1V-3

リンク予測と文章構造を考慮した複数文書からの要約システム

中田真吾† 横山晶一†
山形大学大学院理工学研究科†

1. はじめに

WWW 上では多くの情報が存在し、短時間で必要な情報のみを得る技術が求められている。その方法として、複数文書要約の研究が行われており、重要文抽出の方法や要約の読みやすさなどが重要視されている。

本研究では複数文書要約において、より人が読みやすい文章を保ちつつ、密度の濃い要約を作成するシステムの開発を目標とする。

2. 研究背景

複数文書要約に関連する研究は非常に多く行われている。

Wan ら[1]は文間類似度を利用したリンク解析により、複数文書要約を行った。その際、話題を考慮したクラスタリングにより要約の精度を向上させた。Lapata[2]は複数文書要約において文章を並べ替える手法として、コーパスに出現する内容語を素性とし、連接頻度を用いて 2 文の連接確率を求めて文を順序付けている。

近年の研究では、全工程を統計的な手法のみで処理するものが多い。しかしながら、文章を要約して出力する際、文や文書の構造を考慮しないため、文章のつながりが悪くなり、読みにくい文書になる問題がある。

本研究では、複数文書要約における重要文抽出の際、文章の構造を保持したまま非重要文を除去することで、人が読みやすい要約を作成する方法を提案する。

3. 研究方法

本研究では複数文書要約において、なるべく文章としての形を崩さずに重要な文を抽出し、冗長な文章になることを避けるために類似文(被覆文)を除去することで、可読性の高い要約を作成することを目標とする。文章構造を保持したまま処理をすることにより、従来の手法と比べて人が読みやすい構造化された要約の作成が期待される。

方法は次に示すとおりである。

1:同じ話題の文集合からベース文書を決定
2:重要度の計算
3:被覆する文の判定、除去
4:非重要文の除去
5:被覆しない文の並べ替え
6:ベース文書の変更
7:要約文書の決定

この手法では、元の文書構造をそのまま利用することで、人が読みやすい要約を作成できるのではないか、という仮説を立てて、その検証を行った。なお、対象とする文書はニュース記事とするため、元の文章は適切な形で構造化されていることが保障される。

3.1 重要度の計算

本研究では Wan ら[1]の手法と同様に要約を文をノードとするリンク予測問題とみなし、文間のリンクから重要文抽出を行う。その際、先行研究ではコサイン類似度を用いた Markov Random Walk(MRW)により重要度を算出したが、本研究では文の連接確率を素性とする MRW により重要度を計算する。

$$p(a_i \rightarrow b_j) = \frac{a_i\text{が現れた次の文に}b_j\text{が現れる数}}{b_j\text{が現れる数}} \quad (1)$$

(1)に文の連接確率の計算式を示す。(1)中の $p(a_i \rightarrow b_j)$は i 文目に出現する単語a_iと j 文目に出現する単語b_jの連接確率である。分子はコーパス中で 2 つの単語が連続する 2 文の中に出てくる回数で、分母は後ろの単語がコーパス中で出てくる回数である。今回対象とする単語は、文間の繋がりに強く関係する品詞として動詞、およびサ変名詞を用いた。なお、(1)は Lapata[2]のものを参考にした。

Multi Document Summarization System using Link Prediction and Structure among Sentences
†Singo Nakada, Shoichi Yokoyama
†Yamagata University

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

3.2 被覆率の計算

本研究では被覆する文を判断する方法として、含意関係認識を用いる。含意関係認識とは、ある 2 文に対し、片方の文(仮説文)がもう片方の文(入力文)を含意しているかどうかを判定するものである。

本研究では仮説文をベース文、入力文を非ベース文とすることで、非ベース文がベース文を含意しているかどうかを判定する。含意関係を判定する方法として、n-gram を用いた手法[3]を利用する。n-gram を利用した含意関係の方法を(2)に示す。

$$含意する = \frac{a}{h} \geq t \qquad (2)$$

(2)中の a は入力テキストと仮説の n-gram が一致する数であり、h は仮説の n-gram 総数、t は閾値(0≦t≦1)である。また、閾値を満たさない場合は「含意しない」と判定される。

本研究では自立語の bi-gram と文節の uni-gram のどちらかが、閾値 0.3 以上になったものを「含意する」と判定した。その際、片方の含意率が 0 だったのものは「含意しない」とした。上記の組み合わせは、n-gram 数や閾値を変更して複数回実験を行った際、平均して良い結果であったために今回用いた。

4.実験

前述の手法により得られた重要度と被覆率を用いて、複数文書要約を行った。対象としたデータは NTCIR3[4]で設定された自由作成要約タスクに該当する文書である。この文書はキーワード毎に Topic を作成し、その Topic に該当する 98 年から 99 年の間に刊行された毎日新聞記事データを集めた文書集合である。

得られた要約結果の評価方法として、読みやすさの評価を重視するために、3 人の学生にアンケートを実施した。表 1 にその結果を示す。

「提案」は、今回提案した手法により作成された要約、「ベース」は tf-idf 値が最も高い文を順番に抜き出して作成された要約、「正解」は NTCIR のタスク「NTCIR-3 サブタスク TSC-2」を参考に、NTCIR3 で提供されている文書データを人手により文抽出して作成した要約を表している。それぞれの要約を見せ、「要約として読みやすい」を「良」、「要約として読みづらい」を「悪」、「どちらとも言えない」を「無し」として、最も多い回答がその要約に対する読みやすさとして判断した。なお、3 人とも違う回答であった場合は「不明」とした。

表 1.アンケートによる評価結果

	良	悪	無し	不明
提案	3	7	0	0
ベース	5	3	1	1
正解	6	1	0	3

アンケートの結果、提案手法により「良」と判断された要約の数は 3 と他の手法に比べて悪い結果であった。

提案手法による要約が読みやすいと判断された 3 つの要約は、全てベースとなった文書中の文を多く含んでおり、文章構造が保持されていた。逆に提案手法による要約が読みづらいとなったものは、どれもベース文書の原型をとどめておらず、文章構造が上手く保持できなかったことが読みづらさの原因に考えられる。

5.考察

実験の結果、今回提案した手法では汎用的で読みやすい要約システムを作成することはできなかった。しかしながら、文章の構造を保持したまま要約を作成することができれば、読みやすい文章を作成することが可能であることが確認できた。

このため Entity-Grid や格フレームなど、より文章の構造を考慮した手法を取り入れることで、読みやすい要約の作成が期待できる。具体的な方法として、入力文に対して格解析を行い、コーパスを利用してその格構造の文の次に来易い各構造を探し出し、入力文の後に付属させる、といった方法が考えられる。

参考文献

[1] X.Wan, J.Yang: Multi-Document Summarization using Cluster-based Link Analysis, In Proc. of the 31st ACM SIGIR, pp.299-306, 2008

[2] M.Lapata. Probabilistic text structuring Experiments with Sentence Ordering. In Proc, of the 41st Meeting of the Association for Computational Linguistics. PP545-552(2003)

[3] 宇高邦弘, 山本和英：n-gram を用いた日本語テキスト含意関係の検討. 言語処理学会第 16 回年次大会発表論文集. Pp462-465(2010)

[4] NTCIR3,http://research.nii.ac.jp/ntcir/ntcir-ws3/work-ja.html (アクセス時 2012.01.10)

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.