応用を考慮した包括的な言い換えの類型: 現在の類型と課題

渡邊晃一朗 [1,2] 藤田響平 [3] 永田基樹 [3] 森遼太 [3] 小代義行 [3]
[1] 東京大学大学院教育学研究科 [2] 理化学研究所 [3] 株式会社 pluszero
kouichirou-watanabe495@g.ecc.u-tokyo.ac.jp
{fujita.kyohei,nagata.motoki,mori.ryota,ojiro.yoshiyuki}@plus-zero.co.jp

概要

自然言語処理において、これまで言い換えの重要性は指摘されてきた。そして言い換えを理解するために、文法や意味において同一であると考えられる言語表現の対の関係の類型の作成が行われてきた。しかし、意図まで考慮した上で、実際の応用の中で言い換えとみなしたい言語表現の対を考慮した類型の作成はなされてはこなかった。本研究では、応用において言い換えとみなす言語表現の対を考慮した言い換えの類型の提案を行う。応用を考慮した際、平易化においては難易度の調整を伴った言い換え、質問応答においてはある発話に後続する発話を想定した言い換えがそれぞれ必要であり、これらを包括的に扱うための類型を本研究では提案する。

1 はじめに

自然言語処理の分野では同じ意味を有するが異なる表記である言語表現の対を「言い換え」であると定義していた [1]。しかし、質問応答などの実際の応用を考えた時、厳密には同じ意味を有さないものの応用の観点から言い換えであるとみなしうるような言語表現の対が存在する。意味が同一ではないものの何らかの観点から言い換えであるとみなしうる言語表現の対の間の差異は “pragmatic difference” と呼ばれており [1]、その存在は指摘されてきたものの、Kovatchev ら [2] が “textual paraphrase” という語句を使ったように、意図などのようなテキストそれ自体からは考慮できないものを考慮した言い換えはこれまでの研究では十分に扱われてこなかった。以下、本研究では、意味としては同一でなくとも、何らかの観点から同一であると認識できる言語表現の対が言い換えであると定義し、実際の応用において言い換えとしてみなしうるものを考慮した言い換えの類型を提案する。ここでの応用は質問応答や平易化などを考慮している。質問応答での言い換えを考えた際には、ある発話の意味の同一性のみではなく、その後に続くと想定される発話の同一性を考慮することが必要である。このように、何らかの点で同一性があると考えられる言語表現の対を言い換えとして考慮する。

これまでの自然言語処理の分野において、以下で挙げるように言い換えの類型の提案がなされてきた。まず文法的な観点からの言い換えの類型を提案したものとしては、Dras [3] や Fujita [4] の研究が挙げられる。これらの研究は、文法の観点から考えられうる言い換えの類型を整理したもので、基本的に意味を考慮した言い換えを扱っていない。意味も含めて言い換えの類型を整理した研究として、まず挙げられるのが Bhagat ら [1] の研究である。Bhagat ら [1] は意味の観点から言い換えとみなせるものも含めて、25 個の類型に整理した。Vila ら [5] の研究は、基本的にこの Bhagat ら [1] が提唱した 25 個の類型を基礎として行われ、各類型の定義の精緻化を行った。最新の言い換えの類型として Kovatchev ら [2] が提案した Extended Paraphrase Typology (EPT) は Bhagat ら [1] が意味を考慮しているのに加えて、談話のレベルまで考慮した言い換えの類型である。

本研究は、応用として考えられるデータを出発点に、実際の応用で言い換えとみなした言語表現の対を考慮した言い換えの類型を提案する。既存研究では、実際の応用を考えた上での言い換えは考慮されていなかった。Kovatchev ら [2] が “textual paraphrase” を考慮すると述べていたように、基本的にテキストの表現とそれが表現する意味が同じであるとみなせる言語表現の対のみを既存研究は扱ってきた。本研究は、テキストが表現する意味が同じではないが応用の観点から考えたときに言い換えとみなしたいものを言い換えとして同定することを目的とする。そこで、Microsoft Research Paraphrase Corpus [6] のようにこれまでの言い換えのコーパスが表現として似ているものをある程度自動的に収集

しそれを基に構築されたのに対し、人手でデータを収集することから始め、言い換えとみなしたい言語表現の対を集めその類型を記述する。

2 方法と提案する類型

本研究は、以下の 2 つの段階で構成される:

1. 言い換えの類型の整理
2. データのアノテーション

まず、既存研究で提案された言い換えの類型を整理し、それらに新たな類型を加える形で包括的な言い換えの類型を作成した。この類型を用いて、応用において用いられるデータに対してアノテーションを行った。このアノテーション時にそれまでで作成された類型に該当しないような言い換えが存在した時、それを包括できるような類型を新たに追加した。結果として本研究で提案する類型を表 1 に示している。各層は複数の類型に共通するような区別にはなっておらず、単純に上下関係を意味する。

データとしては、応用を考慮して以下のものを扱った:

1. 質問応答システムにおける応答が同じと考えられる発話
2. 教科書における同一概念の説明

質問応答については、その意図と回答が同一であることが望まれる質問群にある質問がお互いに言い換えであるとみなして収集した。教科書においては、異なる校種間、特に小学校と中学・高校の教科書において同一の概念の説明を行っている部分を主に取り上げた。これは応用として平易化を主に考慮しており、特に同一概念の説明の平易化を主に扱ったデータとして本研究で扱っている。

データのアノテーションは、言い換えとみなしたい言語表現の対をデータから抽出し (これを抽出対と呼ぶ)、一方に対し言い換えの類型を適用して他方に変換するということを複数回行った。この時、1 回の変換で 1 つの類型を適用し、既存の類型で変換が不可能だと判断された場合、新しい類型の追加を行った。例えば、以下のような抽出対を考える[1]:

1a 沿岸や沖合いにおける水産資源管理の試みに加えて、**人工孵化による放流を行う**栽培漁業が注目されるようになった。

1b 沿岸や沖合いにおける水産資源管理の試みに加えて、**魚のたまごから稚魚を育てて放流する**栽培漁業が注目されるようになった。

1a から 1b への言い換えは、以下のようなステップで行われる:

1.「本質的意味の活用」のうち「含意関係の活用 (動詞)」を適用: 沿岸や沖合いにおける水産資源管理の試みに加えて、**魚のたまごから稚魚を育てて**放流を行う栽培漁業が注目されるようになった。
2.「抽象-具体」のうち「既存語彙資源で可」の下位類型「ドメイン適合不要」を適用: 沿岸や沖合いにおける水産資源管理の試みに加えて、魚のたまごから稚魚を育てて放流**する**栽培漁業が注目されるようになった。

以上の手続きを経て結果として、得られた抽出対の数は 179 対、合計の変換の回数は 1074 回であった。本研究では、複数人で協議しつつアノテーションを行った。

3 議論

3.1 類型の敷衍

本研究で特に重要と位置付ける以下の類型を例を挙げながら詳述する:

1. 語義
2. 応答

語義については、ある語句に対して、その類義語や上位・下位語を扱うだけでなく、それの辞書的な語義で置き換えることで言い換えをしていると考えられるような事例があることが実際に教科書で見られた。以下に例を挙げる:

2a 夏の季節風は、とくに東アジアから南アジアにかけて多くの雨をもたらすため、**雨季**そのものをモンスーンということもある。

2b 夏の季節風は、とくに東アジアから南アジアにかけて多くの降水をもたらすため、**夏の雨が多い時期**そのものをモンスーンということもある。

この例は、「雨季」をその語義で置き換えた言い換えである。

この例ではある語句をその語義を変換することなく置き換えれば言い換えが成立するが、加えて前後の文脈に応じて補うように追加で変換を行う必要が

1) 本論文では、言い換えとみなす対を示した時、その間で差異がある部分を太字で示す。

表 1　言い換え類型: Bhagat ら [1] と EPT [2] の列はこれら既存研究に存在したかを示し、新規の列は本研究で新たに提案された類型であるかを示している。

分類 1	分類 2	分類 3	分類 4	分類 5	Bhagat ら [1]	EPT [2]	新規
抽象-具体	既存語彙資源で可	ドメイン適合不要			✓	✓	
		ドメイン適合必要			✓	✓	
	新語彙資源が必要	ドメイン適合不要			✓	✓	
		ドメイン適合必要			✓	✓	
比喩	単語単位				✓		
	句以上						✓
言い換え	表現手段	マルチモーダル	聴覚	音			✓
			触覚	点字			✓
			視覚	絵文字			✓
				アイコン・記号			✓
		文体・話法	話し言葉中心	方言			✓
				スラング			✓
				口語表現			✓
			書き言葉中心	話法	✓	✓	
			両方	敬意の変換			✓
		表記	フォント				✓
			字種				✓
	内容の意味解釈	構成性原理から可	単語レベル	文字の類似性		✓	
				照応	✓		
				語義			✓
				派生	✓	✓	
				用法			✓
				関係性	✓	✓	
			句・節レベル	単語の類似性		✓	
				固有表現			✓
				機能語	✓	✓	
				用法		✓	
				イディオム		✓	
				インターセクティブ			✓
				サブセクティブ			✓
				句読点		✓	
			文・文章レベル	文の主要素	✓		
				文型			✓
				文の種類	✓	✓	
				比較			✓
				例外系	✓	✓	
				論理構造		✓	
		構成性原理から不可	メトニミー	行為者/行為	✓		
				操作者/操作	✓		
				オブジェクト・属性			✓
				部分・全体	✓		
			本質的意味の活用		✓	✓	
			応答				✓
			語用論的言い換え			✓	

ある場合がある。以下に例を挙げる:

3a　温度が一定のとき、一定量の気体の圧力 P は気体の体積 V に反比例する。

3b　温度が一定のとき、一定量の気体の圧力 P が 2 倍、3 倍となっていくとき、気体の体積 V が 2 分の 1 倍、3 分の 1 倍となる。

この例では「反比例」をその語義によって言い換えていると考えられるが、「一定量の気体の圧力 P」や「気体の体積 V」といった表現を補わなければならない。このような例を考慮すると、自動的な言い換えを生成する際には語義に応じた処理が必要であると考えられ、その処理の必要性の有無を類型に反映している。

このような語義を用いた言い換えは異なる校種間の教科書を比較した際に見られ、テキストの難易

度を操作する平易化において有用であると考えられる。

次に、応答については、考慮される対のみだけでなく、その前後の文脈の同一性を考慮したような言い換えを考慮している。以下に例を挙げて示す:

4a 具体的な停止位置**はどこなのか**
4b 具体的な停止位置**を教えて**

この例では一方は疑問文、もう一方は命令文であり文法的な類型とその意味は異なるが、想定される応答が同じであることから言い換えとみなしうる。特に質問応答を応用として考慮すると、異なる発話の対の同一性はそれらに続く応答の同一性から得られうる。このように、本研究で提案する言い換えの類型は、考慮される対のみだけでなく、その前後の文脈の同一性を考慮したような言い換えを考慮している。この応答を考慮した言い換えは質問応答における自動応答において有用であると考えられる。

3.2 今後の展開

現在の類型において、より詳細な類型が必要である部分を議論する。ここでは「例外系」の中でも特に「追加・削除」という類型を取り上げる。この追加・削除という類型は既存研究で提案された類型でも存在し、本研究でもそれを継承したが、明確にどのような類型であるのかについては言及されてこなかった[2)]。

追加・削除と考えられるような対の関係について、まだ十分に議論を尽くしているとは言えない状況である。特に、追加・削除されるものにどのようものがあるのかについての知見はまだ得られていない。追加・削除により生じる対の間の意味の差異は先行研究でも “content loss” と呼ばれて存在が指摘されていた [5] が、言い換えとみなしうる対において起きている content loss の体系的な整理はなされていない。言語表現の変換をする際には content loss は必然的に起きうるものであり、content loss の体系的な整理を目指して、既存研究で提案されてきたコーパスも踏まえた上で、追加・削除とみなされる対のより詳細な類型の構築をすることが今後の課題として挙げられる。この追加・削除という類型には少なくとも主に 2 つの類型が存在する。1 つ目は意味に大きな変更がない場合であり、2 つ目は意味に大きな変更があるものの応用の観点からは言い換えとみなしたい場合である。

まず、追加・削除がなされても意味に大きな変更がない場合の例を挙げる:

5a 近くの植物**など**との距離はどの程度で大丈夫か
5b 近くの植物との距離はどの程度で大丈夫か

この例のように「など」のような語句の追加・削除は対の間で大きな意味の差異を生まないと考えられる。

一方で追加・削除がなされると意味に大きな変更があるものの応用の観点からは言い換えとみなしたい場合は以下のような場合である:

6a **許容範囲内の**-とはどの程度の-を指すか
6b -とはどの程度の-を指すか。

この例の「許容範囲内の」のように、追加・削除と分類される対の中で大きく意味が異なるような対が存在する。

その他、言語特有の問題も挙げられる。以下に例を示す:

7a 点検**の**時の注意点は何か。
7b 点検時の注意点は何か。

この例のような「の」の追加・削除については、対の間で意味内容が大きく変わらない。このような機能語の追加・削除については個別言語に依存する部分があり、英語が主に考慮されていた既存研究では十分に考慮されてきたとは言えない。実際の応用においては個別言語に適した類型を用いる必要があり、本研究においては日本語特有の追加・削除の類型を検討する必要がある。

4 おわりに

本研究では、既存の言い換えの類型を整理した上で、応用を考慮した新たな言い換えの類型を提案した。議論の部分で述べた通り、今後はより詳細な類型を立てることが必要であるため、今後はデータの収集と類型の精緻化を、特に追加・削除の部分について行っていく。本研究で提案した言い換えの類型を基礎として、今後は自動対話や平易化のシステムの構築を行うことを予定している。

2) 類似した類型として同じく「例外系」の 1 つの類型である「省略」が存在するが、与えられた対から変更部分が自明に復元可能であるものを省略、そうでないものを追加・削除として分類した。

謝辞

本研究は株式会社 pluszero との共同研究「言い換えに関する研究」の助成を受けたものです。また、データのアノテーションにおいては井上峻之介氏と杉本智紀氏のご協力を頂きました。

参考文献

[1] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, Vol. 39, No. 3, pp. 205–218, 2013.

[2] Venelin Kovatchev, Antònia Martí, and Maria Salamó. Etpc: A paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 1384–1392, 2018.

[3] Mark Dras. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Macquarie University, 1999.

[4] Atsushi Fujita. *Automatic Generation of Syntactically Well-Formed and Semantically Appropriate Paraphrases*. Nara Institute of Science and Technology, 2005.

[5] Marta Vila, Antònia Martí, and Horacio Rodríguez. Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, Vol. 4, pp. 205–218, 2014.

[6] William Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pp. 9–16, 2005.