複数の電子化辞書に含まれる類似概念の対応付けとその応用

杉本 徹　　篠埜 功
芝浦工業大学 工学部 情報工学科
{sugimoto, sasano}@shibaura-it.ac.jp

1　はじめに

自然言語処理の分野では，EDR 電子化辞書 [1]，日本語語彙大系 [2]，分類語彙表 [3]，WordNet [4] など様々な電子化辞書が利用されている．これらの電子化辞書の多くは，膨大な数の単語や単語が表しうる意味（概念）を階層的に分類，体系化した形で収録している．例えば，EDR 電子化辞書には約 27 万語の日本語単語と約 40 万個の概念が収められ，概念は唯一のルート概念を起点とする巨大な分類体系をなしている．また，日本語語彙大系は約 30 万語の日本語単語と約 3,000 個の意味属性を持ち，意味属性は 12 階層からなる木構造として体系化されている．このような電子化辞書は日常使われる言葉と意味の多くの部分を網羅し，コンピュータによる言語処理において不可欠な資源となっている．

しかしその一方で，完全・完璧な辞書というものは存在せず，既存の辞書は多かれ少なかれ次のような不完全性を持っている．

1. 未収録の単語が存在する（希少語，新語など）．
2. 収録された単語にも，未収録の意味がありうる．
3. 単語や概念の分類方法は 1 通りではないため，その辞書における分類では捉えられない側面がある．

このような辞書の不完全性は，言語処理の精度や汎用性を高める上で障害となる．電子化辞書の中には単語や概念の共起関係を収集した共起辞書や動詞が取りうる格要素の意味的制約を整理した格フレーム辞書を併せ持つものもあるが，それらにおいては不完全性がより顕著な問題となる．

我々は，このような問題に対して複数の電子化辞書を併用し相補的に用いることにより対処するアプローチを検討している．その際に重要となるのは，双方の辞書に含まれる類似した構成要素間の対応付けを行うことである．本稿では，階層的な概念体系を持つ複数の電子化辞書に含まれる類似概念の対応付けを行う手法を提案する．対応付けの手法を一般的に述べた後，EDR 電子化辞書と日本語語彙大系を用いて行った対応付け実験の結果を紹介する．また，本手法を単語辞書の補完や文の意味解析に応用する可能性と，それらの観点からの評価についても議論する．

2　関連研究

複数の階層的分類体系の対応付けは幅広い応用が期待される技術で，これまでに日英シソーラスの対応付け [5]，日本語語彙大系と Wikipedia の結合 [6]，Web ディレクトリの対応付け [7]，オントロジーの対応付け [8] など様々な研究が行われている．

これらの研究と比較すると本研究は，電子化辞書の内容の拡張を目的としており，かつ 1 対多の対応付けを行うことができる，という特徴がある．

3　類似概念の対応付け

3.1　対象とする辞書の構造

本研究では，単語が表しうる意味である概念が階層的に分類された概念体系を持つ電子化辞書を扱う．概念体系には親を持たないルート概念が唯一存在し，その他の概念は 1 個以上の親概念を持つ．また，各概念は任意個の子概念を持つ．さらに，各概念にはその概念を意味として持つ単語の集合が結び付けられている．EDR 電子化辞書，日本語語彙大系，分類語彙表，WordNet はすべてこのような構造を持つ電子化辞書と見なすことができる．また，Wikipedia のカテゴリと記事からなる構造も，同様の構造と見なせる．

次節以降，電子化辞書とその概念体系を同一視して同じ記号（$\mathcal{A}$，$\mathcal{B}$ など）で表すことがある．また，ある概念体系 $\mathcal{A}$ に属す概念 A に対して，A および A の子孫概念に結び付けられたすべての単語からなる集合を $word(A)$ と記す．

3.2 対応付けの定義

一般に，概念をどのような単位で設定するかは概念体系によって異なる．そこで，1 個の概念に対する対応付けの相手を単一の概念と限定するのは必ずしも適切でなく，複数の概念の和となる可能性も考慮することが望ましい．そこで本研究では，ある概念体系 $\mathcal{A}$ に属す概念 A に対して，別の概念体系 $\mathcal{B}$ に属す類似した意味を持つ概念の集合 $\{B_1, B_2, \ldots, B_n\}$ を対応付けることを考える．ある概念集合 $\{B_1, B_2, \ldots, B_n\}$ が概念 A の対応付け相手として最適である条件を直接定義するのは困難なので，ある概念 B_i が概念 A に対応付けられる概念集合の要素となるための条件を定義し，その条件を満たす概念を集めることにより対応付け相手となる概念集合を求めることを試みる．この際の概念 B_i と概念 A の関係は，A が B_i を意味的に包摂するという関係である．

2 つの概念の意味的な関係を推測する方法として，概念名を利用する方法や関連概念（親概念，子概念など）を考慮する方法，概念のインスタンスの性質を用いる方法などがあるが，ここでは 1 個の概念に多くの単語が結び付けられているという電子化辞書の特徴を生かして，単語集合 $word(A)$ と $word(B)$ の重なり具合に基づき概念 A, B の包摂関係を定義するというアプローチをとることにする．

まず，2 つの単語集合の重なり具合を表す指標を定義する．概念 A による概念 B の被覆率 $cover(A, B)$ は，概念 B に結び付けられた単語のうち概念 A にも結び付けられている単語の占める割合を表す．

$$cover(A, B) \stackrel{\text{def}}{=} \frac{|word(A) \cap word(B)|}{|word(Root_{\mathcal{A}}) \cap word(B)|}$$

ここで $Root_{\mathcal{A}}$ は概念体系 $\mathcal{A}$ のルート概念であり，$word(Root_{\mathcal{A}})$ は辞書 $\mathcal{A}$ の全単語からなる集合を表している．この集合を考慮するのは，被覆率の算出においてそもそも辞書 $\mathcal{A}$ に含まれない単語の影響を排除するためである．また，$|S|$ は集合 S の要素数を表す．

次に，概念 A が概念 B を意味的に包摂するという関係 $subsume(A, B)$ を次のように定義する．

$$subsume(A, B) \stackrel{\text{def}}{\Longleftrightarrow} cover(A, B) \geq p \text{ かつ } cover(B, A) \geq q$$

ここで p と q は閾値を表す定数である．包摂関係を単語集合間の包含関係に近づけるには $(p, q) = (1, 0)$ とすればよい．p を 1 より少し小さい値に設定すると，厳密には A に包含されないが意味的に似ているという概念を対応付けに含めることができ，結果として単語集合間の重なりがより大きな対応付けが得られる可能性もある．したがって，p の適切な値は目的に応じて決める必要がある．

一方 q には，概念 A に対応付けられる概念集合が細かい概念の寄せ集めになることを防ぐ意図がある．例えば概念 B が唯一の単語 w と結び付いており，w は概念 A にも結び付けられていると仮定すると，$cover(A, B) = 1$ となる．概念 A に対応付けられる概念集合がこのような概念を多数含むと，対応関係が直感的に把握しづらくなり，応用上も望ましくない場合がある．q を適当な大きさに設定することで，このような細かい概念との対応付けを防ぐことができる．

最後に，概念体系 $\mathcal{A}$ に属す概念 A に対応付けられる概念体系 $\mathcal{B}$ に属す概念の集合を次のように定義する．

$$\{B \mid subsume(A, B) \text{ かつ } B \text{ の任意の先祖概念 } B' \text{に対して} \neg subsume(A, B')\}$$

与えられた概念 A に対して，$Root_{\mathcal{B}}$ を B の初期値として $subsume(A, B)$ を満たす B を $\mathcal{B}$ の枝に沿って再帰的に探索して集めることで，この概念集合を求めることができる．

3.3 外れ概念の除去

前節で述べた手法により概念集合を求めたとき，しばしば他の概念と意味的に離れた概念が概念集合に含まれることがある．このような概念は，多くの場合，複数の意味を持つ単語を介して概念集合に紛れ込んだものと考えられ，除去することが望まれる．そこで，求めた概念集合 $\{B_1, B_2, \ldots, B_n\}$ において残りの概念と意味的に離れた概念を検出し除去することを考える．

まず，各 $B_i\,(1 \leq i \leq n)$ に対して，B_i と $B_j\,(1 \leq j \leq n$ かつ $i \neq j)$ の間の概念体系 $\mathcal{B}$ 上の距離（最短経路に含まれる枝の数）の平均値 $distance_i$ を求める．次に，スミノルフ・グラブス検定により各 $distance_i$ の中に外れ値が存在するか調べる．もし存在した場合は，対応する概念を除去した上で，残りの概念集合に対して再度同じ手続きを実行する．

4 対応付けの応用

4.1 単語の補完

概念の対応関係を利用して，各電子化辞書において概念に結び付けられた単語の集合を相補的に拡張することができる．概念 A に概念集合 $\{B_1, B_2, \ldots, B_n\}$

を対応付けた場合，概念 A に新たに結び付けられる（補完される）単語の集合は

$$\left(\bigcup_{i=1}^{n} word(B_i)\right) - word(A) \qquad (1)$$

で表される．

包摂関係の定義における閾値 p が 1 のとき，補完される単語は辞書 $\mathcal{A}$ に存在しない単語に限られる．p を 1 より小さい値にすると，もともと辞書 $\mathcal{A}$ の別の概念に結び付けられていた単語を補完集合に加えることができるが，意味の似ていない単語との不適切な関連付けがなされる可能性も高くなるため，実験を通して p の適切な値を決定する必要がある．

4.2 文の意味解析への応用

電子化辞書によっては，動詞がとる格要素の種類と意味的制約を整理した格フレーム辞書を備えているものがある．このような格フレーム辞書は，文中の多義語の曖昧性解消や省略された単語の補完など，意味解析や文脈解析に利用することができる．

前節で述べた概念と単語の結び付きの補完により，格フレームの適用範囲を拡大することが可能になる．また，逆方向の対応付けも併用することで，2 つの辞書の格フレームを相補的に利用することも考えられる．これらによって文の意味解析の精度や汎用性を高めることができると期待される．

5 評価実験

提案した対応付け手法の有効性を評価するために，日本語語彙大系（NGT と記す）の意味属性を EDR 電子化辞書（EDR と記す）の概念に対応付ける実験を行った．また，この対応付け結果の妥当性を単語補完および意味解析への応用という観点から検証するために，被験者実験を行った．

5.1 概念の対応付け実験

まず 3.2 節で述べた手法を用いて，NGT の一般名詞属性 2,710 個それぞれに対して，対応する EDR の概念識別子の集合を求めた．この際，閾値 p，q の値の組み合わせを何種類か変えて実験した結果，$(p, q) = (0.8, 0.03)$ という組み合わせを基本とすることにした．

その結果，全 2,710 個の意味属性のうち 2,474 個に対して 1 個以上の要素を持つ EDR の概念集合を対応付けることができた．対応付けられた概念集合のうち約 6 割は要素数が 10 個未満であったが，要素数の最大値は 79 個，平均値は 12.2 個であった．得られた対応関係の例を表 1 に示す．

表 1: 概念の対応関係の例

NGT 意味属性	対応する EDR 概念集合
【植物】	〈植物〉,〈果物〉,〈野菜〉,〈実〉
【贈り物】	〈謝礼〉,〈御土産〉, 〈年の節目に贈る品物〉
【コンピュータ】	〈記憶装置〉,〈電子計算機〉, 〈ハードウェア〉,〈出力装置〉
【抽象物（精神）】	〈学問〉,〈文学〉
【人事】	〈解雇する〉,〈募集する〉

表 2: 外れ概念の例（下線部が外れ概念）

NGT 意味属性	対応する EDR 概念集合
【医師】	〈医師〉,〈歯科医〉,〈骨接ぎ〉, 〈藪医者〉, 〈医学〉
【光学用部品】	〈鏡〉,〈反射鏡〉, 〈鏡物（物語）〉
【勝敗】	〈勝利する〉,〈負け（争い）〉, 〈負け（値段）〉

続いて，3.3 節で述べた手法で外れ概念の検出を行った．その結果 826 個の意味属性に対応する EDR 概念集合において 1 個以上の外れ概念が検出された．外れ概念検出の例を表 2 に示す．

5.2 単語補完の妥当性評価

対応付け実験の結果を利用して，NGT の各意味属性に結び付けられた単語集合の補完を行う．ここでは比較のため $p = 0.8$ の場合に加えて $p = 0.7$ および 0.9 で対応付けを行った場合の結果も求めた．単語補完の結果を表 3 に示す．表において，対応概念存在率は全

表 3: 単語補完率の比較

p	対応概念存在率	単語補完率		
		NGT 内	NGT 外	計
0.7	0.95	0.09	0.90	0.99
0.8	0.91	0.04	0.69	0.73
0.9	0.85	0.01	0.38	0.39

表 4: 補完された単語の例

NGT 意味属性	単語
【子】	やや，童子，チビッコ，×跡職
【出版等】	DTP，委託出版，重刻，×再生
【時】	ひと時，瞬く間，食時，間合い，×ワンタッチ

2,710 個の意味属性のうち 1 個以上の EDR 概念を対応付けることができた意味属性の割合を表す．また，単語補完率は式 (1) で表される補完集合の要素数を元の単語集合 $word(A)$ の要素数で割った値を表す．表には，補完された単語を NGT 内に存在する（他の概念に結び付けられた）単語と存在しない単語に分けて求めた数値も示されている．

次に，補完された単語の妥当性を確認するための被験者実験を行った．任意に選んだ NGT 意味属性 30 個に対して，$p = 0.7, 0.8, 0.9$ のそれぞれの条件の下で補完される単語を求めて，それらの中からランダムに 10 個ずつ選んだ単語群を被験者に提示した．被験者は NGT において各意味属性に結び付けられている単語の一覧を見ながら，提示された各単語に対して，その単語を元の単語一覧に加えても構わないか，加えるべきでないかを回答した．3 名の被験者に実験を行い，2 名以上が「加えても構わない」と回答した単語の割合を求めた．その結果，$p = 0.7, 0.8, 0.9$ それぞれの場合に $0.926, 0, 937, 0.963$ という値が得られた．補完された単語の例を表 4 に示す．表の中で×印がついた単語は，被験者が妥当でないと判断した単語である．

5.3 格フレームの利用における妥当性評価

文の意味解析への応用を目指し，NGT の構文体系から選んだ「N が V」という形の文型パターン 30 個に対して，名詞 N に対する選択制限を満たす補完単語をそれぞれ 10 個選び，計 300 個の文を作成した．また比較のため，同じ文型に対して NGT に既存の単語を用いて同数の文を作成した．これら計 600 文を 3 人の被験者に提示し，各文が意味的に許容されるかどうかを主観的に判断してもらった．

実験の結果，意味的に許容される文の割合は，NGT に既存の単語を用いて作った文で 0.700，補完された単語を用いた文で 0.777 であった．この結果から，対応付けにより補完された単語は格フレームの利用において元からある単語と同等の妥当性を持つことがわかる．補完単語を用いて作った文の例を表 5 に示す．

表 5: 格フレームを用いて作った文の例

文型	単語
【具体物】が砕ける	サンドエッジ 液晶テレビ ×ローファットミルク
【人工物】が作動する	STB，椅子車 ×熱帯果実，×氷詰

6 まとめ

複数の電子化辞書に含まれる階層的に分類された概念間の対応付けを行う手法を提案し，日本語語彙大系の意味属性を EDR の概念集合に対応付ける実験を行うことにより，その有効性を確認した．

今後は，概念の対応付けに基づく格フレームの相補的利用など，文の意味解析への応用を進めるとともに，対象とする電子化辞書や言語資源の種類の拡大にも取り組んでいく予定である．

参考文献

[1] 日本電子化辞書研究所: EDR 電子化辞書 第 2 版, 2001.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦編: 日本語語彙大系 CD-ROM 版, 岩波書店, 1999.

[3] 国立国語研究所編: 分類語彙表 増補改訂版, 大日本図書, 2004.

[4] Fellbaum, C.: *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[5] 麻野間直樹, 松尾義博: 日英間のシソーラス対応と構造比較, 電子情報通信学会技術研究報告 NLC 2001-6, 2001.

[6] 柴木優美, 永田昌明, 山本和英: 日本語語彙大系を用いた Wikipedia からの汎用オントロジー構築, 情報処理学会研究報告 2009-NL-194(4), 2009

[7] 市瀬龍太郎, 武田英明, 本位田真一: 階層的知識間の調整規則の学習, 人工知能学会論文誌, Vol. 17, No. 3, pp. 230-238, 2002.

[8] Euzenat, J. and Shvaiko, P.: *Ontology Matching*, Springer, 2007.