NEXTI at NTCIR-12 IMine-2 Task

Hidetsugu Nanba[1], Tetsuya Sakai[2], Noriko Kando[3], Atsushi Keyaki[4], Koji Eguchi[5],
Kenji Hatano[6], Toshiyuki Shimizu[7], Yu Hirate[8], and Atsushi Fujii[4]

[1]Hiroshima City University, [2]Waseda University, [3]National Institute of Informatics,
[4]Tokyo Institute of Technology, [5]Kobe University, [6]Doshisha University, [7]Kyoto University,
[8]Rakuten, Inc.

ABSTRACT

Our group NEXTI participated in the Query Understanding subtask for Japanese. We extracted subtopic candidates from retrieved web documents. Then, we merged them with query suggestion and query log data. We also identified the vertical intent of each subtopic using a method that combines machine learning- and k-NN-based methods. We conducted experiments and confirmed the effectiveness of our method.

Keywords

k-NN, SVM, clustering

1. INTRODUCTION

Our group started Project Next IR[1] in July 2014. The purpose of this project is to investigate the kinds of technologies that are required to enhance the current IR systems and the related problems that need solving. We conducted failure analysis using various test collections [Nanba 2016] and found the limitation of the bag-of-words approach in IR. We believe that the relations between words in a document should be taken into account in some way, and that the IMine-2 task is a good starting point to address this challenge.

The remainder of this paper is organized as follows. In Section 2, we propose the method for identifying subtopics with their vertical intent. Section 3 reports the experimental results and Section 4 concludes.

2. IDENTIFICATION OF SUBTOPICS

Our method of identifying subtopics consists of the following three steps.

1. Extraction of subtopic candidates from web documents
2. Merging of the candidates with query suggestion and query log data
3. Identification of the vertical intent for each subtopic

We describe the details of each step as follows.

2.1 Extraction of Subtopic Candidates from Web Documents

In this step, we extract subtopic candidates from web documents, which were provided by the organizers. For the extraction of candidates, we focused on some linguistic patterns. We assumed that a subtopic appears with its (main) topic in the same sentence. Let us consider that "*iPhone 6*" is a topic and "レビュー(*review*)" is a subtopic, and there is an expression "*iPhone 6* のレビュー (*review* of *iPhone 6*)" in documents. We call character strings between a topic and a subtopic, such as "の (of)", linguistic patterns. If we apply a pattern "*iPhone 6* の [noun phrase]" to documents, and collect noun phrases, we may find new subtopics in the noun phrases.

To collect these patterns, we used NTCIR-11 IMine data [Liu 2014]. First, we retrieved 500 web documents for each topic using Google. Second, we collected character strings between a topic and a subtopic. Finally, we collected 19 linguistic patterns in Table 1. *S* and *T* indicate a topic and a subtopic, respectively. <NP> and </NP> indicate start and end points of noun phrases. We identified nouns in a sentence using a Japanese morphological analyzer, JUMAN[2], and regarded continuous nouns as a noun phrase. In Table 1, "*TS*" indicates that both topic and subtopic appear in the same noun phrase, while "*T*</NP>の<NP>*S*" indicates that the topic and the subtopic appear in different noun phrases, and "の (of)" appears between the topic and the subtopic. Using these patterns, we extracted subtopic candidates from web documents.

2.2 Mergence of the Candidates with Query Suggestion and Query Log Data

In this step, we merged the candidates extracted in step 1 with the following six data, which were provided from the organizers.

- Query suggestion by Google
- Query suggestion by Yahoo
- Query suggestion by Bing
- Co-click queries
- Co-session queries
- Co-topic queries

[1] http://165.242.101.30/~nanba/task/project_next/ (in Japanese)

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN&alias%5B%5D=%E6%97%A5%E6%9C%AC%E8%AA%9E%E5%BD%A2%E6%85%8B%E7%B4%A0%E8%A7%A3%E6%9E%90%E3%82%B7%E3%82%B9%E3%83%86%E3%83%A0JUMAN

We show the procedure of the mergence in Figure 1. We merged the candidates using equation 1.

Table 1. Linguistic Patterns for Extracting Subtopic Candidates from Web Documents.

pattern	freq.	pattern	freq.
TS	574	*S*</NP>『<NP>*T*	5
T</NP>の<NP>*S*	191	*S*,*T*	5
ST	126	*S*</NP>は<NP>*T*	5
S</NP>「<NP>*T*	45	*T*</NP>は、<NP>*S*	4
S</NP>・<NP>*T*	24	*T*</NP>から<NP>*S*	3
T</NP>を *S*	23	*T*</NP>（<NP>*S*	3
T S	13	*S*</NP>と<NP>*T*	3
T</NP>を<NP>*S*	12	*S*</NP>、<NP>*T*	3
S</NP>　<NP>*T*	9	*T*</NP>の<NP>*S*	3
T</NP>は<NP>*S*	8		

$$Score(candidate) = \sum_{i=1}^{7} W_i \ occur(candidate, d_i) \qquad (1)$$

Here, $occur(candidate, d_i)$ is a 0-1 function that indicates whether $candidate$ occurs in d_i ($occur(candidate, d_i)=1$) or not ($occur(candidate, d_i)=0$). $d_1 - d_7$ indicate candidates from web documents, query suggestions using Google, Yahoo, and Bing, and query logs using co-click, co-session, and co-topic, respectively. $W_1=1/3$, and $W_2 - W_7 = 1/6$. As a result of this mergence, we obtained 1,683,456 candidates (for 100 topics) in total.

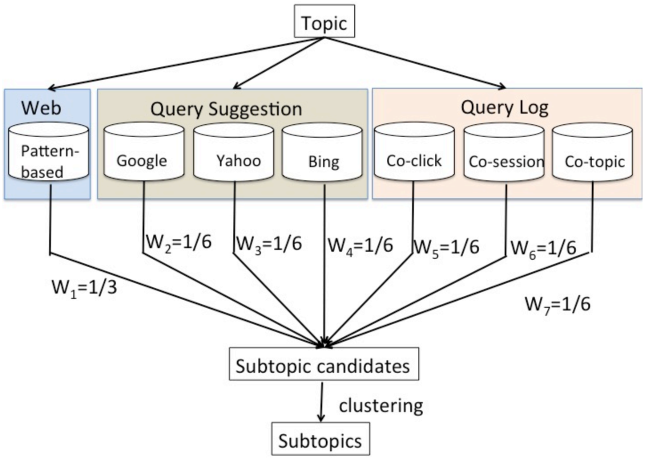

Figure 1. Mergence of Subtopic Candidates.

Then, we conducted clustering for grouping similar subtopics. We assumed that similar subtopics tend to retrieve similar documents. Based on this idea, we retrieved documents using topic-subtopic pairs, and used the top 10 documents with the highest relevance score as features for clustering. We used 44,280 web documents provided by organizers, and constructed an IR system using INDRI[3]. For the clustering, we used bayon[4]. We set the number of clusters as 10, because we were allowed to submit 10 subtopics at most for each topic. Among subtopic candidates in each cluster, we selected the candidate having the highest score calculated by equation 1.

2.3 Identification of the Vertical Intent for Each Subtopic

Our method of identifying the vertical intent for each subtopic consists of the following two steps.

- Identification of content types (vertical intents) of each web page
- Identification of the vertical intent for each subtopic

We describe the details of each step in the following sections.

2.3.1 Identification of Content Types of Each Web Page

Features for machine learning

We identified content types (vertical intents) of each web page using a machine learning-based approach. We used frequencies of the following cue phrases in a document as features for machine learning.

- **Cue phrases for encyclopedia in web pages:** とは (is defined), って (is defined), 事典 (encyclopedia), 辞典 (dictionary), 攻略 (hints-and-tips), 解説 (commentary), 講座 (lecture), 作り方 (how to), すれば良い (recommend to), してください (had better to do), します (do), されます (was done), 手順 (procedure), 編集 (edit), 差分 (view history), 最終更新 (last update), 新規 (new), wiki, 参考 (reference), 出典 (reference), 資料 (material)
- **Cue phrases for encyclopedia in URLs:** jp.wikipedia, dic, weblio.jp, tools
- **Cue phrases for news in web pages:** ニュース (news), 新聞 (news article),トピックス (topics), 日経 (Nikkei), 速報 (breaking news), お知らせ (news), 紹介 (introduction), 公開(release), 配信 (news distribution), 報じる (report), テレビ (TV), 記事 (article) , 更新日 (updated date), Introduction, Staff, Cast, Story, Character, Trailer, スタッフ (staff), キャスト (cast), ストーリー (story), あらすじ (outline), 登場人物 (characters), 予告 (trailer), 試合結果 (result of games), 成績 (score)
- **Cue phrases for news in URLs:** news, headlines
- **Cue phrases for image in web pages:** ムービー (movie), 動画 (movie), 画像 (image), 壁紙 (wallpaper), 待ち受け (standby screen), 画面 (display). 1440[x*]1280,

[3] http://www.lemurproject.org/indri/

[4] https://code.google.com/archive/p/bayon/

1280[x*]960, 1024[x*]768, 800[x*]600, 640[x*]960, 750[x*]1334, Download, ダウンロード(download), 写真 (picture), 素材 (material), 高画質 (high image quality), フリー素材 (free material), DoCoMo, au, Softbank, flickr, フリッカー (flickr), \d 枚, ライブラリー (repository), 撮影 (filmed)

- **Cue phrases for image in URLs:** pixiv, dailymotion, youtube
- **Cue phrases for shopping in web pages:** 販売 (sale), 価格 (price), 消費税 (consumption tax), 商品 (products), 先着 (first .. applicants), 発売 (on sale), チケット (ticket), 期間限定販売 (limited time sale), 営業時間 (open), 定休日 (regular holiday), \d 円 (\d yen), 受付 (reception), 受け付け (reception), 日時 (time and date), 商品番号 (item number), 数量 (quantum), 配送 (shipping), 注文 (order), | 負担 (defrayment), カゴ (shopping cart), お届け (delivery), お試し (trial), 贈答 (present), カート (shopping cart), 営業日 (business day), 通販 (online shopping), 人気 (hot), ショッピング (shopping), 買取 (buy), 買い取り (buy), 宅配 (home delivery), 購読 (subscription), 出荷 (shipping), レストラン (restaurant), ランチ (lunch), グルメ (gourmet), ぐるなび (gurunavi), 本店 (head office), 店舗 (store), メニュー (menu), 入庫 (come in), 値下げ (price cut), アクセス (access), サービス (service), 製品概要 (summary of a product), レンタル (rental), 夏季休業 (non-business day in summer), お任せ下さい (leave it to us), おまかせください (leave it to us), おまかせ下さい (leave it to us), 可能 (possible), メール便 (delivery service), 発送 (shipping), 品名 (name of product), お見積り (estimate), 特価 (bargain price), 修理 (repair), クレジットカード (credit card), ネットショップ (internet shopping), SALE, 安く (discount), 税込 (including tax)
- **Cue phrases for shopping in URLs:** shop, amazon, rakuten.co.jp, gnavi, hotpepper

Data

We manually annotated one of six categories to each web page. The statistics of the data are shown in Table 2.

Table 2. Number of Web Pages for Each Content Type.

Content Type (Vertical Intent)	Number of pages
Encyclopedia	106
News	47
Image	193
QA	35
Shopping	116
Web	244
Average	741

Experimental settings

For classifying multiclasses using binary classifier SVM, we used the one-vs-rest classification method and a linear kernel. We used TinySVM[5] as a machine learning package. We conducted 5-fold cross validation. We used recall and precision for evaluation.

Results

We show the experimental results in Table 3.

Table 3. Evaluation Results for Identification of Content Type for Each Web Page.

Content Type (Vertical Intent)	Recall	Precision
Encyclopedia	0.726	0.794
News	0.319	0.625
Image	0.803	0.994
QA	0.486	0.607
Shopping	0.267	0.492
Web	0.586	0.627
Average	0.591	0.735

We classified 44,280 web documents into six content types, and used them for the identification of the vertical intent for each subtopic.

2.3.2 Identification of the Vertical Intent for Each Subtopic

We identified vertical intents using the k-NN method. We retrieved the top 500 documents using a topic and subtopic pair as a query by the INDRI-based search engine (see Section 2.2). Then, we determined the category that appears most frequently in the top 500 documents as the vertical intent of the subtopic.

3. EXPERIMENTS

3.1 Data and Evaluation

We used 100 topics for the Japanese subtask to evaluate our system. Our system was evaluated by I-rec, D-nDCG@10, D#-nDCG@10, V-score, and QU-score.

3.2 Results and Discussion

We submitted one result provided by our system NEXTI. The experimental results are shown in Table 4. Our system outperformed other systems in all evaluation.

[5] http://chasen.org/~taku/software/TinySVM/

Table 4. Experimental Results.

Evaluation Metrics	Scores
I-rec@10	0.6535
D-nDCG@10	0.5535
D#-nDCG@10	0.6035
V-score(Q-Run only)	0.6507
QU-score(Q-Run only)	0.6271

In the evaluation by QU-score, our system obtained the best scores in 35 topics, while obtained the worst in 8 topics. In the evaluation by I-rec, our system obtained the best scores in 40 topics, while obtained the worst scores in 18 topics.

To investigate the effectiveness of our method, we compared the 15 systems in detail. Figure 2 shows the number of topics over the threshold value (I-rec score) for each system. In the figure, for example, the number of topics over 0.9 I-rec score in our system is 18, which is the highest among all systems. When the threshold value is over 0.6, our system greatly outperformed others.

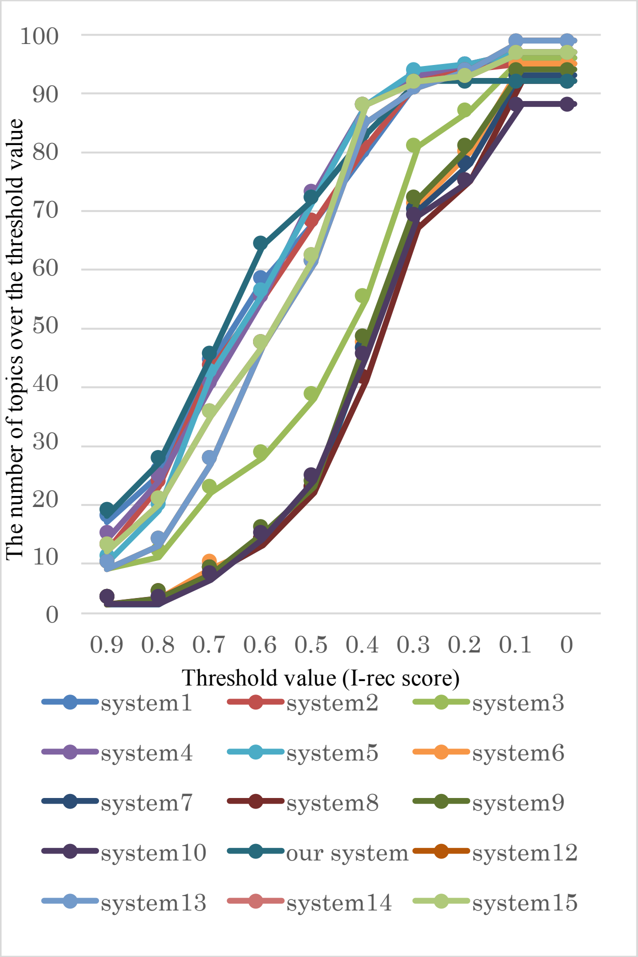

Figure 2. Comparison of 15 systems by I-rec.

4. CONCLUSION

In this paper, we proposed a method that identified subtopics with their vertical intents for a given topic. From the experimental results, our system obtained the best performance among all participant groups.

REFERENCES

[Liu 2014] Liu, Y., Song, R., Zhang, M., Dou, Z., Yamamoto, T., Kato, M., Ohshima, H., and Zhou, K. Overview of the NTCIR-11 IMine Task, *Proc. 11th NTCIR Workshop Meeting* (2014).

[Nanba 2016] Nanba, H., Sakai, T., and Kando, N. Project Next IR: Failure Analysis in Information Retrieval. (Special Issue on “Next NLP: Error Analysis for NLP”), *IPSJ Magazine*, Vol. 57, No. 1 (2016). (in Japanese)

[Yamamoto 2016] Yamamoto, T., Liu, Y., Zhang, M., Zhicheng, D., Zhou, K., Markov, I., Kato, M.P., Ohshima, H., and Fujita, S. Overview of the NTCIR-12 IMine-2 Task, *Proc. 12th NTCIR Workshop Meeting* (2016).