同義扱いされる表現を弁別する

那須川 哲哉　　荻野 紫穂　　西山 莉紗　金山 博
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

何種類もの表現で同じ内容を示すことができる「表現の多様性」は、自然言語の特徴の一つであり、自然言語処理では、この多様性を吸収することが重要な課題となっている。この課題に対し、基本的な意味表現の検討[1]から、同義扱いすべき表現の認識[2,3]に至る様々な取り組みが行われている。

自然言語処理において、多様性の吸収は、同じ内容を同じように扱えるようになるため、工学的なメリットが大きい。例えば、「日本アイ・ビー・エム株式会社」「日本 IBM」「日本アイ・ビー・エム(株)」は同じ法人を示す表現であり、情報検索や機械翻訳、テキストマイニングなどのアプリケーションにおいては、こういった表現を同一視することで、検索漏れを防いだり、対訳辞書等を圧縮(簡素化)したり、法人単位での集計の網羅性を向上させることができる。また、同義性の認識は言い換えや含意関係認識の取り組みにおいても重要な役割を果たしている。

その反面、表現が異なる場合には、書き手が込めた意図やニュアンスが必ずしも一致しない可能性がある。例えば、同じ母親を示すのに「母」「ママ」「お母さん」「お袋」といった表現を状況に応じて使い分けることがあり、そういった意図やニュアンスが多様性の吸収によって落とされてしまう場合がある。

近年、自然言語処理の対象となるテキストデータの量が飛躍的に増大している中、筆者らは、こういった表現の振舞いが異なる現象に着目し、「工学的応用の観点から一般的に同義扱いされている表現」(以下「同義表現」とする)を弁別する試みに取り組み始めた。

本稿では、同義表現にはどのような種類があるか、同義表現間の違いは何か、同義表現を弁別することで何に役立つかを検討する。

2. 取り組みのきっかけ:同義表現の振舞いが異なる現象

自然言語処理のアプリケーションであるテキストマイニングにおいては、特定概念と特定概念との結びつきの分布における偏りの把握が有益な知見につながることが多い[4]。例えば、PC 製品に関する不具合を示すテキストデータが大量に存在する状況で、データ全体では「ハードディスク」に関する言及が 1%程度であるのに対し、製品 A に関するデータのみに限定した場合は 5%に上昇するのであれば、製品 A のハードディスクには、何らかの問題が存在する可能性が考えられる。

その際、ハードディスクを示すのに「HDD」という表現が「ハードディスク」との区別なく用いられる場合、データ量が充分に多いなら、製品 A に関する不具合データにおいては「HDD」に関する言及も 5%程度であることが多い。その場合には、「HDD」と「ハードディスク」を同義表現として扱い、製品 A におけるハードディスクの不具合を分析する上では、「HDD」と「ハードディスク」を含むデータを合わせることで、情報源の充実化を図ることができる。

国土交通省自動車交通技術安全部審査課が収集公開している「自動車不具合情報[1]」のデータのうち、2001 年 4 月から 2010 年 9 月までの 31,255 件を対象として、このような不具合分析を IBM TAKMI ®[5](製品名 IBM® Content Analytics [6]、以下 ICA)で行った結果の一部を図 1 に示す。

サブファセット / キーワード	ブレーキ...効く-ない 523	ブレーキ...利く-ない 288	エンジン...かかる-ない 212	エンジン...始動する-できる-ない 181
車種A 525	11 0.6	3 0.1	3 0.1	4 0.3
車種B 464	5 0.2	1 0	4 0.2	3 0.2
車種C 417	13 0.9	3 0.1	4 0.3	1 0
車種D 375	3 0.1	1 0	3 0.2	4 0.4
車種E 337	5 0.2	4 0.2	1 0	0 0
車種F 335	0 0	0 0	5 0.5	2 0.1
車種G 333	15 1.4	36 3.7	1 0	0 0
車種H 322	2 0	1 0	0 0	1 0

図 1:　日本語自動車不具合データにおける各車種に対する個々の不具合情報の件数分布

図中、一番上の段のセルの数値は、「ブレーキが効かない」「ブレーキが利かない」「エンジンがからかない」「エンジンが始動できない」という表現を含むデータの件数が、各々523 件、288 件、212 件、181 件であることを示しており、左端のセルの件数は各車種に関するデータの件数である。それ以外(内側)のセルにおいては、各車種に関するデータのうち、最上段の不具合表現を含むデータの件数が上の段に、その相関の強さを示す指標が下の段に示されている。ここでは、同義語辞書を適用していないため、「ブレーキが効かない」と「ブレーキが利かない」は別々に集計されている。

[1] http://www.mlit.go.jp/jidosha/carinf/rcl/index.html

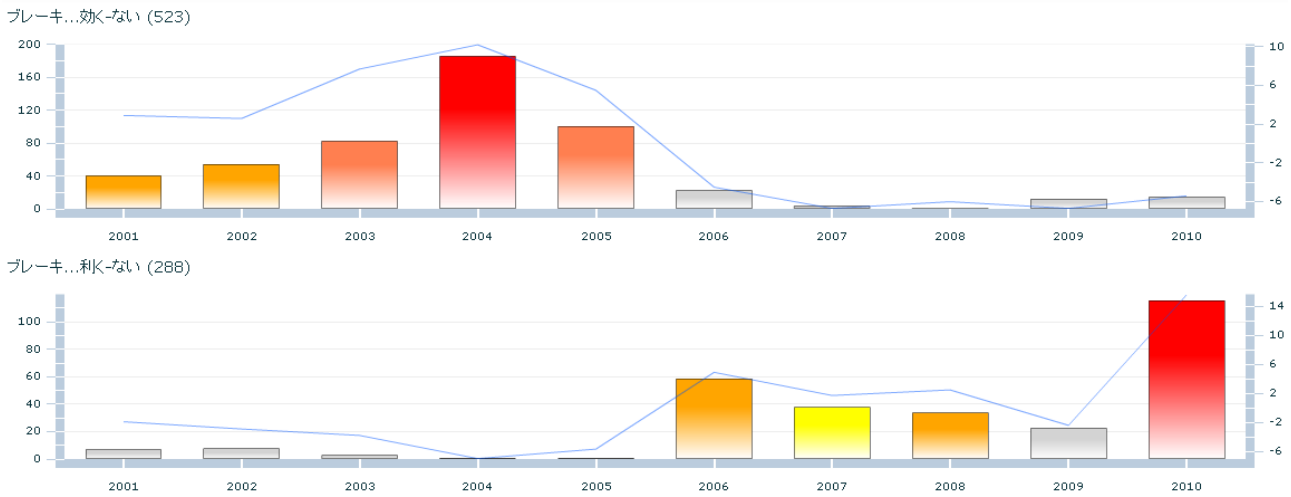

図 2:「ブレーキが効かない」と「ブレーキが利かない」という表現を含むデータの出現年の分布

一般的には「ブレーキが効かない」と「ブレーキが利かない」は同じ現象を示していると考えられるため、両者の出現分布には大きい差が無いと想定される。しかし、実際に図 1 の出現件数を見ると、車種Gに関する不具合情報で「ブレーキが利かない」(36 件)の割合が突出して高く、他の車種における割合に比べて7.7倍程度多いという相関指標が示されているのに対し、「ブレーキが効かない」の割合はそれほど高くない。

この違いがどこから来るのかを調べたところ、図 2 に示すとおり、「ブレーキが効かない」と「ブレーキが利かない」がデータ中に出現する時期が異なっていることが判明した。

すなわち、「ブレーキが効かない」と「ブレーキが利かない」の違いには発生時期の違いが反映されており、「ブレーキが効かない」と相関が高い車種は、概ね 2005 年までに「ブレーキが効かない」不具合が報告された車種であり、「ブレーキが利かない」と相関が高い車種は、概ね 2006 年以降に「ブレーキが利かない」不具合が報告された車種であることが分かる。

このように、一般的に同一視できそうな表現でも、何らかの側面では異なる使われ方をしている場合があり、その異なる振舞いの情報を活用することで、より詳細な分析につなげられる可能性を見出したことが、本研究に取り組んだ背景である。

3. 同義表現のパタン化の試み

同義表現の弁別を考える上では、そもそも同じ内容を示すために多様な表現が使われるケースにどのようなパタンがあるかを把握しておくことが有益ではないかと考えた。

その際、抽象的な概念に関しては、内容の同義性を判断するのが必ずしも容易でないと考えられるため、まずは具象物や固有名(組織名など)に関する表現のみを対象にする。従って、以下では具象的に同じ対象を示していると解釈できる表現を同義表現とする。

同義表現を派生させるケースとして、まず挙げられるのが、省略・短縮及び表記のゆれによるケースである。「マクドナルド」を「マック」や「マクド」と表現したり、「スマートフォン」を「スマフォ」や「スマホ」と表現するケースがこれにあたる。

次に、文字種を変更するケースがある。「Starbucks」と「スターバックス」や、「たんぱく質」「タンパク質」「蛋白質」あるいは、「ひと」「ヒト」「人」のような例を挙げることができる。

また、「お手洗い」と「トイレ」のように外来語を使うことに起因するケースがある。さらには「コルカタ」と「カルカッタ」のような呼称変更や、法人の名称変更に起因するケースもある。

そして、「お手洗い」「便所」「化粧室」「厠」「はばかり」のように、用途や観点から表現が分かれるケースや、書き間違い・言い間違いによるケース、「角帽」で「大学生」を示すような換喩法によるケースなど、具象物や固有名(組織名など)に関する表現に限っても、同義表現が発生する要因には多様性が大きい。そのため、少数のパタンで網羅的に扱うのは困難であるという結論に至った。

4. 同義表現の弁別因子

次に、同義扱いされる複数の表現の違いに関して考察を行なった。

表現の違いに関しては、書き手の特徴に起因するケースと文書の特徴に起因するケースに大別できると考えられる。

表現の違いの要因となる書き手の特徴としては、世代や出身・居住地域、所属組織などを挙げることができる。例えば、「国鉄」や「エベレスト」といった表現は「JR」や「チョモランマ」といった表現を使う世代よりも高い世代の書き手が使う傾向が強く、関東では「マック」と表現される「マクドナルド」が関西では「マクド」と表現される傾向が強いといった形で、書き手の特徴が示されるケース

である。また、特許文書などでは、技術用語の表現が企業によって異なるケースも存在する。

表現の違いの要因となる文書の特徴としては、例えば、公的文書か私的文書か、私的文書であっても目上の人に宛てた文書かどうかといった、表現を使う場の特徴を挙げることができる。

例えば、2011 年 3 月 14 日から 2011 年 3 月 28 日の期間に非連続的に収集した震災に関わる日本語の Twitter データ 1,135,495 件を ICA で分析したところ、「小田急電鉄」は多くの場合「小田急」[2]と表現されている。「小田急電鉄」を含む Tweet が 27 件に対し、「小田急電鉄」ではなく「小田急」のみを含む Tweet は 1,376 件であり、「小田急電鉄」を含む Tweet27 件のうち 19 件は報道記事の内容やタイトルを伝えているものであった。逆に、「小田急」のみを含む Tweet からサンプルとして選んだ 100 件中、報道記事の内容やタイトルを伝える中で「小田急」という表現を使っているのは 1 件のみであった。すなわち、ニュースなど、より公的な性質の高い文書においては、正式名称を省略せずに使う傾向が強い。従って、同義表現の中でも正式名称が使われる文書は公的な性質が高いと考えられる。

また、特定の企業の製品を好んで用いる人々を、企業名を冠して「○○ファン」「○○マニア」などと表現する場合があるが、インターネット上の特定の掲示板では「○○信者」と表現されるケースが目立っており、表現する場によって、このような表現が使い分けられる傾向が強いと考えられる。

さらに、上下関係や親密度といった相手との関係により表現が異なるケースも出てくる。例えば子供相手に幼児語を用いるのも、この範疇に入れることができると考えられる。

同義表現の選択が文書の特徴に起因するケースでは、文書に応じて、同じ書き手が複数の同義表現から適切な表現を選択する形になる。また、その際、対象文書によっては、例えば公的文書で正式名称を用いるように、異なる書き手でも同じ表現を選択する傾向が強いと考えられる。

それに対し、同義表現の選択が書き手の特徴に起因するケースでは、同じ書き手は比較的安定して同じ表現を選択すると考えられる。

5. 応用可能性

本節では、実用的な観点から、同義表現を弁別することで何に役立つかという可能性を議論する。具体的には、主な応用として、テキストマイニング、機械翻訳、および外国語教育を取り上げる。

[2]「小田急」は「小田急百貨店」など他の内容を示している場合も考えられるが、サンプルとして選んだ 100 件中 88 件では「小田急電鉄」の意味で用いられていた。

5.1. テキストマイニングへの応用

取り組みのきっかけで紹介したとおり、テキストマイニングでは表現の分布の偏りを認識することで有用な知見を得られることが多い。その際、統計的に有意なレベルの分析を行うために、ある程度まとまった件数のデータを対象にする必要がある。

テキストマイニングの難しさは、自然言語で記述された構造化されていないデータを対象とする点にあり、基本的な処理の流れとしては、自然言語で記述された内容を認識し、構造化情報として元データに紐付けた上で、他の構造化情報と共に統計的な分析を行うことになる。

その際、例えば、自然言語で記述された内容に出現する概念 A を認識する上で、概念 A を示す多様な表現(「A」「A'」「A''」など)を「A」の同義表現として定義し、概念 A として認識させることで、概念 A に関するデータをより多く集めることができ、分析精度を向上させたり、分析の多様性をひろげたりすることができるようになる。

その反面、実は振舞いの異なる同義表現を同一視してしまうと、振舞いが異なる点に関する情報を欠落させることになってしまう。同じ対象を示している点では同一視できるようにした上で、振舞いの異なる点に関する情報も別途紐付けられることが望ましい。それによって、より詳細な分析が可能になると考えられる。

例えば、法人名に関して正式名称を使っているか略称を使っているかの情報をデータに紐付けておけば、略称を使っているデータはカジュアル度が比較的高いいと判断することができる。また、地域によって異なる表現に関しては、地域の情報を、その書き手のプロファイルの一部としてデータに紐付けておけば、地域性に関する分析につながる可能性が出てくる。

こういった同義表現で振舞いが異なる要因の一つとして丁寧さと乱暴さを挙げることができる。この乱暴さを示す軽卑表現に着目した取り組み[7]においては、ある対象(実験では組織名)に言及しているデータのうち軽卑表現の割合が多いケースでは、その組織に対する非難が示されていることが多いという結果が得られている。

同義表現の使われ方が異なるという、ある意味副次的な情報を活用できるようにすることで、テキストマイニングの分析がより精緻化することが期待される。

5.2. 機械翻訳への応用

機械翻訳においては、同義表現を同一視することで、表記だけの違いを吸収し、対訳辞書等を圧縮(簡素化)することができる点で、翻訳処理の効率を高めることができる。

その反面、表現が異なる点を無視することは、自然言語の多様性やニュアンスを失うことであり、機械翻訳の品質向上の妨げにもなりかねない。逆に言えば、同義表現間で異なる点を認識し、積極的に活用することで機械翻訳の品質向上につなげられる可能性が出てくる。

例えば、「ヒト」も「人」も人間を示すが、「ヒト」は生物学などの分野の文献で使われ易い傾向があり、「人」はより一般的な使い方であると考えられる。この特徴を認識し、データに紐付けた上で他言語に翻訳するならば、例えば英語の場合、“human” と “person” という対訳候補の中から、「ヒト」の対訳として “human” を選択するような形で、他言語でも同じ振舞いをしている同義表現を選択することによって翻訳品質を向上させられる可能性がでてくる。

5.3. 外国語教育への応用

訳語選択の難しさは機械翻訳だけでなく、人間にとっても当てはまる。特に外国語教育においては類義表現・同義表現の使い分けが重要な課題となっている [8] 。例えば、「辞書を引いても、相違点が曖昧で、その具体的な使い方が分からない。」という報告があるように、実際の運用を繰り返さないと習得が困難なことが多い。

大量の文書データから実運用上で同義表現が異なる点を明確にし提示することができれば、語学教育に役立つ可能性が考えられる。

6. おわりに

近年、電子化された膨大な量の文書データへのアクセスが容易になっている中、その環境を活かす試みの一環として、一般的に工学的な観点から同一視される傾向の強い同義表現の違いに関して検討し、その違いを活用する可能性を示した。

報道記事や特許文書に加え Blog や掲示板への書き込みといった多種多様な文書データをより深く分析し活用するためには、表現の微妙な違いを捉えることのできる技術が今後有用性を増していくと考えられる。

今後は表現の同義性を認識した上で、その振舞いの異なる点を自動的に認識する仕組みの実現に取り組んでいきたいと考えている。

謝辞

本研究を進めるにあたり、各種実験をはじめとして、多摩大学の樋浦聖貴氏に多大なご支援をいただきました。ここに記して感謝いたします。

参考文献

[1] 高木朗, 伊東幸宏. 自然言語の処理. 丸善. 1987

[2] Donald Hindle. Noun Classification from Predicate-Argument Structures. Proceedings of the 28th Annual Meeting of ACL, pp.268-275. 1990

[3] Akiko Murakami and Tetsuya Nasukawa. Term aggregation: mining synonymous expressions using personal stylistic variations. International Conference on Computational Linguistics (COLING), pp.806-812. 2004

[4] 那須川哲哉. テキストマイニングを使う技術/作る技術―基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006

[5] Tetsuya Nasukawa and Thoru Nagano. Text analysis and knowledge mining system. IBM Systems Journal, Volume 40, Issue 4, pp.967-984. 2001

[6] Wei-Dong (Jackie) Zhu, Asako Iwai, Todd Leyba, Josemina Magdalen, Kristin McNeil, Tetsuya Nasukawa, Nitaben (Nita) Patel, and Kei Sugano. IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content. ISBN: 0738435287. 2011

[7] 荻野紫穂, 那須川哲哉, 金山博, 榎美紀. 軽卑表現の情報を活用した知識発見. 言語処理学会第 18 回年次大会. 2012

[8] 朱金月. 重要語指導における類義語の指導 －「日本語6α」の授業に参加して－. 日本語教育実践研究 第 5 号, pp.131-138. 2006