不満調査データセットの素性ベクトル化

末廣 駿　　斎藤 博昭
慶應義塾大学 理工学部情報工学科
{suehiro, hxs}@nak.ics.keio.ac.jp

1 はじめに

企業にとって利用者が製品やサービスに対してどのような反応を示しているかはその後の製品開発や経営戦略を考えていく上での重要な情報資源である。とりわけサービスや製品に対する利用者の不満は、企業側がサービス改善を図る際の具体的な助けとなる。また既存のサービスや製品に対して、利用者がどのような不満を抱いているかを調査すれば新しい製品やビジネスチャンスにつながる可能性もある。

こういった背景の中、利用者から不満を買い取りそれを必要とする企業に売るというビジネスが株式会社不満買取センターによって行われている。この不満買取センターに一般ユーザが投稿した様々な不満は、研究者向けに不満調査データセット[1]として提供されている。このデータセットは不満買取センターのオペレータによってタグ付けがなされているが、生データであるため言語処理を施す必要がある。

そこで本稿では、様々な言語処理に対応するための前処理としての素性化の手法を提案する。その際に不満内容だけでなく、タグ付けされた情報やJUMAN[黒橋 05] 付属辞書から得られた意味情報も素性として選択することを考える。また素性化を行ったデータセットに対し、検索タスクにおいて意味情報の使用が場合によっては有用であることを実験を通して示す。

2 提案手法

本節では扱うデータセット、ツールについて説明し、提案手法について述べる。

2.1 不満調査データセット

本稿で扱う不満調査データセットには254,683件の不満が含まれており、json形式で提供されている。不満が投稿された期間は2015年3月18日から2015年9月23日までである。1件の不満は表1のようにタグ付けされた複数の情報を持つ。

表 1: 各不満が持っているタグの一部

タグ名	説明
fuman(必須)	ユーザが投稿した不満そのもの
category	メインカテゴリ
sub_category	サブカテゴリ
company_name	不満の対象となっている企業
product_name	不満の対象

2.2 bag-of-words

一般にテキストのような非構造データは変換処理をかけてベクトルに変換しないと機械学習アルゴリズムなどが適用できない。この変換処理のことを素性化と呼ぶが、素性の選択と符号化の方法は学習結果に大きな影響を与える。

テキストを素性化する方法はいくつかあるが、本稿では基本的なbag-of-wordsを用いる。bag-of-wordsはテキストを単語集合とみなし、その中に単語辞書中の単語が含まれているかのみを考慮するモデルである。bag-of-wordsで素性化すると素性ベクトルの長さが単語辞書のものに固定されるので他の様々なアルゴリズムへ発展しやすくなる。

[1]本稿では、株式会社不満買取センターが国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用させていただいた。

2.3 JUMAN、KNP

本稿では、京都大学の黒橋らが提供している日本語形態素解析ツールの JUMAN 及び構文解析ツールの KNP[Kawahara 06] を使用した。JUMAN は日本語の形態素解析 [保田 06] として分かち書きと品詞タグ付け、見出し語化だけでなく、付属辞書から得た以下のような情報も出力するので、そういった構文・係り受け関係以外の意味情報も素性として選択する。

- 代表表記
  基本語彙の表記揺れをある程度吸収する
  ex.) 子供／子ども／こども → 子供
- カテゴリ・ドメイン
  基本語彙の大まかな抽象化
  ex.) 子供 → カテゴリ:人 ドメイン:家庭・暮らし
- Wikipedia 辞書からの情報
  JUMAN の Wikipedia 辞書に基本語彙について以下の情報があれば出力
  - エントリ
    Wikipedia から獲得した表現
  - リダイレクト[2]
    Wikipedia から獲得した同義表現
  - 上位語
    Wikipedia から獲得した上位語 (一部は JUMAN の意味情報)

  ex.) Ｔｈｉｎｋ Ｐａｄ
  → Wikipedia 上位語:ノートパソコン

2.4 提案システム

1 件の不満を素性化する際のプロセスを図 1 に示す。本稿では、素性として category と sub_category、company_name、product_name、fuman のタグを用いる。各タグについて形態素解析を行ったあと正規化代表表記、カテゴリ・ドメイン、Wikipedia 中のエントリ・リダイレクト・上位語の意味情報を素性として取り出す。ただし fuman に関しては、構文解析の結果から主辞形態素の品詞が

- 動詞または形容詞
- 数詞や代名詞でない名詞

[2]記事によっては略称や表記揺れなどで微妙に異なる場合があるので、正確な名称の記事を表示する必要がある

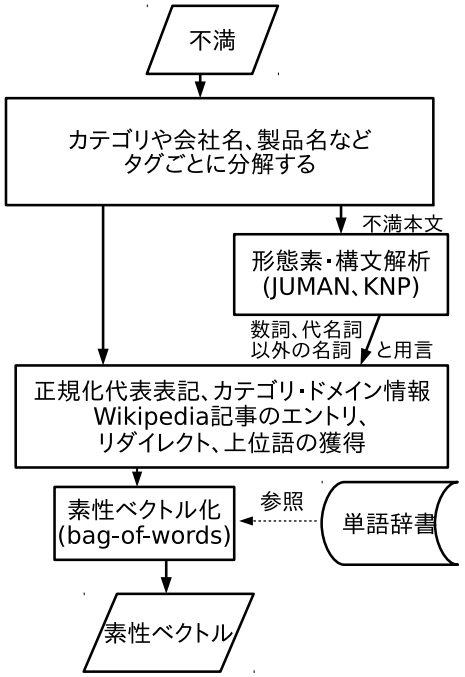

図 1: 素性化のプロセス

である単語についてのみ意味情報を取り出す対象とする。

各タグから取り出した意味情報をまとめ、重複を省いたものを一つの不満の素性とし、それと単語辞書を照らしあわせて素性ベクトルをつくる。単語辞書は全ての不満から得た素性 (単語) をまとめたものである。本稿ではどのタグから得た素性なのかにはこだわらず重複を除いたものを単語辞書として使用した。素性の抽出結果と単語辞書及び素性ベクトルの次元数をまとめたものを表 2 に示す。

表 2: 素性の抽出結果

タグの種類	次元数
category と sub_category	307
company_name	20825
product_name	29371
fuman 中の単語	31032
重複を除いた総数	56388

3 評価実験

本節では素性ベクトル化した不満調査データセットを用いた検索システムについての実験とそれについての評価を述べる。

3.1 実験

本稿で行う検索タスクについて説明する。

1. 入力文の形態素・構文解析を行い、データセットと同様に素性ベクトル化する。
2. 素性ベクトルにおいてビットが立っている要素の集合を検索の候補とする。
3. 検索候補内の要素を全て持つ不満を完全一致としてカウントして出力する。
4. 検索候補内の各要素について、それを持つ不満をカウントしてそれぞれ出力する。

入力は文もしくは単語、また複数の文を検索する場合は半角スペースを区切りとして入力することを想定している。入力に用いられる文字は日本語か英数字で記号類は対象外とした。

3.2 結果

素性化済みの不満調査データセット 25027 件から無作為に抽出した 200 件に対して検索を行い、完全一致の出力結果の再現率、適合率、f 値を算出した結果を表 3 に示す。

表 3: 入力単語と出力結果の評価

入力単語	適合率	再現率	f 値
衣類	0.84	0.61	0.71
スポーツ	0.48	1.00	0.65
金銭	0.93	0.70	0.80
値段 高い	1.00	0.67	0.80
乗り物	1.00	0.93	0.97
時計	1.00	1.00	1.00

3.3 評価

実験結果の f 値は入力によって揺れがあり安定した結果は得られなかった。原因としては意味情報の抽出する段階で間違った情報が素性になってしまうことが考えられる。誤った素性化のパターンとしては

- 単語を必要以上に形態素解析してしまい、余計な意味情報を抽出する。
  - 「時間帯」→「時間」+「帯」→ 意味情報として「衣類」を抽出
- 曖昧性のある単語からの意味情報抽出の失敗。
  - 「時間がかかる」→「時間」+「係る」
- 形態素・構文解析の結果が間違っている。
  - 「インストラクター」→「インス」+「トラクター」→ 意味情報として「乗り物」を抽出
- 未知語に対する意味情報の抽出の失敗。
  - 「ラー麺ずんどう屋」→「ラー麺ズ」→ 意味情報として「お笑いコンビ」を抽出
- 入力自体が間違っている。
  - 登録カテゴリを間違えて関係ない素性が含まれている。

が確認できた。逆にうまく行った場合では

- 不満の内容「ジムで筋肉を見せびらかす人がいて嫌だ」→「ジム」から意味情報として「スポーツ」を抽出
- 対象の製品「ブルゾン」→「ブルゾン」の意味情報として「衣類」を抽出

など正しい素性を生成できている。また意味情報を抽出する際、ドメイン・カテゴリのところで間違った情報を出力してしまうことが多かったのでドメイン・カテゴリ情報を持たない動詞・形容詞を入力したときの検索精度は高かった (表 4)。

表 4: 用言に限った検索結果

入力単語	適合率	再現率	f 値
切れる	1.00	1.00	1.00
汚れる	1.00	1.00	1.00
高い	1.00	0.80	0.89
悪い	1.00	1.00	1.00

4 考察

最後に解析結果および素性化の精度を上げるための方法を考える。

- 別の形態素・構文解析ツールの使用
  本研究では形態素・構文解析ツールとして JUMAN と KNP だけを用いたが、精度や未知語への対応を考えて MeCab[Kudo 05] や KyTea[Neubig 11] のような他の形態素解析システムや、CaboCha[Imamura 07] といった構文解析システムを使用する。

- 外部シソーラスの使用
  本研究では JUMAN がもつ内部辞書のみを使用しているが、他にも日本語語彙体系 [池原 97] や日本語 WordNet[Bond 09] などの他のシソーラスを使用することで抽出する意味情報の幅と精度を上げる。もしくは、Web 上の知識から新しくシソーラスを構築したり [中山 06]、Web 上の知識を用いて既存のシソーラスを拡張したりする [小林 12] という選択肢もある。それでも未知語への対応は完全といえないので、既存のシソーラスから得た汎化推論規則から未知語の属性を推測する [伊東 15] などの必要がある。

- 曖昧性の解消
  本研究では 1 単語ごとに形態素解析・意味情報の抽出をしているため間違った語義で意味情報を抽出している場合があるので、n-gram の情報と Web 資源から語義の曖昧性を解消する [村本 10] ことで精度を向上させる。

5 おわりに

本稿では、不満調査データセットをより扱い易くするための素性の設計と実際に素性化するシステムの提案を行い、単純な検索システムを実装することにより評価を行った。結果、動詞や形容詞などの用言に関してはかなり高い精度で素性化でき、検索タスクにも耐えうるシステムが開発できた。今後は考察で述べた方法の他、処理によっては単語の分散表現なども利用することで、未知語や語義の曖昧性に対して頑強なシステムを構築していく予定である。

参考文献

[Bond 09] Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., and Kanzaki, K., "Enhancing the japanese wordnet", in *Proceedings of the 7th workshop on Asian language resources*, pp. 1–8, Association for Computational Linguistics, 2009.

[Imamura 07] Imamura, K., Kikui, G., and Yasuda, N., "Japanese dependency parsing using sequential labeling for semi-spoken language", in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 225–228, Association for Computational Linguistics, 2007.

[Kawahara 06] Kawahara, D. and Kurohashi, S., "A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis", in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 176–183, Association for Computational Linguistics, 2006.

[Kudo 05] Kudo, T., "Mecab: Yet another part-of-speech and morphological analyzer", *http://mecab.sourceforge. net/*, 2005.

[Neubig 11] Neubig, G., Nakata, Y., and Mori, S., "Pointwise prediction for robust, adaptable Japanese morphological analysis", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 529–533, Association for Computational Linguistics, 2011.

[伊東 15] 伊東 直弘, 吉永 直樹, 鍛治 伸裕, 豊田 正史 "シソーラスと大規模テキストを用いた汎化推論規則の導出", 東京大学 修士論文, 2015.

[黒橋 05] 黒橋 禎夫, 河原 大輔, "日本語形態素解析システム JUMAN version5.1", 2005.

[小林 12] 小林 暁雄, 増山 繁, "日本語版ウィキペディアのカテゴリー階層に着目した日本語 WordNet 上位下位意味体系の拡張手法", 電子情報通信学会論文誌 D, Vol. 95, No. 6, pp. 1356–1368, 2012.

[村本 10] 村本 英明, 鍛治 伸裕, 吉永 直樹, 喜連川 優, "意味カテゴリに基づく語義曖昧性解消における Web 資源の活用について", 情報知識学会誌, Vol. 51, No. 10, pp. 1234–1242, 2010.

[池原 97] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦, "日本語語彙大系", 岩波書店, 1997.

[中山 06] 中山 浩太郎, 原 隆浩, 西尾 章治郎, "Wikipedia マイニングによるシソーラス辞書の構築手法", 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928, 2006.

[保田 06] 保田 明夫, "形態素解析と分かち書き処理", *38p,(PDF), http://wordminer. comquest. co. jp/wmtips/pdf/H15-01*, Vol. 4, 2006.

[Mitsuzawa 16] MItsuzawa, K., Tauchi, M., Mathieu D., Nakashima, M. and Mizumoto, T., "FKC Corpus: a Japanese Corpus from New Opinion Survey Service.", in *proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp.11–18, LREC 2016 Workshop Proceedings, 2016.