A Query Classification System Based on Snippet Similarity for a One-Click Search

Tatsuya Tojima
Nagaoka University of Technology
Nagaoka-shi, Niigata, Japan
tojima@stn.nagaokaut.ac.jp

Takashi Yukawa
Nagaoka University of Technology
Nagaoka-shi, Niigata, Japan
yukawa@vos.nagaokaut.ac.jp

ABSTRACT

This paper proposes a query classification system for a one-click search system that uses feature vectors based on snippet similarity. The proposed system targets the NTCIR-10 1CLICK-2 query classification subtask and classifies queries in Japanese and English into eight predefined classes by using support vector machines (SVMs). In the NTCIR-9 and NTCIR-10 tasks, most participants used complex features or rules that depend strongly on language characteristics. The authors propose a new method that uses feature vectors created by using snippet similarities instead of the above mentioned features. In the proposed method, feature vectors have fewer dimensions, provide better generalization, lower language dependency, and reduced computer resources. This method achieved accuracies of 0.93 for a Japanese task and 0.91 for an English task.

General Terms:

Machine Learning, Web Search

Keywords:

Query Classification, Dimension Reduction, Intent, Mobile

1. INTRODUCTION

The purpose of the 1CLICK search for web searches is to provide users with simple summarized information after clicking the search button. This information is suitable for small screens, such as cellular phones and tablet devices. Providing simple summaries requires these techniques: query classification, information retrieval, summarizing, and information ranking.
The NTCIR (NII Testbeds and Community for Information Access Research) Workshop is a series of evaluation workshops designed to enhance research in information access technologies [19]. The first 1CLICK task (1CLICK-1) began in the NTCIR-9 [15]. The 1CLICK-1 task dealt with four types of queries (CELEBRITY, LOCAL, DEFINITION, and QA) in only the Japanese language. The next 1CLICK task in the NTCIR-10 was "1CLICK-2". 1CLICK-2 consists of a main task and a query classification subtask: The main task is presented with a query and outputs a summarized simple text, and the query classification subtask is presented with a query and outputs a query type as well. Both tasks deal with Japanese and English. In addition, eight query types that are fine-grained compared with 1CLICK-1 are given. The details of the main task can be found in the 1CLICK-2 overview paper [5] and the details of the query classification subtask are described in the next section.
In the NTCIR-9 1CLICK-1, all participants [6, 11, 13] used support vector machines (SVMs) for query classification. Some of the participants achieved good scores. However, the feature vectors of the SVM were high-dimensional and depended strongly on the Japanese language. High-dimensional feature vectors cause overfitting and loss of generalization, and the language dependency is not suitable for applying this task to other language problems. The participants' research also suggested that snippets, which are summaries generated by web search engines, contain a great deal of information that can be used for classification. Moreover, some features were created by manual input from predefined URLs and words, and thus have high updating costs. In the NTCIR-10 1CLICK-2, the KUIDL team [9] achieved accuracies of 0.87 for the Japanese subtask and 0.66 for the English subtask in a way similar to that of the 1CLICK-1. The Japanese subtask result of the KUIDL team was the highest score of the 1CLICK-2 participants, although, the score implied a limit caused by language dependency. The MSRA team [12] and the HUKB team [22] tried to apply rule-based methods using named entity taggers and etc. The MSRA team achieved accuracy 0.83 and the HUKB team achieved 0.80 for the Japanese subtask. These methods had lower accuracy than other methods and they applied only to the Japanese subtask. The ut team [4] used a pre-constructed database made from retrieved documents to create feature vectors of the perceptron classifier in addition to some features used by the KUIDL team and the MSRA team in 1CLICK-1. This team achieved an accuracy of 0.55 on official runs for the English subtask.
In this paper, the authors proposed a query classification system based on snippet similarity for the one-click search and optimize the proposed system empirically.
The remainder of the present paper is organized as follows. Section 2 describe the details of the NTCIR-10 1CLICK-2 query classification subtask. Section 3 describes the proposed system and the method used. Section 4 shows the details of experiments. Section 5 presents the evaluation results. Section 6 describes the discusses the results. Section 7 presents the conclusion.

2. NTCIR-10 1CLICK-2 QUERY CLASSIFICATION SUBTASK

The query classification subtask defined is given a query and outputs a query type. The details of the input and the output are described in the following subsections.

2.1 Input

Queries consist of real web search query strings, which include celebrity, place, definition, and question answering information.
In this task, queries are labeled as one of eight types (followed by the number of queries): ARTIST (10), ACTOR (10), POLITICIAN (10), ATHLETE (10), FACILITY (15), GEO (15), DEFINITION (15), and QA (15). The number of queries is 100 for each task in Japanese and English. The given query format is as follows:

⟨query ID⟩[Tab]⟨query string⟩

2.2 Output

The output is the query type for the given query. The query type is used for judging what is important information for each query of the web search results in the main task. This is a multiclass query classification problem. In this task, the output query type is one of eight types, which are the same as the input types. Each line in the task output files consists of the following format:

⟨query ID⟩[Tab]⟨query type⟩

where ⟨query type⟩ is one of the eight types predicted by the system.

3. SNIPPET-SIMILARITY-BASED SYSTEM

3.1 System Overview

In this section, the details of the proposed system are described. The authors propose a snippet-similarity-based method [21, 20]. The purpose of the proposed system is to use snippet similarities for the feature vector of the SVM. The proposed system has lower language dependency, more flexibility for new words, and lower-dimensional feature vectors.
In 1CLICK-1, the KUIDL team and the MSRA team used snippets to manually create features with predefined characteristic Japanese words. However, the authors decided that snippets have more information. Since snippets contain the essentials of a related web page for each query, the system uses the words in related web pages as the feature vector. Although, the variations of words in snippets are still large and the dimensions of a feature vector becomes high if each word corresponds to each axis in the vector space. High-dimensional feature vectors cause overfitting and loss of generalization. To solve this problem, dimension reduction is required.
Using snippet similarities solves this problem. A high-dimensional word vector of query snippets is replaced by a similarity vector that has only eight features. This is a reduction in the number of dimensions. Therefore, the proposed system was improved through the reduction of overfitting and loss of generalization. In addition, lower-dimensional vectors reduce the amount of computer resources required and improve SVM optimization. As a result, this method provides easier recalculation. This is suitable for web searches in which new words appear.
As shown in Figure 1, the proposed system consists of three components: a search engine, a feature extractor and a classifier. The components of the system are given in the following sections.

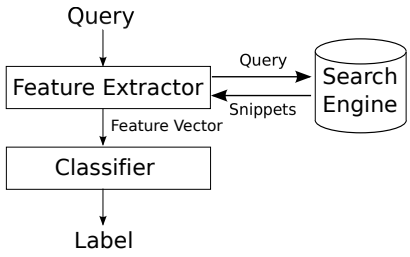

Fig. 1. System overview

3.2 Search Engine

The search engine is constructed from real web search results. The search engine is presented with a query and returns snippets separated in each word. Data source details and the method of processing snippets are described in the following subsection.

3.2.1 Data Source. The NTCIR-10 1CLICK-2 test collection has 500 top-ranked documents returned from the Bing search engine for each query in each language for the main task. This collection was constructed on July 4, 2012. Each document has a page title, a summary (snippet), a URL, and a document rank. In addition, the authors created a document collection from a Yahoo! Japan Web search API[1] [10] with the same structure. This collection was constructed on January 29, 2013.

3.2.2 Snippets. Each snippet was separated into each word. The authors used TreeTagger [17, 18, 16] in English and MeCab [8, 7] in Japanese for morphological analysis. Furthermore, differences of the parts-of-speech (POSs) and the number of snippets were compared by experiments. The details are described in Section 4.

3.3 Feature Extractor

In the vector space model, a vector represents each item or document in a collection. The document vector is defined by $\overrightarrow{d_n}$, $\overrightarrow{D_j}$, and the word value w_i. Word value w_i, which indicates wheather a word exists in each document, is defined by Eqs. (1), (2), and (3).

$$\overrightarrow{d_n} = \{w_1, w_2, w_3, ..., w_i\} \tag{1}$$

$$\overrightarrow{D_j} = \{d_1, d_2, d_3, ..., d_n\} \tag{2}$$

$$w_i = \begin{cases} 1 & (\text{Exists}) \\ 0 & (\text{Does not exist}) \end{cases} \tag{3}$$

Feature vectors are created according to Eq. (4) and Figure 2. First, document vectors $\overrightarrow{d_n}$ are created from each query snippet and $\overrightarrow{D_j}$, which has the same label as that of $\overrightarrow{d_n}$. Second, the similarity between document vectors is calculated with the query document vector. Finally, all similarities are joined as a vector by Eq. (4).

$$\overrightarrow{f} = \{sim(\overrightarrow{D_1}, \overrightarrow{q}), sim(\overrightarrow{D_2}, \overrightarrow{q}), ..., sim(\overrightarrow{D_8}, \overrightarrow{q})\} \tag{4}$$

The methods of calculating document similarity are described bellow. Cosine Similarity, Jaccard index, and Simpson coefficient

[1]This API service ended on August 14, 2013.

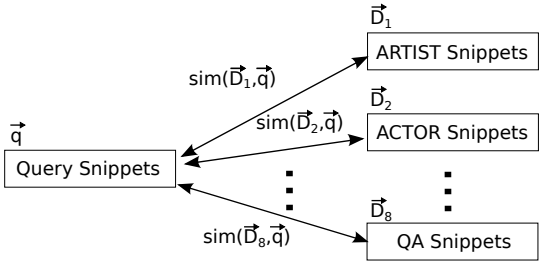

Fig. 2. Calculating the similarity in the feature extractor

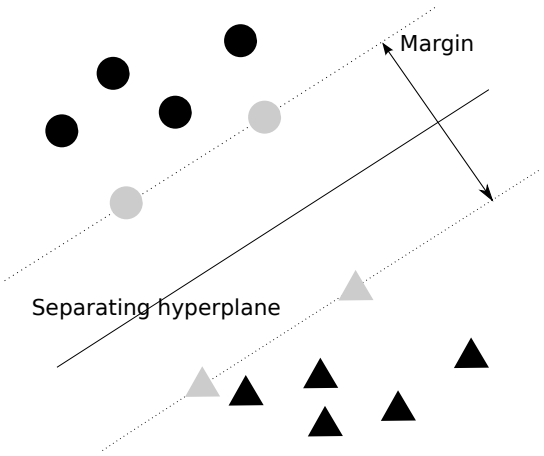

Fig. 3. Strategy of SVMs

[2] between $\overrightarrow{D_j}$ with $\overrightarrow{q}$ are defined by Eqs (5), (6), and (7), respectively.

$$Sim(\overrightarrow{D_j}, \overrightarrow{q}) = \frac{\overrightarrow{D_j} \cdot \overrightarrow{q}}{|\overrightarrow{D_j}||\overrightarrow{q}|} \quad (5)$$

$$Sim(\overrightarrow{D_j}, \overrightarrow{q}) = \frac{|\overrightarrow{D_j} \cap \overrightarrow{q}|}{|\overrightarrow{D_j} \cup \overrightarrow{q}|} \quad (6)$$

$$Sim(\overrightarrow{D_j}, \overrightarrow{q}) = \frac{|\overrightarrow{D_j} \cap \overrightarrow{q}|}{min(|\overrightarrow{D_j}|, |\overrightarrow{q}|)} \quad (7)$$

3.4 Classifier

An SVM is used for the classifier. SVMs, which are also referred to as support-vector networks [3], are supervised learning models in machine learning. SVMs determine the separating hyperplanes, as shown in Figure 3. The support vectors, denoted in gray, define the margin of largest separation between two classes. The advantage of SVMs is their ability to ensure high generalization with a small number of vectors. This is a suitable characteristic for this task. The LIBSVM [1] is used in this task and implements a one-against-one approach for multiclass classification.

The SVM is requires the appropriate optimization of parameters to obtain better results. In this system, the C-support vector classification (C-SVC) and the radial basis function (RBF) kernel were selected for configuring the classifier. SVM optimization of the parameters is described in the next section. The authors apply leave-one-out cross-validation for learning. The feature vectors for learning are created in the following steps.

(1) Create vectors $\overrightarrow{d_1}$ to $\overrightarrow{d_{99}}$ by the snippets of each learning query (i.e., not the test query $\overrightarrow{q}$).
(2) Select one $\overrightarrow{d_n}$ from the 99 $\overrightarrow{d_n}$ vectors.
(3) Create $\overrightarrow{D_j}$ from $\overrightarrow{d_n}$ vectors for each label (j = 1–8).
(4) Calculate the similarities between $\overrightarrow{D_j}$ with the selected $\overrightarrow{d_n}$.
(5) Repeat steps 1–4 for all queries (99 times).
(6) Conduct learning with these 99 feature vectors.
(7) Conduct the tests by the test query $\overrightarrow{q}$.
(8) Repeat steps 1–7 for all test queries (100 times).

The above procedure means $\overrightarrow{q}$ is always excluded from $\overrightarrow{D_j}$ in the learning and testing sequences.

4. EXPERIMENTS FOR OPTIMIZATION

Experiments were conducted to determine the methods and parameters with the best accuracy for each language in the query classification subtask.

In these experiments, the authors assumed the following: Firstly, since the Yahoo! Japan Web search API is a service provided in Japanese, it is assumed that this service would bring better results than would the Bing search results on the Japanese task. Second, higher ranked snippets that contain better information are used as the basis for determining classes. Third, the most important POSs are assumed to bring better results. In addition, methods of calculating the document similarity have different denominators that are believed to produce different results. Finally, the best SVM parameters are determined by grid searches.

The authors counducted additional experiments that have new POSs conditions for verbs and inspected the details of misclassified queries, in addition to the ones conducted in this paper [20].

All combinations of the following factors were investigated experimentally.

—Data source:
Bing search result, Yahoo! Japan Web search API

—Number of snippets:
10, 30, 50, 100, 200

—POSs used (Table 1):
Nouns, Verbs, Nouns & Verbs, NTCIR-10POSs , All POSs

—Method of calculating document similarity:
Cosine similarity, Jaccard index, Simpson coefficient

—SVM parameters:
C (cost parameter): 2^{-5}, 2^{-3}, 2^{-5}, ... , 2^{13}, 2^{15}
Gamma (RBF kernel parameter): 2^{-15}, 2^{-13}, ... ,2^{1}, 2^{3}

5. RESULTS

The Japanese and English query classification subtask experimental results are described separately in this section.

5.1 Japanese Subtask Results

Figures 4 to 8 show how the results differ based on which POSs are selected for the most suitable SVM parameters. Here, “n” refers to the number of top-ranked snippets. “B” and “Y” in the legend mean Bing search results and Yahoo! Japan Web search results,

Table 1. Selected POSs

Name	Language	Selected POSs
Nouns	Japanese	Noun
	English	NN, NNS, NP, NPS
Verbs	Japanese	Verb
	English	VB, VBD, VBG, VBN, VBP, VBZ,
Nouns & Verbs	Japanese	Noun, Verb
	English	NN, NNS, NP, NPS, VB, VBD, VBG, VBN, VBP, VBZ
NTCIR-10 POSs	Japanese	Noun, Verb, Adjective, Adnominal, Auxiliary Verb
	English	CD, FW, JJ, JJR, JJS, NN, NNS, NP, NPS, PP, PP$, RB, RBR, RBS, RP, VB, VBD, VBG, VBN, VBP, VBZ, WDT, WP, WP$
All POSs	Japanese	All POSs
	English	All POSs

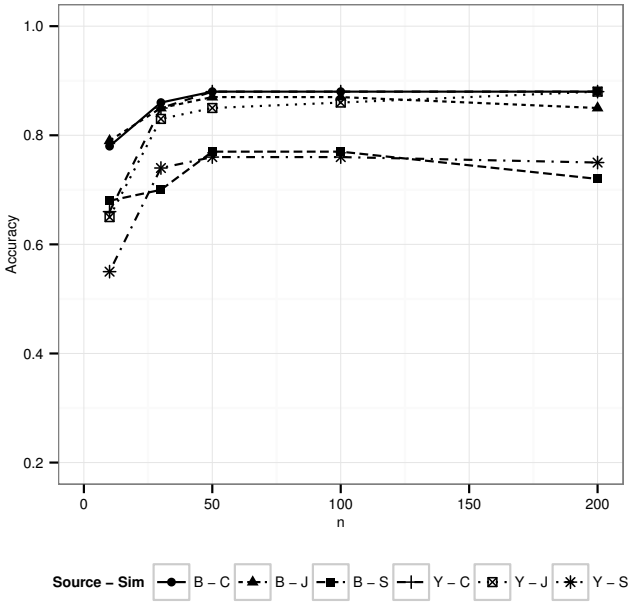

Fig. 4. "Nouns" POS results for the Japanese task

respectively. Also, "C", "J", and "S" means Cosine similarity, Jaccard index, Simpson coefficient, respectively.
Cosine similaritiy achieved the highest score in each POS condition. The highest score is 0.93 for 50 ranked snippets of the Yahoo! Japan Web search results for "All POSs" words. Figure 9 shows the result for the grid search of the cost and the gamma parameters for SVM for the best score. Table 2 shows details of the misclassifying queries (actual queries are represented with Kanji characters but they are converted to Romaji in this paper).

5.2 English Subtask Results

Figures 10 to 14 show how the results differ based on the POSs selected for the most suitable SVM parameters.

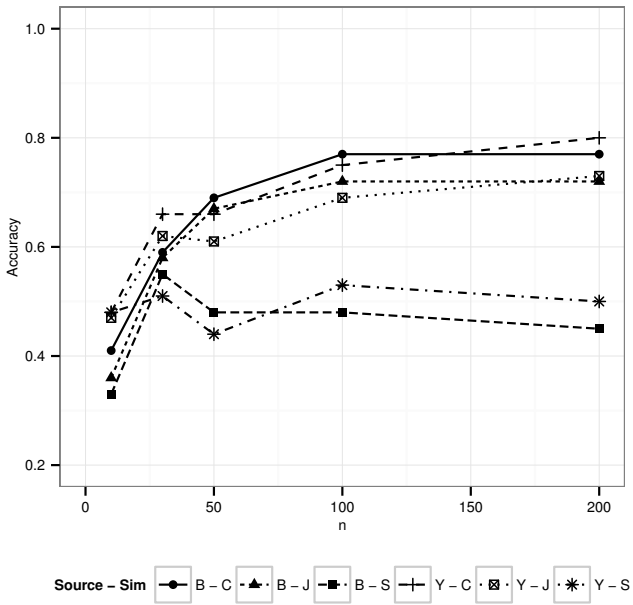

Fig. 5. "Verbs" POS results for the Japanese task

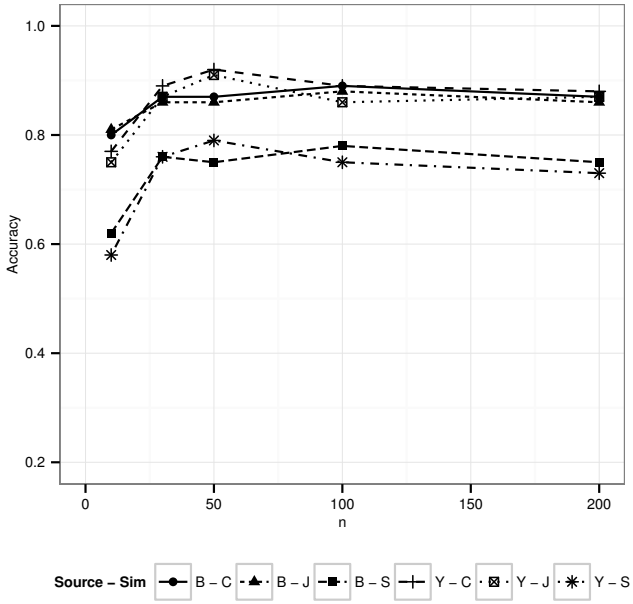

Fig. 6. "Noun & Verbs" POSs results for the Japanese task

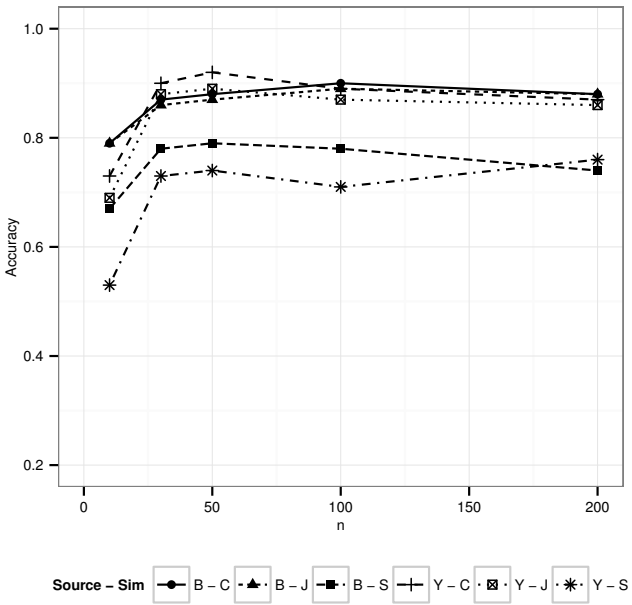

Fig. 7. "NTCIR10 POSs" results for the Japanese task

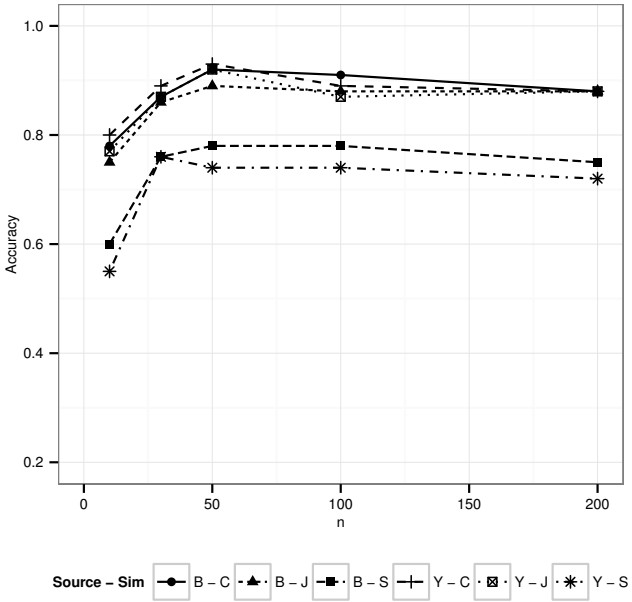

Fig. 8. "All POSs" results for the Japanese task

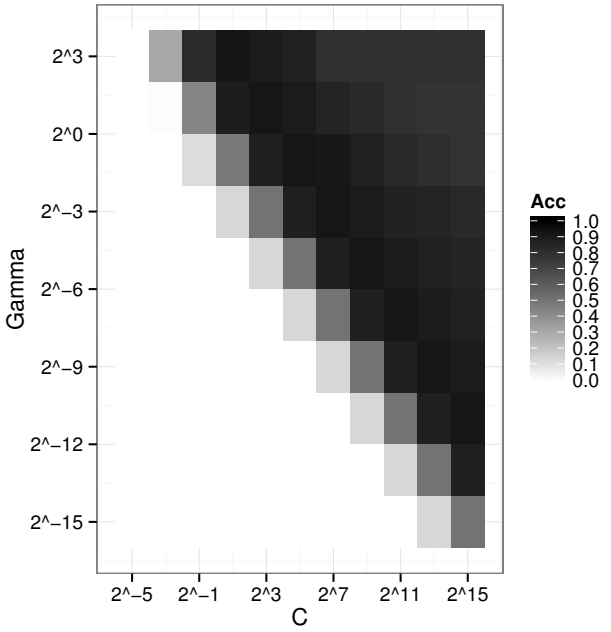

Fig. 9. Grid search result for the Japanese task

Table 2. Misclassified queries for the Japanese task

Query ID	Query	Collect	Result
1C2-J-0024	"bloomberg shichou" (bloomberg mayor)	POL	ART
1C2-J-0043	"hotel ambia shofukaku" (hotel ambia shofukaku)	FAC	GEO
1C2-J-0069	"sapporo biyou senmongakkou" (sapporo beauty college)	GEO	FAC
1C2-J-0072	"inuyasha" (Inu Yasha)	DEF	ART
1C2-J-0082	"parkinson byou" (Parkinson's disease)	DEF	QA
1C2-J-0086	"plasma to ekisyou no chigai" (difference between plasma and LCD)	QA	DEF
1C2-J-0090	"kanchu mimai no bunrei" (example of mid-winter greetings)	QA	DEF

Cosine similarity achieved the highest score in each POS condition as was the same as the Japanese subtask results. The highest score for the smallest number of snippets is 0.91 for the top 50 ranked snippets of the Yahoo! Japan Web search results for "All POSs" words. Figure 15 shows the result of the grid search of the cost and the gamma parameters for the SVM for the best score. Table 3 shows the details of misclassifying queries.

6. DISCUSSIONS

The results shown in Figures 4 and 10 indicate that nouns are the most important POSs for distinguishing query types in each language. Moreover, important information for classification is concentrated in the top 50 ranked snippets for each language.

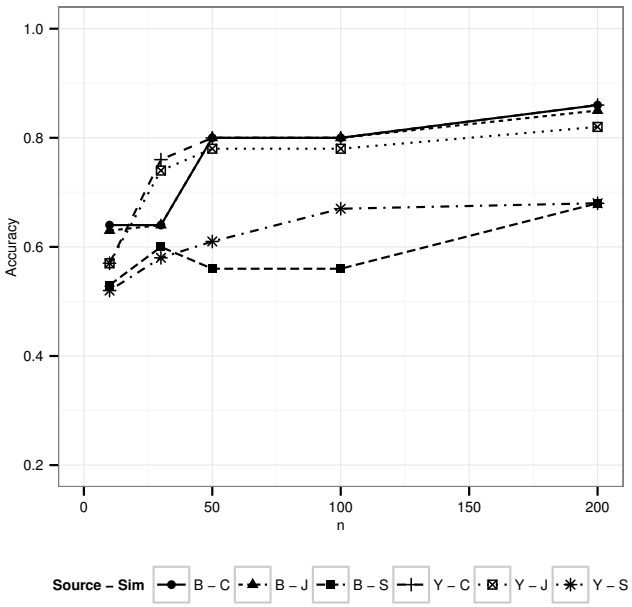

Fig. 10. "Nouns" POS results for the English task

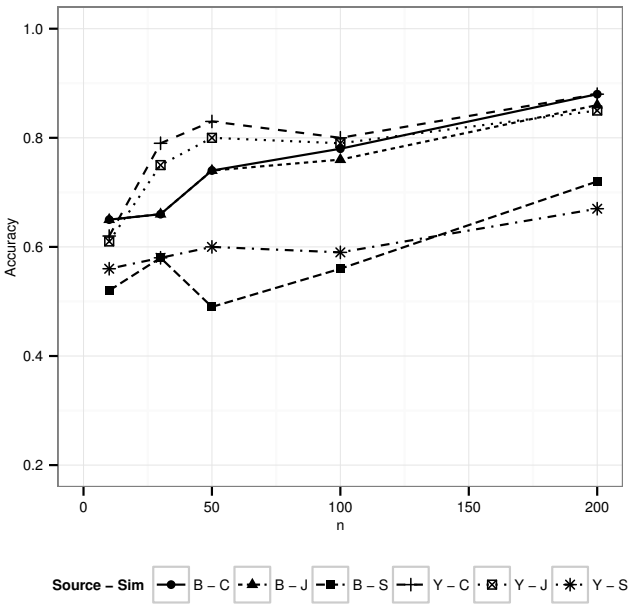

Fig. 12. "Noun & Verbs" POSs results for the English task

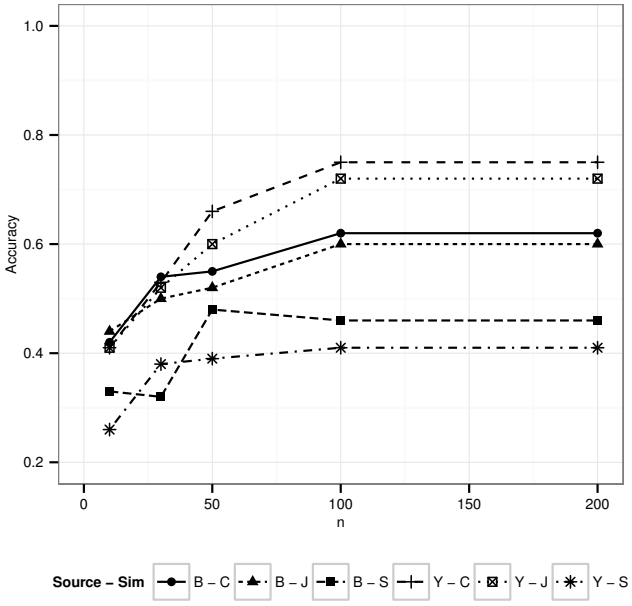

Fig. 11. "Verbs" POS results for the English task

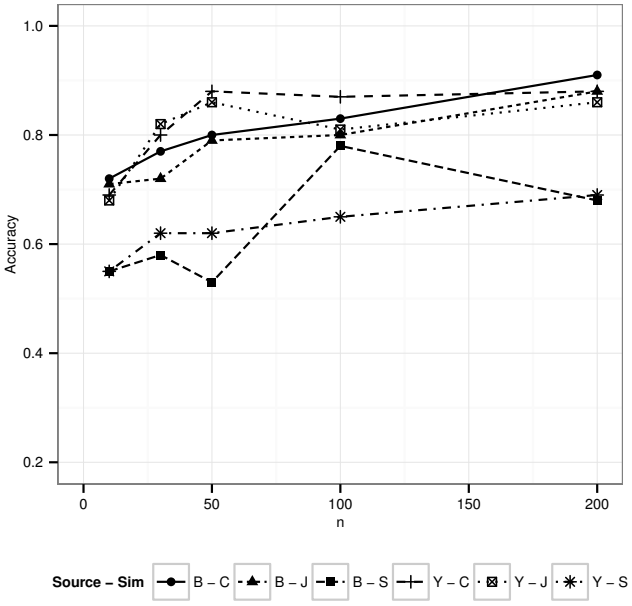

Fig. 13. "NTCIR10 POSs" results for the English task

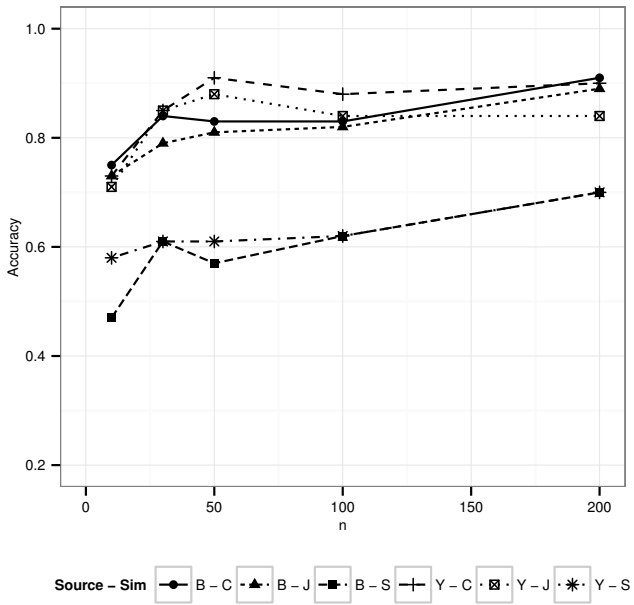

Fig. 14. "All POSs" results for the English task

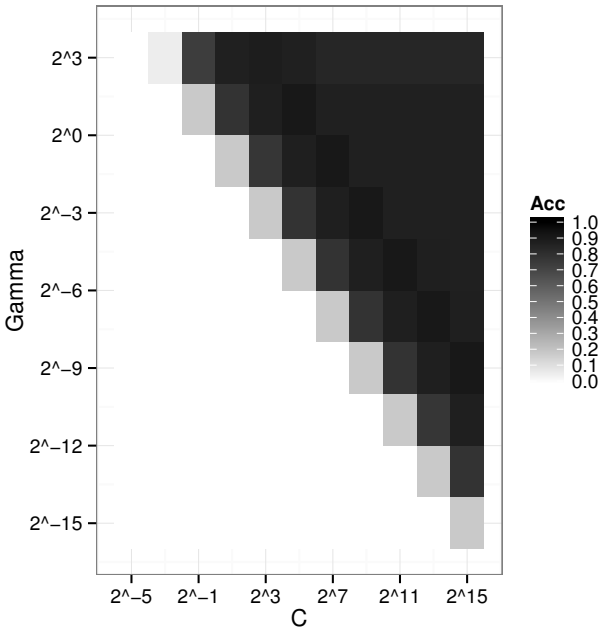

Fig. 15. Grid search result for the English task

Table 3. Misclassified queries for the English task

Query ID	Query	Collect	Result
1C2-E-0007	whitney houston movies	ART	ACT
1C2-E-0121	sears illinois	GEO	FAC
1C2-E-0127	Japan earthquake location	GEO	DEF
1C2-E-0144	geothermal energy	DEF	QA
1C2-E-0146	thanksgiving canada	DEF	QA
1C2-E-0166	rebar	DEF	GEO
1C2-E-0167	big dig	DEF	FAC
1C2-E-0189	why is apple developing maps	QA	DEF
1C2-E-0203	how is trash processed	QA	DEF

This tendency is common for all POS conditions, including nouns. Figures 5, 6, 11, and 12 show the effectiveness of verbs. Verbs are not as effective as nouns, although they contribute by improving the accuracy. In addition, Figures 7, 8, 13, and 14 show that the best POSs are found to be all of the POSs used in these experiments. A bigger variety of POSs can provide more stability than can a smaller variety.

It is also seen in Figures 9 and 15 that the cost and the gamma parameters of the SVM with the RBM kernel have peaks on diagonal lines. Thus, a more sensitive search might yield better results.

From Tables 2 and 3, it is found that the proposed method is suitable for celebrity-type queries. Furthermore, for GEO, DEFINITION and QA type queries, it is found that these query types have a tendency to misclassify each other. However, some of these misclassified queries do not seem to affect the next step (information retrieval, summarizing, etc.) in a significant way.

The proposed method achieves the performance requirements for the NTCIR-10 1CLICK-2 test collection. However, this test collection consists of only a very small number of logs of queries on the web, so its applicability must be confirmed by using other collections for testing. Moreover, the NTCIR-10 1CLICK-2 test collection does not include "navigational" or "resource" queries [14], which are entered with the intention of finding a specific URL or resource. Such queries must be considered in order for the method to be useful for actual queries.

7. CONCLUSIONS

The authors proposed a query classification system based on snippet similarity for one-click search. By selecting an appropriate number of snippets, POSs , and SVM parameters in addition to data sources and methods of calculating document similarity, the proposed system achieves accuracies of over 0.9 for each language (Japanese and English) NTCIR-10 1CLICK-2 query classification subtask without language dependency. Cosine similarity is always the best of the three methods for calculating document similarity, and the use of the top 50 ranked snippets with all POSs yielded good results. In addition, nouns are the most important POSs for classification. Further studies should be conducted to apply queries to other test collection queries.

8. ACKNOWLEDGEMENT

The present study would not have been possible without the NTCIR project and the NTCIR-10 1CLICK-2 test collection.

9. REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[2] Alan H Cheetham and Joseph E Hazel. Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, pages 1130–1136, 1969.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] Dan Ionita, Niek Tax, and Djoerd Hiemstra. An api-based search system for one click access to information. In *Proceedings of the 10th NTCIR Conference*, 2013.

[5] Makoto P Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Overview of the ntcir-10 1click-2 task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[6] Makoto P Kato, Meng Zhao, Kosetsu Tsukuda, Yoshiyuki Shoji, Takehiro Yamamoto, Hiroaki Ohshima, and K Tanakai. Information extraction based approach for the ntcir-9 1click task. *Proceedings of NTCIR-9*, 2011.

[7] Taku Kudo. Mecab:yet another japanese dependency structure analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP*, volume 4, pages 230–237, 2004.

[9] Tomohiro Manabe, Kosetsu Tsukuda, Kazutoshi Umemoto, Yoshiyuki Shoji, Makoto P Kato, Takehiro Yamamoto, Meng Zhao, Soungwoong Yoon, Hiroaki Ohshima, and Katsumi Tanaka. Information extraction based approach for the ntcir-10 1click-2 task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[10] Yahoo Japan Corporation. Yahoo japan web search api. http://developer.yahoo.co.jp/webapi/search/.

[11] Hajime Morita, Takuya Makino, Tetsuya Sakai, Hiroya Takamura, and Manabu Okumura. Ttoku summarization based systems at ntcir-9 1click task. *Proceedings of NTCIR-9*, 2011.

[12] Kazuya Narita, Tetsuya Sakai, Zhicheng Dou, and Song Young-In. Msra at ntcir-10 1click-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[13] Naoki Orii, Young-In Song, and Tetsuya Sakai. Microsoft research asia at the ntcir-9 1click task. *Proceedings of NTCIR-9*, 2011.

[14] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.

[15] Tetsuya Sakai, Makoto P Kato, and Young-In Song. Overview of ntcir-9 1click. *Proceedings of NTCIR-9*, 2011.

[16] Helmut Schmid. Treetagger - a language independent part-of-speech tagger. http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/.

[17] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK, 1994.

[18] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer, 1995.

[19] NTCIR: NII Testbeds and Community for Information access Research. http://ntcir.nii.ac.jp/about/.

[20] Tatsuya Tojima and Takashi Yukawa. Optimization of query classification based on snippet similarities for a 1 click search system. In *Proceedings of 3rd International Symposium on Engineering, Energy and Environment*, November 17-20 2013. accepted to be presented.

[21] Tatsuya Tojima and Takashi Yukawa. Query classification system based on snippet summary similarities for ntcir-10 1click-2 task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[22] Masaharu Yoshioka. Query classification by using named entity recognition systems and clue keywords. In *Proceedings of the 10th NTCIR Conference*, 2013.