定量調査のための意見調査コーパス構築への取り組み

三澤賢祐† 成田和弥† 田内真惟人† 中島正成† 黒橋禎夫‡§
†エン・ジャパン株式会社 ‡京都大学 §科学技術振興機構 CREST
{kensuke_mitsuzawa, maito_tauchi,
kazuya_narita, masanori_nakashima}@en-japan.com
kuro@i.kyoto-u.ac.jp

1 はじめに

SNS はこの 10 年近くで利用者が増え、大きく飛躍した。SNS の中でも特に Twitter はそのユーザー数の多さ、投稿数の多さから、Tweet を元にした統計調査と政策改善 [10] から、消費者行動の予測 [5] や経済指標の予測 [11] の研究も行なわれている。一方で Twitter を始めとする Web 言語資源の利用には、実用上の問題がある。第一に、Web 言語資源はノイズが多く、調査目的・研究目的に合わせて、目的のデータだけを抽出する作業を含めた前処理が必要である。第二に、Web 言語資源は、投稿者や文書作成者の性別や年齢など、デモグラデータが利用できないことが多い。こうした問題に対処するため、Web 言語資源を利用する前に、人手で前処理とラベル付けをしなければいけない状況がしばしば発生する。

三澤ら [4] はこうした問題に対し、FKC コーパスを作成し、研究向けに公開[1] している。FKC コーパスは不満買取センター[2] (FKC) で収集された意見より作成されたコーパスであり、豊富なユーザーデモグラ情報を保有し、基本的にはネガティブ意見の投稿で占められている。ネガティブ意見は改善策を考案するための有益な情報となり得るため、FKC コーパスは「意見分析」に適したデータセットであると考えられる。しかし、一方で FKC コーパスは統計調査を始めとする定量調査には適応が難しい。定量調査のためには、意見フレーズや、意見の対象を集計し、分類をする必要があるが、FKC コーパスには集計に適したメタデータが付与されていない。そのため、FKC コーパスを定量調査目的で利用するには、分析者は目的に応じてメタデータ付与を実施する必要があり、負担が大きい作業である。

本論文では、こうした FKC コーパスの問題点に対し、メタデータの自動付与を紹介する。我々は FKC コーパスのテキストに、意見タグ、*WikiData ID* の 2 種類のタグ付与を、構文・格解析結果に基づくルールベースにより実施した。この 2 種類タグにより、文脈情報を残しつつ、単語以上での集計処理が可能であるため、ユーザーデモグラ情報に加え、FKC コーパスは定量調査を始めとする各種統計調査に寄与するデータであると考える。FKC コーパスには、日用品や食品、小売店を対象とする不満意見から政治や公共施設を対象とする不満まで、拾い範囲の対象が言及されている。そのため、付与された意見情報は、統計調査のための資料、サービス改善のための問題発見の材料として活用が期待できる。または、消費者行動の予測、経済指標を予測するための機械学習モデルの特徴量としても活用が期待できる。

[1] http://www.nii.ac.jp/dsc/idr/fuman/fuman.html
[2] http://www.fumankaitori.com

2 類似データセット

2.1 MPQA Opinion Corpus

MPQA Opinion Corpus [1] は、主にニューステキストを中心として構築されたコーパスであり、人手でアノテーションが実施されている[3]。MPQA Opinion Corpus では意見タグに似たメタデータが付与されているが、コーパスのテキストは静的である。したがって、話題の移り変わりを分析していくような用途には向いていない。一方で、FKC コーパスは、1 年以上にわたる時系列データであるため、こうした話題の移り変わりを分析することも可能である。

2.2 ACP Corpus

ACP コーパス [2] は約 100 万文の日本語テキストから構成されているコーパスであり、タグ付け作業が自動で行なわれている点が特徴的である。しかし、付与されているタグ情報は、文単位での評価極性タグ (肯定的または否定的) だけであり、意見の具体的な内容を知ることは困難である。FKC コーパスのタグ情報は意見単位での集計を実行することが可能である。

3 コーパス属性情報とタグ付与処理

不満買取センターは一般消費者ユーザーから不満意見を収集しているサービスであるが、その特徴として「入力すべき情報のほとんどがテキスト情報」という点が挙げられる。表 1 に投稿形式とその例を示す。ここで、不満内容のテキストに対して、我々は「意見タグ」を定義する。意見タグは意見対象部と意見述部、2 つをつなぐ格情報の 3 つの構成要素から成り、意見対象部と意見述部を独立して集計することも可能である。本論文中では、意見タグを「意見対象部-格-意見述部」の順番で表記する。

不満投稿の中で、Wikipedia に存在するエントリ名が言及されている場合には、意見を補足する情報になり得る。そこで、我々は「テキスト中で言及されている Wikipedia エントリ名」に対しては *WikiData ID* を導入する。WikiData は複数のデータベースを統合し

[3] http://mpqa.cs.pitt.edu/

表 1: 不満投稿のフィールドと投稿例

フィールド	データの種類	投稿例
不満内容（必須）	free text	おしゃれな革の鞄は持ち運びが大変。軽くて丈夫な革製品が買いたいな。
不満の対象（任意）	free text	革のカバン
サービス・製品提供者 (任意)	free text	XX カバン
カテゴリ（必須）	categorical	ファッション
サブカテゴリ（必須）	categorical	鞄

表 2: タグ付与処理の例

対象フィールド	意見タグ	WikiDataID
不満内容	持ち運び-ガ-大変, おしゃれな革の鞄-ガ-大変, 丈夫な革製品-ガ-買いたいな, 軽くて革製品-ガ-買いたいな	おしゃれ, 皮革, 鞄, 丈夫, 製品
不満の対象	-	皮革, 鞄
サービス・製品提供者	-	XX カバン

た 2 次情報データベースであり、多言語の記事を 1 つの ID で管理できる、Wikipedia の Infobox 構造から構築されたネットワークグラフを利用できる、などの点で、Wikipedia よりもデータベースとして優れている。

表 2 に、不満投稿テキストとテキストに付与された 2 種類のタグ例を示す [4]。タグ付与処理では、「不満内容」フィールドには 2 種類のタグを付与し、「不満の対象」と「サービス・製品提供者」には WikiData ID のみを付与する。意見タグは 1 つであると限らず、文中に複数の意見が存在している場合は、複数のタグが付与される。また、WikiData ID に対しては、我々は語義曖昧性の解消を実施しておらず、候補をすべてタグ情報として記録する。

3.1 意見タグの作成

意見述部および意見対象部抽出の流れを図 1 に示す。まず、各不満内容に対して、構文解析および格解析を実行する。そして、その結果を利用して、構成要素の抽出および紐付けを行う。

構文・格解析には、日本語形態素解析システム JUMAN [7]、日本語構文・格・照応解析システムには KNP [8] を用いる。リアルタイムに収集された不満投稿には、より新しい言葉が出現する。JUMAN でも新語の解析を行うことは可能だが、より新しい言葉に対応するため、mecab-ipadic-NEologd [6] を JUMAN 形式の辞書に変換して用いる。

3.1.1 意見述部の抽出

まず、構文解析の結果をもとに、意見述部の抽出を行う。意見述部の候補として、KNP によって、⟨ 用言:動 ⟩⟨ 用言:形 ⟩⟨ 用言:判 ⟩ が feature として付与されている文節を抽出する。形態素ではなく、文節を抽出の単位とするのは、不満投稿においては、述語に付随する機能表現の役割が重要であり、それを情報として意見述部内に残すためである。図 1 の例では、「おしゃれな」「大変」を意見述部として抽出する。

しかしながら、1 文節を意見述部の候補として抽出するだけでは、複数文節にまたがる表現に対応できない。例えば「気になる」「気持ちが悪い」は、それぞれ「なる」「悪い」を一つの意見述部とするのは不十分であり、「気になる」「気持ちが悪い」を一つの意見述部とすることで、最終的に、何に対して「気になる」のか、といった情報として抽出したい。そこで、このような複合表現に対応するため、抽出された意見述部の候補をもとに、意見述部の拡張を行う。

意見述部の拡張を行うために、日本語大シソーラス [9] を利用する。日本語大シソーラス自体は、類語をまとめ上げた辞書ではあるが、非常に多様な表現が収録されており、その中には複合表現も多く含まれているため、複合表現の検出に利用する。

具体的には、抽出された意見述部の候補と、日本語大シソーラス中のエントリへの一致を確認し、一致の場合には、一致する文節に意見述部を拡張する。ここでいう一致とは、JUMAN、KNP が出力する、正規化代表表記、および、格構造が一致している場合のことを指す。そのため、事前に日本語大シソーラス中のエントリすべてに対して、不満投稿と同様の設定で、JUMAN、KNP による解析をしておく。単純な表層形でなく、正規化代表表記や格構造による一致を用いるのは、表記ゆれの影響を加味するためである。例えば、日本語大シソーラス中では、「気持が悪い」という表現は収録されているものの、「気持ちが悪い」「気持も悪い」といった表現は収録されていない。正規化代表表記や格構造を用いることで、このような表記揺れを吸収しながら、複合表現の検出を行うことができる。

また、日本語大シソーラス中に含まれない表現でも、機能的述語である「ある」「できる」が意見述部になっている場合は、その述語項まで意見述部を拡張する。

3.1.2 意見対象部の抽出

意見対象部の候補として、KNP によって、⟨ 体言 ⟩ が feature として付与されている文節から、末尾の助詞を除外した部分を抽出する。また、形式名詞「の」を含む文節には ⟨ 体言 ⟩ が付与されず、候補として抽出できないため、⟨ID:〜の〜⟩ が feature として付与さ

[4] 実際は Q から始まる ID が付与されているが、説明のために記事名を表記した。また「XX カバン」という項目名は実在しないが、説明のために記載した。

れている文節から、末尾の助詞を除外した部分も同様に抽出する。図 1 の例では、「革の」「鞄は」「持ち運びが」が〈体言〉を含む文節であるので、「革」「鞄」「持ち運び」を意見対象部として抽出する。

意見述部と同様に、意見対象部に対しても拡張を行う。拡張を行うことで、例えば、単純に「革製品」ではなく、「丈夫な革製品」「軽い革製品」のように、どんな「革製品」なのか、という情報も含めて意見対象部を作ることができる。そのために、KNP による構文解析結果を利用する。具体的には、意見対象部に係っている文節の内、〈連体修飾〉が feature として付与されている文節に、意見対象部を拡張する。さらに、拡張した文節に対しても同様の拡張を行っていき、〈連体修飾〉を含む文節が係らなくなるまで繰り返し拡張を行う。

図 1 の例では、「革」「鞄」にそれぞれ係っている連体修飾節を辿ることで、「おしゃれな革」「おしゃれな革の鞄」を意見対象部として得ることができる。

3.1.3 意見述部と意見対象部の紐付け

最後に、格解析結果に基づいて、抽出された意見述部と意見対象部との紐付けを行う。意見述部内の述語の格解析結果が、意見対象部内の核となる名詞を示す場合に、それら構成要素を紐付ける。意見述部および意見対象部の組み合わせによって、不満を端的に表現する、というのが目的である。そこで、経験則に基づき、格解析結果の中でも、ガ格、ニ格、ヲ格のみを対象として紐付けを行う。ただし、そのうちヲ格に関しては、単純な肯定文の場合には不満を表していない場合が多く、否定や疑問などの表現が付随している際には、不満を表すことが多いことから、意見述部側に〈否定表現〉などの否定関連 feature、あるいは、〈モダリティ-疑問〉などのモダリティ関連 feature が付与されている場合にのみ、紐付けを行う。また、紐付けによって既に拡張した部分との重複が起こる場合には、紐付けを行わない。

図 1 の例では、「革」が「おしゃれな」のガ格、「運び」および「鞄」が「大変」のガ格であると、格解析によって判定されている。この解析結果から、それぞれを含む構成要素同士を紐付けると、「おしゃれな革-ガ-おしゃれな」「持ち運び-ガ-大変」「おしゃれな革の鞄-ガ-大変」となる。このうち、「おしゃれな革-ガ-おしゃれな」は重複が起こっているため、紐付けを行わない。そのため、最終的には「持ち運び-ガ-大変」「おしゃれな革の鞄-ガ-大変」が得られる。

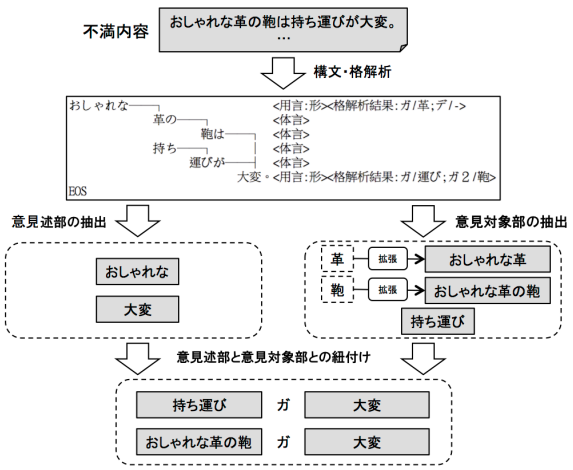

図 1: 意見述部および意見対象部の抽出の流れ

3.2 WikiData ID の付与

我々は WikiData から英語、または日本語を言語として持っている項目だけを選択し、英語と日本語のラベル文字列とリダイレクト文字列[5] を辞書データベースの見出し語として利用する。同じ見出し語に対し項目数が複数存在する場合は、1 対多の関係になるように辞書を構築する。英語を選択したのは、英語のラベル文字列とリダイレクト文字列にも、不満投稿で言及される文字列が頻出するからである。例えば「アップル (企業)」は、不満投稿中では *Apple* と表現されることがしばしばあるが、*Apple* という項目は、英語のラベルにしか記載がない。我々は 2016 年 3 月時点の配布データを利用し、項目数が 28,178,841、見出し語数が 19,877,598 の辞書データベースを構築した。構築した辞書データベースを利用し、最長一致の文字列検索にて、エンティティリンキングを実施した。

3.3 タグ付与における課題点

意見述部と意見対象部の抽出における課題として、拡張範囲の問題があげられる。意見述部の場合は、日本語大シソーラスを用いて拡張を行った。しかしながら、日本語大シソーラスは類語をまとめ上げた辞書であるため、この用途でのカバレッジは明らかでない。本来拡張ができていない割合を検証すると同時に、他のリソースを利用した拡張についても検討する必要がある。意見対象部では、〈連体修飾〉を含む文節を辿ることにより拡張を行った。どこまで連体修飾節を辿るべきなのか、あるいは連体修飾節の拡張が不要な場合があるのか、といった検証は今後の課題である。

また、タグ付与の過程では明示的に意見性の判断を実施していない。したがって、「おしゃれな革の鞄-ガ-大変」のような非文や、意見性が薄いフレーズが抽出される可能性がある。こうしたケースに対しては、小林ら [3] のように、付与されたタグに対して意見性の判断を実施する必要があると考えられる。

WikiData ID の付与においては、語義曖昧性を実施していないため、複数の WikiData ID が記録されてしまうという課題がある。集計する上で曖昧性は解消されている方が望ましいが、分析者は集計後に文脈に適切な記事名を選択できるため、致命的な問題ではないと考える。また、本来は該当する項目が辞書に存在していないにも関わらず、文字列が辞書項目に該当する場合は WikiData ID が付与されてしまうという問題がある。例えば、表 2 では、「おしゃれ」に対する WikiData ID が付与されているが、これはテレビ番組名であり、誤りである。このような問題に対しては、記事の内容を判断し、付与可否を決定するプロセスが必要である。

4 意見タグによる集計

我々は 3 節に記述した方法で、FKC コーパスのテキストにタグ付与を実施した。定量調査における意見タ

[5] WikiData では *Also known as* と表現されている

表 3: 意見タグによる集計と単語集計の比較

	意見タグ (頻度; 割合)	単語 (頻度; 割合)
1	子供-ガ-遊べる (87; 0.9%)	駐車場 (2,106; 2%)
2	公園のトイレ-ガ-汚い (56; 0.7%)	利用 (1,567; 2%)
3	駐車場-ガ-ない (50; 0.6%)	施設 (997; 1%)

表 4: 意見タグと単語の共起比較調査

	タグの共起 (頻度)	単語の共起 (頻度)
1	子供-ガ-遊べる & 遊べる公園-ガ-ない (5)	駐車場 & 駐車 (156)
2	子供-ガ-遊べる & 遊べる場所-ガ-ない (3)	駐車場 & 利用 (113)
3	子供-ガ-遊べる & 掃除-ヲ-してほしい (2)	駐車場 & 郵便局 (77)

グの有効性を示すために、意見タグ集計結果、比較のために単語集計結果を示す[6]。

調査対象は、「公共施設」カテゴリ[7]以下に、2015 年 3 月から 2016 年 12 月 20 日までに投稿された 34,000 件の投稿である。「公共施設」カテゴリは身近な行政サービスが言及される傾向があるため、地域性のある問題発見につながると考える。

表 3 に意見タグ集計と単語集計のうち、上位 3 件の結果を示す。意見タグ集計では、「公園のトイレ-ガ-汚い」のように、意見タグ自体が意見情報として機能しているケースも観察できる。単語集計では「駐車場」が多く言及されていることが観察できるのみで、それ以上の情報を得ることはできない。頻度と割合を比較すると、意見タグ集計は単語集計と比較して低頻度になっているが、意見タグ単位で集計したために、単語よりもさらに疎な情報になってしまったためであると考えられる。

意見タグ集計結果においても「子供-ガ-遊べる」のように、文脈情報が足りないケースが発生している。こういったケースに対し、文脈情報を補うために、共起集計を実施した。表 4 に意見タグと単語の 1 位項目のみについて、共起の集計結果を示す。意見タグ共起の集計結果からは、「遊べる公園-ガ-ない」といった「子どものための場所」を求める声を観察することができる。一方で、単語共起の集計結果からは、駐車場を利用する上での問題を言及していると推測はできるが、これ以上の情報を観察することができない。

この集計結果から、単語集計と単語共起だけでは文脈を捉えることが難しくとも、意見タグ集計では、文脈情報を捉えた分析を実施可能であると言える。

5 おわりに

本論文では、FKC コーパスからのタグ情報付与の取り組みを紹介した。我々の手法で付与されたタグは、4 節の例で示した通り、単純集計でも定量調査に耐えうる情報であると言える。課題点としては、意見タグの拡張範囲と、意見とは言い難いフレーズが誤抽出される問題が挙げられる。これに対しては、統計情報を利用した意見タグの拡張範囲決定と、意見性の判別を行なう必要がある。集計時の課題点としては、意見タグの集計結果が疎情報になってしまう傾向がある。この問題に対しては、密ベクトル情報を利用したクラスタリングの導入が検討できる。

今後は、こうした問題の解決に取り組むと共に、本データセットの研究目的での公開を目指す予定である。

[6]形態素解析には JUMAN を利用し、名詞の一部と形容詞、動詞を集計の対象とした。

[7]カテゴリは投稿時にユーザーが選択する。「公共施設」は「公共・環境」カテゴリ内に存在するサブカテゴリであるが、簡便のため、カテゴリと記述した。

参考文献

[1] Lingjia Deng and Janyce Wiebe. Mpqa 3.0: An entity/event-level sentiment corpus. In *HLT-NAACL*, pp. 1323–1328. The Association for Computational Linguistics, 2015.

[2] Nobuhiro Kaji and Masaru Kitsuregawa. Automatic Construction of Polarity-tagged Corpus from HTML Documents. In *21st International Conference on Computational Linguistics, Poster Sessions (COLING/ACL2006)*, pp. 452 – 459, 2006.

[3] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1065–1074, 2007.

[4] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. FKC Corpus: a Japanese Corpus from New Opinion Survey Service. In *Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp. 11–16, 2016.

[5] Shigeyuki Sakaki, Francine Chen, Mandy Korpusik, and Yan-Ying Chen. Corpus for customer purchase behavior prediction in social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 23–28, may 2016.

[6] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[7] 黒橋禎夫, 長尾真. 日本語形態素解析システム juman version 3.62 使用説明書. 1999.

[8] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学. 構文・述語項構造解析システム knp の解析の流れと特徴. 言語処理学会 第 19 回年次大会, pp. 110–113, 2013.

[9] 山口翼. 日本語大シソーラス類語検索大辞典. 大修館書店, 2005.

[10] 矢野晋哉, 安田幸司. ツイッター情報を利用した道路開通に関する評価分析事例. Technical report, 一般社団法人システム科学研究所調査研究部, 2013.

[11] 光秋迫村, 潔和泉, サンティセーヨー. Twitter のテキストとネットワークの解析による経済動向分析. 第 10 回人工知能学会研究会資料, 2013.