統計的手法に基づくWebページからのヘッドライン生成

廣嶋 伸章　長谷川 隆明　山崎 毅文
日本電信電話株式会社 NTT サイバースペース研究所
{hiroshima.nobuaki, hasegawa.takaaki, yamazaki.takefumi}@lab.ntt.co.jp

現状の検索エンジンが出力する文書リストの概要文は Web ページの先頭数十文字などであるため内容が把握できず、必要な情報に効率よくアクセスできない。これを解決するためには、概要文の代わりに Web ページの内容を簡潔に表したヘッドラインを提示すればよい。そこで本研究は、Web ページからその内容を簡潔に表したヘッドラインを自動生成することを目的とする。ヘッドラインは「(1) 内容網羅性」、「(2) 可読性」、「(3) 高圧縮性」の 3 条件を満たす必要があるが、従来のテキスト要約技術ではこれらの 3 つの条件を同時に満たすことができない。本研究では、2 値分類の機械学習手法である Support Vector Machine(SVM) を用いて、単語がヘッドラインとして必要か不要かに分類することにより重要語の選択を行い、単語 trigram モデルと単語の重要度を組み合わせた Noisy channel model を用いてヘッドライン生成を行う方法を提案する。Web ページを用いた評価実験の結果、提案した重要語選択モデルは TF·IDF モデルより優れていることを検証し、これを用いたヘッドライン生成実験において TF·IDF モデルに基づくベースライン手法よりテキスト全体の内容をより的確に表せることを検証した。

Headline Generation from Web Pages Based on Statistical Method

Nobuaki HIROSHIMA　Takaaki HASEGAWA　Takefumi YAMAZAKI
NTT Cyber Space Laboratories, NTT Corporation
{hiroshima.nobuaki, hasegawa.takaaki, yamazaki.takefumi}@lab.ntt.co.jp

The purpose of this study is to generate a headlines automatically from a given web page. We can define a headline as the sentence which is brief, easy to read and highly compressed. Our method consists of two steps: word selection and headline generation. For word selection, we classify each word into two categories by appling Support Vector Machine(SVM). For headline generation, we use a noisy channel model which is the combination of a word trigram model and a word importance model. The experimental result shows that the generated headline by our method explains the original contents more precisely than that in the baseline.

1　はじめに

近年、インターネットが普及し、Web ページが爆発的に増加しつつある。その大量の Web ページから情報を効率的に収集するための手段としては、検索エンジンを用いるのが一般的である。しかし、現状の検索エンジンでは、検索結果の概要文が Web ページの先頭数十文字であったりキーワード周辺の文の寄せ集めであったりするため、内容が検索要求に適合しているかどうかの判断が困難であり、情報の取捨選択のガイドラインとして十分な役割を果たしているとはいえない。

そのため、Web ページの内容を簡潔に表したヘッドラインを検索結果とともに提示することが望まれる。これにより、そのページがユーザにとって必要な情報かどうかを知ることができ、効率良く情報の取捨選択ができるようになる。しかし、Web ページは膨大な量であるだけでなく日々更新されるため、人手によってヘッドラインを作成するのは現実的でない。また、Web ページは人によって内容の属する分野や書き方のスタイルが異なるため、ヘッドラインを生成するためのルールを記述するのも困難である。

そこで本研究では、統計的手法によって Web ページからヘッドラインを自動生成する手法を提案する。

2　背景

本章では、まず本研究におけるヘッドラインの定義について述べる。次にヘッドライン生成というタスクにおける従来手法の問題点について述べる。

2.1　本研究におけるヘッドラインの定義

ヘッドライン生成は極めて高い圧縮率を必要とするテキスト要約技術であると位置付けることができる。したがって、本研究においては、以下の条件を満たす文をヘッドラインと定義する。

(1) 内容網羅性
　テキスト全体の内容を適切に表している文

(2) 可読性
　単語の羅列ではなく、読みやすい文

(3) 高圧縮性
　十数単語からなる簡潔な文

このうち、(1) および (2) は要約として満たすべき条件である。ヘッドラインであるための付加条件として (3) を定義する。

2.2　従来手法における問題点

2.1 節で述べたように、ヘッドラインはこれらの 3 つの条件を満たす必要があるが、従来手法をヘッドライン生成に対して適用しても、これらの 3 つの条件を満たすことができない。

重要文抽出 [4] をヘッドライン生成に適用すると、「(3) 高圧縮性」を満たすために抽出する文を 1 文または 2 文程度に抑える必要があるが、ヘッドラインに必要な情報は複数の文にまたがって存在することもあるため、「(1) 内容網羅性」を満たせない。

また、文圧縮手法 [7] では圧縮できる長さに限界があり、ヘッドライン生成に適用した場合「(3) 高圧縮性」を満たせない。

一方、TF·IDF モデルなどによりキーワードを求め、抽出された複数のキーワードを抽出して並べたとしてもそれは単語の羅列にすぎず、「(2) 可読性」という点を考えるとヘッドラインとはいえない。

堀ら [6] は、単語の重要度と言語尤度を用いた要約文生成によりニュース音声を要約する手法を提案している。ニュース文 1 文に含まれる N 単語からなる単語列から要約文として $M(M < N)$ 個の単語を抽出し接合した単語列 $V = v_1, v_2, \cdots, v_M$ の要約スコアを次式のように定義している。

$$S(V) = \sum_{m=1}^{M} \{\log P(v_m|v_{m-2}v_{m-1}) + \lambda I(v_m)\}$$

ここで言語スコア P には単語 trigram が、単語重要度スコア I には重み付き TF·IDF 尺度が用いられている。重要語を選択することにより「(1) 内容網羅性」を満たすことができ、単語間のつながりを考慮するため「(2) 可読性」を満たすことができる。さらに、要約文の単語数を任意に設定できるため、「(3) 高圧縮性」をも満たすことが可能であり、ヘッドライン生成に適した手法であるといえる。しかし、人間が重要語を選択する際には頻度以外の素性も考慮すると思われるため、TF·IDF 尺度だけでは重要語の選択に用いる素性としては不十分と考えられる。

3 統計的手法に基づくヘッドライン生成

本研究では、ヘッドラインとテキストの対から統計的に重要語を選択し、その重要語を考慮するとともに、単語間のつながりを単語 trigram モデルで考慮して文生成を行う手法を提案する。要約文生成における先行研究では、重要語の選択に用いられていた素性が不十分であった。そこで本研究では、人間が重要語選択するプロセスを模倣するような重要語選択モデルを構築し、機械学習を用いて統計的に重要語を選択する。

人間がヘッドラインを作成する際、何らかの素性によって重要箇所を抽出し、それをつなぎ合わせてヘッドラインを作成すると考えられるが、重要箇所の判定に人間がどのような素性を用いているかはテキストの分野などによって異なると考えられる。そこで、本研究では、様々な分野のテキストとヘッドラインの対を収集し、それらを用いてテキストから重要語を選択する統計モデルを学習する。これにより、どのような分野のテキストに対してもロバストに重要語を抽出することができる。以下では、重要語選択および文生成におけるモデルについて説明する。

3.1 重要語選択モデル

テキスト中の単語がヘッドラインに必要な単語かどうかは、複数の素性によって決まると考えられる。これらの素性が有効かどうかはテキストの種類などによって異なるため、素性に対して人手で重み付けを行うのは現実的でない。

そのため、複数の素性の組み合わせから効率良くモデルを学習するための機械学習手法が必要となるが、重要語を選択するという問題は、テキスト中の単語がヘッドラインとして”必要”か”不要”かという 2 値分類の問題に置き換えることができる。そこで、本研究では重要語の選択において 2 値分類の機械学習手法を適用する。2 値分類の機械学習手法としては SVM(Support Vector Machine)[3] を用いる。

本研究では、SVM を用いて、以下のようにして重要語選択モデルを構築する。

- 学習フェーズ
  - コーパス中のヘッドラインに含まれる内容語をテキスト中の内容語と対応付け
  - 対応付けの結果をもとに、テキスト中の内容語を”必要”、”不要”の 2 つのクラスに分類
  - 分類済みの各内容語に対して素性の値を求め、素性ベクトルを作成 (利用した素性は表 1 のとおり)
  - SVM を用いて、素性ベクトルに対しその単語がヘッドラインとして必要か不要かを判定する 2 値分類器を学習
- 重要語選択フェーズ
  - テキストから素性ベクトルを作成
  - 分類器から正負の尤度を算出

素性ベクトルの作成に用いた素性は、単語自体とその前後 2 単語の計 5 単語に関する文書内頻度 · 文書間頻度 (素性数 $2 \times 5 = 10$ 種類)、文中 · 文書中での位置 ($2 \times 5 = 10$ 種類)、頻出単語かどうか ($1000 \times 5 = 5000$ 種類)、主品詞 · 副品詞 ($373 \times 5 = 1865$ 種類)、意味的カテゴリ ($1715 \times 5 = 8575$ 種類)、および単語が属する文内に特定の頻出単語が含まれるか (1000 種類)、文内に特定の主品詞 · 副品詞が含まれるか (373 種類)、文内に特定の意味的カテゴリが含まれるか (1715 種類) の計 18548 種類である。また、分類の結果分類器が出力する値は素性ベクトルと分離超平面との距離を表す正負の尤度であるが、その絶対値が大きいほど必要／不要である可能性は高くなると考えられるため、尤度を正規化した値を単語の重要度として利用する。

3.2 要約文生成モデル

本研究では、「(1) 内容網羅性」「(2) 可読性」「(3) 高圧縮性」の 3 条件を満たすヘッドラインを生成可能な要約文生成モデルとして noisy channel model を用いる。noisy channel model は、以下の式で表されるモデルである。

$$\begin{aligned} s &= \arg\max_s P(s|t) \\ &= \arg\max_s P(s)P(t|s) \end{aligned}$$

$P(s)$ を source model、$P(t|s)$ を channel model とよぶ。source model はヘッドラインの文らしさを示すモデルであり、channel model はヘッドラインがもとのテキストの内容をよく表しているかどう

かを示すモデルである。ここで、source model に「(2) 可読性」を表すモデルとして単語 trigram モデル、channel model に「(1) 内容網羅性」を表すモデルとして提案した重要語選択モデルを適用すると、最尤のヘッドライン s^* は、

$$s^* = \arg \max_{w_1,\cdots,w_N} \prod_{j=1}^{N} P_T(w_j|w_{j-2}w_{j-1})P_I(w_j)$$

によって求められる。ここで、N はヘッドラインに含まれる単語数、w_j は j 番目の単語、P_T は trigram 確率、P_I は重要語選択モデルにおける重要度である。w_j が内容語のときは $P_I(w_j)$ には前節で求めた重要度の値を用い、w_j が機能語のときは $P_I(w_j)$ には一定の低い値を割りあてる。N を小さくすることで「(3) 高圧縮性」を満たすことができ、ヘッドラインとしての 3 条件をすべて満たすことができる。

4 評価実験

まず、それぞれのモデルに関する評価実験を行うために Web ページからコーパスを収集した。人手によりヘッドラインが作成されているディレクトリサイトである、Open Directory Project[8] からヘッドラインとテキストを収集した。ディレクトリサイトでは、様々なページへのリンクが張られており、リンクの横にそのサイトの簡単な説明が 1 行程度で記述されている。その説明文をヘッドラインとし、リンク先の Web ページのタグを除いたテキスト部分をテキストとした。テキストに含まれる平均単語数は 327.1 語である。収集した文書のうち 5601 文書を訓練コーパスとし、967 文書をテストコーパスとした。

次に重要語を選択するための重要語選択モデル、単語間のつながりを考慮した文を生成するための要約文生成モデルを実際に構築し、収集したコーパスを用いてモデルの性能評価を行った。

4.1 Web ページを用いた重要語抽出実験

提案した重要語選択モデルの有効性を検証するため、テキスト中の内容語から重要語を抽出する実験を行った。内容語の対応付け方法としては、テキストとヘッドラインを JTAG[5] を用いて形態素解析し、順序に関係なく表記と品詞が完全に一致する形態素同士を対応付けた。この対応づけの結果をもとに重要語選択モデルを学習した。分類器により得られる尤度が非負の内容語を重要語として抽出し、代表的なキーワード抽出手法である TF・IDF モデルと F 値に関して比較評価を行った。

表 1: 重要語抽出実験結果

	モデル	適合率	再現率	F 値
訓練コーパス	本手法	0.188	0.191	0.189
	TF·IDF	0.096	0.244	0.138
テストコーパス	本手法	0.160	0.160	0.160
	TF·IDF	0.100	0.228	0.139

結果を表 1 に示す。適合率が向上し、F 値も TF・IDF モデルを上回る結果となった。F 値が全体的に低い値となったのは、人間がテキストからヘッドラインを作成する過程で頻繁に言い換えを行うため、内容語の対応がうまくとれなかったことが主な原因である。

4.2 Web ページからのヘッドライン生成実験

提案した要約文生成モデルの有効性を検証するため、Web ページからヘッドラインを生成する実験を行った。正解ヘッドラインの平均単語数は約 20 単語であったため、生成するヘッドラインの長さの上限も 20 単語とした。単語 trigram は、訓練コーパスとテストコーパス両方から学習した。ヘッドラインの比較対象としては、TF・IDF モデルを用いて重要語を算出し、要約文生成モデルで生成したヘッドラインをベースラインとし [6]、その他にリード文 (本実験では先頭の 1〜2 文) との比較を行った。

評価は目視により行った。テストコーパスから生成されたヘッドラインのうち、各 50 ずつのヘッドラインに対し、可読性と内容網羅性に関してそれぞれ○・△・×の 3 段階で評価した。表 2 および表 3 に可読性と内容網羅性の評価基準、本手法で生成されたヘッドラインの例、正解ヘッドラインの例を示す。これらの基準で評価した結果を表 4 に示す。

可読性に関しては、本手法とベースライン手法とを比較すると、○・△・×の割合がほとんど変わらないことが表からわかる。本研究のヘッドラインはベースライン手法とほぼ同じ可読性であることが確認できた。要約文生成の手法が同じであるた

表 2: 可読性の評価基準

評価	評価基準	生成されたヘッドライン	正解ヘッドライン
○	文のつながりに問題がない	検索エンジンのしくみ教えます。	サーチエンジンを動作原理から解説している。
△	文のつながりが一箇所おかしい	せかいのお米はいろんな国で、お米もあります。	世界各国の「お米」について解説しているサイト。
×	文のつながりが複数おかしい	新着情報満載で案内・本場所情報インタビュー協会。	番付表や力士へのインタビュー、チケット情報などを掲載。

表 3: 内容網羅性の評価基準

評価	評価基準	生成されたヘッドライン	正解ヘッドライン
○	正解と意味が同じ	名古屋を中心に活動しているアコーディオン奏者・青笹真樹の活動情報など	名古屋を中心に活動中の女性ジャズ・アコーディオニスト青笹真樹に関するサイト。
△	主要なキーワードを含む	レコーディングを体験してオリジナル CD を作ろう	自分だけのオリジナル CD を作ることができるサービスを提供する東京の企業。
×	主要なキーワードを含まない	私のコレクションは、新聞・音楽情報誌	デヴィット・ボウイに関するサイト。新聞・雑誌の切り抜きなどを収集。

め、これは妥当な結果といえる。ただし、約半数のヘッドラインが×判定となっており、可読性については十分読めるレベルに達しているとはいえない。

内容網羅性に関しては、本手法とベースライン手法とを比較すると、○および△の割合は本手法のほうが高く、×の割合が低いことから、内容網羅性では本手法のほうが優れているといえる。また、リード文は内容を正しく表している文が少なかったため、×判定の割合が約半数を占めており、やはり本手法のほうがリード文よりも内容をよく表したヘッドラインであることが検証できた。

5 考察と検討課題

本手法をベースライン手法と比較すると、可読性は同等であるが、内容網羅性では本手法のほうが優れている。よって、ベースライン手法と同程度の可読性を保ちながら、テキスト全体の内容をよりよく表すヘッドラインを生成できる。内容語選択モデルにおいて考慮した大量の素性が有効に働き、内容網羅性が向上したと考えられる。本手法をリード文と比較すると、可読性ではリード文が優れているが、内容網羅性では本手法が優れており、どちらがよいとはいえない。しかし、効率的な Web ページへのアクセスという目的においては、内容を間違えるよりは多少読みにくくても内容をよく表した本手法のほうがヘッドラインとしては有効であると考えている。

今後の検討課題として以下のことが考えられる。

- Web ページの特徴の利用
  重要語抽出モデルの精度を向上させるには、有効と思われる新たな素性を加える必要がある。単語がアンカータグに挟まれれいるかどうかなど、新たな素性としてタグ情報などの Web ページの特徴が利用できると考えている。

- 同義語リストやシソーラスの利用
  Web ページにおいては、テキストから人手でヘッドラインを生成する際に言い換えが頻繁に起こる。同義語リストやシソーラスを用いることで言い換えに対応でき、重要語選択の精度が向上すると思われる。

- 重要語の順序づけをするための尺度の検討
  重要語の選択において分類器が出力する尤度の値を観察すると、そのほとんどが ±1 付近に集中していた。そのため、尤度から算出される重要度がほぼ一定の値となってしまい、この影響で頻度の低い未知語や固有名詞が抽出されにくくなっていた。重要語の順序付けをするための尺度について検討したい。

表 4: 目視による評価結果

本手法		内容網羅性			計
		○	△	×	
可読性	○	7	2	3	12
	△	5	7	3	15
	×	5	7	11	23
計		17	16	17	50

ベースライン		内容網羅性			計
		○	△	×	
可読性	○	3	3	6	12
	△	2	3	9	14
	×	3	6	15	24
計		8	12	30	50

リード文		内容網羅性			計
		○	△	×	
可読性	○	7	10	24	41
	△	0	3	6	9
	×	0	0	0	0
計		7	13	30	50

- 文の結合に関するモデルの検討
  重要語が複数の文にまたがった場合に文の結合がうまくいかず、可読性が低下しているものが多かった。今後は、文の結合時に適切な語句を補うようなモデルについて検討したい。

6 関連研究

以下では、本研究に関連する研究について述べる。

Banko ら [1] は、1 文よりも短いヘッドラインを生成することを目的として、ヘッドライン生成のための統計モデルを構築した。モデルは本研究と同様に「(1) 内容網羅性」「(2) 可読性」「(3) 高圧縮性」が考慮されているが、「(1) 内容網羅性」のモデルとしてテキスト中の単語がヘッドラインに含まれる頻度しか考慮されていない点が本研究と異なる。

Berger ら [2] は、Web ページからのヘッドライン生成を試みている。統計モデルとしては、本研究と同様に noisy channel model を用いている。Berger らの研究では、人間がヘッドラインを作成する際にテキスト中の単語を別の語に置き換えることを考慮し、source model の素性として、テキストとヘッドライン間の単語の類似度を用いている点が本研究と異なる。

7 まとめ

本研究では、SVM を用いた重要語選択モデルにより重要語の選択を行い、単語 trigram モデルと重要語選択モデルを組み合わせた要約文生成モデルによりヘッドライン生成を行う方法を提案した。重要語抽出実験の結果、提案した重要語選択モデルは TF·IDF モデルより重要語の抽出精度が高いことを検証した。また、Web ページを用いたヘッドライン生成実験の結果、本手法はベースライン手法よりもテキスト全体の内容をより的確に表すヘッドラインを生成できることを検証した。

参考文献

[1] M. Banko, V. Mittal and M. Witbrock. "Headline Generation Based on Statistical Translation." In Proc. of the 38th Annual Meeting of the Association of the Computational Linguistics(ACL), 2000.

[2] A. Berger and V. Mittal. "OCELOT: A system for summarizing web pages." In Proc. of the 23rd Annual Conf. on Research and Development in Information Retrieval (ACM SIGIR), pp. 144-151, 2000.

[3] V. Vapnik, "The Nature of Statistical Learning Theory." Springer, 1995.

[4] 平尾 努, 前田 英作, 松本 裕治. "Support Vector Machine による重要文抽出." 情報処理学会情報学基礎研究会報告, pp. 121-128, 63-17, 2001.

[5] 渕 武志, 松岡 浩司, 高木 伸一郎. "保守性を考慮した日本語形態素解析システム." 情報処理学会自然言語処理研究会報告, 119-09, 1996.

[6] 堀 智織, 古井 貞熙. "ニュース音声の自動要約法とその評価法に関する検討." 日本音響学会春季講演論文集, Vol.1, pp.63-64, 2000.

[7] 若尾 孝博, 江原 暉明, 白井 克彦. "テレビニュース番組の字幕に見られる要約の手法." 情報処理学会自然言語処理研究会報告, pp. 83-89, 122-13, 1997.

[8] "Open Directory Project." http://dmoz.org/ .