単語間の意味的類似度に基づく文書検索手法

大井 耕三　　隅田 英一郎　　飯田 仁
ATR 音声翻訳通信研究所

1 はじめに

近年、膨大な電子化された情報の中から必要な情報を適切に抽出する技術が強く求められている中、類似性に基づく検索技術の研究などが行なわれている [1, 2, 3, 4]。本稿では、(1) 階層的シソーラスに基づく質問中の単語と文書中の単語との間の意味的類似度、(2) 質問中の複合語の各単語に類似している文書中の単語間の出現位置の近さ、(3) 文書内の出現頻度と全文書中の出現文書数に基づく単語の重み、の3つの尺度に基づいた質問－文書間の関連度計算に加え、コーパスに基づく単語の意味的曖昧性解消手法を導入した文書検索手法を提案する。英語の標準的テストセットを使って実験を行なったのでその結果を報告する。

2 検索手法

2.1 階層的シソーラスに基づく単語間の意味的類似度

単語間の意味的類似度は、単語に付与されているシソーラス上の概念の間の類似度により求め、質問中の単語の拡張、および、質問－文書間の関連度の計算に用いる。

本稿の実験では、シソーラスおよび単語辞書として、EDR 電子化辞書 [5] の概念体系辞書と英語単語辞書[1]を使用した。概念体系辞書は、約457,000 概念から成り、各概念が上位－下位の関係により階層的に構成されている。英語単語辞書は約213,000 語から成り、各単語には1つ以上の概念が付与されている。

EDR シソーラスは、ルートの概念から末端の概念までの深さは一定でなく、また、各概念からの下位概念への分岐数も一定ではない。このようなアンバランスなシソーラス上での2概念間の類似度を次のように定義した (図1参照)。この定義は、階層的なシソーラスであればどのようなシソーラスでも利用可能である。

- 各概念に対して、その概念より下位に位置する概念の総数に応じた**レベル**[2] を割り当てる。レベルは、下位に位置する概念ほど値が大きくなるように割り当てる。図ではレベルの数 (NL) は9となっている。
- 概念 AB の間の類似度 Sim は、A と B の相対的な位置関係 (3種類) に応じて、次のように定義する。

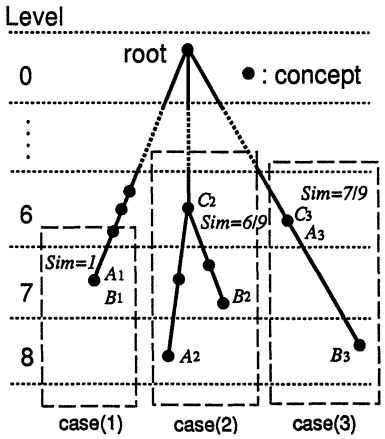

図1: シソーラス上の2概念間の意味的類似度の例 (Sim: 概念 AB の間の類似度, C: A と B の最も近い上位の共通概念)

(1) A と B が同じ場合、$Sim = 1.0$。

(2) A と B の最も近い上位の共通概念 C が A でも B でもない場合、$Sim = LC/NL$。(LC : C のレベル).

(3) $A(B)$ が $B(A)$ の上位概念の場合、$Sim = (LC+1)/NL$。

2.2 類似単語の出現位置の近さ

本手法では、質問中には、単語のほかに複合語 (2単語以上からなるもの) の指定も可能にしている。質問中に複合語が指定された場合は、その複合語中の各単語に類似している文書中の単語の出現位置の近さを、質問－文書間の関連度に反映させる。

出現位置の近さは、例えば、複合語が "parallel algorithm" の場合、"parallel" と "algorithm" それぞれに類似した単語が文書中に出現しているとき、その類似した単語の出現位置が近いほど、質問－文書間の関連度が高くなるように定義する。

出現位置の近さ PN の定義式を次に示す。

$$PN = c_1 \times \frac{1}{\frac{Dis+1-N}{c_2}+1} \qquad (1)$$

ここで、c_1, c_2: 定数 (例: $c_1 = 2$, $c_2 = 10$)。
Dis: 類似単語の出現位置の最小距離。
N: 複合語の単語数。

[1] 評価版第2.1版。

[2] 例えば、総数が6～30の概念にはレベル6を割り当てる。レベルはルート概念からの深さを表す値ではない。

2.3 単語の重み

単語の重みづけに関しては様々な手法が提案されている [6]。本手法では、単語の文書内の出現頻度と全文書中の出現文書数に基づいた重みづけとして [7] で定義された重みづけを使う。文書 D 内の単語 dt の重み w は、次のように定義される。

$$w = \frac{tf}{max\_tf} \times \frac{\log \frac{ND}{f}}{\log ND}$$

ここで、
- tf: 文書 D 内の単語 dt の出現頻度。
- max_tf: 文書 D 内の各単語の出現頻度のうち、最大の出現頻度。
- f: 単語 dt が出現している文書の数。
- ND: 文書数。

2.4 質問－文書間の関連度

2.4.1 質問

本手法では、質問 Q は次に示すようなブーリアンの形式で表現可能なものを前提とする。

$$Q = q_1 \mid q_2 \mid \cdots \mid q_K \quad (2)$$

$$q = \underbrace{qt_{11}, \cdots, qt_{1N_1}}_{\text{質問語 } qc_1} \& \cdots \& \underbrace{qt_{M1}, \cdots, qt_{MN_M}}_{\text{質問語 } qc_M} \quad (3)$$

ここで、'|' と '&' はそれぞれ 'OR' と 'AND' のオペレーションを表す。この形式は、K 個の q が 'OR' オペレーションで結合され、各 q は M 個の質問語 qc(単語または複合語) が 'AND' オペレーションで結合され、各質問語 qc_i は N_i 個の単語 qt から成っている。

2.4.2 関連度

質問 Q 中の 1 つの単語 qt と文書中の 1 つの単語 dt との間の類似度 $Sim(qt, dt)$ は次のように求める。

単語一致 (単語の見出しと品詞が同じ) の場合、$Sim(qt, dt)$ は単語の概念が一致している場合の類似度よりも大きい値に設定する。詳細は 4.2節を参照。単語一致以外の場合、$Sim(qt, dt)$ は、qt の概念と dt の概念との間のすべての組合せにおける 2 概念間の類似度の最大値とする。2 概念間の類似度は 2.1節で述べた定義により求める。

最終的に求める質問 Q と文書 D との間の関連度 $Sim(Q, D)$ は次式のように定義した。

$$Sim(Q, D) = max\{Sim(q_1, D), Sim(q_2, D), \cdots, Sim(q_K, D)\} \quad (4)$$

$$Sim(q, D) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\{Sim(qt_{ij}, D) \times PN(qc_i, D)\}^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} Sim(qt_{ij}, D) \times PN(qc_i, D)} \quad (5)$$

$$Sim(qt, D) = max(w_1, w_2, \cdots, w_L) \times max\{Sim(qt, dt)\}$$

ここで、
- L: qt に最も類似している単語 ($Sim(qt, dt)$ が最大の単語) の数。
- w: qt に最も類似している単語の重み。
- $PN(qc, D)$: 文書 D 内での、qc 中の各 qt に最も類似している単語の出現位置の近さ (2.2節の式 (1))。

$Sim(Q, D)$ の計算は、$q_1, q_2, \cdots, q_K$ のうち、少なくとも 1 つの q に含まれる単語それぞれに対して類似している単語が出現している文書に対してのみ行なう。「類似している単語」とは、$Sim(qt, dt) \geq$ [あらかじめ指定されたしきい値] である単語 dt を意味する。

3 意味的曖昧性解消手法の導入

多くの単語は、文脈を考慮しない場合にはいくつかの意味を持っている。本手法では単語間の類似度を求める際にすべての意味 (概念) を考慮しているが、実際に文書中で使われている意味を同定することができれば、検索の精度はかなり向上することが期待できる。そこで、単語の意味的曖昧性をコーパスに基づいて解消する手法を導入することにした。

3.1 Voorhees の手法の適用

Voorhees は、単語の意味を階層的に構成した WordNet というシソーラスを使って意味的曖昧性解消を行ない、文書検索の実験を行なっている [8]。

文書 D 中の単語 dt の意味の曖昧性解消手法の概要は次のとおりである。単語 dt の各意味に対して、$hood$ (その意味を含み他の意味を含まない最も上位の意味) を求めた後、式 (6) の差異を求め、差異が最大である意味を単語 dt の意味として選択する。

$$\text{差異} = \frac{hood\text{下の意味を持つ内容語の文書}D\text{内出現数}}{\text{内容語の文書}d\text{内出現数}} - \frac{hood\text{下の意味を持つ内容語の全文書内出現数}}{\text{内容語の全文書内出現数}} \quad (6)$$

この手法を我々の検索手法に適用するに際し、次の変更を加えた。(1) 差異が正である概念 (意味) を単語 dt の概念として選択する。(2)$hood$ は最も上位の概念でなく原則としてあるレベル (2.1節参照) 以上の概念に制限する (抽象的すぎる概念まで含むのを避けるため)。

3.2 Yarowsky の手法の適用

Yarowsky は、Roget の階層的シソーラス (カテゴリ数:1024) を用いて曖昧性解消の実験を行なっている [9]。

文書中の単語 dt の意味の曖昧性解消手法の概要は次のとおりである。コーパス中の各単語に対して、前後 50 語ずつ合計 100 語を含む文脈を抽出し、単語 dt が属するシソーラス中のカテゴリ RCat 毎に、単語 dt の文

脈中の単語のうち式(8)がある値以上の単語 ct に対して、式(7)のスコアを求め、最大のスコアとなるカテゴリを単語 dt の意味としている。

$$Score = \sum_{ct} \log \frac{Pr(ct|RCat)}{Pr(ct)} \quad (7)$$

$$\frac{Pr(ct|RCat)}{Pr(ct)} = \frac{RCatに属する単語の文脈中の単語ctの出現確率}{単語ctの全文脈中の出現確率} \quad (8)$$

この手法を我々の検索手法に適用するに際し、次の変更を加えた。(1)Rogetのカテゴリを Voorheesの手法における $hood$ 下の概念に置き換える[3]。(2)前後4語ずつ合計8語を文脈とする[4]。(3)スコアが正である概念を選択する。また、次節の実験ではコーパスとして検索対象文書を用いた。

4 実験

4.1 標準的テストセット

英語の標準的な評価用のテストセットの1つに Fox[10]が作成したものがある。実験では、その中の CACM と呼ばれるセットを使った。CACM には、コンピュータサイエンスに関する 3,204 の文書(タイトル, アブストラクト)、3種類の質問セット(自然言語文からなる NLQ, ブーリアン形式の BLQ1, BLQ2)、質問ごとの関連文書の文書番号が含まれている。各質問セットには64個の質問が含まれている。NLQ はオリジナルの質問セットで、BLQ1, BLQ2 は NLQ を基にして作られている。

実験では、NLQ の質問文中の単語が比較的多く使われている BLQ2 の中で、NOT オペレーションを含むものと正解の関連文書がないものを除く47個の質問に変更を加えたもの(以下では、変更版 BLQ2 と呼ぶ)を用いた。変更版 BLQ2 は、複合語と考えられる部分を複合語として指定したものである。BLQ2, 変更版 BLQ2 の質問番号35の質問内容を下に示す。変更版 BLQ2 において、シングルクウォート(')で囲まれた部分が1つの質問語を表している。

[BLQ2]

```
#q35= #and( 'probabilistic', 'algorithm',
            #or( 'algebraic', 'symbolic'),
            'manipulation');
```

[変更版 BLQ2]

```
#q35= #and( 'probabilistic algorithm',
            #or( 'algebraic manipulation',
                 'symbolic manipulation'));
```

[3]Roget のカテゴリを EDR の概念にそのまま置き換えたとすると、EDR の概念はそのほとんどが非常に具体的な概念であるため、各概念に属する単語の文脈数は非常にスパースとなる。

[4]文脈幅を変えて実験を行なったが、8語以上の場合、結果に大きな差は生じなかった。

4.2 検索処理

実験では、シソーラス上の概念のレベルの数 NL (2.1節参照)を9とし、単語一致の時の $Sim(qt, dt)$ は単語の概念が一致する時の値(1.0)よりも大きくなるように $10/9(=(NL+1)/NL)$ とした。

実際の検索処理に先だって、(a)質問中の単語への概念の付与(手作業)、(b)全文書に対するインデックスファイル(各文書の各見出し語・品詞のペアに対して、文書番号、重み、出現位置が付与されたファイル)の作成、を行なった。

変更版 BLQ2 の各質問に対する検索処理手順は次のとおり。

(1) 質問を 2.4.1節の式(2),(3)の形に変換する。
(2) 式(2)の少なくとも1つの q 中の単語 qt それぞれに対して $Sim(qt, dt) \geq T$(しきい値)を満たす単語 dt が出現している文書を検索する。
(3) 検索された文書毎に、質問－文書間の関連度 $Sim(Q, D)$ を 2.4.2節の式(4)から求める。
(4) 検索された文書を関連度順にランクキングする。

4.3 評価方法

評価は、情報検索の分野で一般的によく使われている再現率(recall)と適合率(precision)を用いた。再現率および適合率は次式で定義される。

$$再現率 = \frac{ランクN位までの検索文書中の関連文書数}{関連文書数} \quad (9)$$

$$適合率 = \frac{ランクN位までの検索文書中の関連文書数}{N} \quad (10)$$

再現率－適合率のグラフは次に示す手順で作成した。

[再現率－適合率グラフの作成手順]

(1) 式(9),(10)の値 N を任意に複数個決める。
(2) 質問ごとに、式(9),(10)により N における再現率と適合率を求める。
(3) N における再現率と適合率の全質問に対する平均を求め、プロットする。

次節の結果では、N を $10, 20, 30, \cdots, 200$ に設定し、次の4種類の検索結果を比較した。このうち WM を比較の基準とした。

WM: 単語一致(T=10/9)+単語の重み(2.3節)。2.4.2節の式(5)の $PN(qc_i, D)$ は1とする。
AM: 2節の検索手法(意味的曖昧性解消なし)。
DM(V): 2節の検索手法+Voorhees の手法。
DM(Y): 2節の検索手法+Yarowsky の手法。

4.4 結果

図2に、しきい値T(2.4.2節参照)が8/9の場合の再現率－適合率グラフを示す。AMの結果は、WMと比べて再現率は向上する一方、適合率は減少している。この原因の1つに単語の多義性がある。単語間の類似度を求める際にすべての概念を考慮しているので、実際に文書中で使われている概念とは異なる概念のために多くの非関連文書が検索されてしまう。

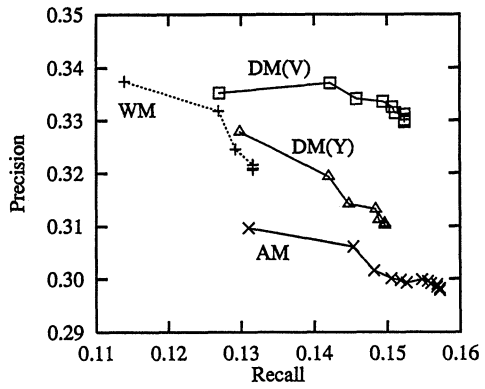

図2: 再現率(recall)－適合率(precision)[T=8/9]

意味的曖昧性解消手法を導入したDM(V)とDM(Y)は、AMと比べた場合、再現率はわずかに減少しているものの、適合率は明らかに向上している。このことは、AMで検索されていた関連文書がDMでは曖昧性解消の失敗により検索されなくなった場合がわずかに生じているものの、それ以上にAMで検索されていた非関連文書がDMでは曖昧性解消の効果により検索されなくなっていることを示している。

一方、WMと比べた場合でも、DM(V)は再現率・適合率ともほぼ向上している。DM(Y)は再現率が向上する一方、適合率が減少しているが、ある再現率(例えば0.13あたり)で比較すると適合率が向上しているのが分かる。

DM(V)とDM(Y)ではDM(V)の方が優れているが、DM(Y)での曖昧性解消に用いたコーパスは検索対象文書(延べ単語数: 約17万語)であり、より大規模なコーパスを用いることにより精度の向上が予想される。

5 関連研究

シソーラスを使った検索手法に関して、Rada[11], 青山[3],Paice[12]などの研究がある。

Radaと青山は、どちらもシソーラス上の概念間の類似度を使っている。しかしながら、Radaはシソーラス上の概念を個々の質問と文書に対して手作業で付与し、文書や質問中の単語に付与していない。よって我々の手法のように質問中の単語に類似した単語を含む文書を検索することはできない。青山は単語に付与された概念間の類似度を、入力文に類似した文の検索に応用している。

Paiceは意味ネットワークからなるシソーラスを使っている。質問単語に対応するシソーラス中のノードから一定の距離内にあるノード(単語)に重み付けを行ない、それを拡張単語として検索を行なうことを提案しているが、実験結果は示されていない。

6 おわりに

階層的シソーラスに基づく単語間の意味的類似度、類似単語の出現位置の近さ、単語の重み、の3つの尺度に基づいた質問－文書間の関連度計算に加え、コーパスに基づく単語の意味的曖昧性解消手法を導入した文書検索手法を提案した。英語の標準的テストセットを使った実験で良好な結果を確認した。

謝辞 本研究ではEDR電子化辞書の英語単語辞書・概念辞書(評価版第2.1版)を使用している。関係各位に深謝する。

参考文献

[1] 隅田, 堤: "翻訳支援のための類似用例の実用的検索法", 信学論, J74-D-II(10), 1991.

[2] Sato, S.: "CTM: An Example-Based Translation Aid System," Proc. of COLING'92, 1992.

[3] 青山, 他: "形態素解析と意味コード化に基づく翻訳支援のための類似例文検索システム", 電子情報通信学会NLC93-68, 1994.

[4] 美馬, 他: "類似検索を用いた情報検索システム", 言語処理学会第2回年次大会, A5-3, 1996.

[5] 日本電子化辞書研究所: EDR電子化辞書仕様説明書, 1993.

[6] Salton, G. and Buckley, C.: "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, 24(5), 1988.

[7] Turtle, H. and Croft W.B.: "Evaluation of an Inference Network-Based Retrieval Model," ACM Transactions on Information Systems, 9(3), 1991.

[8] Voorhees, E.M.: "Using WordNet to Disambiguate Word Senses for Text Retrieval," Proc. of SIGIR'93, 1993.

[9] Yarowsky, D.: "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," Proc. of COLING'92, 1992.

[10] Fox, E.: "Virginia Disk One," Virginia Polytechnic Institute and State University Press, 1990.

[11] Rada, R., et al.: "A Knowledge-Base for Retrieval Evaluation," Annual Proc. of the ACM'85, 1985.

[12] Paice, C.D.: "A Thesaural Model of Information Retrieval," Information Processing & Management, 27(5), 1991.