E-047

単語間の意味カテゴリー距離に基づく用例ベース未知語意味カテゴリー推定
Example-based Inference of Unknown Word Meaning Based on Word Semantic Distance

福岡 知隆† 服部 峻‡ 久保村 千明* 亀田 弘之‡
Tomotaka Fukuoka †, Shun Hattori ‡, Chiaki Kubomura *, and Hiroyuki Kameda ‡

1. はじめに

近年の情報通信技術の進歩により，人間の対話相手が増えた．人間に比べて膨大な情報の保持が可能なコンピュータである．チャットなどでの雑談相手，Web 上での商品の説明，介護における話し相手など，多岐にわたり人間はコンピュータと対話するようになった．

しかし，人間同士の対話と比較すると，コンピュータの返答結果や対話の過程は劣っている場合が多い．その原因の一つが円滑性（発話者の意図に沿い，対話が速やかに行われること）の欠如である．コンピュータの単語辞書内に未登録な単語，即ち未知語に遭遇した場合にその現象は著しい．従来の対話システムでは未知語に対してまったく対応できなかったり，あるいはできたとしても人間への質問や話題転換が頻繁に起こってしまい，対話の円滑さが損なわれる場合がある．

この問題を解決するため，対話システムにおける未知語処理を改善し人間とコンピュータ間の対話をより自然で円滑にする必要がある．一つの解決手法として，システムが自動的に未知語の情報を推定することで，既知語だけの発話と同様に応答することが可能になると考えられる．

本稿では未知語の意味カテゴリーを未知語を含む入力文とシステムが保持する用例文の類似性に基づき推定する手法の提案と評価を行う．

2. 従来研究とその問題点

システムが未知語の品詞や意味を推定する手法の多くは，例えば，その未知語そのものを含む他の文を辞書や新聞コーパスから獲得し，利用している．確かにこのような手法では Web などの膨大な量の情報源を利用して行われるため，システムが未知語の用例を獲得し，情報を推定できる可能性は大きい．下山ら[1]は Web 上に存在する未知語に対して，その出現ホストの類似性を用いて分野を推定している．この手法では，単語のクラスタリングを行い，同じクラスに属する既知語の分野から未知語の分野を推定しており，約 63%の精度で未知語の分野推定が可能となっている．また，浦本[2]は未知語の係り受け関係を利用し，その類似語のシソーラス上の位置から未知語の意味カテゴリーの推定を行う手法を提案している．佐藤ら[3]は用例に基づくのではなく，未知語の説明文を Web 上から参照し分類を行う手法を提案している．

しかし，これらの手法では未知語が極めて新しい造語である場合や，ごく一部の人間の間でしか使われない用語の場合，未知語の用例が情報源に存在しないなど適切に対応ができない可能性がある．また，Web を情報源とする場合は情報の信憑性にも問題がある．さらに一つの未知語に対して，複数の用例を必要とするため，用例が入力文一つしか存在しない対話中での利用は難しい．

そこで我々は，当該の未知語を含まない情報だけに基づいて未知語の情報を推定する手法の可能性を研究している．当該未知語の用例文を用いずに，未知語が含まれている入力文における品詞並びパターンや単語間の共起パターンに着目し，これに類似する類似用例文（ただし，この用例文には当該未知語は含まれていない）から未知語に関する諸情報を推測する．この手法ならば未知語の用例がその入力文一つだけでも対応が可能である．以下，我々が提案する手法について詳しく述べる．

なお，情報源としては，不特定多数者が作成する Web などの情報ではなく，システム管理者が精査した対話用例データなどを用いる．

3. 提案手法

既に述べたように本稿で提案する未知語意味カテゴリー推定手法では，あらかじめ用意した用例の中から入力文と類似した用例を抽出し用いる．入力文と用例の類似度は品詞並びと意味カテゴリーそれぞれの類似度から決定する．この二つの類似度に基づいて類似している用例が一つ選択され，その用例中の単語に着目して未知語の意味カテゴリーを推定する．

3.1 処理手順

提案する手法は以下の手順で意味カテゴリー推定を行う．

Step1：入力文の情報取得 MeCab を用いて入力文の形態素解析を行い，得られた入力文に対して，単語データベースを参照し意味カテゴリーを与え，その集合を入力文の意味カテゴリー群として取得する．

Step2：品詞並び類似用例検索 入力文の中から未知語とその前後の単語の三形態素の品詞並びをクエリとし，表層類似度に基づき用例データベースから類似用例検索を行う．

Step3：意味カテゴリー類似用例選択 入力文の単語の意味カテゴリーと類似用例の単語の意味カテゴリー群から二つの文の意味カテゴリー類似度を計算し，その類似度が最大の用例を選択する．

Step4：意味カテゴリー推定 選択された類似用例の単語の中から入力文と品詞並び上同じ位置にある単語の意味カテゴリーを用いて未知語の意味カテゴリーを推定する．

†東京工科大学大学院バイオ・情報メディア研究科，Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
‡東京工科大学, Tokyo University of Technology
＊山野美容芸術短期大学, Yamano College of Aesthetics

3.2 データベース

本手法においては，単語データベースと用例データベースの 2 種類のデータベースを使用する．単語データベースは ipadic2.7.0 を使用し，各単語に関する意味カテゴリー情報は日本語 WordNet [4]から取得した．すべての単語に対して意味カテゴリー情報を取得することはできなかった．表 1 に WordNet で分類されている 4 品詞それぞれにおいて ipadic にも含まれる単語数と ipadic の単語数を示す．また，本研究では ipadic の基本品詞 13 種類に加え，その中の助詞を細分化した 22 種類の品詞を用いた（表 2）．

一方，用例データベースは，Web 上に公開されている対話コーパスを用いる[5]．対話文章を文単位に分割し，一つのスキーマとしてデータベースに格納する．コーパスから用例を獲得する際には単語データベースに存在しない未知語が出現した文を除いたため，本稿における用例データベースは 1279 件の用例となった．

表 1　品詞別の WordNet と ipadic の共通単語数

品詞	WordNet と ipadic 共通単語数	ipadic 単語数
名詞	33818	229689
動詞	4959	130749
形容詞	949	27565
副詞	1057	3031

表 2　助詞を細分化した品詞の分類

名詞	動詞	助動詞	形容詞
副詞	接続詞	感動詞	接頭詞
連体詞	記号	フィラー	その他
格助詞	**副助詞**	**終助詞**	**係助詞**
接続助詞	**並立助詞**	**副詞化**	**連体化**
副助詞／並立助詞／終助詞			**特殊**

3.3 品詞並びによる類似用例検索

類似用例を用例データベースから検索するクエリは，入力文の未知語とその前後の単語の品詞並びである．未知語をワイルドカードとし，未知語の前後の単語の品詞が一致する品詞並びを保持する用例を類似用例とする（図 1）．その際，各類似用例において，未知語の位置に存在する単語の情報を抽出し，単語の表層文字列を用いた表層類似度を計算する．計算式は以下のダイス係数を用い，w_sを表層類似度，$S(x)$を入力文 x の単語の集合，$S(y)$を類似用例 y の単語の集合とする．一つの文に二つ以上の同じ単語が存在した場合は，それぞれを異なる要素として扱う．

$$w_s(x,y) = 2 \times \frac{|S(x) \cap S(y)|}{|S(x)| + |S(y)|} \quad \cdots (1)$$

また，一つの用例の中で，複数箇所がクエリと一致した場合，それぞれを異なる類似用例とする．なお，未知語が文の先頭，もしくは文末にある場合はそれぞれ未知語の直後，直前の品詞だけを利用する．

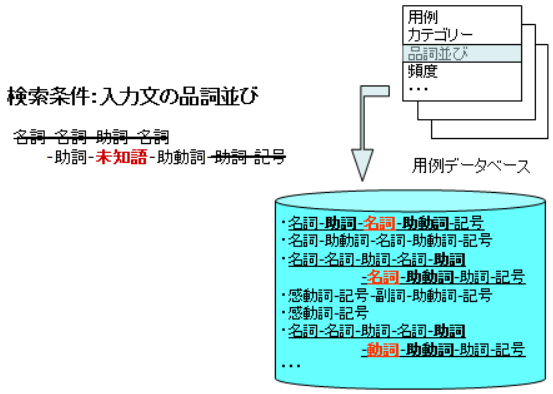

図 1　品詞並びによる類似用例検索

3.4 意味カテゴリー類似度の計算

入力文と 3.3 にて検索された用例の意味的な類似性は文中の単語の意味カテゴリー距離を基に計算される類似度により判断される．類似度は 2 種類あり，一つは入力文，用例それぞれに含まれる意味カテゴリー間の意味カテゴリー距離により求められる類似度である．もう一つは意味カテゴリー間の類似度を用いて計算される入力文と用例の類似度である．

システムは意味カテゴリーを持つ単語から，日本語 WordNet の「上位」関係に基づいた意味カテゴリーの木構造を用いて各単語からその最上位意味カテゴリーまでの上位意味カテゴリー群の塊を一つの要素として保持する．単語によっては複数の意味カテゴリーを持つ場合や最上位意味カテゴリーが異なる場合，同じ最上位意味カテゴリーまでに含まれる上位意味カテゴリーが異なる場合があるが，それらはすべて異なる要素として扱う．システムはこれらの要素を用いて意味カテゴリー距離を取得し，意味カテゴリー間の類似度を求める．ここでの意味カテゴリー距離とは，日本語 WordNet の意味カテゴリーの「上位」関係に基づく木構造においての二つの意味カテゴリー間の枝の数である（図 2）．

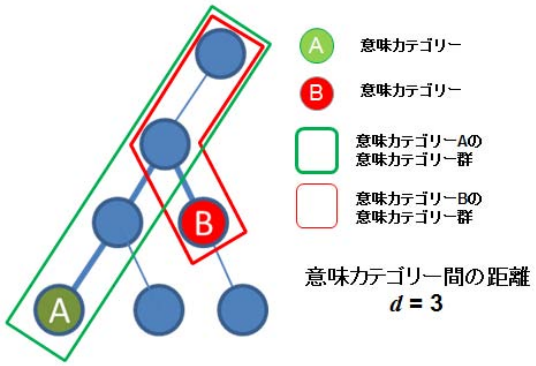

図 2　木構造上での意味カテゴリーの位置

意味カテゴリーa，b の類似度は以下の式(2)で計算される．w を類似度とし，d はそれぞれの意味カテゴリー中の最下位の接点の間にある枝の数である$(0 < w \leq 1)$.

$$w(a,b) = \frac{1}{d(a,b)+1} \cdots (2)$$

入力文と用例の類似度の計算は，異なる二つの手法を用いる．以降は各計算手法について述べる．

3.4.1 すべての意味カテゴリー間の類似度に基づく手法

入力文と用例がそれぞれ保持する意味カテゴリー間の類似度をすべて用いて入力文と用例の意味カテゴリー類似度を計算する．即ち，入力文に存在する意味カテゴリーと用例に存在する意味カテゴリー一つ一つの類似度を計算し，その総和を用いる．例えば入力文と用例が未知語部分以外は全く同じだった場合でも類似度が低くなるが，すべての意味カテゴリーの類似度が大きい用例を選択しやすくすることで，文全体に対して類似する未知語の意味カテゴリーを推定する（図 3）．

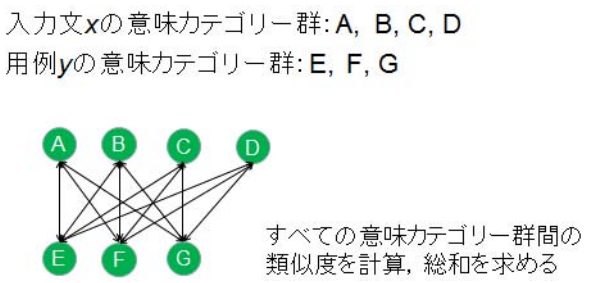

図 3　すべての意味カテゴリー群の類似度

入力文 x と用例 y のすべての単語の意味カテゴリー群間の類似度の総和を入力文 x の意味カテゴリー集合 $C(x)$および用例 y の意味カテゴリー集合 $C(y)$それぞれの要素数の積の平方根で割ることで，入力文 x と用例 y の意味カテゴリー類似度w_c^1を計算する．

$$w_c^1(x,y) = \frac{\sum_{a\in C(x)} w(a,b)}{|C(x)|\cdot|C(y)|} \cdots (4)$$

すべての類似用例の意味カテゴリー類似度を計算し，w_c^1が最も大きい用例を選択する．類似用例には未知語と同じ位置にある単語（以降類似語とする）がそれぞれ定められているので，その単語の意味カテゴリー情報を未知語の意味カテゴリー情報とすることで推定処理を行う．意味カテゴリー類似度が同一の用例が複数存在した場合は，表層文字列の類似度が大きい用例を選択し，表層文字列の類似度も同じだった場合は，先に用例データベースに格納している用例を選択する．また，すべての類似用例が意味カテゴリーを持つ単語を含まなかった場合は，表層類似度の大きい用例を選択する．

3.4.2 意味カテゴリー類似度の最大値に基づく手法

一つの意味カテゴリーの類似度が最大となる組み合わせにより，入力文と用例の類似度を計算する．この計算手法は以下の順に行われる．

① 入力文 x に含まれる意味カテゴリー群 $C(x)$と用例 y に含まれる意味カテゴリー群 $C(y)$それぞれの組み合わせの意味カテゴリーの類似度を計算する．
② 最も類似度 $w(a,b)$が大きい意味カテゴリーa と b の組を一つ選択する．類似度が最大の組が複数あった場合は $C(y)$において先に出現する意味カテゴリーを含む組を優先する．$C(y)$の意味カテゴリーが同じであった場合は，$C(x)$において先に出現する意味カテゴリーを含む組を選択する．
③ 選択された組の意味カテゴリーを含むその他の組の類似度を 0 に変更する．
④ ②，③を入力文と用例それぞれの意味カテゴリー群数の最小値と同じ回数繰り返す．
⑤ 式(5)を用いて入力文と用例の類似度w_c^2を計算し，3.4.1 と同様にして未知語の意味カテゴリーを推定する．

$$w_c^2(x,y) = \frac{\sum_{\substack{a\in C(x)\\ b\in C(y)}} w(a,b)}{\sqrt{|C(x)|\cdot|C(y)|}} \cdots (5)$$

4. 評価実験

実験 I

（1）**目的**

3.4.1 と 3.4.2 で提案した二つの意味カテゴリー類似度計算手法の有効性を確認する．

（2）**方法**

提案手法の意味カテゴリー推定精度を評価するため，未知語を含む入力文を用意し，その意味カテゴリー推定を行った．入力文は新語辞典[6]の用例を用いた．まず，新語辞典に記載されている用例を，MeCab により形態素解析し，その中から未知語を一つだけ含む用例抽出する．なお，形態素解析の結果， 辞典中では新語とされる単語が既知語であり，用例中のその他の単語が未知語扱いされる場合の用例も含める．

次にこれらの用例の中から，未知語の意味カテゴリーを一意に決定できる用例を選択する．本評価実験では，類語辞典[7]において未知語に対する見出し語が存在し，なおかつその見出し語が WordNet 上での意味カテゴリーを有する場合，その見出し語の意味カテゴリーを未知語の意味カテゴリーとする．上記の条件を設定し，84 の未知語を含む入力文セットを用いた（未知語の重複を含む）．

なお，本評価実験においては，システムの処理速度を高めるため，**3.3** の処理が完了した時点で入力文の保持する意味カテゴリーを持たない類似用例は意味カテゴリー類似度の計算を行わないこととする．また，一つの入力文においてその品詞並び上の類似用例の類似語がすべて意味カテゴリーを持たない場合を除くとともに，類似語が意味カテゴリーを持たない類似用例の意味カテゴリー類似度計算を行わない．

（3）**素材**

入力文セットとして類語辞典より 84 文，単語データベースとして類語辞典を用いた（意味カテゴリーは WordNet を利用）．

（4）**結果と考察**

表 3 に二つの類似度計算手法に共通する実験結果を示す．

表 3　両類似度計算手法に共通する結果

意味カテゴリーが推定された入力文数	45
意味カテゴリーが推定されなかった入力文数	39
各入力文の類似用例数の和	671
意味カテゴリーが推定できない平均類似用例数	2.15
クエリとなる品詞並び種類	39

また，表 4，5 にてそれぞれの計算手法に基づいた類似用例の意味カテゴリー類似度の平均と推定された未知語の意味カテゴリーの精度を示す．なお，すべての意味カテゴリー間の類似度に基づく手法を手法(a)，意味カテゴリーの類似度の最大値に基づく手法を(b)とする．

表 4　手法(a)による意味カテゴリー推定結果

平均意味カテゴリー類似度	0.03
平均意味カテゴリー精度	0.00

表 5　手法(b)による意味カテゴリー推定結果

平均意味カテゴリー類似度	0.08
平均意味カテゴリー精度	0.00

表 3 の結果において意味カテゴリーが推定されなかった入力文数は，類似用例が存在しなかった入力文数と，意味カテゴリー類似度が最も大きい用例の品詞並び上で未知語と同じ位置にある単語が意味カテゴリーを持たない入力文数を指す．表 1 で示した通り，今回使用した単語データベースは意味カテゴリーを持っている単語の方が少ない．そのため，類似語が意味カテゴリーを持たない類似用例が存在している．今回用いた入力文は類似用例検索に用いる品詞並びが固有である例は少なく，他の入力文と同じ類似用例が得られた中で，意味カテゴリーの類似度から未知語の正しい意味カテゴリーが取得可能かどうかも判断される．

表 4，5 の結果から両手法による平均意味カテゴリー推定精度は 0 であった．即ち，提案手法は用意した入力文セットに対して，正答と判断できる結果を示していない．この結果から，意味カテゴリー類似度による推定結果の優劣を判断できないため，意味カテゴリー類似度計算手法に対する議論は行わない．この結果を受けて，この問題点の原因としてまず考えられるのはデータベースである．単語データベースにおいては，意味カテゴリーを保持する単語は全体の 10%にすぎないため，ほどんどの用例においては単語の意味カテゴリーを十分に用いることができていない．単語データベースのすべての単語が意味カテゴリーを持つようにすることは難しいので，単語データベース全体のどの程度の単語が意味カテゴリーを持つことができれば，用例の意味カテゴリーの類似性が有効であるかを今後検討する必要がある．また用例データベースの規模も小さいため，意味カテゴリーを保持した類似用例数が少なく，意味カテゴリー類似度を用いた用例のランキングが行えていない．

実験 II

（1）**目的**

品詞並びの類似度に基づいて抽出した用例が意味カテゴリー推定に有効であるか確認する．また，用例データベース内に実験 I で用いた入力文セットの未知語の意味カテゴリーを推定精度が 0 より大きい用例が存在するか確認する．

（2）**方法**

意味カテゴリー推定精度が 0 より大きい用例を抽出し，その用例の意味カテゴリー類似度を評価する．また，用例データベースの中から，意味カテゴリーの推定精度が 0 より大きくなる単語を含む用例の抽出を行う．

（3）**素材**

表 3 の類似用例数の和と，用例データベースの全用例を用いた．

（4）**結果と考察**

すべての類似用例の中から，意味カテゴリー推定精度が 0 より大きくなる類似用例の抽出がされなかった．

また表 6 に用例データベースから直接未知語と推定精度の計算を行い 0 より大きい単語を含む用例が存在する入力文数と，その推定精度の平均値，手法(a)，(b)それぞれにおける類似度の平均値を示す．

表 6　用例データベースからの抽出結果

推定精度の判断可能な用例を持つ入力文数	25
推定精度の判断可能な各入力文の用例数の和	1172
平均推定精度	0.58
手法(a)を用いた平均類似度	0.04
手法(b)を用いた平均類似度	0.10

推定精度の判断可能とは，品詞推定精度が 0 より大きくなる単語を含むということであり，その用例が存在する入力文は 25 文であり，今回の評価実験に用いた 84 の入力文の 7 割は必ず意味カテゴリーの推定結果が誤りとなる．

品詞並びの類似性に基づいた類似用例の中には，意味カテゴリーが類似する用例が存在せず，表 6 より 1172 の用例が 0 より大きい推定精度を示している．また，類似用例の抽出処理を行っていない用例における意味カテゴリー類似度の平均値は表 4，5 の結果より高い数値であり，これは品詞並びによる意味カテゴリー類似用例の抽出は効果がないことを示している．意味カテゴリーを持つ単語の品詞が名詞，動詞，形容詞，副詞であるため，その前後の単語が助動詞などの意味カテゴリーを持たない場合が多く，意味カテゴリー上での類似度に対して効果が低い，もしくはないと考えられる．

5. おわりに

本稿では単語の意味カテゴリーと品詞並びを用いた未知語の意味カテゴリー推定手法の提案と評価を行った．

今回行った評価実験の結果は使用したデータベースの規模のために提案手法による意味カテゴリー推定の精度に関する議論は行えなかった．しかし，その結果品詞並びの類似度を用いることによる意味カテゴリー上の類似用例の抽出は効果がないという結論を得た．今後は単語データベースにおける意味カテゴリーを保持する単語の割合，用例データベースの拡大による意味カテゴリーを持つ用例数の変化と提案手法による意味カテゴリー類似度の評価を行う．

参考文献

[1] 下山 剛司, 秋岡 明香, 村岡 陽一, “単語の出現ホストを利用した未知語の分野推定”, 情報処理学会研究報告, pp71 -- 76 (2009).
[2] 浦本 直彦, “コーパスに基づくシソーラス:統計情報を用いた既存のシソーラスへの未知語配置”, 情報処理学会論文誌, Vol.37, No.12 , pp2128—2189 (1996).
[3] 佐藤 直人, 藤本 浩司, 小谷 善行, “ウェブ上の商品情報を利用した商品のカテゴリ分類”, 人工知能学会知識ベースシステム研究会, Vol.87, pp7—10, (2010).
[4] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/ (2011)
[5] 重点領域研究 音声対話 コーパス, http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/ (2011)
[6] 学研辞典編集部, 用例でわかるカタカナ新語辞典 改定第 2 版, (2007)
[7] 岡部 学, 用例でわかる類語辞典, 学研教育出版 (2009)