3V－11

語句が持つ概念的複雑度に着目した類似内容文章の検索

後藤 正樹　村岡 洋一
早稲田大学 理工学部

1 はじめに

本研究では、ある日本語で書かれた文章から、内容的に類似した文章を検索するシステムを開発した。本研究において検索対象としたのはネットニュースであり、質問を投稿する前に質問文章を本システムによる文章検索にかけ、類似記事を抽出する事を想定する。

2 従来の検索手法の問題点とその解決

文章を入力とした検索システムでは、辞書を使った構文解析を行うもの[1]が一般的であった。しかし、辞書を使った解析には、新語・造語に弱いといった欠点がある。この欠点は、新しく生み出された言葉についての知識を得る事を目的として検索を行った時に問題となる。

本論文では入力文章および蓄積文章から検索に必要な語句集合を切り出して、これらの間の類似度を評価する検索手法を提案する。まず語句を切り出す手法としては、n-gram 概念単位分割を提案する。これは辞書を用いる形態素解析の代わりに、ひらがなを分割子として文章から検索に使う語句を切り出す手法である。

切り出された語句集合間の類似度の評価については、文章の特徴を強く示す語句が一致した場合に文章の類似度が高いと判定した方が検索結果が向上すると思われるため、後述の Particular Topic Search により語句に重みづけをする方法を提案する。さらに、これらの手法を用いた検索システムを開発し、実験によってその有用性を確認する。

Related document search by noticing the complicated words in concept, Masaki Goto, Yoichi Muraoka, School of Science and Engineering, Waseda University

3 検索単位の切り出し

本稿では、WWW 上のキーワード検索システム[2][3]において検索に用いるキーワードのように、他から区別された具体的で固有な意味を持つものを概念と定義する。

ここで、日本語文章における助詞や助動詞、接続詞などは意味的補助を行うものであるため、本手法ではこれを概念を持つ語句ではないと判断し、これら以外の概念部分を切り出して入力文章と蓄積文章の概念部分同士を比較し、検索を行う事にする。

漢字仮名交じり文で日本語を記述した時、各品詞は自立語部分と付属語部分で構成される。本手法においては、自立語部分が概念を表し、助詞や助動詞などで記述される付属語部分は補助的意味を表すとみなし、付属語部分を切り落とし、自立語部分のみを切り出す事を考える。

日本語では品詞間の接続には副用言を除き付属語を介する必要があるのだが、日本語における付属語部分は必ずひらがなで記述される。副用言のうち接続詞・感動詞は概念部分を持たず、ひらがなのみで記述される。連体詞は必ずひらがなで終了する。副詞のみは漢字のみで記述する事ができるが、副詞だけで文章を記述する事はできない。

これらの特性を踏まえ、本手法における n-gram 概念単位分割では日本語文章に存在するひらがな部分または明示的に示された区切り記号を概念的な区切り単位とみなし、文章を漢字、カタカナ、英数字で表される、概念部分の集合に切り分ける。

4 検索手法

本研究では切り出された n-gram 概念単位の、検索の際の重要度を概念的複雑度を基準に判断する手法として Particular Topic Search を提案する。

Particular Topic Search では、切り出されたある n-gram 概念単位を構成する base 概念単位の個数が多いほど概念的複雑度が高く、検索の際に重要であると判定する。base 概念単位は、

(1) 漢字表記のものに関しては漢字 1 文字
(2) 英語などに関しては 1 単語
(3) (2) において、abbreviation もしくは acronym を用いてある場合はアルファベット 1 文字

と定義する。

漢字はその成り立ちを追えばわかるように、当時必要であった概念に対し 1 対 1 対応で文字が割り当てられた。よって、基本的な概念は漢字 1 文字で表現できる。しかし現在、漢字自体が増える事はないので、新たな概念は base 概念単位、つまり個々の漢字の組み合わせで表現する。ここで、表現しようとする概念が複雑であればある程多くの base 概念単位を必要とする。これを応用したのが Particular Topic Search であり、多くの漢字がまとまって表現される複雑な概念は、その文章を特徴付ける、検索に重要なものであると判定する。また、英語などについても同様の考えを適用する事により (1) (2) (3) の定義付けを行った。

5 重要度順リストによる重み付け

システムは、類似度判定のために蓄積文章に入力文章との類似度を反映したスコアを付ける。入力文章と蓄積文章の各 n-gram 概念単位について一致した base 概念単位の個数が多いほど重要なものだとして高いスコアを割り振り、累積スコアにより類似度を判定する。

ここで、検索精度を上げるため、ある n-gram 概念単位が文章内でどれだけ相対的な重要度を持つのかを検索結果に反映させる。そのため、n-gram 概念単位を重要度が高い順に並べた重要度順リストを作成し、リストにおける順位が近い n-gram 概念単位ほど高いスコアを割り振る事にする。

このため、base 概念単位の一致が見られたある n-gram 概念単位の重要度順リスト中の位置が入力文章では p_1,、蓄積文章では p_2 であり、また重要度順リストの長さが入力文章では w_1、蓄積文章では w_2 であった時に、x の絶対値を表す関数を abs (x)、2 値 a,b の最大値を抽出する関数を max(a,b) とし、

$$1-\frac{abs(p_1-p_2)}{\max(w_1,w_2)}$$

と表される勾配式を割り振るスコアに乗算する。

6 評価

これらの手法を用いて、それぞれ分野の異なる 5 つのニュースグループから 500 ずつ抽出した記事を母体集合とし、その中のランダムに抽出した記事から他の関連記事を抽出する作業を 300 回繰り返すという実験を行った。ヒットした検索結果の類似性の判定は、客観性を持たせるために、記事のヘッダ情報から得た実際の記事関連性を用いて行った。検出結果を類似度が高い順に並べた時、30 位以内までに関連記事の検出を行えた確立を母体記事中に存在する関連記事の数に応じて示すと、グループ fj.lang.perl については以下のようになった。

関連記事数	1	2	3	4	5
関連記事の検出確率	80%	90%	82%	100%	96%

また、別に用意した、切り出した n-gram 概念単位に対して特殊度の重み付けを行わない検索システムを用いて同様な実験を行い、類似度順に並べた検出結果を本手法によるシステムと比較したところ、どのグループにおいても本手法によるシステムの方が、実際の関連記事がより高い順位で検出される事が確認され、本手法の有用性が示された。

7 今後の課題

現在、英語を由来とした n-gram 概念単位について、英語表記とカタカナ表記では別の概念だと判定されてしまう。この問題を解決するために、英語表記からカタカナ表記を自動生成する事を考えている。

参考文献

[1] 後藤和之,笹気光一,中山康子, 知識情報共有システムの開発と実践—オフィス知識ベースの構築とノウハウベースとの連携— ,電子情報通信学会研究報告 AI97-12-23 人工知能と知識処理, 1997
[2] 検索エンジン goo, http://goo.ne.jp/
[3] YAHOO!JAPAN, http://www.yahoo.co.jp/