用語間の関連度を測る指標の提案

佐々木　靖弘　　佐藤 理史　　宇津呂　武仁

京都大学大学院 情報学研究科
sasaki@pine.kuee.kyoto-u.ac.jp, {sato, utsuro}@i.kyoto-u.ac.jp

1. はじめに

「ある用語を知る」ということは、その用語が何を意味し、どのような概念を表すかを知ることである。それと同時に、その用語が他のどのような用語と関連があるのかを知ることは非常に重要である。特定の専門分野で使われる用語 (専門用語) は、その分野の中で孤立した用語として存在することはない。その分野で使われる他の用語に支えられ、その関連を土台として、始めて意味を持つ。それらの用語間の関連を把握することは、「その専門分野について知る」ことでもある。

我々はこれまで、ウェブを利用して与えられた用語の関連用語を見つけ出す「関連用語収集システム」を実現してきた [1)~3)]。このシステムでは、ウェブのサーチエンジンのヒット数を利用して 2 つの用語間の関連度を測定していた。しかし、この関連度には、ウェブの大きさの変化に対応できないという問題点があった。本稿では、この問題に対応した新たな関連度を提案し、その有用性について検討する。

2. 用語間の関連度の測定

2.1 従来の指標

関連用語収集システムにシードワード s を入力したときに、システムが収集する関連用語 x の満たすべき条件として、我々は以下の 2 つの条件を考えた [1)]。

条件 1　x は専門用語である。

条件 2　x は s と関連する。

x が条件 1 を満たすかをチェックするための手がかりとして、x のサーチエンジンにおけるヒット数 $H(x)$ を利用した。x が専門用語として使われているならば、x は特定の分野においては広く、または、それなりに使われているが、一般語ほど頻繁には使われていないはずである。サーチエンジン goo* で専門用語および一般語のヒット数を調査したところ、専門用語のヒット数が 100 件から 10 万件の間に分布しているという結果が得られた。よって、goo でのヒット数 $H(x)$ が 100 件未満および 10 万件以上の用語 x は専門用語ではないと判断し、除外した。

また、条件 2 をチェックするための手がかりとして、以下の 2 つの関連度 $R_{s\to x}$、$R_{x\to s}$ を定義し、そのいずれかが十分大きければ、x は s と関連する、とした。

$$R_{s\to x} = \frac{H(s \wedge x)}{H(s)}$$

$$R_{x\to s} = \frac{H(s \wedge x)}{H(x)}$$

ここで、$H(s \wedge x)$ は s と x の AND 検索のヒット数である。

2.2 従来の指標の問題点とその対策

従来の指標の問題点は、条件 1 をチェックするためのヒット数の閾値として、100 件や 10 万件という絶対値を用いていることである。ウェブは日々更新され、その大きさも刻々と変化する。よって、ある時点の観測結果から閾値を決定してしまうのは問題である。

そこで、次式で表される新しい関連度 $R_{\wedge/\vee}$ を定義した。

$$R_{\wedge/\vee} = \frac{H(s \wedge x)}{H(s \vee x)}$$

ここで、$H(s \vee x)$ は s と x の OR 検索のヒット数である。この関連度を用いることにより、条件 2 の「x が s と関連するか」がチェックできるだけでなく、条件 1 のチェックも可能となる。

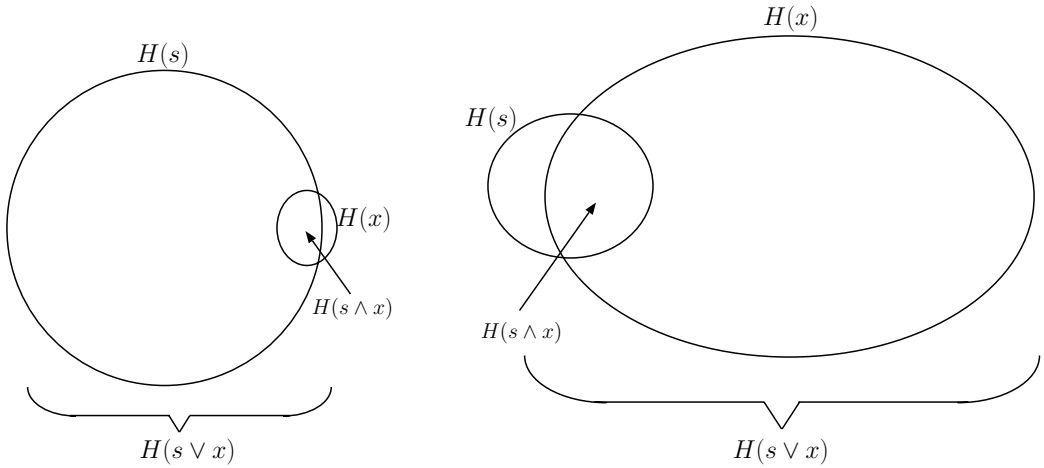
用語 x が専門用語として使われているとは、x が特定の分野において広く、または、それなりに使われているが、一般語ほど頻繁には使われていない、と考えた。もし、x が広く使われていないならば、すなわち、x のヒット数 $H(x)$ が小さいならば、$R_{\wedge/\vee}$ の分子 $H(s \wedge x)$ も小さくなり、$R_{\wedge/\vee}$ の値は小さくなる (**図 1(a)**)。一方、x が頻繁に使われているならば、すなわち、$H(x)$ が大きいならば、$R_{\wedge/\vee}$ の分母 $H(s \vee x)$ も大きくなり、$R_{\wedge/\vee}$ の値はやはり小さくなる (**図 1(b)**)。よって、条件 1 を満たすには、$R_{\wedge/\vee}$ が十分に大きければよい、ということになる。

以上のことから、関連度 $R_{\wedge/\vee}$ を用いることにより条件 1 と条件 2 を同時にチェックすることができる。$R_{\wedge/\vee}$ は s と x のヒット数から相対的に決定されるため、ウェブの大きさに依存しない。また、1 つの指標で評価できるため、システムをシンプルに構成することができる。

3. 関連用語収集システム

指標 $R_{\wedge/\vee}$ を用いて関連用語収集システムを構成した。本システムは、ウェブのサーチエンジンを利用して、システムに入力された用語 s に関連する用語集合 T を収集

* http://www.goo.ne.jp/



(a) $H(x)$ が小さいと $R_{\wedge/\vee}$ は小さくなる　　(b) $H(x)$ が大きいと $R_{\wedge/\vee}$ は小さくなる

図 1　$H(x)$ の大きさにより $R_{\wedge/\vee} = \frac{H(s \wedge x)}{H(s \vee x)}$ は変化する

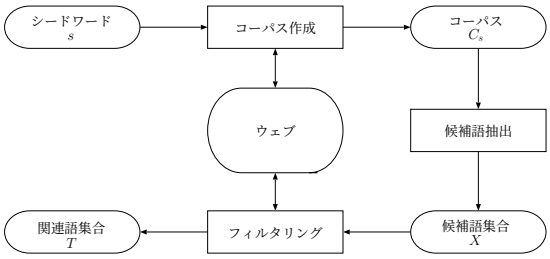


図 2　システム構成

するシステムである。本システムは、(1) コーパス作成、(2) 候補語抽出、(3) フィルタリング、の 3 つのモジュールから構成される。本システムの構成図を**図 2** に示す。

以前のシステムとの変更点は、候補語抽出の際に行っていた重要語抽出をやめたことと、フィルタリングにおいて $R_{\wedge/\vee}$ を用いるようにしたことである。

3.1　コーパス作成

本システムでは、用語 s を与える他は、コーパスを事前に用意しておくということは行わない。よって、関連用語収集の第 1 ステップとして、与えられた用語 s に関するコーパス C_s を作成する必要がある。

本システムでは、ウェブを利用した次のような方法でコーパスを作成する。

(1) ウェブページの収集: 与えられた用語 s に対して、「s とは」「s という」「s は」「s」という 4 種類のクエリを検索エンジンに入力し、得られた URL のそれぞれ上位 100 ページを入手する。さらに、それらのページに、用語 s がアンカーテキストとなっているアンカーが存在する場合は、そのアンカー先ページも入手する。

(2) 文の抽出: それぞれのページを整形して文に分割し、用語 s を (文字列として) 含む文のみを抽出し、コーパス C_s を作成する。

サーチエンジンとしては、goo を利用している。文抽出では、用語 s を含む文のみを抽出しているが、前後 n 文を一緒に抽出するという方法も考えられる。

3.2　候補語抽出

関連用語収集の第 2 ステップは、第 1 ステップで作成したコーパスから s の関連用語の候補となる語を抽出する処理である。

ここでは、複合名詞を候補語とする。コーパス C_s 中のそれぞれの文の文節を認識して、名詞を 2 つ以上含む文節を取りだし、その主要部 (付属語等を除いたもの) を候補語とする。

今回は複合名詞を候補語としたが、入力用語のタイプに合わせて候補語のタイプを変化させてもよい。例えば、ある企業のイメージを知りたいときに、入力用語を企業名とし、関連用語の候補語を形容詞とすると、その企業がどのような形容詞と関連するかを調べることができる。

3.3　フィルタリング

ステップ 2 で得られた候補語に対し、2.2 節で定義した $R_{\wedge/\vee}$ を用いてフィルタリングを行う。サーチエンジン goo を用いて、s のヒット数 $H(s)$、x のヒット数 $H(x)$、s と x の AND 検索のヒット数 $H(s \wedge x)$ を求め、$R_{\wedge/\vee}$ を計算する。ここで、OR 検索のヒット数 $H(s \vee x)$ は以下の式で計算する。

$$H(s \vee x) = H(s) + H(x) - H(s \wedge x)$$

表 1 に、「自然言語処理」をシステムに入力したときに得られた候補語 667 語の内 $R_{\wedge/\vee}$ 上位 30 語を示す。

4.　実　　験

4.1　関連用語収集実験

作成したシステムを用いて、関連用語収集実験を行った。「自然言語処理」、「日本語」、「IT」の 3 つの分野か

表 1 s =「自然言語処理」としたときに得られる関連用語 ($H(s) = 7091$)

$R_{\wedge/\vee}$	x	$H(x)$	$H(s \wedge x)$	$H(s \vee x)$
0.454	自然言語	15416	7031	15476
0.098	自然言語処理技術	683	696	7078
0.095	形態素解析	3365	903	9553
0.067	コーパス	6970	880	13181
0.066	言語処理学会	1150	512	7729
0.065	機械翻訳	9134	997	15228
0.064	自然言語処理研究会	446	452	7085
0.057	構文解析	7026	761	13356
0.049	知識表現	2048	423	8716
0.045	対話システム	2117	397	8811
0.043	機械学習	1939	370	8660
0.041	自然言語処理学	300	294	7097
0.041	意味解析	1428	333	8186
0.040	知識処理	2577	369	9299
0.038	認知科学	13078	732	19437
0.037	言語理解	2762	354	9499
0.035	形態素	2606	332	9365
0.035	人工知能	42448	1692	47847
0.034	情報処理学会	28653	1189	34555
0.030	自然言語処理システム	217	216	7092
0.030	機械翻訳システム	1021	237	7875
0.029	情報抽出	2322	262	9151
0.027	マルチモーダル	3414	276	10229
0.026	形態素解析システム	604	197	7498
0.025	意味論	9755	418	16428
0.024	電子情報通信学会	26166	794	32463
0.024	言語情報	4633	276	11448
0.022	論文誌	18882	565	25408
0.022	音声言語処理	658	165	7584
0.021	自動要約	701	162	7630

表 2 実験結果

入力用語	候補語総数	$X_{R>0.01}$*	正解数	精度 [%]
自然言語処理	667	71	59	83
有限オートマトン	415	50	47	94
構文解析	1021	67	53	79
形態素解析	859	74	71	95
意味解析	334	46	39	84
文脈解析	221	24	21	87
格文法	144	33	33	100
照応関係	267	7	7	100
情報検索	1254	54	26	48
機械翻訳	1171	51	47	92
シソーラス	1106	47	39	82
漢字制限	423	41	36	87
形態論	734	36	21	58
言語教育	774	37	33	89
言語行動	383	21	20	95
言語遊戯	161	3	2	66
構文論	270	24	15	62
混種語	76	19	12	63
常用漢字	756	48	41	85
類義語	444	22	18	81
cdmaOne	1227	74	58	78
Ethernet	1555	46	42	91
QC7 つ道具	157	18	16	88
グローバルアドレス	406	36	29	80
工業所有権	971	82	52	63
字句解析	318	35	32	91
新 QC7 つ道具	113	9	5	55
バッチファイル	645	45	31	68
フラクタル	1607	72	68	94
論理演算	488	55	50	90
合計	18967	1247	1023	82

*$X_{R>0.01}$: $R_{\wedge/\vee}$ が 0.01 以上の候補語数

らそれぞれ 10 語ずつ、計 30 語を入力としてシステムに与え、得られた出力のうち関連度 $R_{\wedge/\vee}$ が 0.01 以上の用語に対して、入力用語の関連用語として適切であるかどうかを人手で評価した。結果を**表 2** に示す。入力の 30 語に対して $R_{\wedge/\vee}$ が 0.01 以上の用語は 1247 語収集され、1023 語 (82%) が入力の関連用語として適切であった。

4.2 $R_{\wedge/\vee}$ の適切さ

$R_{\wedge/\vee}$ が関連度の指標として適切であるならば、$R_{\wedge/\vee}$ が高い用語ほど入力の用語に関連する確率が高くなければならない。$R_{\wedge/\vee}$ の適切さを確かめるために、$R_{\wedge/\vee}$ と精度の関係を調査した。

$R_{\wedge/\vee}$ が上位の用語から見ていったときの、正解用語数を横軸、それぞれの入力用語に対する非補間平均精度 [4] を縦軸としたときのグラフを**図 3** の実線で示す。

図 3 より、グラフが右肩下がりであることから、$R_{\wedge/\vee}$ が高い用語ほど、入力用語の関連用語である可能性が高いことが分かる。よって、$R_{\wedge/\vee}$ は関連度の指標として妥当であると考えられる。

4.3 考　　察

表 2 を見ると、入力用語によって、精度が極端に低かったり、収集された関連用語数が少なかったりする。この理由としては、以下の 3 つの理由が考えられる。

理由 1: 入力用語が表す概念が相対的に大きい

例えば『情報検索 』という用語は、インターネットや検索エンジンなどの広がりにより、一般的な概念となっている。このような語が入力された場合、$H(s)$ が大きくなるため*、$R_{\wedge/\vee}$ は小さくなる (cf. 図 1(a))。

理由 2: 入力用語が表す概念が相対的に小さい

『混種語』とは、和語と漢語、和語と外来語、あるいは漢語と外来語が組み合わされた用語である。しかし、それらの用語に比べて『混種語』は (少なくともウェブの世界では) マイナーな用語である。よって、それらの用語のヒット数に対して $H(s)$ が小さいため**、$R_{\wedge/\vee}$ は小さくなる (cf. 図 1(b))。

理由 3: 入力用語が複数の概念を持つ

例えば『形態論』は語の形態に関する用語であるが、経済学の分野に『価値形態論』や『企業形態論』という用語があり、サーチエンジンではこれらを区別することができない。よって、このような複数の概念を持つような用語を入力すると、うまく関連用語を収集することができない。

4.4 関 連 研 究

自然言語処理において、二言語間の対応推定やコロケーションの抽出などに、*Dice* の相関係数や χ^2 統計量など

* 本実験時における『情報検索』の goo でのヒット数は約 14 万件であった。

** 本実験時における『混種語』の goo でのヒット数は 54 件であった。

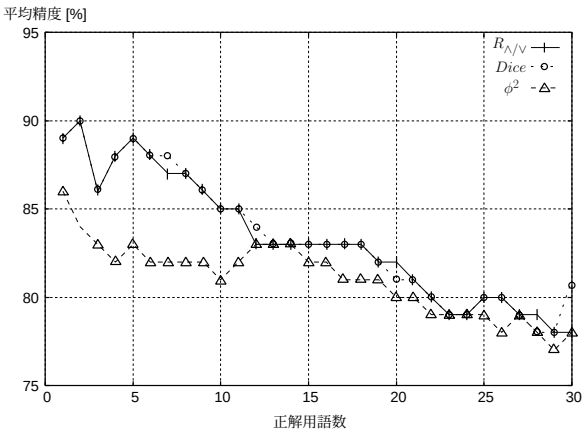


図 3 正解用語数 - 平均精度

の、共起情報を利用した多くの統計的指標が利用されている [5]。ここでは、$R_{\wedge/\vee}$ とそれらの指標との比較を行う。

表 3 のような 2×2 分割表を考えると、本実験では a、b、c はそれぞれ

$$a = H(s \wedge x)$$
$$b = H(x) - H(s \wedge x)$$
$$c = H(s) - H(s \wedge x)$$
$$a + b + c = H(s \vee x)$$

となる。また、s も x も含まないウェブページ数 d は一般に非常に大きいと考えられる*。すなわち、

$$d \gg a, b, c$$

である。

このとき、*Dice* 係数は、

$$Dice = \frac{2a}{2a+b+c} = \frac{2H(s \wedge x)}{H(s \wedge x) + H(s \vee x)}$$

となり、$R_{\wedge/\vee}(= \frac{H(s \wedge x)}{H(s \vee x)})$ の分子・分母にそれぞれ $H(s \wedge x)$ を加えたものとなる。

また、χ^2 統計量に用いられる ϕ^2 係数**は、

$$\begin{aligned}
\phi^2 &= \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \\
&= \frac{(ad-bc)^2/d^2}{(a+b)(a+c)(b+d)(c+d)/d^2} \\
&= \frac{(a-\frac{bc}{d})^2}{(a+b)(a+c)(\frac{b}{d}+1)(\frac{c}{d}+1)} \\
&\simeq \frac{a^2}{(a+b)(a+c)} \quad (\because d \gg a, b, c) \\
&= \frac{H(s \wedge x)}{H(s)} \frac{H(s \wedge x)}{H(x)} \\
&= R_{s \to x} R_{x \to s}
\end{aligned}$$

となり、以前の関連度 $R_{s \to x}$ と $R_{x \to s}$ を掛け合わせたものとなる。

30 語の入力用語に対して、$R_{\wedge/\vee}$ の場合と同様にして、

* goo で検索可能なウェブページ数は約 33 億件 (日本語以外の言語を含む。2004 年 2 月現在)

** $\chi^2 = n\phi^2$ (n は総標本数)

表 3 2×2 分割表

	x を含む	x を含まない
s を含む	a	b
s を含まない	c	d

Dice 係数と ϕ^2 係数の、正解用語数と平均精度のグラフを図 3 の点線 ($\circ$: *Dice*、$\triangle$: ϕ^2) で示す。これを見ると、*Dice* 係数の結果は $R_{\wedge/\vee}$ とあまり変わらず、ϕ^2 係数の結果は、$R_{\wedge/\vee}$ や *Dice* に比べて精度が悪くなっている。

この原因として、ϕ^2 係数では、『自然言語処理関係』や『処理』といった用語がスコア上位に現れていることが上げられる。

『自然言語処理関係』は、『自然言語処理』という専門用語に『関係』という語が付加されて、専門用語性が薄れたにも関わらず、語の中に『自然言語処理』が含まれるために共起頻度が高くなっている用語である。

また『処理』は一般的な用語であるが、『自然言語処理』に『処理』が含まれているため、これもやはり共起頻度が大きくなっている。

このような用語の場合、ϕ^2 係数では分子において共起頻度が二乗されているため、値が大きくなってしまうが、$R_{\wedge/\vee}$ では、値が大きくなるのをある程度抑えることができる。よって、2 節で議論したように、$R_{\wedge/\vee}$ は関連度のチェックと共に、専門用語性のチェックも兼ねた指標となっている。

参 考 文 献

1) 佐藤理史, 佐々木靖弘: ウェブを利用した関連用語の自動収集, 情報処理学会研究報告 NL-153-8, pp. 57–64 (2003).

2) 佐々木靖弘, 佐藤理史: ウェブを利用した関連用語の自動収集, 言語処理学会第 9 回年次大会発表論文集, pp. 278–281 (2003).

3) Sato, S. and Sasaki, Y.: Automatic Collection of Related Terms from the Web, *ACL-03 Companion Volume to the Proceedings of the Conference*, pp. 121–124 (2003).

4) 徳永武伸: 情報検索と言語処理, 東京大学出版会 (1999).

5) Manning, C. D. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (2002).