4N-3

Web を用いた未知語検索キーワードの シソーラスノードへの割付け手法

後藤和人† 渡部広一† 河岡司†
†同志社大学 工学部 知識工学科

1. はじめに

Webには膨大な情報が存在するが，必要な情報を効率的に短時間で的確に得ることは困難である．そこで，キーワードによる検索要求に対して，その語が大局的にどのような意味を持つのかをその語の所属すべきシソーラス[1]のノードを提示することで，未知語の内容を簡明に表示することができると考える．本稿では，シソーラスと概念ベース[2]とWebを用いて，未知語が属すべきシソーラス上に定義されていない場合にも最適なノードを見出す手法を提案する．

2. 使用技術

2.1 シソーラス

シソーラスとは一般名詞の 2710 個の意味属性（ノード）の上位下位，全体部分関係が木構造で示されたものである．ノードに属する名詞として約 13 万語のリーフが登録されている．

本稿では，未知語の属するノードを探す上で必要のないノードを削除している．結果，使用するノード数は 370 個となっている（図 1）．

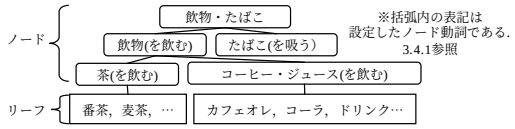

図 1 シソーラス（一部）

2.2 概念ベースと関連度

概念ベースとは語（概念）の特徴を表す語（属性）を大量に集めたものであり，属性には重みが定義されている．本稿では，複数の国語辞書や新聞などから抽出した概念や属性を加えた約 9 万の概念からなる概念ベースを使用する．

関連度[2]とは概念と概念の関連の強さを定量的に評価するものである．各概念を 2 次属性まで展開し，重みを考慮した属性集合の一致度合いを計算する．本稿では未知語属性とノード属性との関連の深さを判断するのに関連度を用いる．

Allocation Method of an Unknown Search Keyword to a Thesaurus Node By Web
†Kazuto GOTO, Hirokazu WATABE, Tsukasa KAWAOKA
Knowledge Engineering and Computer Sciences, Doshisha University

3. シソーラスノードへのマッピング

未知語とノードの性質を比較するために属性を獲得する．次に，獲得した属性列を用いて関連度計算を行い，所属候補ノードを絞り込む．さらに，シソーラスが持つ情報を利用（ノード特定手法）して，未知語が所属すべきノードを特定する．図 2 に未知語をシソーラスノードへマッピングする流れを示す．

図 2 ノード特定の流れ

3.1 Web からの未知語属性の獲得

未知語を用いてGoogle[3]で検索を行い，検索結果ページを取得する．そして，不要な情報を取り除いた文書群の形態素解析を行い，自立語を抽出する．最後に，概念ベースに存在する語のみを未知語の属性とし，頻度情報と特定性情報であるidf値を掛け合わせたものを属性の重みとする．表 1 に獲得した属性列を示す．

表 1 G ショック（未知語）の属性

属性	時計	腕時計	電波	…
重み	358.94	280.64	254.72	…

3.2 シソーラスのノード属性の構築[4]

シソーラスのノードにおいてノードに属するリーフを概念ベース参照して，その 1 次属性と重みを取得する．それをノード内全てのリーフに対して行い，それらを足し合わせてノードの属性とする．この作業を全てのノードに対して行い，シソーラスのノード属性を構築する．なお，属性の重みは取得した全ノード属性の重みに idf 値を掛け合わせたものを属性の重みとする．表 2 に一例として「時計」のノード属性を示す．

表 2 ノード「時計」の属性

ノード属性	懐中時計	掛時計	…
重み	4733.49	3476.39	…

3.3 シソーラスのノード絞込み

3.1 で説明した手法を用いて獲得した未知語属性とノード属性に対して関連度計算を行い，関連度が 0.02 以上のものを所属候補ノードとする．

3.4 所属ノードの決定

先ほど求めた所属候補ノードに対して，以下の手法を用いてノード特定を行う．

3.4.1 ノード動詞

シソーラスは単語を体系的に配置しており，「同一のノードに属するリーフは助詞を伴う動詞の係り受けに同様の語を取る」という関係が存在する．ノード動詞とはこの関係を利用して，ノードに設定したキーワードのことであり，ノード特定の補助に利用する．例えば，未知語が「マイルドセブン」，所属候補ノードが「たばこ」である場合，「マイルドセブンを吸う」というキーワードの HIT 数を求める．

3.4.2 共起ヒット

未知語とノード名の And 検索による HIT 数を調べてノード特定の補助を行う．例えば，未知語が「マイルドセブン」，所属候補ノードが「たばこ」である場合，「マイルドセブン」と「たばこ」で And 検索を行い，HIT 数を求める．

3.4.3 共起率

未知語とノードが共に出現する頻度から関連性を判断する共起率を求めることでノード推定の補助を行う．共起率は以下の式(1)で定義される．Aは未知語，Bは所属候補ノードを表す．

$$kyokiritu=\frac{1}{2}\left(\frac{A\&B\text{のHIT数}}{A\text{のHIT数}}+\frac{A\&B\text{のHIT数}}{B\text{のHIT数}}\right) \quad (1)$$

3.4.4 所属ノードの決定方法

ノードの決定を以下の 5 つの条件で行った．また以下の式(2)，(3)，(4)，(5)は各条件で用いるものであり，所属候補ノード $node_i$ の中でノード得点 $NodeValue$ が最も高いものを所属ノードとする．$RelValue$ が未知語と $node_i$ の関連度，$VerbHit(node_i)$ は未知語にノード動詞を連結したキーワードを Google で検索を行ったときの HIT 数，$KyokiHit(node_i)$ は未知語とノード名の And 検索を Google で行ったときの HIT 数を表す．

① 全てのシソーラスノード（2710 個）を使用

② 不要なノードを削除（使用ノード 375 個）

$$NodeValue(node_i)=RelValue \quad (2)$$

③ ②+ノード動詞を使用

$$NodeValue(node_i)=RelValue\cdot\log VerbHit(node_i) \quad (3)$$

④ ②+ノード動詞と共起ヒットを使用

$$NodeValue(node_i)=RelValue\cdot\log VerbHit(node_i)\cdot\log KyokiHit(node_i) \quad (4)$$

⑤ ②+ノード動詞と共起率を使用

$$NodeValue(node_i)=RelValue\cdot\log VerbHit(node_i)\cdot kyokiritu\cdot 100 \quad (5)$$

4. 評価

評価を行うために，200 個の未知語を用いてシステムの評価を行う．評価に使用したテストセットの一部を表 3 に示す．

表 3 テストセットの一部

FinePix	冬のソナタ	壬申の乱
武蔵坊弁慶	インドガビアル	…

テストセットを使用し，未知語の所属ノードの決定を行い，人手で評価したときの精度と未知語 1 語あたりの平均処理時間を図 3 に示す．評価は 3.4.4 で説明した条件で行った．

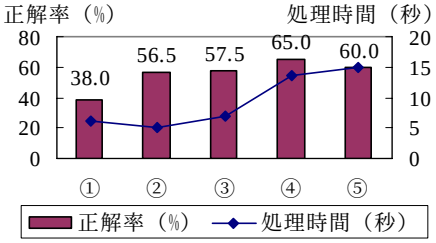

図 3 テストセットの正解率と処理時間

図 3 より，不要なノードを削除し，ノード動詞および共起ヒットを利用したとき（手法④）に 65.0%と最も高い精度を得ることができた．

5. おわりに

本稿では，ユーザが未知の単語に出会い，その単語が一体何なのかを知りたいとき，その語が大局的に見て何なのかをシソーラスのノードを用いて提示する手法を提案した．さらなるシステムの精度向上としては，未知語と所属ノードをより適切に関連づけるために，未知語とノードの属性数を考慮するなどの方法が有効と考えられる．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1] NTT コミュニケーション科学研究所監修，日本語語彙体系，岩波書店，1997

[2] 渡部広一，河岡司：常識的判断のための概念間の関連度評価モデル，自然言語処理，Vol.8, No.2，pp.39-54，2001

[3] Google：http://www.google.co.jp

[4] 伊藤俊介，渡部広一，河岡司：情報検索における未知語理解支援方式～未知語のシソーラスノードへの分類～，情報処理学会自然言語処理研究会資料，2004-NL-159，pp.61-66，2004