3C3-1

単語と意味属性との共起に基づく概念ベクトル生成手法

Concept Vector Generation Method Based on Co-occurrences between Words and Semantic Attributes

別所 克人[*1] 古瀬 蔵[*1] 片岡 良治[*1]
Katsuji Bessho Osamu Furuse Ryoji Kataoka

[*1] 日本電信電話株式会社 NTTサイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

We propose a method generating the concept vectors, which are the semantic representations of words, based on co-occurrence frequencies between words and those semantic attributes in a corpus. Using semantic attributes, concept vectors can represent the meanings of the words more appropriately. The experimental results of document retrieval using the generated concept vectors showed that the proposed method yields a higher retrieval accuracy than the conventional method.

1. はじめに

コーパスにおける単語同士の共起頻度を記録した共起行列に対し特異値分解を行い、単語を次元数の縮退したベクトルで表現したものを概念ベクトルと呼び、単語とその概念ベクトルの対の集合を概念ベースと呼ぶ。概念ベースは、単語の意味的類似性を定量化できるため、情報検索[Schutze 94][Kato 99][熊本 99]や、テキストセグメンテーション[別所 06]等に適用され、効果をもたらしてきた。

しかしながら、特異値分解処理は一般に多量の計算量をとるため、共起頻度をとる単語の集合を制限する必要がある。このため、生成された概念ベクトルの質に問題があった。

本稿では、コーパスにおける単語と、単語に付随する意味属性との共起頻度から、単語の意味表現としての概念ベクトルを生成する手法を提案する。意味属性を用いることにより、従来手法と比べ、計算量を増やすことなく、概念ベクトルが、対応する単語の意味をより的確に表すようになる。

以下、2 章で従来手法のアルゴリズムと問題点を述べ、3 章で提案手法を述べる。4 章で、生成した概念ベクトルを用いた文書検索の評価実験の結果を述べ、5 章でまとめを述べる。

2. 従来手法

2.1 概念ベース生成アルゴリズム

本章では、従来手法である[Schutze 98]における潜在的意味解析の手法を述べる。

まずコーパスを形態素解析し、名詞、用言等の内容語のみを残す。残った異なり単語の集合を G , K ($G=K$)とする。G 中の単語を概念語、K 中の単語を共起語と呼ぶ。任意の概念語と共起語とが 1 文中に共起する頻度をカウントし、各行が概念語に対応し、各列が共起語に対応しているような共起行列を作成する。共起行列から零ベクトルである行ベクトルを削除する。共起行列の各行ベクトルは、対応する概念語の共起パターンを表しており、この行ベクトルを共起ベクトルと呼ぶ。ある 2 単語に対応する共起ベクトルが近ければ、共起パターンが似ているので、この 2 単語は意味的に近いということが推測される(図 1)。

但し、このままではデータのスパースネス性があることを始めとして、テキストデータから抽出される単語の情報には常に欠落

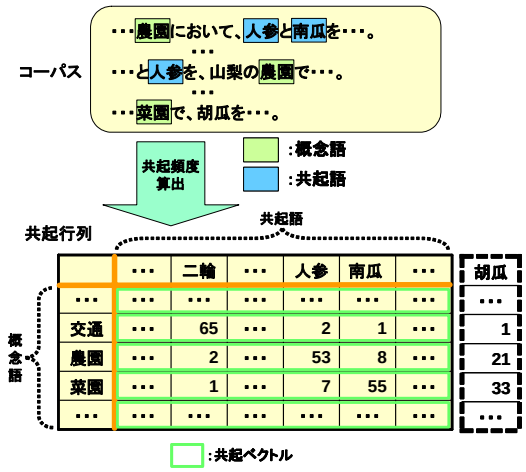

図 1: 従来手法における共起行列

があると予想されるため、ベクトル間の類似度の精度は低いと考えられる。また、一般に共起ベクトルの次元数は非常に大きなものとなるため、共起ベクトルを利用した言語処理の計算量も無視できないものとなる。このため共起行列を特異値分解により、次元数を縮退させた行列に変換する。

G , K の要素数が多いと、特異値分解の計算量は多量になるため、低コストで実行することが不可能となる。そこで、G , K を、高頻度語の集合に限定した上で特異値分解を実行する。ここで、精度向上のため、共起行列中の各成分をその平方根に変換して得られる行列 X に対し特異値分解を実行する。

X を $p\times q$ の行列としたとき、特異値分解により X は、以下のように分解できる。

$$\underset{p\times q}{X} = \underset{p\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times q}{V^t} \qquad (1)$$

ここで、添字 t は行列の転置を表す。$r = rank\,X \le \min(p,q)$, $U^tU = V^tV = I$ (I :単位行列)であり、$\Sigma = (\delta_{ij})$ としたとき、$\delta_{ii} \ge \delta_{jj} > 0 \left(1 \le^{\forall} i \le^{\forall} j \le r\right)$, $\delta_{ij} = 0\ (i \ne j)$ である。$\delta_{ii}\ (1 \le i \le r)$ を X の特異値と呼ぶ。

ここで、$1 \le^{\forall} r' \le r$ に対し、U の最初の r' 列、V^t の最初の r' 行、Σ の最初の r' 行、r' 列をとり、

$$\underset{p\times q}{X'} = \underset{p\times r'}{U'}\ \underset{r'\times r'}{\Sigma'}\ \underset{r'\times q}{V'^t} \qquad (2)$$

連絡先:別所克人, 日本電信電話株式会社 NTT サイバーソリューション研究所, 神奈川県横須賀市光の丘 1-1, Tel: 046-859-2525, Fax: 046-855-1730, bessho.katsuji@lab.ntt.co.jp

とする。U' の行ベクトルを長さ 1 に正規化したものを単語概念ベクトルと呼び、概念語とその概念ベクトルの対の集合を単語概念ベースと呼んでいる。

2.2 従来手法の問題点

従来手法では、共起行列の列となる単語の中に同一のカテゴリに属するものがあり、それらの単語との共起頻度が別々にカウントされるため、共起ベクトルが適切なものでなくなるという問題がある。例えば、図 1 の”人参”と”南瓜”は同一のカテゴリ”野菜”に属するが、それらとの共起頻度が別々にカウントされるため、”農園”と”菜園”の共起ベクトルが適切なものでなくなり、”農園”と”菜園”は意味的に近いにも関わらず、対応する共起ベクトルは遠くなる。

また、従来手法では、共起行列の列となる単語から漏れる単語が多数あり、そのような単語との共起頻度は考慮されないという問題がある。例えば、図 1 の”胡瓜”との共起頻度が考慮されない。このような情報の欠落により、共起ベクトルの質が低下する。

3. 提案手法

従来手法の問題点を解決するため、提案手法では、コーパスにおける単語同士の共起頻度ではなく、コーパスにおける単語と、単語に付随する意味属性との共起頻度をとる。この意味属性とは、日本語語彙大系[池原 97]における一般名詞意味体系の意味属性を意味している。

日本語語彙大系における一般名詞意味体系は、名詞と用言の意味を体系立てたシソーラスであり、各ノードを意味属性と呼ぶ。このシソーラスは 12 階層であり、2715 個のノードからなる。

本提案手法では、形態素解析プログラムとして JTAG[Fuchi 98]を用いているが、JTAG が参照する単語辞書では、各名詞と用言に意味属性が付与されている。一つの単語に複数の意味属性が付与されていることもあるが、これらの意味属性は、使用される局面が高いと思われる順に順序付けられている。形態素解析結果において、各単語には、対応する意味属性の情報が付随している。

任意の概念語と意味属性とが 1 文中に共起する頻度をカウントし、各行が概念語に対応し、各列が共起語に対応しているような共起行列を作成する(図 2)。

このように単語ではなく、意味属性との共起頻度をとることにより、同一の意味属性をもつ個々の単語との共起頻度は、該意味属性との共起頻度に含まれるため、共起ベクトルが、より適切なものとなる。例えば、図 1 における”二輪”の意味属性は”車”で、”人参”、”南瓜”の意味属性は”野菜”であるため、”人参”、”南瓜”それぞれとの共起頻度は、”野菜”との共起頻度に含まれる。これによって意味的に近い”農園”と”菜園”の共起ベクトルは値が近くなる。

また、意味属性の数は高々2715 であるため、全意味属性を共起行列の列として採用することができる。このため、従来手法で、共起行列の列となる単語から漏れていた単語との共起頻度も、該単語の意味属性との共起頻度に含まれるため、共起ベクトルが、より豊富な情報をもつようになる。例えば、図 1 における”胡瓜”の意味属性は”野菜”であるため、”胡瓜”との共起頻度が”野菜”との共起頻度に含まれる。従来手法では、考慮されなかった”胡瓜”との共起頻度が、提案手法では考慮されるようになる。

概念語・意味属性間共起行列中の各成分をその平方根に変換して得られる行列 X に対して特異値分解を実行する。その結果得られる(2)式における U' の行ベクトルを長さ 1 に正規化したものを提案手法の単語概念ベクトルとし、概念語とその概念ベクトルの対の集合を提案手法の単語概念ベースとする。

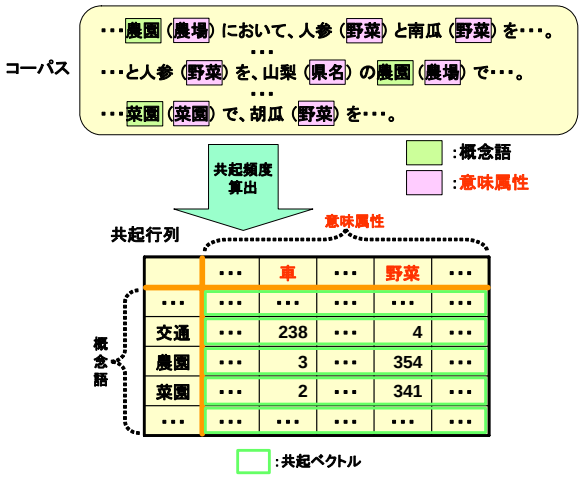

図 2: 提案手法における共起行列

4. 評価実験

従来手法と提案手法の精度比較を、各手法で得られた概念ベクトルを用いた文書検索の精度比較により行った。

単語概念ベース生成用コーパスとして、110,000 個の Q&A 文書と、それを包含する 1,874,553 個の Q&A 文書の 2 つを用いた。名詞、用言等の内容語の異なり数は、それぞれ 85,687、154,195 であった。共起行列のサイズは、従来手法と提案手法とで条件を揃えるようにした。共起行列の行数は 26,900 とし、列数は、2715 とした。共起行列の行数をこのようにとったのは、与えられたメモリ(8GB)内で特異値分解を実行できる行数の上限がこの値であったからである。今回の実験では、形態素解析結果中の一単語の意味属性が複数ある場合は、それらの中で最も使用される局面が高いと思われる意味属性のみを使用することとした。特異値分解により、200 次元の概念ベクトルを生成した。

検索アルゴリズムは、以下のとおりである。各検索対象文書を形態素解析し、名詞、用言等の内容語のみ残す。各検索対象文書において、残った単語の概念ベクトルの和を長さ 1 に正規化したものを、該検索対象文書の概念ベクトルとする。各検索対象文書の概念ベクトルは、あらかじめ生成しておきインデックスに格納しておく。検索キーとなる入力文書に対しても同様の手順で、その概念ベクトルを生成し、入力文書概念ベクトルと各検索対象文書概念ベクトルとの距離の近い順に、検索対象文書集合をランキングして検索結果とする。

あらかじめ一つの検索対象文書と文意が同じで異なる表現の入力文書を作成する。入力文書を検索キーとして検索を実行し、得られた検索結果における、該入力文書に対応する検索対象文書の順位を n としたとき、$1/n$ の平均値(平均逆順位と呼ぶ)を精度の指標とする。

検索対象文書集合としては、単語概念ベース生成用コーパスとは共通部分をもたない 110,000 個の Q&A 文書を用いた。これを用いて 6,068 個の入力文書を作成した。

各手法の、単語概念ベース生成用コーパスごとの平均逆順位は表 1 のようになった。

表 1: 検索精度

手法＼コーパス文書数	110,000 個	1,874,553 個
従来手法	0.307	0.326
提案手法	0.329	0.338

実験で用いた単語概念ベース生成用コーパスの量に関わりなく、提案手法の方が従来手法よりも精度が高いという結果となった。提案手法により単語概念ベクトルの質が向上していることが伺える。

5. まとめ

単語と意味属性との共起頻度から単語の概念ベクトルを生成する手法を提案し、従来の単語・単語間共起に基づく手法と比べ、検索精度が向上することを確認した。今後は、提案手法で生成された単語概念ベースに含まれない単語の概念ベクトルを推定することにより、検索精度のさらなる向上を目指していく。

参考文献

[Shutze 94] H. Schutze, and J.O. Pedersen, A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Proc. RIAO'94, pp.266-274, 1994.

[Kato 99] T. Kato, S. Shimada, M. Kumamoto, and K. Matsuzawa, Idea-Deriving Information Retrieval System, Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193, 1999.

[熊本 99] 熊本睦, 島田茂夫, 加藤恒昭: 概念ベースの情報検索への適用―概念ベースを用いた検索の特性評価, 情報処理学会研究報告, Vol.SIG-ICS 115, pp.9-16, 1999.

[別所 06] 別所克人: クラスタ内変動最小基準に基づくテキストセグメンテーション, 情報処理学会論文誌, Vol.47, No.3, pp.957-967, 2006.

[Shutze 98] H. Schutze, Automatic Word Sense Discrimination, Computational Linguistics, Vol.24, No.1, pp.97-123, 1998.

[池原 97] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: 日本語語彙大系, 岩波書店, 1997.

[Fuchi 98] T. Fuchi, and S. Takagi, Japanese Morphological Analyzer using Word Co-occurence-JTAG, COLING-ACL, pp.409-413, 1998.