論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Enhanced models for query-oriented extractive summarization
著者(和文)	森田一
Author(English)	Hajime Morita
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9659号, 授与年月日:2014年9月25日, 学位の種別:課程博士, 審査員:奥村　学,新田　克己,山田　誠二,本村　陽一,高村　大也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9659号, Conferred date:2014/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（論文博士）

論　文　要　旨（和文2000字程度）

（Summary）

報告番号	乙 第　　号	氏　名	森田　一

（ 要　旨 ）

本論文ではクエリ指向の抽出的複数文書要約のための新しい手法の提案を行った．クエリ指向要約は要約元文書に対して，ユーザの与えるクエリに基づいた要約を行うタスクであり，情報検索や質問応答等，多くの応用先を持つ．クエリ指向の要約モデルには一般的に，クエリとの関連の強さを計算するモデルとその関連度の強さに基づき要約を生成するモデルが含まれる．本論文では要約モデルのその両側面に対して，クエリの表現に対する拡張と要約元文書から内容を抽出する自由度の向上により、それぞれの側面を改善する手法を提案する．まず，クエリの表現方法の拡張として，要約元文書集合上での単語の共起に基づくクエリ関連度を計算する手法である**Query Snowball**の提案を行った．この手法では，クエリを文書中の各単語に対するクエリ関連度として表現することで，ユーザから与えられたクエリ内の語が直接含まれない文についてもクエリとの関連度を計算することが可能となる．本手法では，最初に要約元文書集合中での単語の共起関係をグラフとして表現し，クエリ語との直接的な共起，および，他の単語を経由した間接的な共起を元に，クエリ語と直接共起した語，間接的に共起した語の順にスコアを伝播させていくことによりクエリ関連度を計算する．伝播の際に，どの程度スコアを伝播させるかは２つの語の共起頻度に基づき計算を行う．また，文のクエリ関連度を計算するための手法として，語の組み合わせに対してクエリとの関連度を計算する手法を提案した．語の組み合わせをスコアリングおよび冗長性を計算する単位とすることで，特定のキーワードに関連する複数の話題を要約に含めやすくすることができる．この語の組み合わせに対するクエリ関連度をスコアとし，**Maximum coverage problem with knapsack constraint (MCKP)**として要約の定式化を行った．この手法を評価するため文抽出要約として質問応答用のデータセットである，**ACLIA2 Japanese test collection**上で評価を行い，**Maximal Marginal Relevance**を用いたベースラインに対して**Pyramid F3**スコアで３６％の向上を達成した．評価では単語に対するクエリ関連度と，単語の組み合わせに対するクエリ関連度がどちらも要約の質を向上させることを明らかにした．次に，ユーザの与えるクエリに対して，文の選択と文の圧縮を同時に行い，必要な個所のみを要約として提示するための部分木抽出要約モデルの提案を行った．一般的な文抽出要約では，長い文の一部のみが重要な情報を含み，他の部分は冗長あるいは重要でない場合にも，不要な部分を含めて文を選択する必要がある．しかし，文圧縮を要約を行う前あるいは後に行っても，すべての不要な箇所を選択する文

から取り除くことはできない．圧縮をあらかじめ行ってから要約をする場合には，重要でない情報は要約前に捨てることができるが，複数の文に共通して出現する重要な情報は，要約の際には冗長となるが，あらかじめ圧縮をすることにより除くことはできない．このため，圧縮と選択を同時に行いすべての圧縮された文の候補を考慮し，必要な個所を選択して要約を作る必要がある．依存構造木から部分木を抽出するモデルでは，部分木を抽出することが文の選択と圧縮を同時に行うことに相当し，この文選択と文圧縮を行う順番に関する問題の影響を受けずに要約を行うことが出来る．また，抽出する部分木の重要度を計算するために前述のクエリ関連度を用いているため，クエリが直接含まれなくとも語単位でクエリとの関連性を計算することができ，文の中からユーザにとって必要な一部のみを抽出して要約を生成することが可能となる．この要約モデルを新しいクラスの劣モジュラ最大化問題である **Budgeted monotone non-decreasing submodular function maximization with a cost function** として定式化した．最大化問題を近似的に解くための貪欲アルゴリズムの提案，及び貪欲アルゴリズムの性能について理論的な解析を行い，この近似アルゴリズムが最適解に対して一定の近似率が保証されることを示した．また，実際にすべての部分木の候補を列挙し選択することは，ある文に対応する部分木が非常に多く困難であるため，貪欲アルゴリズムで必要とされる，長さあたりのスコアが最大となる部分木を探索するための動的計画法に基づくアルゴリズムを提案した．劣モジュラ最大化問題としての定式化およびこれらのアルゴリズムにより，提案モデルはユーザのクエリに対して実時間で応答する上で十分に高速に動作することが可能となる．同様に **ACLIA2 Japanese test collection** 上で行った実験により，本手法と同じく劣モジュラ最大化として定式化され，文圧縮を扱う要約を含め，最も良い結果を示していたベースラインの要約手法（**Hui Lin**ら，**2012**）と比較し、**Pyramid　F3**スコア上で９．２％の向上を示した．

（論文博士）

論　文　要　旨（英文）　（300語程度）

報告番号	乙 第　　号	氏　名	森田　一

（要　旨）

We propose new methods for query-oriented extractive multi-document summarization. A query-oriented summarization model consists of calculation of query relevance and generation of a summary based on the relevance. We improve both parts of the summarization model by refinement of query representation and sophistication of compressive extraction. For each model improvement, we propose a model of query relevance based on word co-occurrence in the source documents, and a method for subtree extraction from dependency trees of sentences. 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. The word pairs word pairs remedy the problem that answers for user's information need have word overlaps. Then, we formalized a query-oriented compressive summarization that is a task in which one simultaneously performs sentence compression and extraction, as a new optimization problem: budgeted monotone non-decreasing submodular function maximization with a cost function. We translate the the task into a problem of extracting a set of dependency subtrees in the document cluster. We also encode obligatory case constraints as must-link dependency constraints in order to guarantee the readability of the generated summary. Leveraging dynamic programming based approaches, we developed a densest subtree extraction algorithm in order to extract key subtrees in source sentences. By using the formulation, our model can respond to a user's query efficiently enough for real-time processing. Experiments show that our model outperforms a state-of-the-art baseline method.