InteractiveMediaMINE at the NTCIR-11 IMine Search Task

Shohei MINE
Graduate School of Engineering, Kogakuin University
em14013@ns.kogakuin.ac.jp

Takuma MATSUMOTO
Faculty of Informatics, Kogakuin University
j111098@ns.kogakuin.ac.jp

Tomofumi YOSHIDA
Faculty of Informatics, Kogakuin University
j111114@ns.kogakuin.ac.jp

Takuya SHINOHARA
Faculty of Informatics, Kogakuin University
j111053@ns.kogakuin.ac.jp

Daisuke KITAYAMA
Faculty of Informatics, Kogakuin University
kitayama@cc.kogakuin.ac.jp

ABSTRACT

The InteractiveMediaMINE team participated in the Task Mine subtask of the NTCIR-11 IMine Search Task. Our framework consists of three steps. First, we extend the query entered by the user in order to optimize the search engine. Second, we extract candidates of tasks from "Yahoo! Chiebukuro" with the extended search query. Here, we use the top 10 pages of the search results. Finally, we calculate the score of the extracted tasks by the words frequency of each sentence; our system outputs tasks in the descending order of the score. This paper describes our approach to solving the Task Mine problem and discusses its results.

Team Name

InteractiveMediaMINE

Subtasks

Task Mine (Japanese)

Keywords

Morphological Analysis, Dependency Parsing, Web Search

1. INTRODUCTION

The InteractiveMediaMINE team participated in the NTCIR-11 IMine Search Task Mining (TaskMine) subtask. This paper describes our approach to solving the Task Mine problem and discusses its results. We use Yahoo! Chiebukuro[2], a web-based Q&A service, as our system resource. In general, Q&A services aim to collect answers that solve the user's problems. Consequently, we expect Yahoo! Chiebukuro to be useful as our system resource for mining tasks. In our system, first, the user inputs a query that shows the problem she/he wants to solve. Second, our system extends queries by using a morphological analysis and retrieves the top 10 pages of the search result from Yahoo! Chiebukuro. Third, considering that "を (wo)" means particles function as a direct object in Japanese, we extract candidate tasks using the syntax pattern of "noun + "を (wo)" + verb." Finally, we calculate the score of the extracted tasks by using the word frequency of each sentence; our system outputs tasks in the descending order of the score.

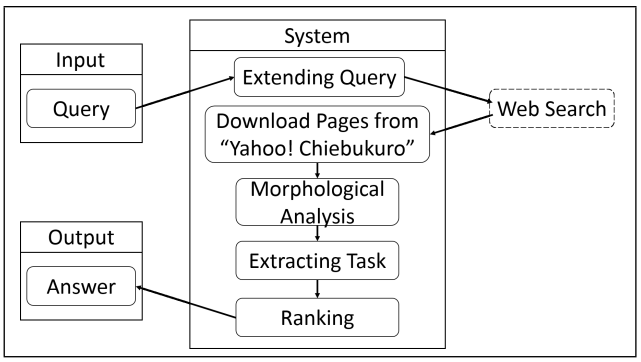

Figure 1: Our Framework

2. FRAMEWORK

Figure 1 shows our framework. In this section, we describe our framework; it consists of three sections, extending query and retrieving, extracting tasks, and ranking. The implementation details are as follows:

2.1 Extending Query and Retrieving

First, our system executes a morphological analysis for the query that the user inputs by natural language and then, extracts nouns and verbs from the query. In this paper, we use Mecab[5] as the morphological analysis tool. Then, the system joins the extracted nouns, a single-byte blank, and verbs to make a sentence that will be actually used as a query for retrieving information from the web. "方法" is a word that means "method" or "way" in Japanese; therefore, we think that adding "方法" at the end of the query is effective in retrieving pages that include questions about methods used for solving certain problems. In this study, our system used Yahoo! Chiebukuro and retrieves the top 10 pages of the search results.

2.2 Extracting Tasks

Considering "wo" is a Japanese particle pointing to a direct object, it is useful to extract tasks that can solve the user's problem. Therefore, we extract the answer text from each searched web page and then, extract chunks that contain the following pattern: "Noun + "を (wo)" + the chunk that includes certain verb + chunks that depend on the

Table 1: Top 5 result for the query “ご飯を炊く”

Rank	Extracted Task	Score
1	炊き方は、洗った米を“ザルにあげて水を切って”30 分したら、分量の水を加える	250
2	鍋を中火にかけて沸騰すれば１～２分間キープして直ぐに火を最低限まで絞って１５分間、最後に一瞬強火にして火を止める	238
3	お米２合を普通にといで鍋に入れて水を２カッププラス大さじ２杯入れる	195
3	言われるように、鍋に米を入れて指の第一関節や手首までお水を入れる	195
5	米を炊く時は、基本的に吸水させたあとに、最大火力になるんですが、火を止める	184

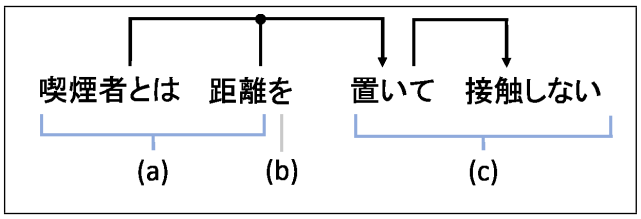

Figure 2: Dependency Parsing: “喫煙者とは距離を置いて接触しない” means “Keep a distance from smoker and don’t meet him.” “喫煙者” means “smoker.” “とは” means “from.” “距離を置いて” means “keep a distance.” “接触しない” means “don’t meet him”

verb.” Our system performs dependency parsing for all the extracted sentences in order to extract a pattern of “chunks that depend on noun + noun + を (wo) + the chunk that includes certain verb.” At this point, the extracted verb end of the sentence are converted into its dictionary form. As a result, our system extracts sentences as tasks. In this paper, we use “Yahoo!日本語係り受け解析 API” (Yahoo! Japanese Dependency Parsing API)[3] as the dependency parsing tool. Figure 2(a) shows “chunks that depend on noun + noun”, and Figure 2(c) shows “the chunk that includes certain verb.”

2.3 Ranking

We define the evaluation scores of the i th extracted task t_i as follows:

$$Score(t_i) = \sum_{noun \subseteq nouns_{t_i}} \sum_{ans \subseteq A} NounFreq(noun, ans) + \sum_{task \subseteq T} VerbFreq(verb_{t_i}, task) \quad (1)$$

when A denotes a set of all text that represents the answer information. T represents a set of all tasks that have been extracted. $nouns_{t_i}$ refers to a noun set included in front of the “を (wo)” in t_i. $verb_{t_i}$ denotes a verb that is extracted from the next chunk of which include “を (wo)” in t_i. $NounFreq(noun, ans)$ represents the number of occurrences of the noun $noun$ in the answer ans by some respondents. $VerbFreq(verb_{t_i}, task)$ refers to the number of occurrences of the verb “$verb_{t_i}$” in the task $task$. Based on the supposition that the words that appeared frequently are important in solving problems, we define that tasks including many of these words are also important. For example, Table 1 shows the ranking result of “ご飯を炊く”, and Table 2 shows a part of the lists of nouns and verbs that are used for calculating the score. From Table 1, we can see that the task “炊き方は、洗った米を“ザルにあげて水を切って”30 分したら、分量の水を加える” includes some frequently appearing words such as “水,” “分,” and “米”; therefore, it has the highest score in this list.

Table 2: Top 9 results of frequently appearing words of the query “ご飯を炊く”

Noun	Frequency
水	66
火	53
飯	49
分	49
鍋	48
米	43
ん	41
め	38
炊飯	22

Verb	Frequency
溜める	14
離さない	10
抑える	6
戻す	6
変える	4
飛ばす	2
買う	2
入れ直す	2
入れる	2

3. RESULTS AND ANALYSIS

We mined tasks for the query set of the TaskMine subtask. Figure 3 shows experimental results[4]. TM-019, “歯周病を治療する”, has one of the highest scores for all metrics, namely nDCG@1, @5, @10, and @50. Table 3 shows the extracted tasks for TM-019 and match gold standards. TM-023, “レーザーカッターを使う,” has one of the lowest scores for all metrics, namely nDCG@1, @5, @10, and @50. Table 4 shows the extracted tasks for TM-023 and match gold standards. Our system depends on answers extracted from Yahoo! Chiebukuro. Yahoo! Chiebukuro is a general Q & A service that is not specialized in any specific field; therefore, our system makes it easy to collect answers accurately for ordinary questions such as TM-019. On the other hand, the accuracy of the extracted tasks decreases for queries such as TM-023 that are not the type of questions that ordinary users ask. We mined tasks for the query set of the TaskMine subtask. From Table 4, we observe that only four tasks that were extracted for TM-023 matches the gold standard tasks. To solve this problem, we plan to use other Q&A services that are specialized in some specific fields. For example, “teratail [1]” is a Q&A service specialized in computer science. We expect that we can extract more accurate tasks for problems about computer science when we use this service. In addition, although we use “使う” as a verb of “レーザーカッターを使う,” in the future, we plan to use synonyms of “使う” or words that co-occur with “使う.” In fact, “カット” and “切断”, words which mean “cut” in Japanese, are often used with “レーザーカッター” in Yahoo! Chiebukuro. According to Table 3, we can see that tasks that are extracted by our system tend to be long because the extracted tasks include chunks that depend on nouns and chunks that depend on verbs. In addition, tasks that have a large num-

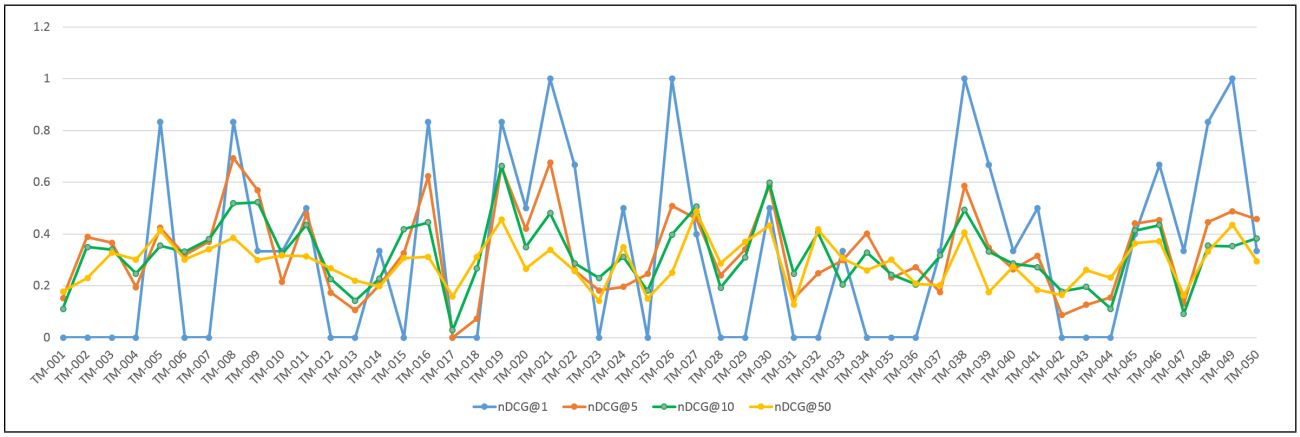

Figure 3: Experimental Results

Table 3: Ranking table of the top 10 results of TM-019

Rank	Extracted Task	Matched gold standard task
1	歯周病治療は、歯周病の原因となる汚れを除去した後、毎日の歯磨きを妨げる歯周ポケットを除去する	炎症を引き起こす細菌を徹底的に除去する
2	歯がグラグラする原因として、歯周病でなくても、その歯だけ強く当たっていたりしても歯の周りの骨が減っていくこともありますので、歯医者に行ってレントゲンを撮る	噛みあわせを調整する
3	信用に足るのは、歯周病学会や臨床歯周病学会のＨＰで専門医や認定医を探す	
4	歯周病の治療は、歯がグラグラしているくらいだと、歯の周りに歯石やプラークがたくさん付いていると思いますので、クリーニングをする	歯科衛生士に専門的なクリーニングをしてもらう
5	普通は歯の頭に深いむし歯ができた場合に神経を取りますが、重度の歯周病で歯茎の奥底の根っこ側から神経に細菌感染した場合も神経を取る	歯の神経を取り除き痛みをなくす
6	歯が揺れると骨がどんどん溶けていくので、銀歯などで複数本の歯を連結固定	歯のぐらつきを抑えるため歯を連結する
7	歯周病の原因である歯垢と歯石を除去し、進行を止める	歯肉のなかまで入っている歯石を取り除く
8	歯周病で歯の神経を取る	歯の神経を取り除き痛みをなくす
9	歯茎など組織の状態が正常に戻るのを確認した後に、新たに差し歯を作り替えることで健全な歯周組織を取り戻せる	歯周組織再生法をする
10	現在装着されている差し歯の適合不良、お手入れ不足が原因と考えられますので、差し歯を一旦外して仮歯を装着し歯周組織の予防処置を行う	簡単に治る病気ではないため予防を徹底する

ber of characters tend to include many nouns and therefore, tend to have a high score. To prevent this, it is necessary to normalize a task's score by its number of characters. Furthermore, this decreases the effect of the verbs on the task's score as the term frequency of the nouns tends to be higher than that of the verbs. Thus, we plan to be in balance with the weights of the noun's frequency and the verb's frequency.

4. CONCLUSION

In this paper, we proposed a method for the TaskMine subtask. We used Yahoo! Chiebukuro as our system resource because we believe that a Q&A service has many answers that can solve a user's problems. Further, we experimentally confirmed that a Q&A service is an effective resource for extracting tasks. In the future, we will add some Q&A services and Internet bulletin boards specialized in some specific fields to the proposed system's information resource. We also plan to improve the ranking method and the precision of the extracting task.

5. REFERENCES

[1] Teratail. https://teratail.com/.
[2] Yahoo! Chiebukuro. http://chiebukuro.yahoo.co.jp/.
[3] Yahoo! Japanese Dependency Parsing. http://developer.yahoo.co.jp/webapi/jlp/da/v1/parse.html.
[4] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the NTCIR-11 IMine task. In Proceedings of the NTCIR-11, 2014.
[5] Taku Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/, Feb 2013.

Table 4: Tasks for TM-023 and match gold standard

Rank	Extracted task	Matched gold standard task
1	取り扱い上の防護処置をとる	カット開始前に排気用の煙突を窓の外に出す
2	加工機では金属の塊を切断するものもあり、そのような物にはより高出力のレーザを使用しない	カットしてはいけない材料を知る
3	w 以上の出力がレーザーポインターでは許可されなくて加工機は許可される事に関して法的にはどのように定められているのかという事を聞く	彫刻できるものを知る
4	レーザ加工機は、樹脂や金属など様々な物体にマーキングを付けたり、削ったり、切ったり、紫外線等により樹脂を硬化する	彫刻できるものを知る