FOR THE PROPOSED METHOD FOR THE QUERY "NINTENDO DS" (NUMBERS IN THE PARENTHESES INDICATE ANSWER SCORE).

rank	$\beta = 0$	$\beta = 0.3$
1	work	game device
2	appearance work	home computer game
3	game	peripheral device
4	game device	portable game device
5	home computer game	game hardware that uses ROM software
6	product	game software
7	game work	portable game device
8	biggest-selling computer game	Nintendo hardware
9	song content	computer software · game
10	Gundam series game	consumer game
nDCG@10	0.571	0.837

rank	$\beta = 1$	answer data
1	computer software · game	portable game device (2.00)
2	peripheral device	game device (1.91)
3	game software	computer game (2.00)
4	Nintendo hardware	Nintendo hardware (1.91)
5	game hardware that uses ROM software	game (1.91)
6	available terminal	home computer game (1.91)
7	brain training game	portable video game player (1.91)
8	goods · service	product (1.80)
9	portable video game player	Nintendo software (1.73)
10	portable game device	consumer game (1.70)
nDCG@10	0.703	1.0

REFERENCES

[1] K. Shinzato and K. Torisawa, "Acquiring hyponymy relations from web documents." in *HLT-NAACL*, 2004, pp. 73–80.

[2] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proc. of the 14th Conference on Computational Linguistics - Volume 2*, ser. COLING '92, 1992, pp. 539–545.

[3] H. Kawai, H. Mizuguchi, and M. Tsuchida, "Cost-effective web search in bootstrapping for named entity recognition," in *Database Systems for Advanced Applications*, ser. Lecture Notes in Computer Science, vol. 4947, 2008, pp. 393–407.

[4] H. Ohshima, S. Oyama, and K. Tanaka, "Searching coordinate terms with their context from the web," in *Proc. of the 7th International Conference on Web Information Systems*, ser. WISE'06, 2006, pp. 40–47.

[5] A. Ritter, S. Soderland, and O. Etzioni, "What is this, anyway: Automatic hypernym discovery," in *AAAI Spring Symposium: Learning by Reading and Learning to Read*, 2009, pp. 88–93.

[6] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," in *Advances in Neural Information Processing Systems 17*, 2005, pp. 1297–1304.

[7] R. C. Wang and W. W. Cohen, "Language-independent set expansion of named entities using the web," in *Proc. of the 2007 Seventh IEEE International Conference on Data Mining*, ser. ICDM '07, 2007, pp. 342–350.

[8] M. Yamaguchi, H. Ohshima, S. Oyama, and K. Tanaka, "Unsupervised discovery of coordinate terms for multiple aspects from search engine query logs," in *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '08, 2008, pp. 249–257.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Mit Press, 1998.

[10] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[11] Z. Chen, S. Liu, L. Wenyin, G. Pu, and W.-Y. Ma, "Building a web thesaurus from web link structure," in *Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03, 2003, pp. 48–55.

[12] M. Ito, K. Nakayama, T. Hara, and S. Nishio, "Association thesaurus construction methods based on link co-occurrence analysis for wikipedia," in *Proc. of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08, 2008, pp. 817–826.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, pp. 604–632, September 1999.

[14] R. Lempel and S. Moran, "SALSA: the stochastic approach for link-structure analysis," *ACM Trans. Inf. Syst.*, vol. 19, pp. 131–160, April 2001.

[15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proc. of the 7th International Conference on World Wide Web*, ser. WWW'07, 1998, pp. 107–117.

[16] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.