IPSJ Transactions on Databases Vol.5 No.2 11–16 (June 2012)

Regular Paper

Query Snowball: A Co-occurrence-based Approach to Multi-document Summarization for Question Answering

Hajime Morita[1,a)] Tetsuya Sakai[2] Manabu Okumura[3]

Received: December 20, 2011, Accepted: April 10, 2012

Abstract: We propose a new method for query-oriented extractive multi-document summarization. To enrich the information need representation of a given query, we build a co-occurrence graph to obtain words that augment the original query terms. We then formulate the summarization problem as a Maximum Coverage Problem with Knapsack Constraints based on word pairs rather than single words. Our experiments with the NTCIR ACLIA question answering test collections show that our method achieves a pyramid F3-score of up to 0.313, a 36% improvement over a baseline using Maximal Marginal Relevance.

Keywords: question answering, query-oriented, multi-document summarization, information need representation

1. Introduction

Automatic text summarization aims at reducing the amount of text the user has to read while preserving important contents, and has many applications in this age of digital information overload [18]. In particular, *query-oriented multi-document summarization* is useful for helping the user satisfy his information need efficiently by gathering important pieces of information from multiple documents.

In this study, we focus on *extractive* summarization [16], in particular, on sentence selection from a given set of source documents that contain relevant sentences. One well-known challenge in selecting sentences relevant to the information need is the vocabulary mismatch between the query (i.e., information need representation) and the candidate sentences. Hence, to enrich the information need representation, we build a co-occurrence graph to obtain words that augment the original query terms. We call this method *Query Snowball*.

Another challenge in sentence selection for query-oriented multi-document summarization is how to avoid redundancy so that diverse pieces of information (i.e., *nuggets* [24], [28]) can be covered. For penalizing redundancy across sentences, using single words as the basic unit may not always be appropriate, because different nuggets for a given information need often have many words in common. Thus, if we use single words as the basis for penalising redundancy in sentence selection, it would be difficult to cover both of these nuggets in the summary because of the word overlaps. We therefore use *word pairs* as the basic unit for computing sentence scores, and then formulate the summarization problem as a Maximum Cover Problem with Knapsack Constraints (MCKP) [5], [25]. This problem is an optimization problem that maximizes the total score of words covered by a summary under a summary length limit.

Figure 1 shows examples of the vocabulary mismatch problem and the word overlap problem from the NTCIR-8 ACLIA2 Japanese question answering test collection. Here, three gold-standard nuggets for the question "*Sen to Chihiro no Kamikakushi (Spirited Away)* is a full-length animated movie from Japan. The user wants to know how it was received overseas" (in English translation) are shown. Each nugget represents a particular award that the movie received. It can be observed that, while Nugget example 2 have a few words in common with the question, Nugget example 1 has no overlap. Thus, to capture nuggets such as Nugget example 1, we need to enrich the information need representation. On the other hand, Nuggets example 3 has three words in common with Nugget example 2 (underlined). Therefore, we need to accept such word overlap to capture both of Nugget examples 2 and 3.

We evaluate our proposed methods using Japanese complex question answering (QA) test collections from NTCIR ACLIA — Advanced Cross-lingual Information Access task [20], [21]. However, our method can easily be extended to other languages. It should be noted that our methods are *components* useful for complex QA, and that we treat the QA test collections as those for query-biased extractive summarization. Other standard components of QA such as question classification, document retrieval and answer extraction may be combined with our proposed methods to build an end-to-end complex QA system, but this is beyond the scope of our study.

2. Related Work

Much work has been done for generic multi-document summarization [3], [12], [14], [25], [26]. Carbonell and Goldstein [2] proposed the Maximal Marginal Relevance (MMR) criteria for non-redundant sentence selection, which consist of document

[1] Tokyo Institute of Technology, Yokohama, Kanagawa 226–8503, Japan
[2] Microsoft Research Asia, Haidian District, Beijing 100080, P.R. China
[3] Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, Kanagawa 226–8503, Japan
[a)] morita@lr.pi.titech.ac.jp

- Question
  千と千尋の神隠しは日本の長編アニメーション映画であるが、この映画の海外での評価について知りたい。
  Sen to Chihiro no Kamikakushi (Spirited Away) is a full-length animated movie from Japan. The user wants to know how it was received overseas.
- Nugget example 1
  ドイツ で グランプリ を 受賞
  Awarded a Grand Prize in Germany
- Nugget example 2
  全米 映画 批評 会議 の アニメ 賞
  National Board of Review of Motion Pictures Best Animated Feature
- Nugget example 3
  ロサンゼルス 批評 家 協会 賞 の アニメ 賞
  Los Angeles Film Critics Association Award for Best Animated Film

Fig. 1 Question and gold-standard nuggets example in NTCIR-8 ACLIA2 dataset.

similarity and redundancy penalty. McDonald [19] presented an approximate dynamic programming approach to maximize the MMR criteria. Yih et al. [29] formulated the document summarization problem as an MCKP, and proposed a supervised method, while, our method is unsupervised. Filatova and Hatzivassiloglou [5] also formulated summarization as an MCKP, and they used two types of concepts in documents: single words and *events* (named entity pairs with a verb or a noun). While their work was for generic summarization, our method is designed specifically for query-oriented summarization.

MMR-based methods are also popular for query-oriented summarization [6], [7], [11], [15]. Moreover, graph-based methods for summarization and sentence retrieval are popular [1], [22], [27]. Unlike existing graph-based methods, our method explicitly computes indirect relationships between the query and words in the documents to enrich the information need representation. To this end, our method utilizes within-sentence co-occurrences of words.

Recently, new summarization methods that adopt monotone submodular object function have been proposed [12], [13]. Lin and Bilmes [13] proposed a new monotone submodular function for query-oriented summarization. Instead of assigning a query relevance score to a sentence, our method first assigns a query relevance score to each word scores. Thus our method can easily be extended to sentence compression, which aims to remove unimportant words or clauses from the original sentences. Moreover, while Lin and Bilmes try to enrich the information need representation using Wordnet, our method relies only on co-occurrences within the source documents.

The approach taken by Jagadeesh et al. [7] is similar to our proposed method in that it uses word co-occurrence and dependencies within sentences in order to measure relevance of words to the query. However, while their approach measures the generic relevance of each word based on *Hyperspace Analogue to Language* [17] using an external corpus, our method measures the relevance of each word within the document contexts, and the query relevance scores are propagated recursively.

3. Proposed Method

Section 3.1 introduces the Query Snowball (QSB) method which computes the query relevance score for each word. Then, Section 3.2 describes how we formulate the summarization problem based on word pairs.

3.1 Query Snowball Method (QSB)

The basic idea behind QSB is to close the gap between the query (i.e., information need representation) and relevant sentences by enriching the information need representation based on co-occurrences. To this end, QSB computes a *query relevance score* for each word in the source documents as described below.

Figure 2 shows the concept of QSB. Here, Q is the set of query terms (each represented by q), $R1$ is the set of words ($r1$) that co-occur with a query term in the same sentence, and $R2$ is the set of words ($r2$) that co-occur with a word from $R1$, excluding those that are already in $R1$. The imaginary root node at the center represents the information need, and we assume that the need is propagated through this graph, where edges represent within-sentence co-occurrences.

3.1.1 Preliminary Analysis

While, in theory, the propagation process can be iterated to further enrich the information need representation, our preliminary analysis showed that it is not useful to go beyond $R2$, say, to '$R3$.' **Table 1** shows the total number of word overlaps between the answer nuggets and the original query terms, $R1$, $R2$ or $R3$ for the NTCIR-7 ACLIA1 collection, which we use as development data. It can be observed that $R3$ is not very useful as it drifts away from the original information need.

3.1.2 Query Relevance Score

Our first clue for computing a word score is the query-independent importance of the word. We represent this *base word score* by $s_b(w) = \log(N/ctf(w))$ or $s_b(w) = \log(N/n(w))$, where $ctf(w)$ is the total number of occurrences of w within the corpus and $n(w)$ is the document frequency of w, and N is the total number of documents in the corpus. We will refer to these two versions as *itf* and *idf*, respectively. The reason why we consider

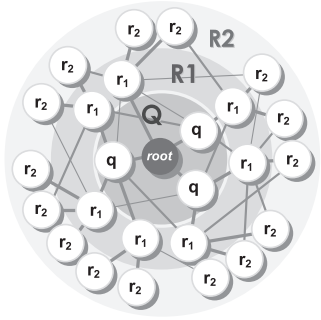

Fig. 2 Co-occurrence graph (Query Snowball).

Table 1 Size of word overlap between word sets and answer nuggets in ACLIA1 test dataset.

	size
Query terms	325
R1	6471
R2	623
R3	34

itf as well as idf is that we found in a preliminary analysis that idf tends to assign high scores to non-topical words. Our second clue is the weight propagated from the center of the co-occurrence graph shown in Fig. 2. Below, we describe how to compute the word scores for words in $R1$ and then those for words in $R2$.

As Fig. 2 suggests, the query relevance score for $r1 \in R1$ is computed based not only on its base word score but also on the relationship between $r1$ and $q \in Q$. To be more specific, let $freq(w, w')$ denote the within-sentence co-occurrence frequency for words w and w', and let $distance(w, w')$ denote the *minimum dependency distance* between w and w': A dependency distance is the path length between nodes w and w' within a dependency parse tree; the minimum dependency distance is the shortest path length among all dependency parse trees of source-document sentences in which w and w' co-occur. A low value indicates that the word pair has a strong connection within a context of source-documents. We use the minimum dependency distance rather than the mean as we do not require the word pair to have a strong connection in every co-occurrence. Then, the query relevance score for $r1$ can be computed as:

$$s_r(r1) = \sum_{q \in Q} s_b(r1) \left(\frac{s_b(q)}{sum_Q} \right) \left(\frac{freq(q, r1)}{distance(q, r1) + 1.0} \right) \quad (1)$$

where $sum_Q = \sum_{q \in Q} s_b(q)$. It can be observed that the query relevance score $s_r(r1)$ reflects the base word scores of both q and $r1$, as well as the co-occurrence frequency $freq(q, r1)$. Moreover, $s_r(r1)$ depends on $distance(q, r1)$, the minimum dependency distance between q and $r1$, which reflects the strength of relationship between q and $r1$. This quantity is used in one of its denominators in Eq. (1) as small values of $distance(q, r1)$ imply a strong relationship between q and $r1$. The 1.0 in the denominator avoids division by zero. The itf score of a very frequent word can be negative. In such a case, we reset the score to 0. This prevents propagation of negative scores, and also ensures the monotone submodularity of the objective function.

Similarly, the query relevance score for $r2 \in R2$ is computed based on the base word score of $r2$ and the relationship between $r2$ and $r1 \in R1$:

$$s_r(r2) = \sum_{r1 \in R1} s_b(r2) \left(\frac{s_r(r1)}{sum_{R1}} \right) \left(\frac{freq(r1, r2)}{distance(r1, r2) + 1.0} \right) \quad (2)$$

where $sum_{R1} = \sum_{r1 \in R1} s_r(r1)$.

3.2 Score Maximization Using Word Pairs

Having determined the query relevance score, the next step is to define the summary score. To this end, we use word pairs rather than individual words as the basic unit. This is because word pairs are more informative for discriminating across different pieces of information than single common words (Recall the example mentioned in Section 1). Thus, the word pair score is simply defined as: $s_p(w_1, w_2) = s_r(w_1) s_r(w_2)$ and the summary score is computed as:

$$f_{QSBP}(S) = \sum_{\{w_1, w_2 | w_1 \neq w_2 \text{ and } w_1, w_2 \in u \text{ and } u \in S\}} s_p(w_1, w_2) \quad (3)$$

where u is a textual unit, which in our case is a sentence. Our problem then is to select S to maximizes $f_{QSBP}(S)$. Let $l(u)$ denote the length of u. Given a set of source documents D and a length limit L for a summary, we used **Algorithm 1** to produce a multi-document summary S. The above function based on word pairs is still monotone submodular, and therefore we can apply the greedy approximate algorithm with a performance guarantee of $1 + 1/\sqrt{e}$ as proposed in previous work [8], [25]. That is, the algorithm has a guarantee that its output S always satisfies $f(S^*) \leq 1 + \frac{1}{\sqrt{e}} f(S)$, where S^* is the optimal solution. The algorithm iteratively selects a sentence u that maximizes the score difference $f(S \cup \{u\}) - f(S)$ and adds the sentence to a summary S, then outputs a summary that has the maximum score within the generated summary and every sentence in source documents. In our experiments, we used f_{QSBP} and other variants we will show later. We call our proposed method QSBP: Query Snowball with Word Pairs.

Algorithm 1 Algorithm for summarization.

```
Require: D, L
  W = D, S = φ
  while W ≠ φ do
    u = arg max_{u∈W} (f(S∪{u}) − f(S)) / l(u)
    if l(u) + Σ_{u_S∈S} l(u_S) ≤ L then
      S = S ∪ {u}
    end if
    W = W \ {u}
  end while
  u_max = arg max_{u∈D} f(u)
  if f(u_max) > f(S) then
    return u_max
  else return S
  end if
```

4. Experiments

4.1 Experimental Environment

We evaluate our method using Japanese QA test collections from NTCIR-7 ACLIA1 and NTCIR-8 ACLIA2 [20], [21]. The collections contain complex questions and their answer nuggets with weights. **Table 2** shows some statistics of the data. We use the ACLIA1 development data for tuning a parameter for our baseline as shown in Section 4.2 (whereas our proposed method is parameter-free), and the ACLIA1 and ACLIA2 test data for evaluating different methods. In this paper, we only discuss the results for the ACLIA2 test data, but those for the ACLIA1 test data were very similar. As our aim is to answer complex questions by means of multi-document summarization, we removed factoid questions from the ACLIA2 test data.

Although the ACLIA test collections were originally designed for Japanese QA evaluation, we treat them as query-oriented summarization test collections. That is, in our problem setting, the

Table 2 ACLIA dataset statistics.

	ACLIA1		ACLIA2
	Development	Test	Test
#of questions	101	100	80*
#of avg. nuggets	5.8	12.8	11.2*
Question types	DEFINITION, BIOGRAPHY, RELATIONSHIP, EVENT		+WHY
Articles years	1998–2001		2002–2005
Documents	Mainichi Newspaper		

*After removing the factoid questions.

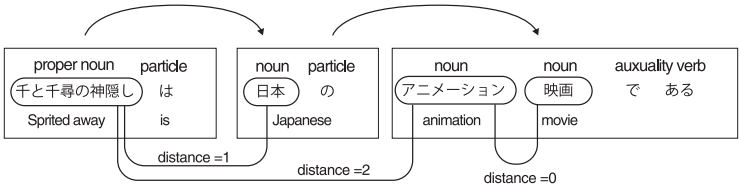

Fig. 3 Dependency distance: square and arrow indicate clause and dependency between clauses respectively.

candidate documents are already given. We use all of these documents provided by ACLIA as input to the multi-document summarizers, even though some of the documents do not actually contain any answer nuggets [20], [21].

We preprocessed the Japanese documents basically by automatically detecting sentence boundaries based on Japanese punctuation marks, but we also used regular-expression-based heuristics to detect a glossary of terms often provided at the end of articles. As these glossaries are usually very useful for answering BIOGRAPHY and DEFINITION questions, we treated the entire description of a term (generally multiple sentences) as a single sentence.

We used MeCab [10] for morphological analysis, and calculated base word scores $s_b(w)$ using Mainichi articles from 1991 to 2005. We also used MeCab to convert each word to its base form and to extract content words using POS tags. As for dependency parsing for distance computation, we used CaboCha [9]. In the case of Japanese, dependency is defined between clauses that contain several words (morphemes). That is, we cannot obtain a dependency relation between words within a clause. Therefore, we define the dependency distance between words within a clause as 0, and define the distance between words across clauses as the distance between these clauses, as shown in **Fig. 3**. We did not use a stop word list or any other external knowledge.

Following the NTCIR-9 one click access task setting [*1], we aimed at generating summaries of Japanese 500 characters or less. To evaluate the summaries, we followed the practices at the TAC summarization tasks [4] and NTCIR ACLIA tasks, and computed pyramid-based precision with an allowance parameter of C, recall, $F\beta$ (where β is 1 or 3) scores. The value of C was determined based on the average nugget length for each question type of the ACLIA2 collection [21]. Precision and recall are computed based on the nuggets that the summary covered as well as their weights. The first author of this paper manually evaluated whether each nugget matches a summary. The evaluation metrics are formally defined as follows:

$$precision = \min\left(\frac{C \cdot (\sharp \text{ of matched nuggets})}{\text{summary length}}, 1\right),$$

$$recall = \frac{\text{sum of weights over matched nuggets}}{\text{sum of weights over all nuggets}},$$

$$F\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}.$$

4.2 Baseline

MMR is a popular approach in query-oriented summarization. For example, at the TAC 2008 opinion summarization track, a top performer in terms of pyramid F score used an MMR-based method. Our own implementation of an MMR-based baseline uses an existing algorithm to maximize the following summary set score function [12]:

$$f_{MMR}(S) = \gamma\Big(\sum_{u \in S} Sim(u, v_D) + \sum_{u \in S} Sim(u, v_Q)\Big) - (1-\gamma) \sum_{\{(u_i,u_j)|i \neq j \text{ and } u_i,u_j \in S\}} Sim(u_i, u_j) \quad (4)$$

where v_D is the vector representing the source documents, v_Q is the vector representing the query terms, Sim is the cosine similarity, and γ is a parameter. Thus, the first term of this function reflects how the sentence represents the entire documents; the second term reflects the relevance of the sentence to the query; and finally the function penalizes redundant sentences. The algorithm maximizes the function f_{MMR} by iteratively adding $\arg\max_u \frac{f_{MMR}(S \cup \{u\}) - f_{MMR}(S)}{l(u)^r}$ to output S, where r is a parameter that determines the balance between cost and gain to add a new sentence to a summary. We set γ to 0.8 and the scaling factor r used in the algorithm to 0.3 based on a preliminary experiment with a part of the ACLIA1 development data. We also tried two variants of Eq. (4): The first one is incorporating sentence position information [23] to our MMR baseline but this actually hurt performance in our preliminary experiments; The second one is that completely disregards the similarity with v_D, to avoid creating summaries that are too generic (as opposed to query-biased). We refer to this variant as "baseline (no generic)."

4.3 Variants of the Proposed Method

To clarify the contributions of each components, the minimum dependency distance, QSB and the word pair, we also evaluated the following simplified versions of QSBP. (We use the itf version by default, and will refer to the idf version as QSBP(idf).) To examine the contribution of using minimum dependency distance, we remove $distance(w, w')$ from Eqs. (1) and (2). We call the method QSBP(nodist). To examine the contribution of using word pairs for score maximization (see Section 3.2) on the performance of QSBP, we replaced Eq. (3) with:

$$f_{QSB}(S) = \sum_{\{w|w \in u_i \text{ and } u_i \in S\}} s_r(w). \quad (5)$$

We will refer to this simply as QSB. Also, to examine the contribution of the QSB relevance scoring (see Section 3.1) on the performance of QSBP, we replaced Eq. (3) with:

$$f_{WP}(S) = \sum_{\{w_1,w_2|w_1 \neq w_2 \text{ and } w_1,w_2 \in u_i \text{ and } u_i \in S\}} s_b(w_1) s_b(w_2). \quad (6)$$

We will refer to this as WP. Note that this relies only on base word scores and is query-independent.

[*1] http://research.microsoft.com/en-us/people/tesakai/1click.aspx

Table 3 ACLIA2 test data results.

Method	Precision	Recall	F1 score	F3 score
Baseline	0.076*	0.370*	0.116*	0.231*
Baseline (no generic)	0.080	0.274	0.108	0.186
QSBP	**0.107**$^{\ddagger\dagger\bullet\star\sharp}$	0.482$^{\ddagger\dagger\bullet\star\sharp}$	**0.161**$^{\ddagger\dagger\bullet\star\sharp}$	0.312$^{\ddagger\dagger\bullet\star\sharp}$
QSBP(idf)	0.106$^{\ddagger\dagger\bullet\star\sharp}$	**0.485**$^{\ddagger\dagger\bullet\star\sharp}$	**0.161**$^{\ddagger\dagger\bullet\star\sharp}$	**0.313**$^{\ddagger\dagger\bullet\star\sharp}$
QSBP(nodist)	0.083‡*	0.396*	0.125*	0.248*
QSB	0.086‡*	0.400*	0.129‡*	0.253†*
WP	0.053	0.222	0.080	0.152

Table 4 F3-scores for each question type (ACLIA2 test).

Type	BIO	DEF	REL	EVENT	WHY
Baseline	0.207*	0.251*	0.270	0.212	0.213
QSBP	**0.315****	**0.329**‡*	**0.401**†	0.258$^{\dagger\star\sharp}$	0.275$^{\star\sharp}$
QSBP(idf)	0.304$^{\star\star\sharp}$	0.328†*	0.397†	**0.268**†*	**0.280***
QSBP(nodist)	0.255	0.281*	0.329	0.196	0.212*
QSB	0.245*	0.273*	0.324	0.217	0.215
WP	0.109	0.037	0.235	0.141	0.161

4.4 Results

Tables 3 and **4** summarize our results. We used the two-tailed sign test for testing statistical significance. Significant improvements over the MMR baseline are marked with a † (α=0.05) or a ‡ (α=0.01); those over QSBP(nodist) are marked with a ♯ (α=0.05) or a ♯♯ (α=0.01); and those over QSB are marked with a • (α=0.05) or a ⁚ (α=0.01); and those over WP are marked with a ⋆ (α=0.05) or a ⁎ (α=0.01). From Table 3, it can be observed that both QSBP and QSBP(idf) significantly outperform QSBP(nodist), QSB, WP and the baseline in terms of all evaluation metrics. Thus, the minimum dependency distance, Query Snowball and the use of word pairs all contribute significantly to the performance of QSBP. When we compare the two baselines, it can be observed that the use of the similarity with v_D (Eq. (4)) boosts recall and thereby improves the F3-score. On the other hand, it can also be observed that the recall of the query-independent WP is also low. These results suggest that both generic and query-biased information are useful and complementary.

QSBP and QSBP(idf) achieve 0.312 and 0.313 in F3 score, and the differences between the two are not statistically significant. Table 4 shows the F3 scores for each question type. It can be observed that QSBP is the top performer for BIO, DEF and REL questions on average, while QSBP(idf) is the top performer for EVENT and WHY questions on average. It is possible that different word scoring methods work well for different question types. Recall that we are using the ACLIA data as summarization test collections where candidate documents are already given, and that the official QA results of ACLIA are not directly comparable with ours.

To further investigate the effect of using word pairs rather than words as the basis for selecting novel sentences, we plotted each question from the ACLIA2 test set as shown in **Fig. 4**. Here, the x-axis represents the questions ranked by the number of word overlaps between a nugget pair averaged across all nuggets for a question. Thus, this represents how different nuggets for a question resemble with each other. Whereas, the y-axis represents the same questions ranked by the F3 score difference between QSBP and QSB, i.e., the gain in F3 as a result of using word pairs instead of single words. There is a correlation between the two rankings (0.307 in Kendall's τ, p-value < 0.0001), which suggests that our word-pair based method is effective especially for questions whose nuggets have high word overlaps.

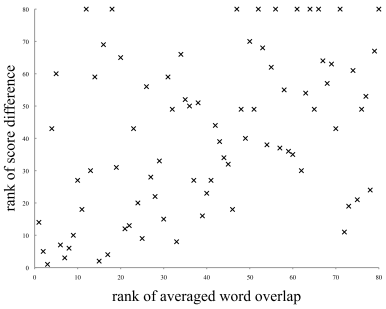

Fig. 4 Word overlap and score difference.

5. Conclusions and Future Work

We first proposed the Query Snowball (QSB) method for query-oriented multi-document summarization. To enrich the information need representation of a given query, QSB obtains words that augment the original query terms from a co-occurrence graph. We then formulated the summarization problem as an MCKP based on word pairs rather than single words, in order to select novel sentences that cover different nuggets. Our combined method, QSBP, achieved a pyramid F3-score of up to 0.313 with the ACLIA2 Japanese test collection, a 36% improvement over a baseline using Maximal Marginal Relevance. An analysis showed each part of our method, Query Snowball and Word pairs, contribute to the improvement and word pairs remedy the problem that answer nuggets have word overlaps.

Moreover, as the principles of QSBP are basically language independent, we will investigate the effectiveness of QSBP in other languages. Also, we plan to extend our approach to abstractive summarization.

References

[1] Bosma, W.: Contextual Salience in Query-based Summarization, *Proc. International Conference RANLP-2009*, Borovets, Bulgaria, pp.39–44, Association for Computational Linguistics (2009).

[2] Carbonell, J. and Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries, *Proc. 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98*, New York, NY, USA, pp.335–336, ACM (1998).

[3] Celikyilmaz, A. and Hakkani-Tur, D.: A hybrid hierarchical model for multi-document summarization, *Proc. 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, Stroudsburg, PA, USA, pp.815–824, Association for Computational Linguistics (2010).

[4] Dang, H.T.: Overview of the tac 2008 opinion question answering and summarization tasks, *Proc. Text Analysis Conference* (2008).

[5] Filatova, E. and Hatzivassiloglou, V.: A formal model for information selection in multi-sentence text extraction, *Proc. 20th international conference on Computational Linguistics, COLING '04*, Stroudsburg, PA, USA, Association for Computational Linguistics (2004).

[6] Hasegawa, T., Nishikawa, H., Imamura, K., Kikui, G. and Okumura, M.: A Web Page Summarization for Mobile Phones, *Trans. Japanese Society for Artificial Intelligence*, Vol.25, pp.133–143 (2010).

[7] Jagadeesh, J., Pingali, P. and Varma, V.: A relevance-based language modeling approach to duc 2005, *Proc. Document Understanding Conferences* (along with *HLT-EMNLP 2005*), Vancouver, Canada, Citeseer (2005).

[8] Khuller, S., Moss, A. and Naor, J.S.: The budgeted maximum coverage problem, *Information Processing Letters*, Vol.70, No.1, pp.39–45 (1999).

[9] Kudo, T. and Matsumoto, Y.: Japanese dependency structure analysis based on support vector machines, *Proc. 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very*

large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Vol.13, pp.18–25, Association for Computational Linguistics (2000).
[10] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proc. Conference on Emprical Methods in Natural Language Processing* (*EMNLP 2004*), Vol.2004 (2004).
[11] Li, W., Ouyang, Y., Hu, Y. and Wei, F.: PolyU at TAC 2008, *Proc. Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing* (*HLT/EMNLP*), Vancouver, BC, Canada (2008).
[12] Lin, H. and Bilmes, J.: Multi-document summarization via budgeted maximization of submodular functions, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, Stroudsburg, PA, USA, pp.912–920, Association for Computational Linguistics (2010).
[13] Lin, H. and Bilmes, J.: A class of submodular functions for document summarization, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — Volume 1, HLT '11*, Stroudsburg, PA, USA, pp.510–520, Association for Computational Linguistics (2011).
[14] Lin, H., Bilmes, J. and Xie, S.: Graph-based submodular selection for extractive summarization, *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009.*, pp.381–386, IEEE (2010).
[15] Lin, J., Madnani, N. and Dorr, B.J.: Putting the user in the loop: interactive Maximal Marginal Relevance for query-focused summarization, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, Stroudsburg, PA, USA, pp.305–308, Association for Computational Linguistics (2010).
[16] Liu, F. and Liu, Y.: From extractive to abstractive meeting summaries: can it be done by sentence compression?, *Proc. ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09*, Stroudsburg, PA, USA, pp.261–264, Association for Computational Linguistics (2009).
[17] Lund, K. and Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods*, Vol.28, pp.203–208 (online), DOI: 10.3758/BF03204766 (1996).
[18] Mani, I.: *Automatic summarization*, John Benjamins Publishing Co. (2001).
[19] McDonald, R.: A study of global inference algorithms in multi-document summarization, *Proc. 29th European conference on IR research, ECIR'07*, Berlin, Heidelberg, pp.557–564, Springer-Verlag (2007).
[20] Mitamura, T., Nyberg, E., Shima, H., Kato, T., Mori, T., Lin, C.-Y., Song, R., Lin, C.-J., Sakai, T., Ji, D. and Kando, N.: Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access, *Proc. 7th NTCIR Workshop* (2008).
[21] Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C.-Y., Song, R., Lin, C.-J. and Lee, C.-W.: Overview of the ntcir-8 aclia tasks: Advanced cross-lingual information access, *Proc. 8th NTCIR Workshop* (2010).
[22] Otterbacher, J., Erkan, G. and Radev, D.R.: Using random walks for question-focused sentence retrieval, *Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, Stroudsburg, PA, USA, pp.915–922, Association for Computational Linguistics (2005).
[23] Radev, D.R.: Experiments in single and multidocument summarization using MEAD, *In First Document Understanding Conference* (2001).
[24] Sakai, T., Kato, M.P. and Song, Y.-I.: Click the search button and be happy: evaluating direct and immediate information access, *Proc. 20th ACM international conference on Information and knowledge management, CIKM '11*, New York, NY, USA, pp.621–630, ACM (2011).
[25] Takamura, H. and Okumura, M.: Text Summarization Model Based on Maximum Coverage Problem and its Variant, *Proc. 12th Conference of the European Chapter of the ACL* (*EACL 2009*), Athens, Greece, pp.781–789, Association for Computational Linguistics (2009).
[26] Takamura, H. and Okumura, M.: Text summarization model based on the budgeted median problem, *Proc. 18th ACM conference on Information and knowledge management, CIKM '09*, New York, NY, USA, pp.1589–1592, ACM (2009).
[27] Varadarajan, R. and Hristidis, V.: A system for query-specific document summarization, *Proc. 15th ACM international conference on Information and knowledge management, CIKM '06*, New York, NY, USA, pp.622–631, ACM (2006).
[28] Voorhees, E.M.: Overview of the TREC 2003 question answering track, *Proc. Twelfth Text REtrieval Conference* (*TREC 2003*), Vol.142, pp.54–68 (2003).
[29] Yih, W., Goodman, J., Vanderwende, L. and Suzuki, H.: Multi-document summarization by maximizing informative content-words, *Proc. 20th international joint conference on Artifical intelligence*, San Francisco, CA, USA, pp.1776–1782, Morgan Kaufmann Publishers Inc. (2007).

Hajime Morita was born in 1984. He received his Master's degree from Tokyo Institute of Technology in 2009. He is a Ph.D. student at the Department of Computational Intelligent and Systems Science in Tokyo Institute of Technology. He has been engaging in the research areas of Natural Language Processing, and Machine Learning.

Tetsuya Sakai received his Master's degree and Ph.D. from Waseda University in 1993 and 2000, respectively. He was a visiting researcher at the University of Cambridge Computer Laboratory from 2000 to 2001. After working for Toshiba Corporation and NewsWatch, Inc., he joined Microsoft Research Asia as a lead researcher in 2009. His academic responsibilities include: NTCIR programme co-chair, Asia Information Retrieval Societies (AIRS) conference steering committee chair and ACM SIGIR 2013 conference chair. He is on the editorial board of Information Processing and Management and that of Information Retrieval, and is an editor-in-chief of IPSJ TOD. He has received several awards, mostly from IPSJ. He is a member of ACM, IEICE and IPSJ.

Manabu Okumura was born in 1962. He received his B.E., M.E. and Dr.Eng. from Tokyo Institute of Technology in 1984, 1986 and 1989, respectively. He was an assistant at the Department of Computer Science, Tokyo Institute of Technology from 1989 to 1992, and an associate professor at the School of Information Science, Japan Advanced Institute of Science and Technology from 1992 to 2000. He was also a visiting associate professor at the Department of Computer Science, University of Toronto from 1997 to 1998. He is currently a professor at Precision and Intelligence Laboratory, Tokyo Institute of Technology. His current research interests include natural language processing, especially automatic text summarization, computer assisted language learning, sentiment analysis, and text data mining.

(Editor in Charge: *Akiko Aizawa*)