大規模テキストから位置情報および特徴語を抽出するルールの検討

松川淑子† 小林正博‡ 永井洋一† 木内直人† 今野清孝† 山田洋志† 亀井真一郎†
日本電気株式会社† 株式会社日本システムアプリケーション‡

1．はじめに

情報推薦システムで利用する特徴語を抽出するためのルールについて述べる。今回試作した情報推薦システムでは、ユーザの滞在位置とテキスト閲覧履歴を利用して、ユーザに適したテキストを推薦する。

本稿で述べるルールで抽出する特徴語は、あらかじめテキストと対応付けて保存され、推薦対象テキストを選択する際と、推薦対象テキストをユーザに提示する際に利用される。具体的には、ユーザがテキストAを閲覧した場合、情報推薦システムは次に、テキストAと類似したテキストを推薦したいと考える。そこで「類似」の基準を「テキストAの特徴語と同じ特徴語をもつテキスト」と定め、テキストAと対応付いた特徴語を求め、同じ特徴語をもつテキストBを推薦対象として選択する。そしてテキストBをユーザに推薦する際に、その特徴語を推薦理由として提示する。なおここで述べる特徴語には位置情報も含めるものとする。

テキストに、そのテキストを特徴付ける語句を対応付ける手法は、従来からさまざまに提案されている[1, 2]が、一般には単語を特徴語として扱うことが多い。しかし単語または単語の羅列をキーとして推薦対象テキストを求めると、ヒット件数が多くなりすぎたり絞り込まれすぎたりしてしまうし、該当単語が異なる文脈で使われているテキストが選択される可能性も高くなる。さらに推薦理由として提示する際にも、単語だけではその単語が使われている文脈がわからないため、ユーザは、推薦されたテキストが自分にとって興味ある内容なのか否かが判断できない。

そこで、テキストをより具体的に特徴付けられる語句を抽出することを目的としたルールを検討し、既存のツールを使って、多種大量のテキストから特徴語を抽出した。

2．対象テキスト

以下のような、合計7ジャンルの大量なテキスト群から特徴語を抽出した。

例：グルメ店情報（約30,000件）
　　Q&A（約2,500,000件）
　　ブログ（約236,000件）

一つのテキストは複数のフィールドにわかれており、フィールドはジャンルごとに異なっている。すべてのフィールドに特徴語が含まれているわけではないので、ジャンルごとに、特徴語を抽出すべきフィールドを定義した。

3．特徴語抽出方法

特徴語の抽出には既存の弊社開発技術を利用した。テキストを言語解析し、解析結果から、あらかじめ記述しておいた用語（句）抽出ルールによって該当部分を抽出する技術である（図1参照）。

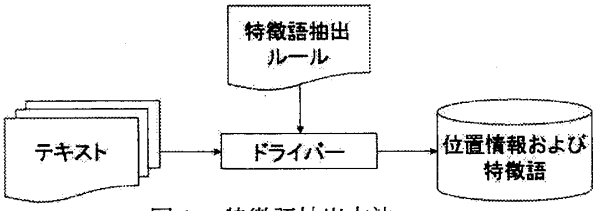

図1　特徴語抽出方法

この技術を用いて、特徴語抽出対象フィールドを形態素解析し、解析結果に、4節で説明する特徴語抽出ルールを適用して、特徴語を抽出した。

本技術では構文意味解析を行って情報抽出することもできるが、今回は大規模なテキストを短時間で処理するという制約の下で、形態素解析のみを行い、目的の達成度と課題を抽出することとした。

4．特徴語抽出ルールと抽出結果

大きくは位置を表す情報を抽出するルール、嗜好を表す情報を抽出するルール、基本的な語彙を抽出するルールを検討した。適用にあたっては、ジャンルごとおよび特徴語抽出対象フィールドごとに、記述された内容の特性をふまえて使用するルールの組み合わせを変えた。

以下では、位置を表す情報を抽出するルールと、嗜好を表す情報を抽出するルールの中でも特に効果的な特徴語を抽出できたルールについて説明する。

4.1. 位置情報抽出ルール

テキストを、ユーザの滞在位置と結びつけるための

Examination of rules that extract location information and characteristic words from large-scale text resources
† Yoshiko Matsukawa, Yoichi Nagai, Naoto Kiuchi, Kiyotaka Konno, Hiroshi Yamada, Shinichiro Kamei (NEC Corporation)
‡ Masahiro Kobayashi (Japan System Applications Co., Ltd.)

情報を抽出するルールである。位置を表す情報として地名、駅名、建物名、住所を抽出することとした。駅名抽出ルールでは、単独の駅名だけでなく、「路線名＋駅名」の組み合わせも抽出するようにした。住所抽出ルールでは、単なる地名の羅列ではなく、その構成要素に Part-Of（全体と部分）の意味属性をもつ地名の連続を抽出するようにした。そのため「世田谷区港区」などは抽出されない。このようなルールで下記のような特徴語を得た。

＜抽出結果の例＞

琴似駅、東西線琴似駅

渋谷、東京都千代田区神田神保町、静岡県沼津市

4.2. 嗜好情報抽出ルール

テキストを、ユーザの嗜好と結びつけるための情報を抽出するルールである。

4.2.1. 「修飾表現＋名詞」抽出ルール

名詞（例：「カレー」）だけを抽出するより、「カレー」が「おいしい」と書かれているのか「まずい」と書かれているのかまでわかったほうが、テキストをより具体的に特徴付けられる。そこで名詞の前後に、物事の性質や状態を示す形容詞・形容動詞が出現する場合には、それらの修飾表現と名詞を合わせて抽出することとした。細かくは次の４種類のルールを定義し、下記のような特徴語句を得た。

(1) 形容詞＋名詞

(2) 名詞＋が, は, も＋形容詞

(3) 形容動詞＋な, の＋名詞

(4) 名詞＋が, は, も＋形容動詞

＜抽出結果の例＞

柔らかい肉、コーヒーもおいしい、

オリジナルの焼酎、デザートメニューが豊富

4.2.2. 「修飾表現＋"雰囲気"」抽出ルール

グルメ、ホテル、ショッピングなどのジャンルにおける店や物の雰囲気は、店や物を決める際の大きな手がかりになる。しかもテキスト内では、「落ち着いた雰囲気」のように直前に修飾表現をともなって書かれることが多い。そこで名詞の中でも「雰囲気」および、「雰囲気」を示すような用語として「ムード」と「店内」に絞り込んで、修飾表現と合わせて抽出することとした。4.2.1 で述べたルールの名詞部分を「雰囲気」「ムード」「店内」に置き換えたルールのほかに、主に次の５種類のルールを定義し、下記のような特徴語句を得た。

(1) 名詞＋的＋な＋雰囲気, ムード, 店内

(2) 名詞＋のような＋雰囲気, ムード, 店内

(3) 名詞＋の＋雰囲気, ムード, 店内

(4) 動詞＋た＋雰囲気, ムード, 店内

(5) 形容動詞＋に＋動詞＋雰囲気, ムード, 店内

＜抽出結果の例＞

隠れ家的な雰囲気、船上のような店内、

和のムード、落ち着いた雰囲気、

気軽に入れるムード

4.2 のルールで得られる特徴語句は、情報推薦システムで利用する上で非常に有効であると考える。「コーヒーもおいしい」,「落ち着いた雰囲気」という特徴語句が推薦理由として提示されれば、コーヒー好きのユーザや落ち着いた雰囲気のお店を探しているユーザは、「もっと詳しい情報を知りたい」と考えて、推薦されたテキストの詳細にアクセスすることが十分予想される。また単語の羅列とは異なり意味をもつフレーズなので、同じ特徴語句をもつテキストは、似たような文脈をもつ類似性の高いテキストに絞り込まれる可能性が高くなる。

5．今後の課題

日本語の地名は人名や一般名詞と同じ表層をもつものが多い。そのため例えば「渋谷」が、地名の場合も人名の場合も抽出されてしまう。今後は単語の使われている前後の文脈や、使われているフィールドなどの情報を最大限に活用して、抽出精度を上げてゆく必要がある。

また「修飾表現＋名詞」抽出ルールでは、

(1) 少なくヘルシー

　原文：赤身モモ肉は脂肪が少なくヘルシーで

(2) メニューが豊富

　原文：産直メニューが豊富

など、具体的な特徴となっていない語句も抽出されていた。このような語句を絞り込むには、次のような対応案が考えられる。

＜(1)への対応案＞

「名詞＋が, は, も＋形容詞＋名詞」というルールを追加する。これにより「脂肪が少なくヘルシー」が抽出でき、これに包含される「少なくヘルシー」は採用しないようにする。

＜(2)への対応案＞

「名詞＋名詞＋が, は, も＋形容動詞」のような、名詞の連続と修飾表現を合わせて抽出するルールを追加し、包含されるルールで抽出された語は採用しないようにする。名詞を３つにするルールや、名詞間に助詞を含むルールも考えられるが、パターンを膨らませすぎると、対応付くテキストが少なくなってしまうので注意が必要である。

以上のような対応策を施し、特徴語句の精度を上げることが今後の課題である。

謝辞：本研究は、経済産業省「情報大航海プロジェクト」における(株)ＮＴＴドコモを中心とした「マイ・ライフ・アシストサービス」実証実験の一環として実施した。推薦対象テキストは、(株)角川クロスメディア、ＮＴＴレゾナント(株)、(株)関心空間の各社からの提供を受けた。

参考文献

[1] 徳永健伸『言語と計算５ 情報検索と言語処理』, 財団法人 東京大学出版会, 1999

[2] 奥村学, 難波英嗣『知の科学 テキスト自動要約』, 株式会社 オーム社, 2005