ウィキペディア記事に対する文書構造を利用したクエリ依存要約

西川 仁 貞光 九月 宮崎 千明 浅野 久子 牧野 俊朗 松尾 義博
NTT メディアインテリジェンス研究所
{ nishikawa.hitoshi, sadamitsu.kugatsu, miyazaki.chiaki
asano.hisako, makino.toshiro, matsuo.yoshihiro } @lab.ntt.co.jp

あらまし

本稿ではウィキペディア記事を対象とするクエリ依存要約を扱う．特に，専門的なドメインに関する質問応答器の内部装置としてのクエリ依存要約器を考え，比較的長い入力文書の中から，ピンポイントに応答として適切な文を抽出する課題を扱う．ピンポイントに適切な文を抽出するため，本稿ではウィキペディア記事の文書構造に着目する．この構造を利用するため，クエリと文とを細かい特徴量へと分解し，あるクエリに対する適切な応答となる文が，ウィキペディア記事中においてどのような位置に表れやすいか考慮できるようにした．また，ウィキペディア記事に散見される冗長な文については文短縮を用いて短縮した．ある特定の観光地に関する質問応答を想定したデータを用いて評価を行ったところ，上述の工夫により，提案手法はベースラインとなる手法に比べて高精度に適切な文を抽出できることがわかった．

1 はじめに

本稿では，ウィキペディア記事に対するクエリ依存要約を扱う．また，それを質問応答へと応用することを考える．本稿では，ある特定のドメインにおける，専門的知識に関する質問応答を想定する．例として，以下のような質問と応答を考える．

質問	この建物はずいぶん古いね.
応答	この山門は 1785 年に，このお寺の当時のリーダーが再建したものだと言われています.

上の例はある特定の観光地に関する質問応答である．ある特定の観光地に関する情報などは専門的な情報であり，万人が有するものではあるとはいえないため，質問応答の有用なドメインである．そのようなドメインにおける質問応答器の知識源としては様々なものが考えられるが，本稿ではウィキペディアを取り上げる．

上述したように，ウィキペディアの記事の中から，与えられた質問に対する応答を見つけ出すため，本稿では質問応答器の内部装置として，クエリ依存要約を扱う．その際，さらに以下のような仮定を置く．

- 質問に対する応答となる言語表現の抽出先となる文書は要約器に対して与えられるものとする．
- 質問となるクエリはあらかじめ定められたものとする．例えば，特定の観光地に関する,「概要」「名物」「名前の由来」などである．

すなわち，質問応答器における文書の検索およびクエリの解釈は本稿で扱う問題の外とし，1 つの文書とある定められた種類のクエリが要約器に与えられる，クエリ依存要約の問題として定式化する．

このとき，本稿で扱う課題には大きくわけて 2 つの困難がある．

1 つは，質問として与えられるクエリと，また別途与えられる，質問に対する応答を含むと思われる文書から，ピンポイントに応答として適切な文あるいは文集合を抽出するという問題である．本稿で取り上げるクエリ依存要約課題の要約率は約 1.8%と非常に低く，自動要約課題として難しい．

もう 1 つは，質問応答器の内部装置として，短く端的な応答を作成するという課題である．ウィキペディアの記事を構成する文は長いものが多いため，そのままでは質問応答器の出力として不適切なものがままあり，これを何らかの方法で書き換える必要がある．

本稿では，前者の問題に対しては，クエリと文をより細かい特徴量に分解し，またウィキペディアの記事が持つ，文書構造に関する豊富な情報を利用して，ピンポイントに適切な文を抽出することを試みる．後者の問題に対しては，文短縮を用い，長い文を短く書き換える．

以降，2 節ではクエリに応じて文の重要度を変化させる具体的な方法を述べる．3 節でては提案する手法を評価するための実験の設定について述べる．4 節は実験の結果について述べ，またそれを考察する．5 節では本稿をまとめる．

2 クエリ依存要約モデル

本稿では，要約器として西川らによる要約器 [4] を用いる．ただし，西川の要約器はクエリを考慮することができないため，これをクエリを考慮できるように拡張する．

2.1 クエリに応じた文の重要度の計算

基本的な考え方は以下の通りである：

1. クエリ q の特徴量の集合 $f(q)$ を考える．
2. 文 s の特徴量の集合 $g(s)$ を考える．
3. 集合 $f(q)$ と 集合 $g(s)$ の直積集合を考え，それを文 s の特徴量の集合とする．

例えば，クエリ q の特徴量として文字ユニグラムを考え，文 s の特徴量として内容語を考えると，クエリ q の文字ユニグラムのそれぞれと文 s の内容語のそれぞれの組み合わせの全てが文 s の特徴量となる．この特徴量による特徴ベクトルと，予め学習したパラメタから文 s の重要度 w を定める．

2.2 特徴量

2.2.1 クエリの特徴量

以下の特徴量を用いてクエリを抽象化する．

- **表記** クエリそのものの表記を特徴量として用いる．
- **単語** クエリを構成する単語を特徴量として用いる．クエリが 1 語であれば，この特徴量は上の**表記**と同じものとなる．
- **文字ユニグラム** クエリを構成する文字それぞれを特徴量として用いる．

クエリが複数の単語からなる場合は，その中の内容語それぞれについて上述の特徴量を抽出し，全てを足し合わせたものをクエリの特徴量とする．

	入力文書	参照要約
平均文字数	7251.9	127.7
平均単語数	3313.6	49.7
平均文数	146.2	2.3

表 1: コーパスの統計量．

2.2.2 文の特徴量

西川らの要約器 [4] が用いる特徴量とは別に，以下の特徴量を追加した．

- **文が含まれている節の見出し** ウィキペディアの記事では，多くの場合，節に見出しがついている．例えば，ある寺社の歴史に関する記述を含む節には，「歴史」という見出しがついていることが多い．これらの見出しに含まれる表記，単語および文字ユニグラムを特徴量として用いた．
- **節の内部での位置** ある節の内部における文番号および段落番号を特徴量として用いた．

3 実験

3.1 データ

実験のため，鎌倉市内の，寺社仏閣をはじめとする観光地に関する質問と，それに対する応答，および応答の元となったウィキペディア記事の組を 151 組用意した．実験においては，質問をクエリ，応答を参照要約，応答の元のなったウィキペディア記事を入力文書として，クエリ依存要約課題として評価を行う．質問は 11 種類のクエリのうちのいずれかとして表現される．例を以下に示す．

クエリ	名前の由来
参照要約	鎌倉を代表する 5 つの氏族の霊を祀った神社であることから，御霊神社と呼ばれるようになりました．

クエリ	名物
参照要約	6 月から 7 月の梅雨の時期にかけてはあじさいが咲き誇り観光客の目を楽しませます．

これら応答の長さはまちまちであるため，要約器を動作させる際には応答と同じ長さを要約長として与えた．コーパスの統計量を表 1 に示す．文字数に基づく要約率は約 1.8%であり，この値は TSC-3 の複数文書要約課題より低い．本課題が単一文書要約であるにもかかわらずこのような低い要約率となる理由は 2 つある．1 つは入力文書となるウィキペディア記事が総じて長いため，もう 1 つは応答となる参照要約が短いためである．そのため，ピンポイントで応答となる部分を特定する必要が生じる．

3.2 評価尺度

要約の内容性の評価には ROUGE-1 [1] を用いた．評価に際しては平尾らの知見 [6] に従い，内容語 [1] のみを利用した．

3.3 比較手法

我々は以下の 5 手法を比較した：

- **RANDOM** 与えられた要約長を満たす範囲でランダムに文を選択．
- **tf-idf** tf-idf で文の重要度を計算し，文を選択する．idf は収集したウィキペディア記事から求めた．
- **tf-idf + Query** tf-idf で文の重要度を計算するが，クエリに含まれる語の重みは 2 乗される．これによってクエリを含む文が選ばれやすくなることを期待した．
- **Proposed w/o Compression** 提案手法．文短縮は用いない．
- **Proposed w/ Compression** 提案手法．文短縮も用いる．

提案手法が用いるパラメタの推定の際には 5 分割交差検定を実施した．パラメタの推定の方法については西川らによる方法 [4] に従った．

4 結果と考察

結果を表 2 に示す．

まず RANDOM の結果をみると，著しく悪い値を示している．要約率を考えると，ランダムに文を選択しても正しい文が選択される確率は低い．したがって参照要約に含まれる単語が要約に含まれる確率も低く，これが RANDOM が低い値を示す理由である．

Method	ROUGE-1
RANDOM	0.030
tf-idf	0.041
tf-idf + Query	0.103
Proposed w/o Comp.	0.653
Proposed w/ Comp.	0.716

表 2: ROUGE による評価の結果．

次に tf-idf の結果をみると，これも同様に低い値を示している．今回の課題はクエリ依存要約であるが，この手法はクエリを考慮せず，文書集合中において希少で，入力文書中において頻出する単語を多く含む文を抽出するものであり，このような手法では芳しい結果は期待できない．

tf-idf に基づくが，クエリとなっている単語の重要度を高めた方法をみると，いくらか値は改善されたものの，依然として芳しい結果ではない．これは，参照要約の多くはクエリとなっている単語を含んでおらず，単にクエリとなっている単語の重要度を高めても参照要約に含まれる文を抽出できないためである．例えば，「概要」というクエリに対する参照要約は全て「概要」という単語を含んでおらず，従ってクエリとなっている単語の重みを高めたとしても無意味である．

これらのベースラインに比べ，提案手法は大きな改善を示した．本稿で提案する手法が大きな改善を示した理由は 2 つあると考えられる．まず，今回対象とした文書は全てウィキペディア記事であり，さらに全て鎌倉市内の観光地に関する記事であるから，文書の構造は似通っている．そのため，節の見出しなどウィキペディア記事固有の情報を有効に利用できたということが考えられる．もう 1 つの理由として，今回の用いたデータにおいてはクエリの種類が多くなく，したがって汎化が容易であったと考えられる．クエリが多種にわたればそれだけ多くのデータが必要となるが，今回はクエリの種類が少ないため，少数のデータでも，ウィキペディアの記事中の適切な部位を特定するように学習ができたものと思われる．

最後の，文短縮を用いることで更に若干の精度の向上が見られた．これはウィキペディアに頻出する長い文を文短縮によって短縮することによって，そのままでは選択できない文が選択できるようになったためである．

[1] 名詞，動詞，形容詞および未知語．

5 関連研究

本稿で提案した手法をクエリ依存要約としてみた場合，特徴は 2 つある．1 つは明示的に文書の構造を特徴量として利用する点であり，もう 1 つは文短縮を利用する点である．

段落などの情報が自動要約に重要であることは以前から知られている [5]．本稿では特に日本語版ウィキペディアの記事を要約の対象として利用したため，記事に含まれる，節などの情報を利用することができた．

文短縮を用いるクエリ依存要約としては長谷川らによる方法 [7] や Morita らによる方法 [2, 3] がある．これらはクエリとして与えられた語と共起しやすい語の重要度を高めることでクエリ依存要約を実現している．本稿では，共起を，語より細かい特徴量として扱うことで，そのようなクエリに対する応答として適切な文が保持する性質をあらかじめ学習しており，この点で前述の方法とは異なる．

6 おわりに

本稿では，特定の専門的知識に関する質問応答器の内部装置としてのクエリ依存要約課題を扱った．ピンポイントで入力文書の中から端的な応答を得るという課題に対して，豊富な特徴量を利用し，また文短縮技術を用いることで対処した．

今後の課題として，文短縮に限らず，ウィキペディアの記事を構成する文をより口語的に書き換えることを検討している．今回対象とした鎌倉市内の観光に関するウィキペディアの記事の中には，仏教用語などをはじめとして難解な単語が多数含まれており，これらの言い換えは重要な課題である．特に，質問応答器の出力を音声合成器に与え，音声にて応答を出力する場合を考えると，これは重要な課題である．

参考文献

[1] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop Text Summarization Branches Out*, pp. 74–81, 2004.

[2] Hajime Morita, Tetsuya Sakai, and Manabu Okumura. Query snowball: A co-occurrence-based approach to multi-document summarization for question answering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 223–229, 2011.

[3] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Subtree extractive summarization via submodular maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1023–1032, 2013.

[4] Hitoshi Nishikawa, Kazuho Arita, Katsumi Tanaka, Tsutomu Hirao, Toshiro Makino, and Yoshihiro Matsuo. Learning to generate coherent summary with discriminative hidden semi-markov model. In *Proceedings of the 25th International Conference on Computational Linguistics (Coling)*, pp. 1648–1659, 2014.

[5] 奥村学, 難波英嗣. テキスト自動要約. オーム社, 2005.

[6] 平尾努, 奥村学, 磯崎秀樹. 拡張ストリングカーネルを用いた要約システム自動評価法. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1753–1766, 2006.

[7] 長谷川隆明, 西川仁, 今村賢治, 菊井玄一郎, 奥村学. 携帯端末のための web ページからの概要文生成. 人工知能学会論文誌, Vol. 25, No. 1, pp. 133–143, 2010.