共起語の包含関係を用いた分野固有知識の獲得

山本英子　井佐原均

独立行政法人　情報通信研究機構

{eiko, isahara}@nict.go.jp

1. はじめに

本研究では，我々が提案した語彙の自動階層構築方法[Yamamoto et al., 2004; Yamamoto et al., 2005]の応用として，分野特化した文書集合からの知識発見を試みる．これまでは提案手法の階層構造構築問題における適用可能性や有用性を示すために，新聞記事から抽出した分野を限定しない言語データから一般的な階層構造を抽出したが，言語データを取り出す元の文書を変更することにより，この手法が分野に特化した知識を抽出することにも適用できるのではないかと考えた．ここでは例として，医学分野の Web 文書集合からそのデータ中の語の出現状況を用いて医学用語間の階層構造を抽出することを試みる．そして，その抽出された用語の階層関係は知識としてどのように解釈できるかを分析し，我々の提案手法の知識発見問題における適用性を検証する．

実験では，Web 文書集合から抽出した 2 種類のデータを用いた．1 つは，Web 文書集合における名詞間の共起関係に基づくデータであり，もう 1 つは，名詞と動詞の係り受け関係に基づくデータである．これらのデータから我々の提案する自動階層構築方法によって抽出した階層（知識）と，共起情報のみによって得られる知識と，医学用語のシソーラスの 1 つである MeSH シソーラス[MeSH thesaurus]の階層とを比較する．その結果，提案する手法が分野特化した文書集合から階層的な意味的関係や因果関係などの分野固有知識を獲得しうることを報告する．

2. 実験データ

本研究では，例として，医学分野に関連する Web 文書集合(10,144 ページ，225,402 文，37M バイト)からその分野固有の知識を獲得することを試みた．まず，文書集合中の文を構文解析し，「の，を，が，に，は」の 5 つの格助詞による係り受け関係を収集する．図 1 に示すように，各文から，「A<の>B」，「P<を>V」，「Q<が>V」，「R<に>V」，「S<は>V」のパターンにあてはまるものを収集する．ここで，<X>は格助詞，A，B，P，Q，R，S は名詞，V は動詞を表す．他の助詞による係り受け関係（図 1 では「光子から聞いた」）は収集しない．次に，収集した関係データから，2 種類の実験データを作成する．1 つは，各文について，上記のパターンに含まれる名詞(すなわち，A，B，P，Q，R，S)を全て集めたものであり，名詞間の共起関係に基づくデータ(データ①)である．もう 1 つは，動詞 V を含むパターンについて，動詞 V と共起している名詞(すなわち，P，Q，R，S)を共起する助詞ごとに集めたデータであり，係り受けに基づくデータ(データ②)である．

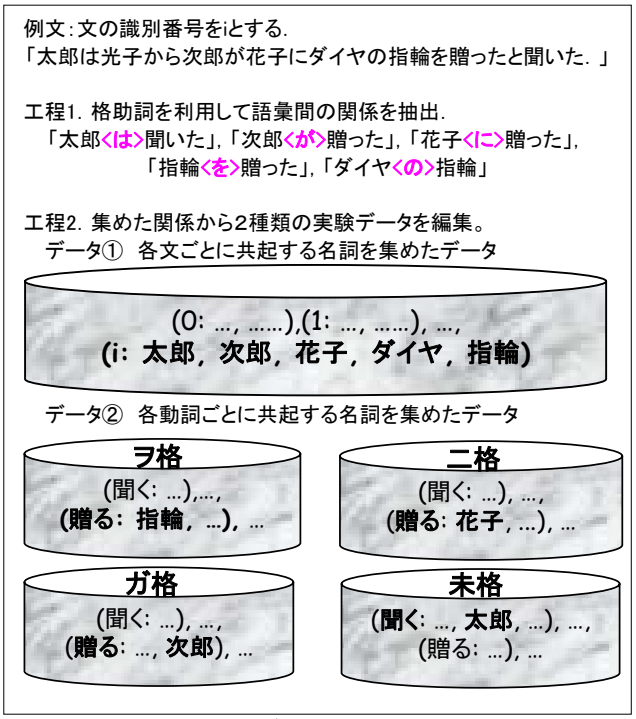

図 1：実験データ作成のイメージ

3. 自動階層構造構築方法

本稿では，これまでに提案した自動階層構造構築方法を用いて，知識獲得を試みる．この提案手法は，補完類似度[Hagita and Sawaki, 1995]という二つのベクトル間の包含関係を測る尺度に基づく手法である．これまでに，補完類似度は形容詞・形容動詞の分類のために，自己組織化マップへの入力に利用できることも示されている [Kanzaki at al., 2004].

一般に二つのベクトル $F = (f_1, ..., f_n)$ と $T = (t_1, ..., t_n)$において，「F が T を包含する度合い」と「T が F を包含する度合い」は異なる．補完類似度は非対称性を持つ尺度であり，このようなベクトル間の包含関係を表すことができる．我々は，対象とする文書

集合中の単語の出現状況をベクトルで表しており，より広い状況で出現する単語はより広い意味を持つといえることから，補完類似度によって決定された 2 つのベクトル間の包含関係をその文書集合における二単語間の上下関係とした．その二単語間の関係を連結していくことで階層構造を構築する．

3.1. 出現状況のベクトル表現

本稿では，名詞の出現状況を二値ベクトルで表し，２つの名詞間の関係を推定する．図 2 に出現状況をベクトルで表したイメージを示す．データ①における名詞の出現状況は，各次元が文に相当し，その名詞が各文に出現するかどうかを 0，1 で表したベクトルとなる．データ②における名詞の出現状況は，各次元が動詞に相当し，その名詞が各動詞と共起するかどうかを 0，1 で表したベクトルとなる．

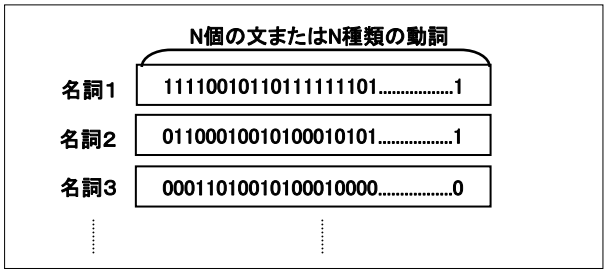

図 2：出現状況のベクトル表現

このような出現状況を表すベクトル間の包含関係を測る補完類似度 CSM(**F**,**T**)は，ベクトル **F** = (f_1, ..., f_n) と **T** = (t_1, ..., t_n)について，次のように定義される．パラメータはそれぞれ図 3 に示す状況にある次元の数である．

$$CSM(\mathbf{F},\mathbf{T}) = \frac{ad-bc}{\sqrt{(a+c)(b+d)}}$$

f_i \ t_i	1	0
1	a	b
0	c	d

図 3：パラメータ

3.2. 階層構造の構築

本手法では，上下関係にある単語対を連結していくことによって階層構造を構築する．図 4 に階層構造の構築過程を示す．補完類似度の値が低ければ，その単語間の関係は信頼性が低い．このため，閾値以上の値を持つ単語対に限り，階層構造を構築する．閾値は実験的に定めた．

4. 実験

実験では，2 節に示す実験データから医学用語に関する知識を獲得することを試みた．対象としたのは医学用語のシソーラスである 2005 MeSH シソーラス[MeSH thesaurus]に記載されている見出し語で

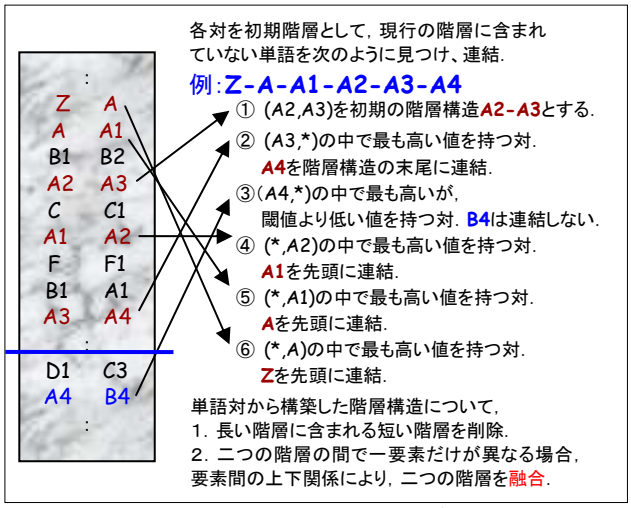

図 4：階層構造の構築

役割-細胞-遺伝子-染色体-受精
分泌-胃酸-胃粘膜-十二指腸潰瘍
皮膚-アトピー性皮膚炎-ヘルペスウイルス-抗ウイルス薬
卵巣-脾臓-触診
疲労-子宮筋-妊娠中毒症
水-鉄-トランスフェリン-ヘモクロマトーシス
疑いを招くようなデータ-原因-うつ病-減少-血小板数-骨髄検査

図 5：データ①から得た結果の一部

ある．実験データには，その見出し語のうち 2,557 語が含まれていた．これらの医学用語について，補完類似度に基づく手法によって，実験データにおける用語の階層構造を構築し，3 つ以上の用語からなる階層を検討対象とした．図 5 にデータ①から得られた結果の一部を示す．

5. 比較・考察

この節では，図 5 に示すような結果を文書集合から獲得した知識と考え，この結果と，共起頻度を用いた手法による結果と，MeSH シソーラスに記述されている知識とをそれぞれ比較し，補完類似度に基づく手法の知識獲得への応用可能性を考察する．

5.1. 共起頻度を用いた手法との比較

コーパスから知識を獲得するための典型的な方法は，データ中での共起頻度を用いて用語のリストを抽出することである．一方，提案手法では，二つの用語間の関係は補完類似度(CSM)によって推定される．これら二つの手法による結果の違いを見るために，データ①から得られた結果について，共起頻度の降順に並べた用語の対と，CSM 値の降順に並べた対を比較した．表 1 に高い CSM 値を持つ上位 10 個の用語対を示す．各対において，提案手法は左列の単語を上位語，右列の単語を下位語と推定した．

表 1 において，CSM 値で 1 位に位置する(投与，処置)は共起頻度も，もっとも高い対である．これは，

共起頻度が高ければ，CSM 値も高いことを示し，基本的にこのような対は上位に位置する．一方，CSM 値で 4 位に位置する(鉄，トランスフェリン)は共起頻度は低い．この 2 つの用語に関して，医学辞書から“Iron is taken into the body with the molecule called Transferrin.”という文を見つけることができる．これは，この対は共起頻度だけでは抽出できないが，有意味な情報であることを示している．

たとえば出現頻度の高い単語 M と低い単語 N があって，N が出現する場合には，ほとんど M も出現するといった状況を考えてみよう．N の出現頻度が低いために，M と N との共起頻度は低くなるが，M の出現する文の集合が N の出現する文の集合を（ほとんど）包含することになり，CSM 値は高くなる．補完類似度を用いた手法では，このように共起頻度が低い対でも重なりが大きければ，その用語間の関係は強い，つまり有意味と捉えうる．

表 1：データ①における高い CSM 値を持つ上位 10 個の用語対

推定された用語対の上下関係	
投与	処置
娘	保育園
注意	紹介
鉄	トランスフェリン
森	オランウータン
娘	息子
役割	サイトカイン
発作	てんかん
分泌	糖質コルチコイド
自然	権利

次に，実験の結果得られた 3 つ以上の用語からなる階層について考察する．提案手法は補完類似度によって関係付けられた用語対を連結し，階層構造を構築する．たとえば，用語対(A,B)と(B,C)を補完類似度によって得た場合，A と B，B と C の間にそれぞれ方向性があるので，A と C が共起しなくても，これらの対を連結できる．しかし，共起頻度に基づく手法の場合，A と B，B と C には上下関係が規定されていないので，共起しない用語 A と C を関連付けるには共起関係以外の根拠が必要となる．

5.2. MeSH シソーラスとの比較

次に，得られた階層を人手によって構築された MeSH シソーラスに含まれる見出し語の木構造(MeSH Tree)と比較した． 図 6 に MeSH Tree の一部を示す．()内にシソーラスに含まれる語の日本語訳を表す．MeSH シソーラスは，最上位で 15 のカテゴリに分類され，以下順次詳細に分類されていく．なお，2 つ以上のカテゴリに分類されている見出し語もある．各見出し語は，この分類を示す識別番号を持つ．たとえば，図 6 に示す「親指」は識別番号「A01.378.800.667.430.705」を持つ．先頭の「A01」は「身体の部位」を表すカテゴリ番号で，「親指」は「身体の部位」に分類されることを示している．識別番号の下 3 桁を削ると，一つ上位の語を知ることができる．この例では，上位語は識別番号「A01.378.800.667.430」を持つ「指」である．これを繰り返すと，「A01.378.800.667」は「手」，「A01.378.800」は「上肢」，「A01.378」は「手足」と，「親指」の階層構造を上に辿ることができる．

このように整理された医学用語シソーラスと実験結果を比較する．提案手法で得られた階層が医学分野における用語間の階層的な意味関係を表す知識であるならば，その階層を構成する用語は MeSH シソーラスにおいて，同じカテゴリに分類されているであろう．表 2 に実験データから得られた階層構造を構成する用語が MeSH シソーラスにおける見出し語の分類に対して，どのように分布するかを示す．ここで，データ①を Co-data，データ②をそれぞれ Wo-data, Ga-data, Ni-data, Ha-data と表す．

A01	body region	（身体の部位）
	:	
A01.378	extremity	（四肢）
A01.378	limb	（手足）
A01.378.610	lower limb	（下肢）
A01.378.610.250	foot	（足）
A01.378.610.250.149	ankle	（足首）
A01.378.610.250.510	heel	（踵）
	:	
A01.378.800	upper limb	（上肢）
A01.378.800.075	arm	（腕）
A01.378.800.420	elbow	（肘）
A01.378.800.585	forearm	（前腕）
A01.378.800.667	hand	（手）
A01.378.800.667.430	finger	（指）
A01.378.800.667.430.705	thumb	（親指）
A01.378.800.667.715	wrist	（手首）
A01.378.800.750	shoulder	（肩）
A01.456	head	（頭）
A01.456.505	face	（顔面）
A01.456.505.580	forehead	（額）
	:	

図 6: MeSH Tree の一部

表 2: 階層構造を構成する用語のカテゴリ分布

	Co-data	Wo-data	Ga-data	Ni-data	Ha-data
得られた階層構造の数	594	194	62	37	85
用語のカテゴリ分布					
1 つ	24	35	12	3	6
2 つ	169	42	19	14	26
3 つ	116	34	10	5	14
階層構造の割合					
1 つまたは 2 つに分布	.32	.40	.50	.46	.38
3 つ以下に分布	.52	.57	.66	.59	.54

表 2 から，実験データから得られる階層構造を構成する用語がすべて同じ 1 つのカテゴリに分布する

場合と，２つのカテゴリに分布する場合を合わせると，データの種類によって 32% から 50%，さらに，用語が３つのカテゴリに分布する場合を合わせると，52% から 66%を占めることがわかった．

実験では，「ガ格」に関するデータから抽出された語彙の階層が MeSH シソーラスのカテゴリ分類にもっとも合致しているという結果を得た．これは，「ガ格」によって表現される主格が他の格と比べて多義性が少ないためであると思われる．実験結果と MeSH シソーラスとの違いを議論するために，階層を構成する用語が１つのカテゴリに分布する構造の例を図７に，用語のうちの１つが他の用語とは異なるカテゴリに分類される構造の例を図 8 に示す．()内にその構造を得た実験データを示す．図 8 に示す下線は，異なるカテゴリに分類される用語を示している．

手-口-耳-指（ni-data）
皮膚-腹部-頚部-口腔-胸部（co-data）
水疱-鼓腸-腰痛-尺骨神経麻痺-脳内出血
　-閉塞性黄疸（wo-data）
心疾患-冠動脈疾患-気管支炎-血栓性静脈炎-鼓腸
　-高尿酸血症-腰痛-尺骨神経麻痺-脳内出血
　-閉塞性黄疸（wo-data）
貧血-嘔吐-腰痛-尺骨神経麻痺-脳内出血
　-閉塞性黄疸（wo-data）
膵炎-狭心症-ネフローゼ症候群-高血圧性脳症（wo-data）
頭痛-関節痛-チアノーゼ-血尿-紫斑-白血球減少-発汗
　-腹水（wo-data）
疲労-ストレス-十二指腸潰瘍（co-data）

図 7：用語が１つのカテゴリに分布する構造の例

アイスクリーム-チョコレート-ワイン（ni-data）
薬-本草学-生薬学（co-data）
卵巣-脾臓-触診（co-data）
変化-反応-発生-分泌（wo-data）
出血-発熱-血尿-意識障害-めまい-高血圧（ga-data）
疲労-子宮筋-妊娠中毒症（co-data）
受胎能力-アクリル樹脂-強心薬-人工血管（ga-data）

図 8：用語の１つが他の用語と異なるカテゴリに分類される構造の例

たとえば，図 8 にある「アイスクリーム - チョコレート - ワイン」は自然言語の観点から見ると，明らかに何らかの関連を持つ塊（コロケーション）であるように思われる．これらの用語について，MeSH シソーラスでは，「アイスクリーム」と「ワイン」は「食物」と分類されているが，「チョコレート」は「原料」と分類されている．しかしながら，実際の文章中では，これらはすべて「食べることができるもの」と考えることができる．このように，提案する手法は，より現実の文章表現（医学知識が書かれた文章表現）を反映した関係を抽出しうる．また，「卵巣 - 脾臓 - 触診」は「卵巣や脾臓の病気は触診による診断される」という医学的知識と見ることができる．同様に，図 5 にある「疑いを招くようなデータ - 原因 - うつ病 - 減少 - 血小板数 - 骨髄検査」は「骨髄の疾患はうつ病や血小板の減少を引き起こす原因となるので，骨髄検査は必要である．」を表す医学的知識と見ることができる．一方，MeSH シソーラスにおいて，この関係を構成する用語は３つ以上の異なるカテゴリに分類されており，このような知識は明示的には表現されていない．

6. まとめ

本稿では，これまでに提案した語彙の自動階層構築方法を知識獲得に応用することを考え，例として，医学分野の Web 文書集合を用い，分野固有知識を獲得できるかどうかを検討した．実験には，文書集合から格助詞を利用して単語間の関係を収集し，その関係集合から編集したデータを用いた．実験結果から，提案する手法によって，階層的な意味的関係や因果関係などの分野固有知識を獲得しうることがわかった．今後，実験データごとに得られた知識を比較し，さらに詳しく分析する．また，得られた階層中の語が検索などに有効に利用できることを検証する予定である．

参考文献

[Hagita and Sawaki, 1995] Hagita, N. and Sawaki, M. Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning, In *Proceedings of the SPIE – The International Society for Optical Engineering*, 2442: pp. 236-244, 1995.

[Kanzaki et al., 2004] Kanzaki, K., Yamamoto, E., Ma, Q. and Isahara, H. *Construction of an objective hierarchy of abstract concepts via directional similarity*. In Proceedings of the 20th International Conference on Computational Linguistics, Vol.2, pp. 1147-1153, 2004.

[Yamamoto et al., 2004] Yamamoto, E., Kanzaki, K. and Isahara, H. Hierarchy Extraction based on Inclusion of Appearance, In *ACL04 Companion Volume to the Proceedings of the Conference*, pp. 149-152, 2004.

[Yamamoto et al., 2005] Yamamoto, E., Kanzaki, K. and Isahara, H. *Extraction of hierarchies based on inclusion of co-occurring words with frequency information*. In Proceedings of the 19th IJCAI, pp. 1166-1172, 2005.

[MeSH thesaurus] The U.S. National Library of Medicine created, maintains, and provides the Medical Subject Headings (MeSH®) thesaurus.