レビュー記事コーパスを利用した英語量化詞の強意性の分析

志築 諒太朗
京都大学 人間・環境学研究科
79steek@gmail.com

1. はじめに

強意 (intensity) について、Labov (1984) は 次のように定義している。

(1) 強意の定義

'Intensity' is defined here as the emotional expression of social orientation toward the linguistic proposition: the commitment of the self to the proposition.

Selting (1994) によれば、強意的 (emphatic) な発話スタイルを構成する手がかりは (i) 音調の有標性 (ii) 特定の語彙選択 (iii) 特定の統語選択 の３要素にまとめられる。本稿が扱うのは、(ii) の語彙選択に関してである。主な関心が話し言葉にある Selting (1994) では、(i) 音調の有標性 を最も重要視しているが、本稿の対象である書き言葉においては、(ii) (iii) の重要性はより高まるものと考えられる。

また、Labov (1984) は、強意が表現される言語形式について次のように述べている。

> Most discussions of intensity involve specialized prosodic contours and a set of adverbs that code intensity directly: *really*, *so*, and *very* in English. *But* intensity is most often expressed through linguistic forms that are normatively devoted to logical relations and conceptual categories. (Labov 1984: 43)

そこで本研究では、論理的関係と結びつく言語形式の１つである量化詞というカテゴリーに注目してその強意用法についての分析と考察を行う。ここで「量化詞」と呼んで扱うのは、池内 (1985: 3) で数量詞 (quantifier) と呼ばれるものと同様である。限定詞 (determiner) の下位区分であり、数量形容詞とも呼ばれる。

2. 研究手法

2.1. データ

ここで使用する "UMass Amherst Sentiment Corpora" (Constant et al. 2009a) は、英語、日本語、中国語、ドイツ語のレビュー記事を含むコーパスである。そのうち本研究では、英語のオンライン書店 Amazon.com からとられた書籍のカスタマーレビューの部分のみを使用した。その総レビュー数は約５万、総語数は約800万語である。それぞれのレビューには、レビュアーによって、レビュー対象に対して付与された５段階の得点があり、それらは星の数で示される。したがって、使用したデータは５段階の得点がタグ付けされた、約800万語のコーパスである。コーパスから得られるデータの処理手法は、Potts and Schwarz (2008) に依拠している。

2.2. 得点と話者の感情の関係

Potts and Schwarz (2008) では、レビューの得点と話者（レビュアー）の感情の関係について、およそ次のような一般化を行っている。

> 極端な得点（１点、５点）のレビューを書く話者は、より高まった感情状態にある。強意表現 (exclamative) を用いる話者は、より高まった感情状態にある。したがって、強意表現は極端な得点のレビューで、より頻繁に生起する。(Potts and Schwarz 2008: 16)

本研究でもこの仮定は妥当と考え、採用する。これは (1) で示した強意の定義とも矛盾しない。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

2.3. 分析手法

語の分布を計測する手法は、Potts and Schwarz (2008) ら同様、対数オッズ (log-odds) を用いるロジスティック回帰分析 (logistic regression analysis) によってグラフの作成を行う。

グラフのx軸には、レビューに付与された５段階の得点をとる。ただし、後の分析を簡便にするため１点から５点の得点を、それぞれ−２点から２点に読み替える。グラフのy軸には、得点に応じた語の頻度を対数オッズの形でとる。対数オッズは一般に次の式で示される。

(2) 対数オッズの定義

$$\ln\left(\frac{N(x)}{N(total) - N(x)}\right)$$ （ただし、$N(x)$ は語 x の語数、$N(total)$ は総語数）

次に行うロジスティック回帰分析は、得点毎の語の分布の異なりを定式化するために行われる。Potts and Schwarz (2008) によれば、このような分析に適した回帰式は次のような２次式である。すなわち、ここで行うロジスティック回帰分析とは、ある語の頻度に関する対数オッズと、その語が生起するレビューの得点の関係を最もよく表す、下式で表される２次関数の回帰係数（とそのp値）を求めることに他ならない。

(3) 用いる回帰式

$$y=\beta 0 +\beta 1x+\beta 2x^2$$

次に、「強意表現は極端な得点のレビューにより頻繁に生起する」という仮定を、回帰分析の結果で解釈すると、次のようになる。

表１：回帰係数による分布の特徴づけ

グラフの形	$\beta 2$	$\beta 2$ の p値	$\beta 1$	$\beta 1$ の p値	強意	評価極性
U字型	正	有意	—	有意でない	有	なし
J字型	正	有意	正	有意	有	肯定
逆J字型	正	有意	負	有意	有	否定

(Potts and Schwarz 2008: 21, 一部改変)

２次回帰係数は強意性の有無に関わる。２次係数が正であれば、そのグラフは下に凸であり、したがって軸から離れるほど（極端な値の得点ほど）高頻度に生起する。すなわち、その項目は強意性を有する。１次係数は評価極性[1] の有無に関わると考えられる。（２次係数が正のとき）１次回帰係数が正であれば、グラフの軸の位置は左に偏り、したがって高得点のレビューでの頻度が相対的に高くなる。１次回帰係数が負のときはその逆で、低得点のレビューにおいてより高頻度の語となる。

以上をまとめると、ある語彙項目の強意性の存在を確かめるためには、ロジスティック回帰分析における２次回帰係数が正であることを確かめればよい。次節で英語の量化詞に対して分析を行った結果を示す。

3. 分析

3.1. 対象

前節で述べた分析を、以下の語彙項目について行った。主に池内 (1985: 4) で挙げられた量化詞を分析の対象としたが、同じ表記形に副詞などの用法が多いもの（no, much, either など）は対象から外している。

1 個人の評価に関する情報を評価情報、評価情報の良い/悪いに関する軸を評価極性と呼ぶ。ある評価情報が良い評価をもつことを肯定極性をもつと呼び、逆に悪い評価をもつことを否定極性をもつと呼ぶ。(乾、奥村 2006: 203)

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

(4)　分析を行った量化詞[2]

every, some, any, each, many, most, both, neither, few, several

3.2.　結果

分析の結果、強意性を持つのは any と every となった。それぞれ、図 1 のようなグラフが得られた。回帰係数はともに正の 2 次係数 (0.072, 0.117) をとっており、強意性の存在を示している。ただし 1 次係数はその正負が異なり、any は負 (-0.078) の、every は正 (0.054) の 1 次係数をとった。すなわち、any は否定の、every は肯定の評価極性を持つと考えられる。

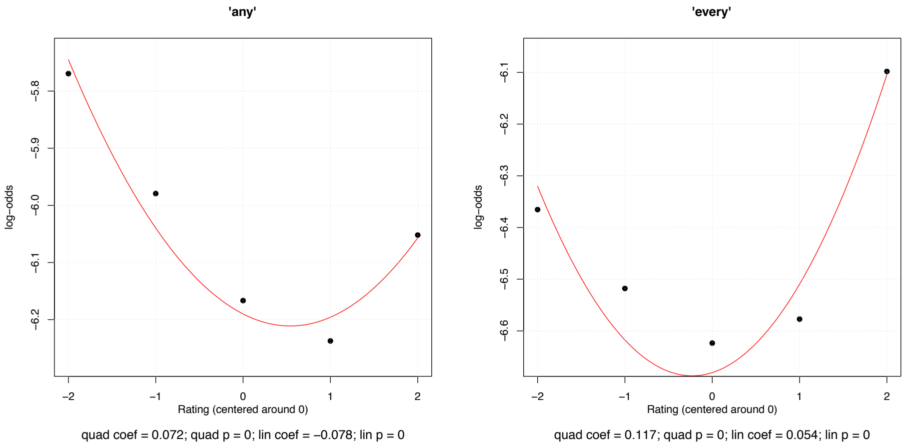

図 1 ：強意性を持つ量化詞 any, every

一方で、意味論的には同じ全称である each については、２次回帰係数は有意でなく、したがって強意性は確認されなかった。その他の項目については、有意な負の２次係数をとった (some, few) か、あるいはその２次回帰係数に有意性はみられなかった（その他）。これらのグラフについては割愛する。

4.　考察

4.1.　全称量化と強意

Labov(1984) は、意味論的な全称量化表現にみられる強意用法について指摘している。every と any に強意性が確認されるという今回の分析は、その指摘を裏付ける結果となった[3]。

ここで、強意性に対する量化詞の意味論的基盤についての問いが生じる。

(5)　全称量化と強意に関する問い

a. 全称量化表現 any, every が強意性を持つのはなぜか。
b. （それをふまえて）全称量化表現 each が強意性を持たないのはなぜか。

まず、全称量化表現が強意性を持つ理由について、Labov(1984) は全称量化の意味と「認知的矛盾 (cognitive contradiction)」という概念を結びつけた説明を試みている。「認知的矛盾」とは、「通常理解される認知的意味とその文脈が一致しない場合」(Labov 1984: 47) である。まず全称の意味を、可能な例外を無視する

---

2 every, some, anyについては、それらが量化詞として働いているとみなされる複合語を合わせて計測した。したがって every は “every”, “everybody”, “everyone”, “everything” の 4 語の総計であり、any, some についても同様である。

3 any を意味論的に全称量化詞と扱うべきかどうかに関しては、議論の余地がある（Horn 2005 など）が、本稿では any の意味論にはこれ以上立ち入らず、Quine(1960) らにならい、any を全称の量化詞と扱う。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

「非厳密解釈」(loose interpretation) と、例外を排除する「厳密解釈」(strict interpretation) に分類する(Labov 1984: 50)。その上で、典型的な話者においては「全称量化表現の強意用法の使用は、その陳述が矛盾する可能性に比例」(Ibid.: 59)し、非厳密解釈の全称量化表現に強意用法がより頻繁に観察されると述べている。非厳密解釈においては全称の認知的意味と、例外が存在する文脈とが認知的矛盾を起こすためである。

また量化詞 each については、「厳密解釈しか持たないように思われる」とされ、「非厳密解釈」の不在によって強意用法も持たないことが予測されている (Ibid.: 66) 。池内 (1985) も、every と each をともに全称量化詞としながら、その違いについて「each は個別的読みを強調する場合に用いられ、一方、every は個体に言及しながら集団全体に言及するといった所があるという点で異なっている。」(池内 1985: 130) と指摘している。これは Labov (1984) の定義による「（非）厳密解釈」に関する傾向と同様といえるだろう。

4.2. 外延の主観的拡張

Labov (1984) は全称量化詞の強意用法が「認知的矛盾」に起因するとした。しかしこの概念は、全称量化詞のみが高い強意性を示すという観察を説明してはいない。たとえば、実際の文脈が厳密解釈を要求する全称命題であり、かつ認知的意味が「例外の存在」である場合においても、定義からすれば認知的矛盾は生じることになり、したがって “some” などの存在量化詞も同様に強意性を持つことが予測されるためである。

本稿ではより精密な発展的概念として、「（外延の）主観的拡張」を提案する。それは、「ある命題に対して、実際には例外が存在するにも関わらず、その例外を無視して言語表現の外延を拡張させることで真の全称命題として認知すること」である。これはいわば、認知的矛盾の概念に単方向性を加えるものである。

5. おわりに

本研究の結論は以下のとおりである。 Amazon.com のレビュー記事を利用したコーパス調査によると、全称量化詞の any と every には強意性が観察されたが、同じく全称量化詞の each に関しては強意性が確認されなかった。このことは Labov(1984) が提案した全称の「非厳密解釈」と強意性の関係を裏付けるが、Labov (1984) の認知的拡張だけでは根拠として不十分である。したがって、発展的概念である「（外延の）主観的拡張」を提案する。

本研究で不十分な点として、用法・品詞に基づいたよりきめ細かい分析や、量化詞以外の量化表現の強意表現に対する意味論的基盤の分析が挙げられる。これらの点は今後の課題としたい。

参考文献

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, England: Cambridge UP.

Constant, N., Davis, C., Potts, C., and Schwarz, F. (2009a). *UMass Amherst sentiment corpora*.

Constant, N., Davis, C., Potts, C., and Schwarz, F. (2009b). "The pragmatics of expressive content : Evidence from large corpora." *Sprache und Datenverarbeitung*, 33(1-2), 5-21.

Davis, C., and Potts, C. (2010). "Affective demonstratives and the division of pragmatic labor." *Proceedings of the 17th Amsterdam Colloquium*, M. Aloni and K. Schulz, eds., 32-41.

Horn, L. R. (2005). “Airport ’86 revisited: Toward a unified indefinite any.” *Reference and Quantification: The Partee Effect*, G. N. Carlson and F. J. Pelletier, eds., CSLI, Stanford, 113-142.

Labov, W. (1984). "Intensity." *Meaning, Form, and Use in Context: Linguistic Applications*, D. Schiffrin, ed., Georgetown UP, Washington, DC, 43-70.

Potts, C., and Schwarz, F. (2008). "Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora." Ms., UMass Amherst, 1-29.

Quine, W. V. O. (1960). *Word and Object*. Cambridge: MIT Press.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Selting, M. (1994). "Emphatic speech style : with special focus on the prosodic signalling of heightened emotive involvement in conservation." *Journal of pragmatics*, 22, 375-408.

池内 正幸. (1985). 『名詞句の限定表現（新英文法選書第6巻）』. 大修館, 東京.

乾 孝司, 奥村 学. (2006). 「テキストを対象とした評価情報の分析に関する研究動向」 自然言語処理, 13(3), 201–241.

Copyright(C) 2011 The Association for Natural Language Processing. All Rights Reserved.