E-028

発話のための Web を用いた背景的知識の構築手法

The Approach to Creating Web-based Background Knowledge for Conversation

林 輝大† 奥村 紀之‡
Kidai Hayashi Noriyuki Okumura

1．はじめに

計算機の飛躍的な発達により，世の中はいよいよ便利になってきた．しかしながら計算機に搭載されているシステムは複雑化し，操作には高い技術が求められるようになってきた．そのため，より使いやすく，より便利にシステムを操作するためには，人間と計算機が互いにコミュニケーションを図ることが重要と考える．

そこで我々は，人間が会話をする際に用いている大量の経験や知識を「背景的知識」として定義し，その背景的知識を用いることで自然な会話が実現で可能であると推定し，背景的知識について実験を行い評価している[1].

本研究では，計算機と人間との対話が十分にできないために不足しがちな背景的知識を，Web を通じて自動的に構築することにより，常に新しく豊富な知識で充実させることを目的とする．具体的な背景的知識としてキーワードと話題語を定め，*tf*・*idf* と相関係数を用い Web から収集したデータをクラスタリングすることにより， 知識の獲得を自動で行う．同じクラスタと類別する際に，基準となる相関係数の値により結果が異なるため，最適となる相関係数の閾値を実験により定めている．

2. 関連研究

我々が会話をする際は，ある話題を掘下げ，関連のある話題に発展させることがある．自然な会話が，話者間の話題の自然な繋がりであるとするならば，背景的知識は話題の繋がりを記録したものであると考えることができる．単語と単語の関連の度合いを計算するためのデータベース(概念ベース)を構築する研究を奥村らが行っている[2].

また，Web からの属性獲得方法については，波多腰らが時間的要因に注目した手法を提案している[3].

クラスタリングの類似手法として，語の共起情報を基に文書内からのキーワード抽出を松尾らは行っている.[4]

3. 提案手法

本研究では，背景的知識の自動構築を目的とし，その手法について以下に詳細を記す.

3.1 背景的知識

我々が会話を行う際，話題や状況に合わせ自らの持つ知識を参照する．その知識は，時に人物のプロフィールであったり，または体験した思い出であったり，時間や場所に関する情報であったりと多岐に渡る．この会話の背景に存在する膨大な知識のことを本研究では「背景的知識」と呼称する．この背景的知識を計算機に蓄積させ，また会話の際に参照することにより，自然な流れの会話が可能になると推測する．

† 長野工業高等専門学校 専攻科 電気情報システム専攻
‡ 長野工業高等専門学校 電子情報工学科

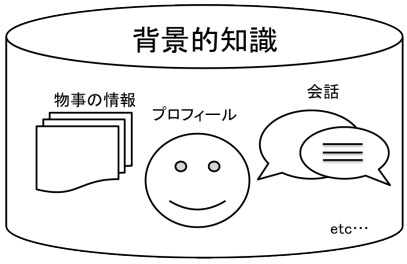

図 1 背景的知識のイメージ

3.2 背景的知識の自動構築

背景的知識は本来ならば計算機と人間との対話で蓄えられるものである．しかしながら対話は時間と労力を大量に必要とするため，簡単には背景的知識を蓄積することができない.そこで Web に溢れる情報を利用し，普段我々が知らないことを Web で検索するように，大量の情報を背景的知識として獲得することで，より豊かな話題提供が可能なのではないかと考える．

以下は背景的知識の自動構築を可能とするための手法を提案していく．

3.2.1 背景的知識の定義

ここでは背景的知識の具体的な定義を行う．まず，語句の混同を防ぐために以下の二つを定義する．

(1) キーワード
会話の中で特徴的な単語

(2) 話題語
キーワードから連想される話題候補の単語

話題語は話題の移り変わりによってキーワードと成り得る．背景的知識はこれら 2 種類の要素からなり，キーワードと話題語の話題として繋がりの強さを記録する．また，構造としては図のようにキーワードと話題語は一対多の関係にあり，ここで言う単語とは名詞・形容詞のことである．

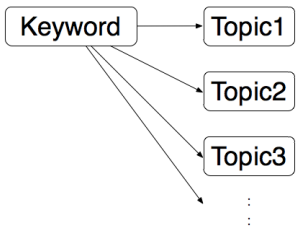

図 2 キーワードと話題語の関係

3.2.2 Web からの話題語獲得 [3]

検索エンジンサイト Google で話題語を検索ワードとして検索を行い，検索結果の上位 100 件すべての Web ページの内容を取得する．取得した Web ページの HTML タグを除去した後，形態素解析を行う．この際，獲得する話題語の精度を向上させるため，ルールに従って形態素解析結果の補正を行う．その後得られた語に対して *tf*･ *idf*

による重み付け, 重み順にソートを行い, 上位の話題語を結果として獲得する.

3.2.3 tf・idf による重み付け

重みには, 局所的重み (local weight) と大域的重み (global weight) がある. 局所的重みは索引語頻度 (*tf*) とも呼ばれ,語の出現頻度が使われる. *tf* では, 文書において出現回数が多い語ほど大きな重みを与えられる. 大域的重みは文書集合全体を考慮しているものであり, 文書頻度逆数 (*idf*) とも呼ばれる. *idf* は, すべての文書に現れる語よりも, 一部の文書のみに現れる語の方が重要であるという考え方である.

情報検索における重み付け手法の 1 つとして幅広く使われている *tf*・*idf* は前述の *tf* と *idf* の積 により与えられる重みである. *tf*・*idf* は, 次の様に定義される.

$$w(t,d) = tf(t,d) \cdot idf(t) \quad \text{(式. 1)}$$

tf(t,d) は文書 *d* における索引語 *t* の出現頻度を表す. (式 .1)における *idf (t)* は検索対象文書数 *N* と,索引語 *t* が出現する文書の数 *df* (*t*) によって(式 .2) のように定義される.

$$idf(t) = \log_2 \frac{N}{df(t)} + 1 \quad \text{(式. 2)}$$

Web を用いた話題語獲得では, 重み付けの際に *Web* − *idf* と *SWeb* − *idf* の 2 種類の *idf* を使用している.以下,これらについて詳細を述べる.

3.2.4 Web - idf

Web – *idf* (Web Inverse Document Frequency) は Web 上での語の大局的な重みである. *idf* の式は (式.2) を用い, 検索対象文書数 *N* を Google が保有している日本語のページ数, 索引語 *t* が出現する文書の数 *d* は, 索引語 *t* を Google で検索を行った時のヒット件数とする.

なお, Google が保有している日本語のページ数は公開されていないため, 日本語の文書の中で最も使用されている「は」で検索を行ったヒット件 数 1,560,000,000 件 (2010 年 7 月 15 日現在) とする.

3.2.5 SWeb - idf

Web − *idf* は Web 上での語の大局的な重みを知ることができるが, Web 上の文章の中で頻出している語を知ることができない. そこで Web 上で統計的に調べた *idf* 値である *SWeb* − *idf* を定義 する.

方法としては, まず偏りなく無差別に選んだ固有名詞のリストを作成する. 全ての語において Google で検索を行い, Web ページを取得し, 得られたページ数を検索文書の数 *N* とする. *idf* の式は *Web* − *idf* と同様に(式.2)を用いる. 索引語 *t* は取得した文書を形態素解析後にルール により補正した語とし, 索引語 *t* が出現する文書の数 *df(t)*とする.

3.2.6 重み付けの手順

Web − *idf* と *SWeb* − *idf* を用いた重み付けの手順について述べる. データベースを参照し, 取得した属性の語が *SWeb* − *idf* に登録されていれば, *idf* 値を *SWeb* − *idf* とする. 登録されていなければ *Web*−*idf* を参照し, 登録されていれば *idf* 値を *Web*−*idf* とする. *SWeb*−*idf* と *Web* − *idf* ともに登録されていなかった属性については, 重みは *SWeb* − *idf* データベースの *idf* の最大値 (max = 22.141976…) とする. *Max* − *idf* とは, *SWeb* − *idf* の最大値の *idf* のことである.

3.2.7 話題語と時間的要素

先述してきた手法により重み付けをした話題語を Web から獲得することは可能である. その重みを利用して, キーワードと話題語の関連の基準としてもよいが, ここで考慮したい点がある. 例えば, 「某芸能人が結婚した」というニュースが大々的に取り上げられた場合, 「結婚」というキーワードのもとに「芸能人の名前」と「結婚相手の名前」とは一時的に話題として関連を持つことになる. また「衣類」をキーワードにした際, 春ならば「春物」, 夏ならば「夏物」というワードがその時々により関連が強くなると考えることができる. つまり, 話題には時間的要素が絡んでいると推測される. 単純な語と語の関連ならば求めた重みを用いればよいが, 本研究では話題の繋がりを主として研究を進めているため時間的要素を考慮する必要がある. 本実験では常に情報が動き, 変化していく Web からの話題語獲得なので, この点に関して考慮しやすい. 以下に具体的にどう考慮するか, その手法を記述していく.

3.2.8 相関係数による話題語クラスタリング

Web から取得してきた話題語は, 取得時間ごとに重みが変化している. そこで横軸を話題語取得時間, 縦軸を話題語の重みとすると, 重みの時間推移が抽出できる. そして重みの時間推移に注目した時, ある 2 つの話題語が似た推移を辿る, または全く正反対の推移を辿るようであれば, これらは互いになんらかの関係あると推測できる.

そこで推移の類似性を推し量る指標と相関係数を用いる. 話題語 *X, Y* の時間 *t* における重み x_t, y_t とすると, 相

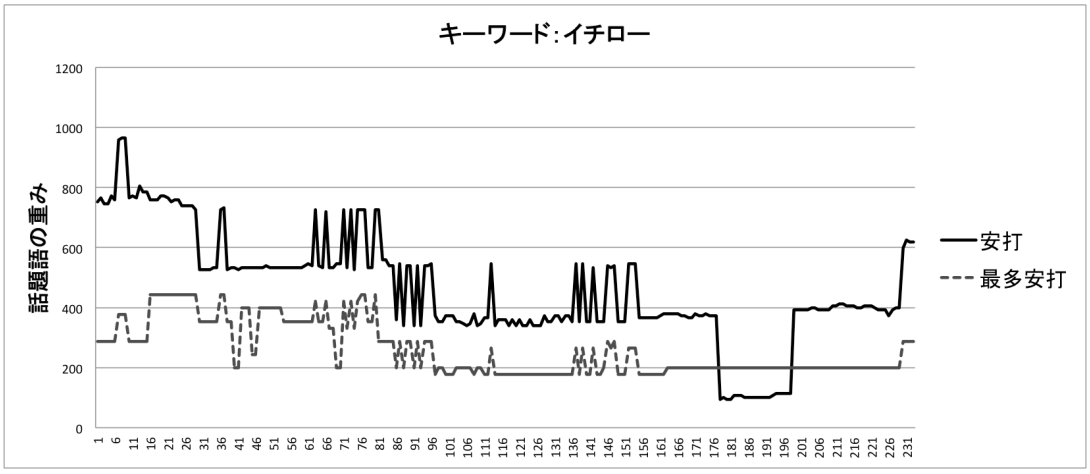

図 3 相関係数の高い、二つの話題語の時間推移

関係数は式. 3 で求められる.

$$Correl(X,Y)=\frac{\sum_t (x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_t (x_t-\bar{x})^2 \sum_t (y_t-\bar{y})^2}}$$

式. 3　使用した相関係数を求める式

相関係数は-1 ～ 1 までの連続値をとり, 1 に近い値であれば正の相関, -1 に近い値であれば負の相関となる. 0 に近い値は相関がないとされる.

図 3 はキーワードを「イチロー」とし, 獲得した話題語のうち「安打」と「最多安打」について, 横軸を時間, 縦軸を話題語の重みとして表したグラフである. この 2 つの話題語の重みの推移は相関係数 0.8675 という強い正の相関を持つ.

3.2.9　基準語とクラスタについて

相関係数でわかるのは, ある 2 つの話題語間にある相関の強弱だけである. したがって, 話題語 A と話題語 B に強い正の相関があり, 話題語 B と話題語 C に強い正の相関があったとしても, 話題語 A と話題語 C に強い正の相関があるとは言えない.

従って相関係数によるクラスタリングでは, 基準となる話題語を設定し, それとある閾値以上の相関を持つ話題語が複数個集まることでクラスタを形成することになる(図 4). このとき, 基準となる話題語を「基準語」と定義し, 以降本稿で使用していく.

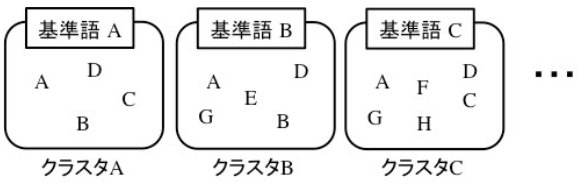

図 4　クラスタのイメージ

4. 評価実験

獲得した話題語をクラスタリングするためには相関係数を用いる. しかし相関係数のどの値までを 1 つのクラスタとみなすかによってクラスタの中身は変化する. そのため, 閾値とクラスタの間に生じる性質と傾向の評価実験を行う.

実験に用いるデータは, 単語 49 語をキーワードとし, 1 ヶ月間 (2010 年 11 月 11 日 ~2010 年 12 月 10 日) 話題語獲得を行った結果を用いる. 話題語獲得の間隔は, 1 日あたり 3 時間間隔で 8 回の話題語獲得を行った.

4.1　実験

本研究の目的は背景的知識の自動構築であるため, 話題語のクラスタリング計算を定期的に繰り返すことになる. それを想定した上で, この実験では 60 時間ごと(話題語獲得回数 20 回)の話題語の重みの時間推移を求め,それぞれに相関係数を計算し, 計 10 回の学習を行ったデータを用いる.

また閾値は, -0.9, -0.7, -0.5, -0.3, -0.1, 0.1, 0.3, 0.5, 0.7, 0.9, ±0.1, ±0.3, ±0.5, ±0.7, ±0.9 の 15 個を用い, 閾値以上(閾値が負の場合は閾値以下)の相関係数をもつ話題語クラスタとして扱った.

4.1.1　実験内容

被験者 4 名に対し, それぞれの閾値で算出したクラスタの基準語の並びに注目させ, キーワードに対し, 話題として繋がりをもつ話題語が上位に出現しているかを判断させた.

基準語の並びは, クラスタ内の重みの平均値, クラスタ内の単語数の平均と分散, 基準語が他のクラスタに出現した総回数の値を降順にしたときのものである.

4.1.2　結果

表 1 は評価実験の結果をまとめた表である. 表内の値は, 話題として繋がりをもつ話題語が上位に出現していると判断した人数である. 列は基準語の並び方別で見たときの人数である. 表内に入る値が大きいほど, 多くの人に話題として賛同される話題語のクラスタを生成する閾値であると言える. また, 列を眺めたときに読み取れる事柄は,並びの基準となる値と深く関係する閾値が読み取るということである.

図 5 は, 表 1 の値を各並び方にて繋がりがあると判断した人数を積み上げ棒グラフにしたものである. 横軸が閾値で, 縦軸がキーワードと基準語の並びに繋がりがあると判断した人の累計である. 縦軸の値を大きくとるほど, 総合的に見て繋がりがあると判断された閾値だと読み取ることができる.

また図 5, 表 1 にある A, B, C, D は基準語の並び方であり, それぞれの対応は以下のようになっている.

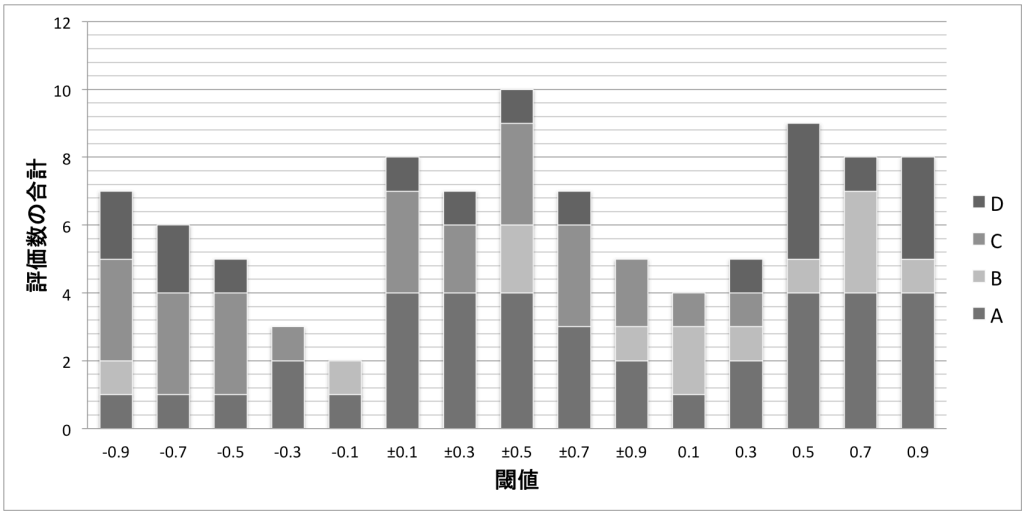

図 5　各閾値における繋がりがあると判断された評価の合計

A：クラスタ内の重みの平均値
B：他のクラスタに出現した総回数
C：クラスタ内の単語数の分散
D：クラスタ内の単語数の平均

表 1　閾値別繋がりがあると判断した人数(被験者 4 名)

閾値	A	B	C	D
-0.9	1	1	3	2
-0.7	1	0	3	2
-0.5	1	0	3	1
-0.3	2	0	1	0
-0.1	1	1	0	0
±0.1	4	0	3	1
±0.3	4	0	2	1
±0.5	4	2	3	1
±0.7	3	0	3	1
±0.9	2	1	2	0
0.1	1	2	1	0
0.3	2	1	1	1
0.5	4	1	0	4
0.7	4	3	0	1
0.9	4	1	0	3

5. 考察

表 1 をみると，クラスタ内の重みの平均値で並べた場合（表 1 の A）に，閾値 15 個中 10 個において，2 名以上の人がキーワードと基準語の並びが話題として繋がりを持つと判断した．これはクラスタ内の話題語の重みが大きいほど，キーワードと話題語間の話題としての繋がりが大きくなっていると判断できる．つまり，クラスタ内の重みの平均値は，キーワードと基準語である話題語との関連の度合いとして利用出来るのではないかと推測する．

次に,図 5 をみる．傾向として，負の相関は相関の度合いが強くなるほど，評価が増えていることがわかる．また，正の相関においては，相関の度合いが強くなるほど増える傾向にはあるが，閾値 0.7，0.9 において評価の伸びが芳しくない．これは閾値を高く設定しすぎたために，クラスタ内の話題語が十分な数が得られず，本来ならキーワードと繋がりの持つような話題語が上位に現れなかったためだと推測する．これらの結果より，相関が強まるにつれ，評価が上がる傾向にあることがわかり，つまり重みの推移が類似する話題語をクラスタ化するという方向性は間違っていないと判断する．

続いて，評価が一番多く得られた閾値は絶対値 0.5 のときであることがわかる．これは各並び順において最も繋がりがあると判断された閾値であることを示唆している．よって，キーワードと話題として繋がりを持つ話題語を相関係数の絶対値 0.5 以上でクラスタリングした結果，キーワードと基準語で最も繋がりがあるクラスタが生成されることが推測される．つまりキーワードと話題語の関係，およびその関係の強さが求まり，本研究の目的である話題の繋がりを記録する背景的知識の構築を行えたと判断する．この背景的知識の特徴としてはキーワードと話題語間の関係だけでなく，基準語を介してキーワードとクラスタ内の話題語にも繋がりが生まれるということである．本来なら重みが小さく，無視されてしまうような話題語であっても何らかのクラスタに属していれば，埋もれずに使用することが可能となる．

表 2 は，実際に閾値を絶対値 0.5 でクラスタリングを行い，クラスタ内の重みの平均値で降順に並び替えた上位 10 個の話題語と，各取得時間における *tf・idf* の重みを単純に平均して降順に並び替えた上位 10 個の話題語の比較である．わずかではあるが，クラスタ内の重みの平均値で選出した話題語の方が，*tf・idf* で選出した話題語より，話題として繋がりがあると評価された人数の平均が上回った．

表 2　話題として繋がりあると判断した人数の比較
キーワード「インフルエンザ」(被験者 4 名)

tf・idf で話題語を選出	[人]	A で話題語を選出	[人]
新型インフルエンザ	4	インフルエンザワクチン	4
インフルエンザウイルス	4	強い	1
ウイルス	4	患者	3
多い	0	ウイルス	4
高い	0	効果	2
地域全体	1	こまめ	1
ウイルス粒子	4	ヒト	2
医療機関	4	インフルエンザウイルス	4
症状	2	新型インフルエンザ	4
重症化	3	マスク	3
平均	2.6	平均	**2.8**

6. おわりに

本研究では，Web からの話題語獲得による背景的知識の自動構築の手法を提案した．背景的知識にキーワードと話題語を定義し，Web から獲得した話題語の重みの時間推移から相関係数を用いてクラスタリングし，その閾値を決める評価実験を行い，結果として相関係数の値を絶対値 0.5 以上でクラスタリングを行うと，キーワードと基準語が最も繋がりのあるクラスタが生成できることがわかった．このキーワードと話題語（基準語）の関係が得られたことにより，背景的知識の自動構築が可能であるとした．

今回，調査を行った中で全く同じクラスタを持つ基準語が存在する場合があった．基準語の集合、つまり話題語のクラスタとみなし，それに関して実験や傾向調査を行うことが今後の課題としてあげられる．また，今回構築した背景的知識を実際の人間と計算機の会話に用いるシステムでこの背景的知識の有用性を調べることも今後の展望として挙げられる．

謝辞

本研究の一部は科研費（23720222）の助成を受けたものである。

参考文献

[1] 「相手の嗜好にあった話題を提供する自動発話システムの開発」林輝大，奥村紀之： 情報処理学会第 72 回全国大会，6X-7，2010 年 3 月
[2] 「概念間の関連度計算のための大規模概念ベースの構築」奥村 紀之, 土屋 誠司, 渡部 広一, 河岡 司 ： 自然言語処理 Volume14 Number5 p.41-64,2007.
[3] 「時間的要因を考慮した属性獲得手法」波多腰 優斗 ： 第 73 回情報処理学会全国大会 4S-4 2011 年 3 月
[4] 「語の共起の統計情報に基づく文書からのキーワー ド抽出アルゴリズム」松尾 豊, 石塚 満：人工知能学会論文誌, 2002