E-009

多対多の関係性を持つ多言語用例対訳のデータ構造の提案と評価
Proposal and Evaluation of Data Structure for Multilingual Parallel Texts Having Many-to-many Relationship

福島　拓 †
Taku Fukushima
吉野　孝 †
Takashi Yoshino

1. はじめに

現在，世界的なグローバル化により，日本国内でも多言語間コミュニケーションの機会が増加している．しかし，在日外国人や訪日外国人の中には，日本語を理解できない人が多数存在している [1]．一般に母語以外の言語によるコミュニケーションは困難である [2]．このため，正確なコミュニケーションが求められる分野の支援には，用例を正確に多言語へと翻訳した多言語コーパスである「用例対訳」が利用されている [3, 4]．

我々は正確性が求められる医療分野の用例対訳の収集，共有を目的とした，多言語用例対訳共有システム TackPad の開発を行っている [5]．しかし，用例対訳には各言語間で一対一に対応しない言葉の組み合わせが存在している．また，使用する相手やニュアンスの違いによって同じ意味の複数の言葉が存在する場合も多い．このような多言語間の言葉の多様性への対応が求められているが，単純な用例間の意味のつながり情報のみでは対応することができない．そこで本稿では，一対多，多対多の関係にある用例対訳に対応するために，メタデータを利用したデータ構造を提案する．本稿では，メタ用例ノード作成手法と用例対訳への適用評価について述べる．

2. 多言語用例対訳データ構造

2.1 多言語用例対訳の一対多，多対多における問題点

本稿で提案する多言語用例対訳のデータ構造は，一対多，多対多の関係性を持つ用例対訳を，多言語対応システムで利用可能にするために用いる．用例対訳の例を図 1 に示す．図 1 中の四角で囲まれた文字は用例を示す．また，二重線は用例作成者が同一意味と判定したことを示す．本稿では，図 1 中の二重線を「用例間リンク」とする．図 1 では，一つの韓国朝鮮語の用例に，複数の日本語の用例が「用例間リンク」で結合されている．仮に，この状態の用例対訳を多言語対応システムに提供した場合，複数の日本語が含まれているため，用例対訳の利用を簡単に行うことができない．

そこで，本稿では システムが自動で，同一意味の用例対訳であることを意味するメタノードの作成 (図 1 中の黒丸) を行う．本稿では「メタ用例ノード」とする．メタ用例ノードは次の条件を満たしている．

- メタ用例ノードは同じ意味の多言語用例を含む
- 一つのメタ用例ノードには，同一言語の用例が複数含まれない

このため，メタ用例ノードを作成することで，同一言語の用例が含まれない用例対訳を提供することができる．

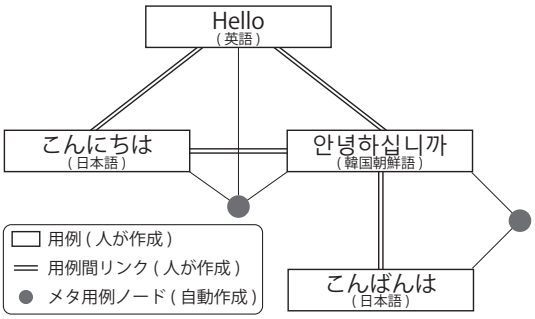

図 1: 多言語用例対訳のデータ構造例

2.2 メタ用例ノード作成手法

本節では，用例作成者が用例 α と用例 β の間に用例間リンクを作成した時を例に，システムが行うメタ用例ノードの作成について述べる．本手法は次の手順で行う．

(1) 用例 α と用例 β が元々持っている,「用例間リンク群」をそれぞれ抽出する．用例 α の「用例間リンク群」と用例 β の「用例間リンク群」の間で重複していた用例を選ぶ．本稿では重複していた用例群を「重複用例群」とする．
(2) 重複用例が 1 つもない場合は，(7)-(a) の処理を行う．
(3) 用例 α，用例 β からつながっている,「メタ用例ノード群」をそれぞれ抽出する．
(4) 用例 α の「メタ用例ノード群」の，各メタ用例ノードに対して，下記の操作を行う．
  (a) (3) で抽出した「メタ用例ノード」に含まれる各用例が，(1) で抽出した「重複用例群」に含まれているかどうかを調べる．
  (b) すべての用例が用例 α，もしくは,「重複用例群」に含まれる用例であった場合，走査していた「メタ用例ノード」に用例 β を追加する．
(5) 用例 β の「メタ用例ノード群」についても，(4) と同様の操作を行う．
(6) (4) または (5) で「メタ用例ノード」へ用例の追加が行われていた場合，下記の操作を行う．
  (a) 用例の追加が行われた「メタ用例ノード」間で，完全に一致する用例群を持つものができていた場合，それらの「メタ用例ノード」を統合する．
  (b) 「メタ用例ノード」に追加された用例を,「重複用例群」から削除する．
(7) この時点での「重複用例群」の数に合わせて，下記の操作を行う．
  (a) 「重複用例群」の数が 0 個の場合，かつ，(4) または (5) で「メタ用例ノード」への用例の追加が行われて いない 場合，用例 α と用例 β を新規の「メタ用例ノード」に追加する．

†和歌山大学

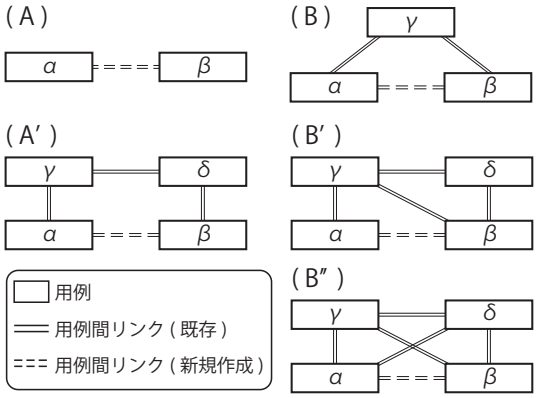

図 2: 多言語用例対訳の結合例

(b) 「重複用例群」の数が 1 個以上の場合,「重複用例群」間で「用例間リンク」が存在しているかを調べる．存在している場合，用例 α，用例 β と「重複用例群」の用例を新規の「メタ用例ノード」に追加する．存在していない場合，用例 α，用例 β と「重複用例群」の用例それぞれを新規の「メタ用例ノード」に追加する．

2.3 メタ用例ノード作成手法の網羅性

本節では，メタ用例ノード作成手法の網羅性について議論する．多言語用例対訳の結合例を図 2 に示す．図 2 中の二重破線は新規作成された用例間リンクを，二重実線はすでに作成されていた用例間リンクを示す．

用例間リンクによる結合の形は，図 2-(A) と図 2-(B) が基本形となり，複雑な用例間リンクが存在した場合も，この二つを組み合わせた形となる．例えば，図 2-(A') では，用例 α と用例 δ，用例 β と用例 γ の間に用例間リンクがないため，メタ用例ノード作成に関係する用例は用例 α と用例 β のみとなる．このため，図 2-(A') は図 2-(A) へ帰着する．また，図 2-(B') も，メタ用例ノード作成に用例 δ は関係しないため，図 2-(B) へ帰着する．同様に，図 2-(B") は用例 α，β，γ の組み合わせと，用例 α，β，δ の組み合わせそれぞれが図 2-(B) へ帰着する．このように，用例間リンクのパターンは図 2-(A) と図 2-(B) に帰着する．なお，図 2-(A) は，前節のメタ用例ノード作成手法の (7)-(a) の処理に当たる．また，図 2-(B) 中の γ は，メタ用例ノード作成手法の「重複用例群」に当たる．これらのことから，メタ用例ノード作成手法はすべてのパターンの用例間リンクを網羅している．

3. 多言語用例対訳共有システムへの適用

多言語用例対訳データ構造を，多言語用例対訳共有システム TackPad で収集済みの用例対訳に適用した．TackPad の収集言語は日本語，英語，中国語，韓国朝鮮語，ポルトガル語，スペイン語，ベトナム語，タイ語，インドネシア語の 9 言語である．収集済みの用例数は 14,487 件，用例間リンク数は 18,285 件であった．

メタ用例ノードに結合された用例数を表 1 に示す．3 つもしくは 4 つの用例が結合したメタ用例ノードが少ないという結果になった．これは，用例の新規作成を伴わない用例間リンク作成機能の提供を，現時点で行っていないためであると考えられる．また，5 つの用例が結合したメタ用例ノードが一番多いという結果になった．これは，収集済みの用例のうち，1 万件あまりの用例はあらかじめ用意した用例対訳をデータベースに直接挿入していることが影響していると考えられる．

表 1: メタ用例ノードに結合された用例数

結合された用例数	メタ用例ノード数
2	1399
3	37
4	75
5	4317
合計	5828

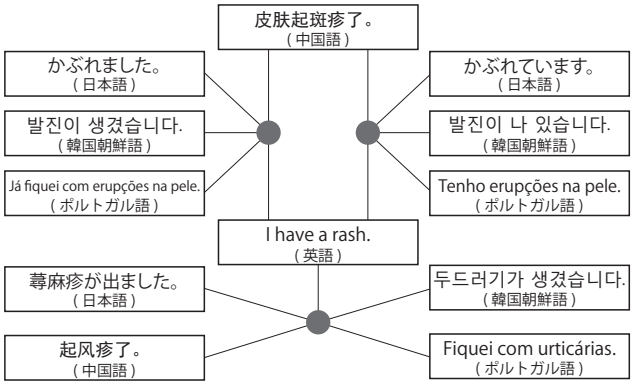

図 3: 作成されたメタ用例ノードの例

作成されたメタ用例ノードの一例を図 3 に示す．この用例対訳群は，英語 1 用例，中国語 2 用例，日本語，韓国朝鮮語，ポルトガル語が各 3 用例が結合されていた．図 3 の下のメタ用例ノードには病名が含まれた用例が結合されている．また，上の二つのメタ用例ノードにはニュアンスが多少異なる症状の用例が結合されており，メタ用例ノードによる分類を行えていることが分かる．

4. おわりに

本稿では，多対多の関係性を持つ多言語用例対訳のデータ構造の提案と，その構築手法について述べた．今後は，実システム上での継続的な運用と，メタ用例ノードで結合された用例対訳の提供を行う．

謝辞

本研究に関して，ご助言をいただいた大阪大学林良彦教授に感謝する．また，本研究の一部は，科研費基盤研究 (B)(22300044) の助成を受けたものである．

参考文献

[1] 田村太郎：多民族共生社会ニッポンとボランティア活動，明石書店 (2000).
[2] Takano, Y., et al: A temporary decline of thinking ability during foreign language processing, Journal of Cross-Cultural Psychology, 24, pp.445-462(1993).
[3] 宮部真衣ほか：外国人患者のための用例対訳を用いた多言語医療受付支援システムの構築，信学論，Vol.J92-D,No.6，pp.708-718(2009).
[4] 福島拓ほか：用例対訳を用いた多言語問診票作成システムの開発と評価，情処研報，2011-GN-78(14)，pp.1-7(2011).
[5] 福島拓ほか：医療分野を対象とした多言語用例対訳収集 Web システム TackPad の開発，情処，DICOMO2008 シンポジウム，pp.1030-1036(2008).