ARG WI2 No.9, 2016
WI2-2016-26

単語の分散表現及び tf-idf 法を用いた自動要約システム

原田 大地　　荒木 健治

北海道大学大学院情報科学研究科

dharada@eis.hokudai.ac.jp　　*araki@ist.hokudai.ac.jp*

概要　本手法では，tf-idf 値と単語の分散表現の類似度を用いてスコアを計算することにより文の重要度を決定し，要約を行う．具体的には，tf-idf 値により重要と判定された単語に加えて，この重要語と類似した単語もまた重要語とみなし，重要語の tf-idf 値と分散表現のコサイン類似度との積を重要語のスコアとする．次に，文中に含まれる重要語のスコアの総和を文長により正規化することにより得られた文の重要度スコアを用いて重要文抽出による単一文書要約を行う．本発表では，本手法に基づく実験システムを作成し，日本語のニュース記事を対象として評価実験および考察を行った結果について述べる．

キーワード　文書要約, tf-idf, Word2Vec

1 はじめに

文書自動要約の代表的な手法として，文や文節などの言語単位を選択，抽出し，要約を生成する手法がある．文選択による要約手法では，各文に何らかのスコアを付与し，スコアの高い順に選択し要約するものが多い．従来のスコア付与の方法には文書内の単語の tf-idf 値など表層的な情報を用いるものがある[1]．本稿では，tf-idf に加えて Word2Vec[2]による単語の分散表現の類似度を組み合わせてスコアを付与する重要文抽出手法を提案する．

日本語文書を対象とした重要文抽出による要約において分散表現を用いた手法として，単一文書要約では野口らの研究[3]，複数文書要約では住田らの研究[4]がある．野口らの研究では，含まれる単語の分散表現の和から文書と文の分散表現を作成し，分散表現のコサイン類似度をスコアとしている．住田らの研究では文書間関連度を最大化する整数線形計画問題で，分散表現のベクトルの近さを文書間関連度とするものである．本研究は tf-idf 値と分散表現を組み合わせてスコアを算出している点で野口らの研究，住田らの研究とは異なっている．

Word2Vec によって得た単語の分散表現を用いて，tf-idf によって決定した重要語に加え，重要語に類似する単語が含まれるかどうかも文のスコアに反映させることにより，表層的な重要語を含む文だけではなく，重要語と関連する単語を考慮に入れ，より要約にふさわしい文を抽出することができると考えられる．本稿では，手法の概要を述べ，この仮説に基づき仮説の妥当性を検証するために行った実験結果について述べる．

2 処理過程

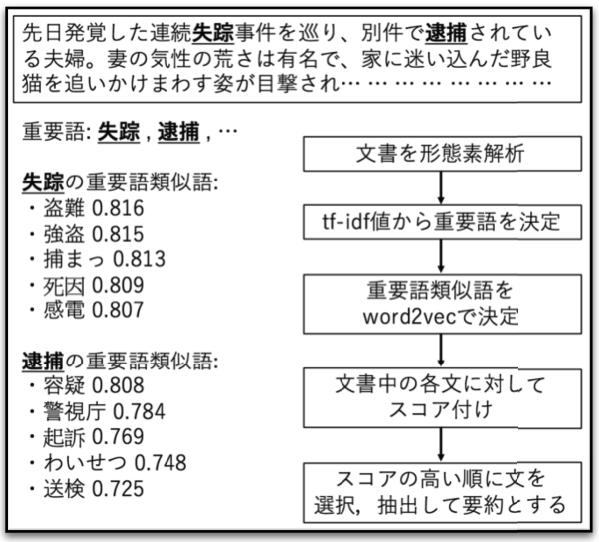

図 1 処理過程および重要語，重要語類似語の例

要約対象文書を Mecab によって形態素解析する．記事中に出現する単語すべての tf-idf 値を算出し，値の大きい順に上位 10 単語を選択し重要語とする．

本稿では重要語と文脈すなわち周辺単語の出現状況が似ている単語を重要語類似語と定義し，Word2Vec で作成した分散表現からコサイン類似度を用いて重要語類似語を決定する．tf-idf 値によって決定した重要語ごとに，分散表現のコサイン類似度が 0.6 以上である単語を値の高い順に最大 20 個取得し，重要語類似語とする．

次に，以下の方法で文にスコア付与を行う．

文sに含まれる重要語wの tf-idf 値の総和を式(1)に示す.

$$Score_1(s) = \sum_i tf\text{-}idf(w_i) \quad (1)$$

文sに含まれる重要語wの tf-idf 値と, 重要語の分散表現$vec(w_i)$と重要語類似語の分散表現$vec(w'_j)$のコサイン類似度の積の総和をとり, 式(2)に示す$Score_2(s)$とする.

$$Score_2(s) = \sum_i \sum_j tf\text{-}idf(w_i) \cdot \mathrm{Cos}\left(vec(w_i), vec(w'_j)\right) \quad (2)$$

$Score_1(s)$と$Score_2(s)$の和を, 文sの形態素数$len(s)$で割ることによって正規化し, 提案手法で用いる文sのスコア$Score_A(s)$を式(3)に示す.

$$Score_A(s) = \frac{Score_1(s)+Score_2(s)}{len(s)} \quad (3)$$

3 評価実験

livedoor NEWS [5] から取得した国内-社会カテゴリに属するニュース 38,677 記事を実験に使用した.

38,677 記事のうち 10,036 記事には人手で作成された要約データが付与されており, 残りの 28,641 記事には要約データは付与されていない. このため, 要約無しの 27,641 記事を学習セットとし, Word2Vec で学習させて単語の分散表現を作成した. また, 要約データ付きの 10,036 記事をテストセットとし, ランダムに 200 記事を抽出した. この 200 記事に対して(3)式の$Score_A(s)$およびベースラインとして tf-idf 値のみを利用した式(4)に示す$Score_B(s)$を用い, それぞれスコアの高い順に 4 文を選択する重要文抽出により要約を作成した.

$$Score_B(s) = \frac{Score_1(s)}{len(s)} \quad (4)$$

ベースラインで作成した要約を TFIDF, 提案手法による要約を TFIDF*W2V, 人手で作成された要約を HUMAN と呼ぶことにする.

要約を作成した結果, TFIDF と TDIDF*W2V の要約結果が異なるものは 200 記事中 90 記事となった.

アンケートによる評価実験を行った. 被験者にはどの手法で作成された要約かを教えずに, TFIDF, TFIDF*W2V, HUMAN の要約記事を表 1 に示す 4 項目について 1 点(悪い)〜5 点(良い)の 5 段階で評価した. この評価項目は livedoor NEWS の記事を自動要約する田中らの研究 [6] で用いられた評価項目を使用しており, DUC における QualityQuestion [7] を参考に作成されている.

被験者は 20 代の理系高専生 A, 理系大学生 B, および第一著者の 3 名である. 被験者はそれぞれランダムに選択された重複しない 20 記事について評価を行った. これにより, 合わせて 60 記事の要約の評価を得た.

表 1 評価実験における評価項目

文法性	(日本語として正しい要約記事か)
冗長性の少なさ	(同じような内容を何度も繰り返していないか)
指示内容の明確さ	(要約文中の名詞および代名詞が何を指しているのかはっきりと分かるか)
総合的な要約の良さ	(総合的な要約記事の良さ)

表 2 に 60 記事の評価値の平均を示す. 冗長性以外の項目において, 提案手法はベースラインを超える結果となった.

表 2 評価実験の結果

	文法	冗長	指示	総合
TFIDF	3.05	3.52	3.35	3.07
TFIDF*W2V	3.12	3.47	3.42	3.23
HUMAN	4.12	4.37	4.25	4.17

TFIDF の結果と TFIDF*W2V の結果を比較すると, 選択文の変更が 113 文存在した. 提案手法における Word2Vec 使用の効果を確認するために, 提案手法において選択されなかった文, 新たに選択された文それぞれに, $Score_2(s)$の項の$Score_A(s)$への寄与率を調べた. 選択されなかった文の平均寄与率は 14.9%であり, 新たに選択された文では 50.9%であった.

4 おわりに

Word2Vec を用いて重要語の類似語もスコアに反映させたことで選択された文が, 要約の評価を高めたと考えられる. これにより, 提案手法がより良い要約を作成するために有効であると示された.

参考文献

[1] Zechner, K.: Fast generation of abstracts from general domain text corpora by extracting relevant sentences, Proc. of the 16th conference on Computational linguistics, Volume 2, pp. 986-989, 1996.

[2] Thomas Mikolov.: word2vec: Tool for computing continuous distributed representations of words, https://code.google.com/p/word2vec/, 2013.

[3] 野口正樹, 谷塚太一, 小林隼人:分散表現を用いたヤフー知恵袋の要約, 言語処理学会第 21 回年次大会発表論文集, pp.1084-1087, 2015.

[4] 住田恭平, 二宮崇:言語概念ベクトルを用いた文書間類似度に基づく複数文書自動要約, 言語処理学会第 22 回年次大会発表論文集, pp.513-516, 2016.

[5] livedoor NEWS, http://news.livedoor.com/.

[6] 田中駿, 笹野遼平, 高村大也, 奥村学:要約長,文長,文数制約付きニュース記事要約, 言語処理学会第 22 回年次大会発表論文集, pp.342-345, 2016.

[7] DUC quality questions, http://duc.nist.gov/duc2007/quality-questions.txt.