様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２２ 日現在

機関番号：８２６３６
研究種目：若手研究(B)
研究期間：2013～2015
課題番号：２５７３０１３９

研究課題名（和文）多言語の対照分析に基づく言語横断的な言い換え分類体系の構築

研究課題名（英文）Building Paraphrase Typology through Comparative Linguistic Analysis

研究代表者
藤田　篤（Fujita, Atsushi）

国立研究開発法人情報通信研究機構・ユニバーサルコミュニケーション研究所多言語翻訳研究室・主任研究員

研究者番号：１０４０２８０１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究課題では，同じ意味内容を表す同一言語の異なる言語表現(言い換え)を幅広く収集，整理し，類型化した．分析対象の事例は，言語学，自然言語処理，言語教育等の多分野の文献からの抽出，含意関係認識のベンチマーキング用に作成されたデータ中の事例の人手による分解，機械翻訳の質が向上するように人間によって試みられた原文の書き換え，の3種類の方法で収集した．そして，従来研究と同様に言い換え対象の範囲，同義性を説明するための言語学的な要素に関する情報のほか，言い換えによって生じうる言外の意味の違い，言い換え生成時に評価すべき文脈要素などの情報を付与しつつ，類型化の基準の外在化を試みた．

研究成果の概要（英文）：We aimed to build a language-independent typology of paraphrases through collecting and analyzing a wide variety of paraphrase examples. We collected the examples in three ways: (i) extracting those illustrated in the books and articles in the related research fields, including linguistics, natural language processing, and language pedagogy, (ii) manually analyzing existing dataset created for benchmarking the systems for recognizing textual entailment, and (iii) incrementally pre-editing sentences so that they can be translated at a higher quality by a machine translation system. We also made an attempt to find a set of descriptive features toward explicit definition of each type of paraphrases. In addition to the features proposed in previous work, i.e., the surface-level scope of paraphrases and linguistic features associated with the paraphrases, the connotational differences that paraphrases incur and contextual elements that could be affected by paraphrasing, were examined.

研究分野：自然言語処理

キーワード：言い換え　パラフレーズ　類型化　自然言語処理

１．研究開始当初の背景

同じ意味内容を表わす同一言語の異なる言語表現，すなわち言い換え(パラフレーズ，換言)を計算機によって自動的に認識したり生成したりする技術は，様々な自然言語処理技術の根幹をなす基本的な技術である．言い換えに関する既存の研究は，次の 5 段階のレイヤに分けられる．

(a) 言い換えの類型化: 様々な種類の言い換えを分類・整理する研究である．ある程度多様な現象を列挙・整理した例は存在する[Fujita, 2010][Vila et al., 2011]ものの，網羅性，言語横断性，類型化の観点の妥当性，客観性等に関する議論はなされてきていない．

(b) 特定の類型の言い換えの詳細な分析: 言語学的研究，工学的研究のいずれも存在する．ただし，各類型について独立に論じられることが多く，異なる類型間の共通性に関する議論は，特定の理論の視点でその有用性を論じたもの[Mel'čuk and Polguèle, 1987][影山, 2001]以外には存在しない．

(c) 言い換え知識の自動獲得: 大規模なテキストデータから，「刺激を受けた」と「刺激された」や「という」と「そうだ」のような同義の表現の対を自動的に獲得する研究が盛んに行われている．ただし，得られる言い換えの種類は，テキストの解析レベルおよび表現間の同義性を近似するヒューリスティクスによって制限される．

(d) 言い換え生成・認識の計算機構: 上記(c)で構築される知識や機械学習等に基づく様々な手法が検討されている．ベンチマーキング用のデータを作成する際，効率化のために，上記(c)と同様にヒューリスティックに候補を得た後に人間が精査する方法を取ることが多く，多様な言い換えのうちどのような現象をどの程度解けるかを，適切に評価することができていない．

(e) 応用技術への適用: 情報検索，質問応答，機械翻訳等の応用技術に言い換え技術を導入し，その効果を調査した研究はある．しかし，どの類型の言い換えがどの程度有効であるかを調査した例はない．

このように，最も基本的な(a)のレイヤにおいて，言語横断的かつ網羅的かつ見通しのよい言い換えの分類体系を構築できていないことが，その他のレイヤの研究の発展の障害となっている．

２．研究の目的

言い換えに関する上述のすべてのレイヤの研究の高度化のために，その道標としての言い換えの分類体系が不可欠である．そこで本研究では，日本語，英語，仏語の 3 言語を対象として，言い換えの類型を網羅する言語横断的な分類体系および事例集を構築することを目的として研究を実施した．

３．研究の方法

網羅性の高い分類体系を実現するには，前提となる多様な事例の効率的な収集が鍵となる．また，言語横断的な分類体系を実現するには，複数の言語の事例を収集するとともに，言語に固有な現象および共通点を見出す必要がある．1 つ目の課題，すなわち事例の多様性の確保については，次の 3 種類の方法を検討した．

(A)言語学，計算言語学，自然言語処理，言語教育等に関する既刊の文献(図書，論文)から，言い換え事例を網羅的に収集する

(B)従来手法(上述の背景の(d))と同様にして候補となる事例を自動生成し，それらを人間が精査する

(C)内省に基づいて新規の事例を作例する

一方，2 つ目の課題に対しては，海外の研究者の協力を仰ぎ，文法性や意味の同義性に関して厳密な判断をするのに必要十分な言語能力を有する母語話者を雇用して作例に従事してもらうことにした．また，その際に，個々の事例に対して下記に示す 4 種類の特徴量を付与し，類型の定義を特徴量の組み合わせによって外在化することにした．最初の 2 件は従来研究で扱われていた情報である．

- 言い換え対象の範囲(単一の語，節内，節よりも大きな表現) [Fujita, 2010]
- 同義性を説明するための言語学的な要素(形態論，統語論等) [Vila et al., 2011]
- 言い換えによって生じうる言外の意味の違い(形式性，感情極性等)
- 言い換え生成時に評価すべき文脈要素(語の共起，文脈上の首尾一貫性等)

特定の言語に依存しないよう注力しつつ，事例の収集および類型化の作業を循環的に実施することによって，言語横断的な分類体系を確立することを目指した．

４．研究成果

事例収集手法(A)については，文法理論や意味論等の言語学関連図書，機械翻訳や質問応答等の応用技術に関する図書，ライティング教育等の言語運用技術関連図書合計約 20 冊，ならびに主に計算言語学に関する論文約 400 件から，約 2000 事例を抽出して電子化した．当初は，言語にかかわらず可能な限り事例を抽出し，類型ごとの言語固有性・共通性を判断する際に参照する予定であった．しかしながら，日本語，英語，仏語以外の言語については言い換えの適否の判断が困難であったため，原文中の記述に基づいて判断できなかった一部の事例については電子化を断念した．また，上述の 4 種類の特徴量を付与し，既存の分類体系[Fujita, 2010]における類型の定義としての有用性を確認した．

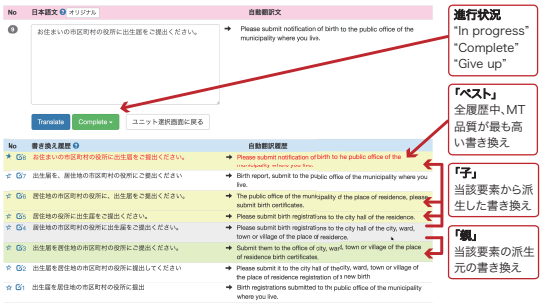
事例収集手法(B)については，含意関係認識に関するプロジェクト(RITE2) [Watanabe et al., 2012]において人手で収集された，複雑

な現象を含む言い換え事例 163 件を，国内の研究者と分担してプリミティブな言い換え事例，非言い換え事例に分解・分類する作業を実施した．この作業を通じて，同義と認められる事例を分解することによってより広範な現象が観察可能であること，ただし分解前の言い換え事例を収集する際のアルゴリズムおよび作業者のバイアスの影響で多様性に限界があること，人間による分解作業の実現可能性の 3 点が明らかになった．

事例収集手法(C)については，所与の文に対して可能な言い換えを付与するというパイロット作業を事前に実施し，制約のない条件下の作業者は一般的に，極めて限定された種類の言い換え(例えば名詞や動詞から同義語への置換)のみを繰り返すことが明らかになっていた．そこで，多様な言い換え事例を効率的に作例できるように，(i)既知の言い換え事例における表層的な差分を抽出してパターン化する手法を開発し，この手法を用いて(ii)人間による作例の新規性を収集済みの事例群に照らして評価する手法，および(iii)この機能を利用しながら作例を行うための作業環境を作成した．しかしながら，既知の類型の作例を許さないという極めて厳格な制約によって，作業者がほとんどまったく作例できない状況に陥ってしまった．

上記の検討をふまえ，より多様な事例を効率的に収集するために，目的指向型の課題を新たに設計した．具体的には，所与の機械翻訳システムによって得られる翻訳の質が向上するように，機械翻訳の入力文を，意味を変えないように漸次的に書き換える(翻訳前編集)課題を考案し，作業者を制御するためのプロトコル，システム，作業環境（図 1)を作成した．そして，これらを用いて日本語の言い換え事例を約 13,000 件，英語の言い換え事例を約 2,300 件，実際に収集し，それらの分解・類型化作業を経て，53 種類からなる類型にまとめた．文法性や意味の同一性に関する厳密な判断基準を事前に指定してあったにもかかわらず，作業者によっては，意味が変化してしまう事例や複数の言い換えが組み合わさった非プリミティブな事例を作例してしまうため，事例収集手法(B)と同様に，事例の選別および分解の作業が別途必要である．ただし，事例そのものは作業時間に応じて増やすことができるため，本研究課題の目的である多様な事例を収集する上では問題にならない．

事例収集手法のうち(B)および(C)においては，文法性や意味の同義性に関して厳密な判断ができる者を作例に従事させる必要があるため，初年度から学生アルバイトの人選を進めていた．しかしながら，研究代表者の所属機関変更後は雇用・支払い手続き上の制約により，特定の個人への作業依頼が不可能となり，配分された予算規模では公募によって作業者を雇用することもできなかった．やむを得ず，(C)の作業は外注によって進めたが，



計画時よりも高額となったため仏語には着手できなかった．また，実際の作業者の人選が不可能であり，研究代表者が作業者を直接監督することもできなかったため，文法性や意味の同一性に関する厳密な判断基準を事前に指定してあったにもかかわらず，意味が同義でない事例も多く含まれてしまった．

言い換えの体系については，各年度に 1 度ずつ，各時点の成果を発表した(学会発表(1), (4), (7))．また，文法性や意味の同義性に関して厳密な判断をするための基準を整理し，その客観性についても分析した(学会発表(6))．その過程で，人間による作例時の内省の限界について明らかになったため，新たに翻訳前編集という目的志向型の課題を考案した．本研究課題では，起点言語における言い換え事例の観察を目的としていたが，収集した事例は自動前編集技術の研究開発に資する．

＜引用文献＞

(1) Atsushi Fujita. Typology of Paraphrases and Approaches to Compute Them. Invited Talk, Workshop on Corpus-based Approaches to Paraphrasing and Nominalization, 2010.
(2) 影山太郎(編)．日英対照 動詞の意味と構文．大修館書店, 2001.
(3) Igor Mel'čuk and Alain Polguèle. A Formal Lexicon in Meaning-Text Theory (or How to do Lexica with Words). Computational Linguistics, Vol. 13, Nos. 3-4, pp. 261-275, 1987.
(4) Marta Vila et al. Paraphrase Concept and Typology: A Linguistically Based and Computationally Oriented Approach. Procesamiento del Lenguaje Natural, Vol. 46, pp. 83-90, 2011.
(5) Yotaro Watanabe et al. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, pp. 385-404, 2013.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 7 件）

(1) 宮田玲, 藤田篤, 内山将夫, 隅田英一郎. 機械翻訳向け前編集の事例収集と類型化. 言語処理学会第 22 回年次大会発表論文集, E5-4, pp. 869-872, 2016 年 3 月 10 日. 東北大学, 宮城県・仙台市
(2) 宮田玲, 藤田篤. 機械翻訳向け前編集に有効な書き換えルールに関する調査. 自然言語処理若手の会第 10 回シンポジウム, 2015 年 9 月 4 日. 和倉温泉 ホテル海望, 石川県・七尾市 【奨励賞受賞】
(3) Atsushi Fujita and Pierre Isabelle. Expanding Paraphrase Lexicons by Exploiting Lexical Variants. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 630-640, 2015 年 6 月 1 日. Denver (United States) 【査読あり】
(4) 藤田篤, 柴田知秀, 松吉俊, 渡邉陽太郎, 梶原智之. 言い換え認識技術の評価に適した言い換えコーパスの構築指針. 言語処理学会第 21 回年次大会ワークショップ『自然言語処理におけるエラー分析』発表論文集, ID:20, 11 pages, 2015 年 3 月 21 日. 京都大学, 京都府・京都市
(5) 藤田篤, Pierre Isabelle. 語彙的対応関係の一般化に基づく言い換え知識の拡張. 言語処理学会第 21 回年次大会発表論文集, D1-5, pp. 321-324, 2015 年 3 月 17 日. 京都大学, 京都府・京都市
(6) Atsushi Fujita. A Consideration on the Methodology for Evaluating Large-scale Paraphrase Lexicons. 情報処理学会研究報告, NL-214-21, pp. 1-8, 2013 年 11 月 15 日. 屋久島環境文化村センター, 鹿児島県・屋久島町
(7) 藤田篤. 言い換え技術の研究動向: 分類体系, 知識獲得, 応用. 情報処理学会研究報告, NL-212-6, p. 1, 2013 年 7 月 18 日. 公立はこだて未来大学, 北海道・函館市 【招待講演】

〔その他〕
ホームページ等
『言い換え技術に関する研究動向の包括的調査』http://paraphrasing.org/

６．研究組織
(1)研究代表者
藤田 篤 (FUJITA, Atsushi)
国立研究開発法人情報通信研究機構・ユニバーサルコミュニケーション研究所多言語翻訳研究室・主任研究員
研究者番号：10402801