# 認識誤りに頑健な重要語抽出

松尾義博　　林 良彦
日本電信電話株式会社 NTT サイバースペース研究所
{matsuo.yoshihiro, hayashi.yoshihiko}@lab.ntt.co.jp

## 1 はじめに

ブロードバンドネットワークの普及に伴って、音声や動画像を含むマルチメディアコンテンツの流通が広がりつつある。流通量の増大によって、その分類・検索への需要も高まり、コンテンツに付与されるメタデータの重要性が高まっている。特に、コンテンツ内の音声や文字情報のインデクシングは、テキストコンテンツの全文検索に相当すると言え、その内容記述の網羅性から考えて極めて重要なメタデータとなる。そのため、我々は、マルチメディアコンテンツの全文検索を可能とする音声インデクシングシステム [1] の開発を進めている。

増大するマルチメディアコンテンツの音声や文字情報を効率的にインデクシングするためには、音声情報、映像情報の自動認識が必要になるが、現状の認識技術では認識結果に誤認識が混入することが避けられない。コンテンツによっては、誤認識部の人手修正のコストをかけることが困難な場合も十分に考えられる。しかし、コンテンツ流通の手助けを目的としたメタデータの場合、分類や検索への悪影響が除去されれば、多少の誤りが含まれていても許容することが可能である。

本論文では文書中の単語への重要度付与、重要語抽出を取り上げ、認識誤りに対する頑健性を検討する。重要度付与は、分類や検索、自動要約などの基本となる技術であり、付与される重要度が認識信頼度を考慮したものであれば、後段の検索等の精度向上も期待できる。さらに、重要度上位語を抽出する重要語抽出タスクは、少数の語をもって文書全体を代表させる、いわば文書の特徴量抽出と言え、選択誤りは対象文書の特徴量の品質を大きく損ねるものとなる。少数の代表語への誤認識語の混入は文書の特徴を全く違った方向へ持っていく可能性があることから、単に非重要語が選ばれてしまったのとは質の異なる悪影響であり、これを排除することが重要である。本論文では、入力中に認識誤りが含まれている場合でも頑健に動作可能な重要語抽出手法を提案する。

## 2 誤りを考慮した重要語抽出

### 2.1 認識誤りの検出

音声認識や文字認識の結果を見ていると、時折、何の脈絡もなく誤認識語が混入していて違和感を持つことが多い。これは、自動認識結果がまさに「脈絡」を十分に考慮しきれていないことに起因していると言える。音声認識の場合、語の選択は一般に (1) 音響的な類似度に基づくパタンマッチングと (2) 言語モデルや制約規則による言語的な選好によってなされる。文字認識の場合も前者が画像の類似度に基づくだけで、基本的には同様である。

選択基準のうち前者は入力信号と音素の対応を求めるだけであり、文脈に依存した要素はせいぜい3音素程度の連鎖関係である。さらに語の意味といった要素は全く考慮されないので、何ら脈絡に応じた出現傾向は含まれていない。

後者は、処理単位が単語であるから、語の意味に応じた選好が期待できるが、考慮される文脈は、計算量とスパースネスの問題からこちらもせいぜい trigram 程度が一般的であり、文書全体の脈絡の反映度は極めて限定的である。言語モデルへ長距離依存関係を導入する試みもキャッシュモデル [2] を始めとして多数行

なわれているが [3]、探索中の適用であるから計算量の制約が大きく適用可能な情報も限られる。

また、認識誤りを考慮した言語処理手法としては、探索過程で得られる尤度を選好スコアに導入して低尤度語を排除する音声要約手法 [4] などがあるが、ここで利用するスコアは認識で用いたモデルに依存した値であり、一般に長距離依存情報を導入することは容易ではない。

そこで本稿で提案する手法では、自動認識によって得られたテキストを対象として、その結果を長距離依存関係を考慮して再評価し、重要度付与へ反映させることを考える。この手法であれば、探索中の計算ではないため計算量への余裕があり、また、特定の言語モデルに依存した処理ではないため、広く適用可能な手法となる。

### 2.2 重要度付与手法

本手法の目的は正しい重要語を抽出することよりも、重要語への誤認識語の混入を抑制することである。重要度の低い語が重要語として抽出されていても、そのことによって文書全体の意味をゆがめてしまうわけではなく、悪影響は限定的である。逆に誤認識語の混入は、文書全体の意味を大きくゆがめてしまうため、優先して排除すべきである。

文書全体の意味をゆがめないためには、全体からかけ離れた語を除去する尺度を導入すればよい。本稿では、単語間の相互情報量を尺度に用いることとする。各単語が文書全体の意味に近い度合 $f$ を、

$$f(w_i) = \sum_{j \in \text{文書}, j \neq i} I(w_i, w_j)$$

と、文書中の他の語との相互情報量の積算で定義する。相互情報量 $I(w_i, w_j)$ は新聞記事等の別コーパスからあらかじめ算出しておくが、各単語のコーパス中での出現確率 $P(w_i)$, $P(w_j)$ と共起出現確率 $P(w_i, w_j)$ を用いて、

$$I(w_i, w_j) = \begin{cases} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} & (P(w_i, w_j) > 0) \\ 0 & (P(w_i, w_j) = 0) \end{cases}$$

と計算できる。

この $f(w_i)$ を重要度算出式に付加することにより、文書の話題から大きく外れた語の出現を抑制する。誤認識を考慮しない重要度を $S_0(w_i)$ とすると、単語 $w_i$ の重要度を

$$S(w_i) = S_0(w_i) f(w_i) \quad (1)$$

とする。重要度算出に tfidf を用いる場合、

$$S_0(w_i) = tf(w_i) idf(w_i)$$

であるから、

$$S(w_i) = tf(w_i) idf(w_i) f(w_i)$$

となる。

また、$f(w_i)$ の影響度、効果を見積もるために、$f(w_i)$ をより強調した、

$$S(w_i) = S_0(w_i) f^2(w_i) \quad (2)$$

と、影響度を弱めた、

$$S(w_i) = S_0(w_i) \sqrt{f(w_i)} \quad (3)$$

の二条件でも実験を行なう。

## 3 実験

実験では毎日新聞 2001 年版を用い、半年分 50000 記事 (学習コーパス) で共起情報と df を取得し、残りの半年から 1000 記事 (評価コーパス) を用いて評価を行なった。実験では、評価コーパスの本文から重要語を抽出し、その重要語が新聞記事の見出しに含まれていれば正解、含まれていなければ不正解とする。学習コーパス中の内容語の頻度上位 10000 語を対象にし、各記事の見出しに含まれていた語数と同数を抽出することにする。評価コーパスの各記事の平均単語数は 94.8 語/記事、見出し語の平均語数は 6.5 語/記事であった。

注意すべきは、本実験では学習コーパスの正解重要語 (見出し) を用いて、重要語抽出パラメータを学習しているわけではない点である。学習コーパスは誤認識語の抽出のための共起情報取得に用いられ、重要語抽出自体は単純な tfidf によっている。つまり正解見出しは学習には何ら寄与していない。さらに、可能な要約

表 1: 等確率誤りモデルの場合の結果

| 認識誤り率 (%) | | 0 | 2 | 5 | 7 | 10 | 20 |
|---|---|---|---|---|---|---|---|
| tfidf | 正解率 (%) | 30.8 | 30.3 | 29.5 | 28.8 | 28.1 | 25.3 |
| | 混入率 (%) | 0.0 | 1.7 | 4.6 | 6.2 | 9.1 | 17.6 |
| tfidf $\times$MI | 正解率 (%) | 30.1 | 30.0 | 29.2 | 29.0 | 28.5 | 26.5 |
| | 混入率 (%) | 0.0 | 0.3 | 1.3 | 1.6 | 2.3 | 5.9 |
| tfidf $\times \mathrm{MI}^2$ | 正解率 (%) | 28.7 | 28.6 | 28.1 | 28.0 | 27.3 | 25.7 |
| | 混入率 (%) | 0.0 | 0.4 | 1.4 | 1.7 | 2.4 | 6.3 |
| tfidf $\times \sqrt{\mathrm{MI}}$ | 正解率 (%) | 30.6 | 30.4 | 29.7 | 29.5 | 28.9 | 26.9 |
| | 混入率 (%) | 0.0 | 0.3 | 1.4 | 1.5 | 2.4 | 6.0 |

表 2: 頻度比例誤りモデルの場合の結果

| 認識誤り率 (%) | | 0 | 2 | 5 | 7 | 10 | 20 |
|---|---|---|---|---|---|---|---|
| tfidf | 正解率 (%) | 30.8 | 30.4 | 29.9 | 29.5 | 29.1 | 27.0 |
| | 混入率 (%) | 0.0 | 0.8 | 2.2 | 2.9 | 4.1 | 9.4 |
| tfidf $\times$MI | 正解率 (%) | 30.1 | 30.0 | 29.3 | 29.3 | 28.5 | 27.1 |
| | 混入率 (%) | 0.0 | 0.4 | 0.9 | 1.3 | 1.8 | 4.7 |
| tfidf $\times \mathrm{MI}^2$ | 正解率 (%) | 28.7 | 28.7 | 28.2 | 28.0 | 27.6 | 26.4 |
| | 混入率 (%) | 0.0 | 0.4 | 0.9 | 1.3 | 1.8 | 4.6 |
| tfidf $\times \sqrt{\mathrm{MI}}$ | 正解率 (%) | 30.6 | 30.2 | 29.8 | 29.6 | 29.1 | 27.4 |
| | 混入率 (%) | 0.0 | 0.4 | 1.0 | 1.3 | 1.9 | 5.0 |

は多数あるにもかかわらず、この評価基準では単一の正解を設定しているため、正解精度自体は高い値を期待できない。これは、本実験の目的がここで設定した正解精度を高めることではなく、抽出語への誤り語の混入を抑制することにあるためである。正解精度は、抑制を導入することによる精度への悪影響がないことを確認するために算出する。

実験では、評価コーパス記事中の単語をランダムに置き換えることにより認識誤りをシミュレートする。ここでは、置換誤りのみを考え、挿入誤り、削除誤りは考えない。置き換える単語 (誤認識語として混入する単語) の選択であるが、

- 等確率でランダムに選択
- 学習コーパス内の出現頻度に比例した確率でランダムに選択

の 2 つの条件で実験を行なった。これは、パタン認識が主要因の誤りは言語的情報に非依存であるから前者の等確率に近い傾向を持つと考えられ、言語モデルの特性によって生じた誤りは一次近似としては出現確率に比例すると考えられるためである。実際の認識処理は両要因を反映していることから考えると、実際の誤認識は両モデルの中間の傾向を持つと考えられる。

### 3.1 結果

実験の結果を表 1 と 表 2 に示す。表 1 は等確率で選択した場合の結果で、表 2 は頻度比例確率で選択した場合の結果である。実験では認識誤り率 0% ~ 20% をシミュレートした。両表とも上から順に単純な tfidf による抽出結果、tfidf に相互情報量を掛け合わせた場合 (式 1) の結果、tfidf に相互情報量の 2 乗を掛けた

結果 (式 2)、tfidf に相互情報量の平方根を掛けた結果 (式 3) である。正解率は、

$$\text{正解率} = \frac{\text{正解語と一致した抽出語数}}{\text{抽出語総数}}$$

であり、混入率は

$$\text{混入率} = \frac{\text{抽出された語のうちの誤認識語数}}{\text{抽出語総数}}$$

である。なお、正解語数と抽出語数は同数であるから正解率=再現率=適合率である。

まず抽出された重要語の正解率を検討する。認識誤り率が大きくなるにつれて正解率が低下しているが、これは、誤認識語が選択されることにより正解重要語が抽出語から押し出されたことに加えて、そもそも、抽出されるべき正解重要語がランダムな置換により入力文書から排除されていることに起因している。単純な tfidf と相互情報量を追加したものとを比較すると、相互情報量を掛けあわせることにより、正解率は 0.7% の低下 ~ 1.2% の上昇となっている。この結果から、正解率への大きな影響はないことがわかる。

また、相互情報量の 2 乗の結果と平方根の結果を見てみると、総じて平方根の方の正解率が良く、2 乗の場合 2.1%の低下 ~0.4%の上昇、平方根の場合 0.2%の低下 ~1.6%の上昇となっている。この結果から、あまり相互情報量を強調しすぎると多少悪影響を生じるものの、適度な重みで加えることにより、条件によっては抽出精度の向上にも寄与できることがわかる。

次に認識誤り語の混入率を見てみると、等確率の誤りモデルの場合は、単純な tfidf に比べて 70% 程度混入率が抑制されていることがわかる。頻度比例の誤りモデルの場合は、単純な tfidf に比べて 50% 程度混入率が減少している。いずれも、もともとランダムに置換した認識誤り率の 1/4 ~ 1/7 程度の混入率であり、誤認識語の影響が抑制されていることがわかる。

また、相互情報量の重みづけ度合は混入率にはほとんど影響がない。このことは、記事内共起の情報によって排除可能な語はほぼ排除されており、残った語の排除にはさらに別の情報源が必要なことを示していると考えられる。

これらの結果から、別コーパスから学習した共起情報を重要語抽出に導入することにより、抽出正解率を保ったまま、誤認識語の混入を半数以下に抑制することが可能なことがわかった。

## 4 おわりに

本稿では認識誤りに頑健な重要語抽出手法について述べた。シミュレーションによる実験では、抽出正解率を保ったまま、誤認識語の混入を 50~70%程度抑制できることがわかった。

今後は、実際の音声認識等に適用した場合の効果を検証するとともに、本重要度付与および重要語抽出手法を検索や要約等のアプリケーションに適用した場合の改善効果の検証を進める。

## 参考文献


[1] 林他. 映像コンテンツのインデクシングのための音声・言語処理. 情報処理学会全国大会, 2003.

[2] Roland Kuhn. Speech recognition and the frequency of recently used words: a modified markov model for natural language. *COLING*, pp. 348–350, 1988.

[3] 黒橋他. 文脈共起ベクトルに基づく大域的言語モデル. 情報処理学会研究会 自然言語処理 139-11, 2000.

[4] 堀他. 信頼尺度を用いた音声自動要約の改善. 日本音響学会秋季講演論文集, pp. 79–80, 2000.