Exploring Semi-Automatic Nugget Extraction for Japanese One Click Access Evaluation

Matthew Ekstrand-Abueg
Northeastern University, USA
mattea@ccs.neu.edu

Virgil Pavlu
Northeastern University, USA
vip@ccs.neu.edu

Makoto P. Kato
Kyoto University, Japan
kato@dl.kuis.kyoto-u.ac.jp

Tetsuya Sakai
Microsoft Research Asia, PRC
tetsuyasakai@acm.org

Takehiro Yamamoto
Kyoto University, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

Mayu Iwata
Osaka University, Japan
iwata.mayu@ist.osaka-u.ac.jp

ABSTRACT

Building test collections based on *nuggets* is useful evaluating systems that return documents, answers, or summaries. However, nugget construction requires a lot of manual work and is not feasible for large query sets. Towards an efficient and scalable nugget-based evaluation, we study the applicability of semi-automatic nugget extraction in the context of the ongoing NTCIR *One Click Access* (1CLICK) task. We compare manually-extracted and semi-automatically-extracted Japanese nuggets to demonstrate the coverage and efficiency of the semi-automatic nugget extraction. Our findings suggest that the manual nugget extraction can be replaced with a direct adaptation of the English semi-automatic nugget extraction system, especially for queries for which the user desires broad answers from free-form text.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

evaluation; information units; NTCIR; nuggets; summaries; test collections

1. INTRODUCTION

For over half a century, information retrieval research has focussed on *document* retrieval. However, in many search tasks, what the user wants is *information* rather than a list of documents. Accordingly, the task of returning relevant *information* in response to a query, and methods for evaluating such tasks based on *nuggets* have received attention recently (e.g. [2]). Building test collections based on nuggets is useful not only for evaluating systems that return *direct answers* or *summaries* (e.g. [7, 9]), but also for handling novelty and redundancy in document retrieval [1] and for efficient relevance assessments [8].

The ongoing NTCIR *One Click Access* (1CLICK) task is an example of an *information* retrieval task, which can be described as: *Given a search query, provide a short summary that fits a mobile phone screen. Put relevant pieces of information first so as to minimise the amount of text the user has to read.*[1] Unlike previous summarization and question answering evaluation, the task was novel in that the *positions* of nuggets found in the system output were leveraged to evaluate systems [9]. In the evaluation of the first *Japanese* 1CLICK task at the NTCIR-9 conference, two types of *manual* effort were required: (a) extracting nuggets from relevant documents; and (b) identifying the matches (and their exact positions) between the system output and the list of gold-standard nuggets. While the manual efforts enable highly reliable and robust evaluation, measures for (semi-)automating some of the processes are essential for enhancing evaluation efficiency and scalability. While Task (b) has been tackled to some extent in the summarization and question answering communities using automatic text segmentation techniques such as N-grams [5, 6], the present study addresses the question of whether Task (a), i.e., extracting nuggets, can be done semi-automatically and effectively.[2]

Recently, Rajput *et al.* [8] proposed a framework for conducting document relevance judging and nugget judging simultaneously. The framework is based on an online *mutual reinforcement* algorithm designed to dramatically reduce the assessor effort. That is, the judged nuggets are used for selecting new documents to be judged; the judged documents are used for selecting new nuggets to be judged. However, their work considered *English* information access only, and it is an open question whether their approach transfers well to other languages. In particular, the Japanese language is radically different from European languages: there are no spaces between words; several different character sets (both ideograms and phonograms) are used together; and the grammar generally allows more flexible word ordering within a sentence than those of European languages. Therefore, in this study, we address the question of whether the mutual reinforcement framework of Rajput *et al.* can be extended for the purpose of semi-automatic extraction of nuggets for the *Japanese* 1CLICK task. We are currently running the second 1CLICK (1CLICK-2) task at NTCIR-10, which comprises English and Japanese subtasks. In this paper, we leverage the data constructed at 1CLICK-2 to pursue this research question.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

[1] 1CLICK homepage: `http://research.microsoft.com/en-us/projects/1click/`

[2] We gratefully acknowledge support provided by NSF IIS-1256172.

Specifically, we examine which types of queries are suited to the system as a direct adaptation from English versus fully manual extraction and conclude by hypothesizing changes to improve general applicability.

2. JAPANESE ONE CLICK ACCESS

For the NTCIR 1CLICK-2 task, the organizers provided a set of queries and the baseline search results, which consisted of top-ranked Yahoo! API search results returned in response to each query, and expected participants to generate system output based on the provided search results for each query. Having received participants' system outputs, the organizers evaluated them based on nuggets prepared for each query in advance. In the following subsections, we explain the queries and nuggets used for the NTCIR 1CLICK-2 task.

2.1 Queries

The NTCIR 1CLICK-2 test collection includes 100 Japanese and 100 English queries. Based on a study on mobile query logs [4], eight query types were considered: ARTIST, ACTOR, POLITICIAN, ATHLETE, FACILITY, GEO, DEFINITION, and QA. For each query type, it was assumed that the user has the following information needs:

ARTIST, ACTOR, POLITICIAN, ATHLETE (10 each) user wants important facts about celebrities;

FACILITY (15) user wants access and contact information for a particular landmark, facility etc.;

GEO (15) user wants access and contact information for entities with geographical constraints, e.g. sushi restaurants near Tokyo station;

DEFINITION (15) user wants to look up a phrase, etc.;

QA (15) user wants to know factual (but not necessarily factoid) answers to a natural language question.

The number of queries for each type is shown in parentheses.

To allow for cross-language comparison, we created 15 queries which overlapped between Japanese and English. Table 1 shows the overlap query set, which comprises one query from ARTIST and ACTOR, two queries from POLITICIAN and ATHLETE, and three queries from FACILITY, DEFINITION, and QA. Note that there is no overlap GEO query as it was difficult to find GEO queries that are used both in English and Japanese. The Japanese version of those 15 overlap queries were used in our evaluation.

2.2 Nuggets

A nugget at 1CLICK-2 is defined as a sentence relevant to the information need for a query, and was used to evaluate the quality of system output by identifying which nuggets are present in its content. At NTCIR 1CLICK-2, native Japanese speakers in the organizer team identified relevant documents from the provided baseline search results, and manually extracted relevant sentences as nuggets. For example, a sentence "Ichiro Suzuki (born October 22, 1973) is a professional baseball player" was extracted as a nugget for query "ichiro suzuki." In total, 3,927 nuggets were extracted for 100 Japanese queries (39.2 nuggets per query on average).

Section 3 describes how the nugget extraction can be semi-automated, and Section 4 then discusses the performance of the semi-automatic nugget extraction.

Table 1: NTCIR 1CLICK-2 overlap queries.

ID	query type	query
1C2-E-0001	ARTIST	michael jackson death
1C2-E-0026	ACTOR	jennifer gardner alias
1C2-E-0042	POLITICIAN	robert kennedy cuba
1C2-E-0045	POLITICIAN	mayor bloomberg
1C2-E-0070	ATHLETE	ichiro suzuki
1C2-E-0071	ATHLETE	fabio cannavaro captain
1C2-E-0092	FACILITY	hawaii pacific university
1C2-E-0093	FACILITY	atlanta airport
1C2-E-0095	FACILITY	american airlines arena
1C2-E-0144	DEFINITION	geothermal energy
1C2-E-0146	DEFINITION	thanksgiving canada
1C2-E-0150	DEFINITION	cubic yard
1C2-E-0178	QA	why is the sky blue
1C2-E-0180	QA	why do cats purr
1C2-E-0184	QA	why is the ocean salty

3. SEMI-AUTOMATIC NUGGET EXTRACTION

As described in [8], the nuggets system uses a mutual, iterative reinforcement feedback system between documents and nuggets automatically extracted from the documents. An assessor judges the relevance of documents and the system updates beliefs about the relevance of unjudged documents and nuggets. This procedure is illustrated in Figure 1.

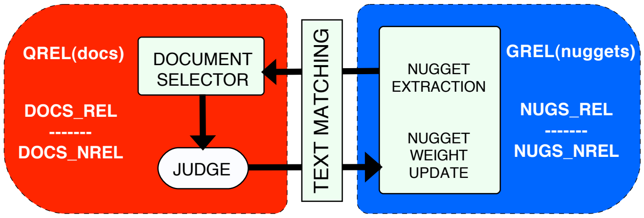


Figure 1: The iterative assessment procedure: documents are selected and assessed, nuggets are extracted and [re]weighted.

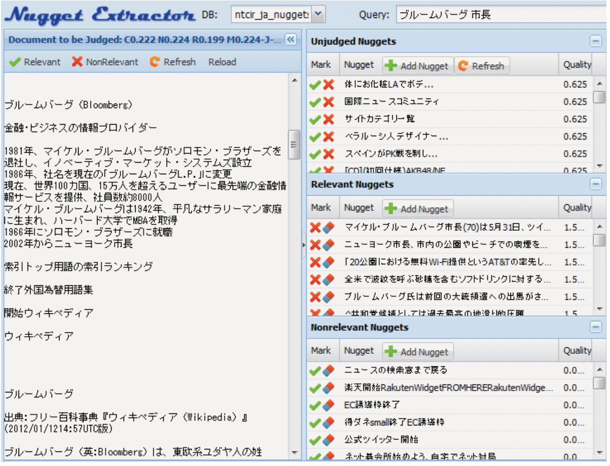
This procedure has been modified to both allow for processing of Japanese text and to allow for judgments on nuggets in addition to documents. As the nugget system is highly modular, the English-specific text processing tasks can be directly substituted with those of another language. Specifically, the two instances in which text is automatically processed are when nuggets are extracted from documents and when text is matched between nuggets and documents.

For nugget extraction, we maintain sentences as the text unit. For Japanese, we use a regular expression to match sentence endings, as these patterns are more well defined than in English. Documents are segmented into sentences and all sentences from relevant documents are used as nuggets in the learning procedure.

For matching, we maintain the shingle matching system previously described. Briefly, for each shingle, a sequence of k consecutive words, in a piece of text and minimum span S of words in a document which contains all shingle words in any order, we calculate the shingle score as

$$\mathit{shingleMatch} = \lambda^{(S-k)/k}$$

where λ is a fixed decay parameter, $\lambda = 0.95$ in our case. However, this requires some knowledge of Japanese morphemes to segment text into word units. Raw text is split into these word units



ブルームバーグ（Bloomberg）

金融・ビジネスの情報プロバイダー

1981年、マイケル・ブルームバーグがソロモン・ブラザーズを
退社し、イノベーティブ・マーケット・システムズ設立
1986年、社名を現在の「ブルームバーグL.P.」に変更
現在、世界100カ国、15万人を超えるユーザーに最先端の金融情
報サービスを提供、社員数約8000人
マイケル・ブルームバーグは1942年、平凡なサラリーマン家庭
に生まれ、ハーバード大学でMBAを取得
1966年にソロモン・ブラザーズに就職
2002年からニューヨーク市長

索引トップ用語の索引ランキング

終了外国為替用語集

開始ウィキペディア

ウィキペディア

ブルームバーグ

出典：フリー百科事典『ウィキペディア（Wikipedia）』
(2012/01/1214:57UTC版)

ブルームバーグ（英：Bloomberg）は、東欧系ユダヤ人の姓

Unjudged Nuggets

Mark	Nugget	Quality
✓ ✗	体にお化粧LAでボデ...	0.625
✓ ✗	国際ニュースコミュニティ	0.625
✓ ✗	サイトカテゴリ一覧	0.625
✓ ✗	ベラルーシ人デザイナー...	0.625
✓ ✗	スペインがPK戦を制し...	0.625
✓ ✗	[CD](初回仕様)AKB48/NE	0.625

Relevant Nuggets

Mark	Nugget	Quality
✗ ◆	マイケル・ブルームバーグ市長(70)は5月31日、ツイ...	1.5...
✗ ◆	ニューヨーク市長、市内の公園やビーチでの喫煙を...	1.5...
✗ ◆	「20公園における無料Wi-Fi提供というAT&Tの率先し...	1.5...
✗ ◆	全米で波紋を呼ぶ砂糖を含むソフトドリンクに対する...	1.5...
✗ ◆	ブルームバーグ氏は前回の大統領選への出馬がさ...	1.5...
✗ ◆	今共和党候補としては過去最高の地滑り的圧勝	1.5

Nonrelevant Nuggets

Mark	Nugget	Quality
✓ ◆	ニュースの検索窓まで戻る	0.0...
✓ ◆	楽天開始RakutenWidgetFROMHERERakutenWidge...	0.0...
✓ ◆	EC誘導枠終了	0.0...
✓ ◆	得ダネsmall終了EC誘導枠	0.0...
✓ ◆	公式ツイッター開始	0.0...
✓ ◆	ネット碁会所始めよう。自宅でネット対局	0.0

Figure 2: Screenshot of the Nugget Extractor for Japanese.

using a Japanese morphological analyzer, *MeCab*[3]. Only nouns, verbs, adjectives, and adverbs were used in the shingle matching, while the other part-of-speech, such as particles and conjunctions (or function words), were excluded. From here, the shingle matching system proceeds as before, in which we accumulate all shingles for a nugget-document pair (n, d):

$$M(n, d) = \frac{1}{\#shingles} \sum_{s \in shingles(n)} shingleMatch(s, d)$$

The adaptability of the Hedge algorithm[3], used by the nugget's iterative update procedure, leads to a nugget score being *increased* as long as it matches relevant documents (positive feedback) and *reduced* when the nugget matches non-relevant documents (negative feedback). For a nugget n and new document d:

$$q_n^{new} = q_n^{old} * \beta^{M(n,d)}, \beta = 1.3 \text{ if } d \in \text{Rel}, 0.5 \text{ if } d \in \text{Nonrel}$$

This is highly beneficial for finding new and diverse relevant documents, but sometimes problematic for finding the global set of relevant information.

In the document judgment interface shown in Figure 2, we display nuggets sorted by their quality score q_n to give the user an idea of what information is deemed relevant at this stage of the nugget procedure. We allow the user to mark nuggets as relevant or nonrelevant at any point in the process in order to explicitly overwrite any implicit feedback from the algorithm and to select what portion of information is relevant to that user. This explicit information can be incorporated into the feedback system, allowing for more targeted document evaluation. This is done by automatically assigning the maximum score given to any nugget to all judged relevant nuggets, and the minimum score to all judged nonrelevant nuggets:

$$M_{\text{judged}}(n, d) = \begin{cases} \max\limits_{\forall d'; n' \in \text{Unjudged}} M(n', d'), & \text{if } n \in \text{R} \\ \min\limits_{\forall d'; n' \in \text{Unjudged}} M(n', d'), & \text{if } n \in \text{NR} \end{cases}$$

where R and NR are the set of judged relevant and nonrelevant nuggets respectively. Explicit nuggets judgments have a strong influence on the update procedure, but they do not entirely dominate, as compared to assigning a fixed score or one outside the ranges automatically given.

Additionally, as shown in our previous work, we can evaluate global relevance by examining the total ability of a nugget to bring in relevant vs nonrelevant documents during the iterative procedure. This provides evidence as to the general ability of the chosen nuggets and our matching system to automatically sort information.

Finally, we have modified the system to allow a user to inject manually-created nuggets into the system. These nuggets are automatically matched to all documents, and function just as automatically extracted nuggets in the update procedure. This allows primarily for the injection of seed information to improve the diversity of the results, for instance if an aspect of a query is underrepresented. This can be thought of as a form of query expansion, but requires no modification to the nuggets procedure.

4. RESULTS AND DISCUSSIONS

We manually extract nuggets for the Japanese version of the 15 overlap queries listed in Table 1, and independently semi-automatically extract nuggets for the same query set as described in Section 3. It can be difficult to judge which nuggets contain the same information when they were obtained using different methods as nuggets can include multiple pieces of information and the granularity of each nugget highly depends on the document from which the nugget was extracted.

Thus, nuggets were manually broken down into smaller units, known as information units (or *iUnits*). An iUnit is a factual statement which is *relevant* (satisfies the information need behind the query partially or wholly) and *atomic* (cannot be broken down into multiple iUnits without the components losing meaning). For example, a nugget "Ichiro Suzuki (born October 22, 1973) is a professional baseball player" was broken down into two iUnits: (1) "born October 22, 1973," and (2) "a professional baseball player." In fact, for the NTCIR-10 Japanese 1CLICK-2 subtask, iUnits were utilized instead of nuggets for evaluating system output due to their maximal granularity.

Once the nuggets from both the manual and semi-automatic system are broken down into iUnits, they can be directly compared to assess the commonalities and differences between the two sets. We compare the two sets for all queries, grouped into four meta-categories based on the type of query and desired answers. FACILITY queries look for exact facts about location and contact information, and CELEBRITY queries (ACTOR, ATHELETE, MUSICIAN, and POLITICIAN) look for specific facts about a celebrity, while DEFINITION and QA queries allow for broader answers.

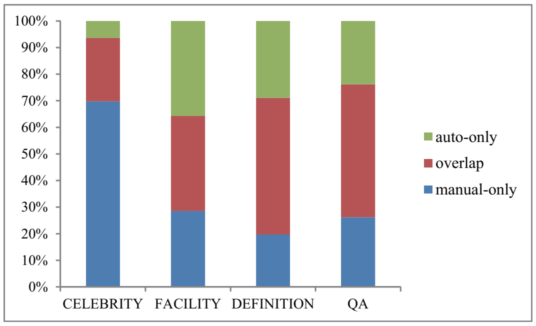


Figure 3: Distinct and overlapping iUnits extracted manually vs semi-automatically for the different query types.

[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

Figure 3 shows the proportion of overlap of manual vs semi-automatic iUnits for the different queries. Manual iUnits are clearly superior for the fact-based queries, while semi-automatic iUnits perform better for the broader queries.

Two primary assumptions of our current system are that nuggets are sentences and that partial sentence matches imply partial information overlap. These assumptions were made specifically for broader informational-based queries, but can plainly cause problems when searching for specific types of answers. A human, for instance, is good at picking out dates and other numbers from the text, but under our first assumption, small facts are not easily represented by nuggets in the form of an entire sentence. The automated system, however, is designed to branch out based on similar information and similar contexts, so it finds a much larger space of information. For instance, with CELEBRITY query "ichiro suzuki", some example iUnits are "No. 31" and "height: 180.3cm," which do not have correct partial matches and are not commonly found as sentences, whereas with DEFINITION query "geothermal energy", iUnits such as "the first geothermal energy plant was developed in Italy" and "considered to be sustainable energy" match exactly the target form.

These assumptions could be modified based on query categorization to search for and extract alternate candidate nuggets and to perform text matching using differing unit sizes based on the type of nugget, both of which are planned areas of future study. For now, however, these results help classify times when manual extraction is necessary versus when the semi-automatic method is beneficial.

In order to evaluate the extraction efficiency of the system, we examine the number of iUnits found over time as document judgments occurred. Figure 4 shows this information averaged over the four meta-categories. There is a clear trend of finding a great deal of relevant information early in the process, with an expected diminishing return as more documents are judged. This is especially true for definition and facility queries, where the system is either especially good at finding information, or there is a small amount of primary relevant information. Using these findings, we can assess useful heuristics for stopping conditions during the judging process based on how much information is believed to be left or to allow the user to decide if continuing is likely to be worth the effort.

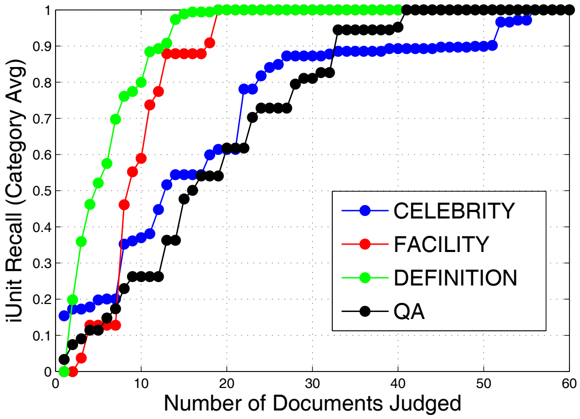


Figure 4: iUnit set growth VS effort on document assessing, averaged over queries for each category.

Finally, as with English queries in our previous work [8], we can examine the ability of the system to accurately infer the quality of a particular nugget. We use the same scoring method based on the ability of a nugget to correctly rank the judged documents by its match to these documents. The MAP value of the inferred nugget scores vs the judged relevance is produced for each of the meta-categories and shown in Figure 5. However, in this task, the average number of documents judged was only 39, so a lower overall score is expected, especially for queries categories with few relevant documents. The score remains useful enough to provide a great reduction in the number of nuggets requiring manual assessment in order to find the relevant information.

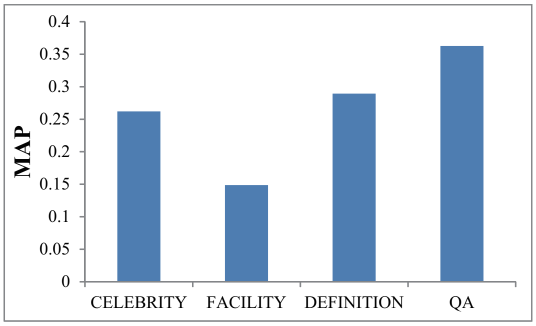


Figure 5: MAP value of the inferred nugget score vs the judged relevance for each category.

5. CONCLUSIONS AND FUTURE WORK

We have introduced an adaptation of our nugget extractor system for Japanese, demonstrated its ability to extract information on Japanese, and examined the cases under which it performs better and worse than manual text extraction. We have seen that, as designed, the nugget system is effective at finding larger quantities of information written as free-form text.

A primary area for future work is examining the effects of using varying sizes of nuggets and the ability of both oracle and automated query classifiers to choose the desired types of nuggets. Additionally, we plan to examine the utility of these nuggets in directly creating summaries. We believe that the redundancy and overlap of information will help a summarizer with both flow and importance of information, two important facets of abstractive summarization.

6. REFERENCES

[1] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *ACM SIGIR 2008*, pages 659–666, 2009.

[2] J. A. et al. Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum*, 46(1):2–32, 2012.

[3] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. COLT '95.

[4] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. In *ACM SIGIR 2009*, pages 43–50, 2009.

[5] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL 2004 Workshop on Text Summarization Branches Out*, 2004.

[6] J. Lin and D. Demner-Fushman. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587, 2006.

[7] T. Mitamura, H. Shima, T. Sakai, N. Kando, T. Mori, K. Takeda, C.-Y. Lin, R. Song, C.-J. Lin, and C.-W. Lee. Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access. In *NTCIR-8*, pages 15–24, 2010.

[8] S. Rajput, M. Ekstrand-Abueg, V. Pavlu, and J. A. Aslam. Constructing test collections by inferring document relevance via extracted relevant information. In *ACM CIKM 2012*, pages 145–154, 2012.

[9] T. Sakai, M. P. Kato, and Y.-I. Song. Click the search button and be happy: Evaluating direct and immediate information access. In *ACM CIKM 2011*, pages 621–630, 2011.