言語横断テキストマイニングのための翻訳対抽出

那須川 哲哉†　Daniel Andrade††　海野 裕也†　村松 祐希†††　山本 和英†††
†日本アイ・ビー・エム株式会社 東京基礎研究所
††東京大学　†††長岡技術科学大学

1. はじめに

企業活動のグローバル化が進む中、世界各地において多様な言語で記述された業務関連のテキストデータが日々蓄積されている。この膨大なデータを活用することで、品質と生産性の向上につなげたいという期待が高まっている。この期待に応えるためには単言語のテキストデータで実現されているテキストマイニングを多言語対応に拡張するだけではなく、複数の言語を横断的に分析する仕組みが必要である。

例えば、世界中に商品を提供している企業のコールセンターには各国の顧客から問い合わせやクレームが寄せられる。多くの企業において、その内容をオペレータが要約し、自由記述形式のテキストで何らかのシステムに入力し蓄積している。その際、顧客が用いた言語で入力されることが多く、基本的には、その言語を母語とする現地法人の担当者のみがデータにアクセスしているのが現状である。この多言語データを一元的に管理して網羅的に分析できれば、多様な地域で自社商品や競合商品がどう評価されどういった不具合を生じているかなどをいち早く把握でき、より早い対応が可能になり、企業の競争力向上に結び付くと考えられる。このような顧客の声のデータを分析するのに有効な技術がテキストマイニング[1]である。

テキストマイニングは、データ全体における内容の偏りや変化を捉えることで、個々のデータを読むだけでは認識できない気づきに結びつけることを可能にする。例えば「消耗が早すぎる」という声が、類似商品に比べて特定の商品に目立っていれば、その商品には何らかの不具合が存在する可能性が高い。また、「味がおかしい」という声が急増すれば、製造工程や流通過程で何らかのトラブルが発生した可能性がある。このような偏りや変化をいち早く検出し、適切に対応することで損失やイメージダウンを抑えることができる。

多言語で記述されたデータをテキストマイニングで横断的に分析するためには、分析対象となる概念が各言語においてどのように表現されているかの知識が必要となる。逆に言えば、分析対象概念の表現さえ認識できれば、各テキストデータの全文を翻訳しなくとも、テキストマイニングにより業務上有益な気づきを得られる可能性がある。但し、分析対象となる概念が多言語のデータでどう表現されているかを把握するのは容易でない。例えば、自動車の不具合に関するデータの場合、米国の顧客から寄せられる英語のデータにおいて、*“driver’s side”*という表現が頻出するが、日本の顧客から寄せられる日本語のデータにおいては、同じ概念を示すのに、「運転手側」と表現されることは少なく、「運転席側」と表現されることが多い。さらに運転席の位置が米国の車と日本の車で通常左右逆のため、*“left”*が「右」に対応する状況も出てくる。したがって、既存の対訳辞書に頼らずに、各表現がデータ中でどう使われているかを把握する必要がある。

本稿では、多言語で記述されたテキストデータを母国語で分析することを可能にする言語横断テキストマイニングの概要と、それを実現する上で必要となる翻訳対抽出の特徴を示した上で、テキストマイニングの機能を利用した翻訳対抽出手法を提案し、実データを用いての実験結果と課題を示す。

2. 言語横断テキストマイニング

多言語で記述されたテキストデータの実例として、以下では日米において自動車の不具合を報告しているデータを用いる。日本語のデータとしては、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 2007 年 7 月までの 20,269 件を用い、米国における英語のデータとしては米国政府組織に属する National Highway Traffic Safety Administration (NHTSA) が収集公開している “Consumer Complaints[2]” のデータのうち、2008 年 5 月 14 日までの 525,055 件を用いた。

日米のデータとも基本的に運転手から報告された不具合の概要が自由記述形式のテキストで入力されており、そこに車名や走行距離といった定型データが紐付いている。テキスト部分には、例えば「アイドリングの際にエンジンから異音がする」「走行中にエアコンの吹き出し口から煙が出た」といった内容が記述されており、個々の自動車会社が収集・蓄積している顧客の声のデータとの類似性が高い。

英語のデータ 525,055 件を筆者らが開発したテキストマイニングシステム IBM TAKMI®[1] の製品版である IBM® Content Analyzer（以下 ICA）で分析すると、例えば、図1に示す相関分析機能により、どの車に関する不具合の報告でどの部位が目立つかを把握できる。図中、最上段のセルに部位を示す表現と、その表現もしくはその同義表現をテキスト中に含むデータの件数が示されており、左端のセルに車名と各車名に結びついたデータの件数が示されている。内側のセルは各車名のデータのうち各部位の表現をテキストに含むデータの件数と、その相関指数が示されている。この相関指数は、直感的な解釈としては、分布に偏りがないと仮定する場合の何倍程度の件数が報告されているかを示す値である。例えば、Model E という車に関する不具合の報告のうち door を示す表現を含むデータが 813 件存在し、その割合が他の車種に比べ 2.3 倍高いということが表示されている。相関値が高いセルは自動的にハイライトして目立たせている。

[1] http://www.mlit.go.jp/jidosha/carinf/rcl/index.html
[2] http://www-odi.nhtsa.dot.gov/complaints/

サブカテゴリ／キーワード	brake 63302	engine 47561	tire 40179	transmission 34307	light 30825	seat 22596	gear 19719	door 18925
Model A 15124	1309 0.7	881 0.6	3761 3.1	1166 1.1	692 0.7	503 0.7	656 1.0	640 1.1
Model B 13574	1051 0.6	1461 1.1	1147 1.0	1402 1.5	560 0.6	205 0.3	540 0.9	237 0.4
Model C 10058	1029 0.8	824 0.8	1061 1.3	346 0.5	320 0.5	220 0.4	328 0.8	485 1.2
Model D 9405	987 0.8	498 0.5	233 0.3	958 1.4	660 1.1	422 0.9	379 0.9	581 1.5
Model E 9167	1710 1.5	739 0.8	282 0.3	627 0.9	312 0.5	295 0.6	620 1.6	813 2.3
Model F 8808	935 0.8	1343 1.6	415 0.5	1373 2.2	700 1.2	187 0.4	523 1.4	546 1.5

図 1: 車名とテキスト中に出現する部位関連表現との相関

各セルをクリックすることで、分析対象をそのセルに対応するデータに絞込み、より詳細な分析を行うことが可能になる。例えば、図 1 の範囲外にあるセルでは、Model X が tank という部位と相関が強いことが示されており、そのセルをクリックして、不具合情報のテキストに tank の表現が含まれている Model X のデータ 721 件を対象とした分析に移行できる。この 721 件のテキスト中に出現する不具合現象を示す表現の一覧を相関値順に出力させた結果が図 2 である。

キーワード	頻度	相関値
fire hazard	20	11.7
leakage	29	6.9
fuel leak	9	3.8
leaking	16	2.9
fire	74	1.9
malfunction	14	1.3

図 2: テキスト中に tank の表現が含まれている Model X のデータ 721 件に相関の高い不具合現象を示す表現

以上の情報から Model X では tank に不具合が発生している可能性があり、しかもそれが燃料漏れや火災に結びついている可能性が考えられる。これがテキストマイニングによって得られる「気づき」である。Model X の tank に構造上の問題が実際に存在するかは、車を調べてみなければ分からないが、少なくとも調査をする価値がありそうなことに気づかせてくれるのがテキストマイニングの効果である。筆者らの経験上、こうした気づきがトラブルの発見につながるケースは実際に多い。

図 1 の最上段や図 2 の左の欄には英語の表現が示されているが、これはラベルに過ぎないため、対応する日本語に置き換えることができれば、英語の知識が無くとも、英語のデータを日本語で分析することが可能になる。そこで、分析対象となる表現が他言語ではどう表現されるかの知識を用い、表示言語を分析したい言語に統一することで言語横断テキストマイニングを実現することができる。

3. テキストマイニングのための翻訳対抽出

以上で示したテキストマイニングの対象となるデータ、特に顧客の声を含むテキストデータにおいては、一般的に

a) 比較的短く簡潔にまとめられている
b) 日々データが追加更新される
c) 内容は特定企業の活動や商品に基本的に関連しているため語彙が比較的限定されている
d) 同じ内容が多様な表現で記述される
e) 走り書きのラフな文体で、スペルミスなどの誤りが多い
f) 略語や特殊な用語が含まれる

といった特徴が見られる。自然言語処理の観点からは、c) d) の特徴のため多義性の解消より同義性の認識が重要な場合が多く、e) f) の特徴のため正確な構文解析を望めない。例えば、かつて筆者らが分析に取り組んだ米国 IBM の PC ヘルプセンターのデータでは、以下のような英文が珍しくない。

- *cust wants memory upgrade info*
- *install sftware...reboot of haRDdrive...make backup*
- *cus said the cdrom wont open*
- *cus cant click on anything*
- *Told cu to C + A + D the sys*

翻訳対抽出を行う上では、このようなテキストデータを対象とするのに加え、テキストマイニングに活用するという観点から

A) 偏りや変化を捉えるため、対象となる概念には一定以上の出現頻度を見込むことができる
B) 分析対象の主な概念はユーザが予め定義する
C) 急増している内容など、未定義の概念をインタラクティブに翻訳対象にしたい場合がある

というアプリケーション上の特徴を考慮した手法が望ましい。

基本的には、一般的な対訳辞書の見出しには含まれていない表現を対象とする上、分析対象となるテキストデータが揃っているため、コーパスから学習するアプローチが妥当と考えられる。但し、データの性質上、対訳コーパスを前提としたアプローチ[3]を取る事はできない。例えばニュースの記事が対象であれば、同じ出来事を対象にした多言語の記事を対象とすることで対訳コーパスを前提とした手法に近いアプローチ[4]を適用することも可能であるが、顧客の声の場合、対象となる出来事が全て異なる。また、翻訳対象表現と原言語コーパス中で共起する表現を要素とする特徴ベクトルを作り、一般対訳辞書を用いて特徴ベクトルを対象言語のベクトルに変換した上で、対象言語コーパス中の各表現から特徴ベクトル同士の cosine 距離が近いものを導き出すアプローチ[5]は対訳コーパスを必要としない点で有望であるが、対象言語の全表現の特徴ベクトルとの距離の比較は計算コストが大きく、日々更新されるデータを用いてのインタラクティブな処理を実現する上では、計算コストをより低くできる手法が望ましい。

以上の観点から、筆者らは、翻訳対象表現に対してテキストマイニングシステムが出力する相関の高い共起表現を翻訳表現に結びつけるピボット表現として利用する翻訳対抽出手法を開発した。相関関係には対象性があるため、原言語表現 $\boldsymbol{S}_i$ に対し相関の高い表現の中に $\boldsymbol{S}_{HC}$ が含まれるなら、$\boldsymbol{S}_{HC}$ に対し相関の高い表現の中には $\boldsymbol{S}_i$ が含まれる。そこでもし、$\boldsymbol{S}_{HC}$ が

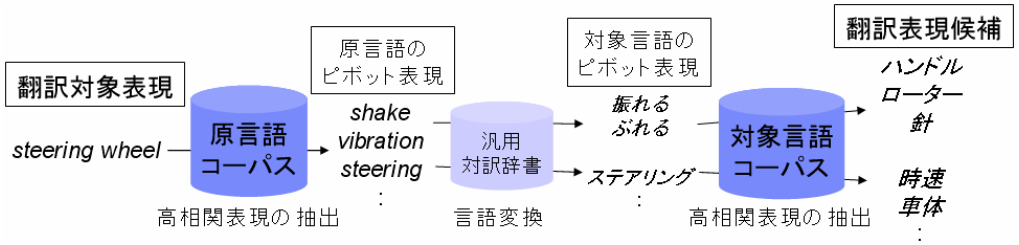

図 3: 翻訳表現候補リスト作成ステップの処理の流れ

対象言語の表現 $\boldsymbol{T}_{\mathrm{HC}}$ に翻訳できるなら、$\boldsymbol{T}_{\mathrm{HC}}$ に相関の高い表現の中には $\boldsymbol{S}_{\mathrm{i}}$ の翻訳表現 $\boldsymbol{T}_{\mathrm{o}}$ が含まれる可能性が高いだろうという考えがベースになっている。このアプローチには、

- 相関の高い表現を抽出する部分で既存のテキストマイニングシステムの API を利用できる
- 特徴ベクトルのインデックスの構築が不要
- 比較対象が限定される

という特長があり、計算コストの削減を図ることができる。

4. 翻訳対抽出アルゴリズム

筆者らが開発した手法は二つのステップから構成される。まず翻訳表現候補のリストを作成し、次に、そのリスト中の各表現の翻訳表現としての妥当性を評価する。

4.1. 翻訳表現候補リスト作成ステップ

本ステップの処理の流れの概要を具体例を用いた形式で図 3 に示す。このステップは三段階に分かれており、第一段階では、与えられた翻訳対象表現に対し、原言語コーパスにおいてその表現と相関の高い表現のリストを抽出する。その際、相関の高さを測る指標として、以下の式で与えられる値を表現 A と表現 B の相関値と呼ぶことにする。

$$\frac{(\text{表現 A と B を両方含む文書数})/(\text{表現 A を含む文書数})}{(\text{表現 B を含む文書数})/(\text{全文書数})}$$

これは、翻訳対象となる表現 A を含む文書集合において表現 B が出現する割合と、原言語コーパス全体において表現 B が出現する割合との比を取ったものであり、その値が 1 を超えれば相関が強いということになる。本ステップにおいては、予め設定された閾値を越える相関値を取る表現が全て出力される。例えば、“*steering wheel*”が翻訳対象の場合、図 3 に示すように、“*shake*” や “*vibration*” といった表現が相関の高い表現として抽出される。この処理は前述のテキストマイニングシステム ICA が図 2 のリストを出力する処理と同等であり、ICA には、原言語コーパスをテキストマイニングの対象とした状態である表現を与えると、その表現と共起する表現のリストを出現頻度と相関値付きで返す API が用意されている。

第二段階では、汎用対訳辞書を用いて、第一段階で得られた相関の高い表現を対象言語に変換する。その際、汎用対訳辞書の見出しに含まれない表現は単純に対象外とし、訳語候補が複数存在する場合には、全ての候補を出力する。

第三段階では、第二段階で得られた各表現に対し、第一段階と同様に対象言語コーパス中で相関の高い表現のリストを抽出する。その結果得られる各リストの表現をマージしたリストが、翻訳表現候補リストとして出力される。

4.2. 翻訳表現としての妥当性評価ステップ

本ステップでは、前ステップで得られた各翻訳表現候補と翻訳対象表現との意味的類似性を測る尺度として、各表現に相関の高い語が汎用対訳辞書を介してどれだけ対応しているかを調べる。例えば、翻訳対象表現が“*steering wheel*”で翻訳表現候補が「ハンドル」の場合、図 4 に示す通り、前ステップの第一及び第三段階と同様にして、各表現に対し予め設定された閾値よりも相関の高い表現のリストを抽出し、各リスト中の表現で汎用対訳辞書を介して対応する表現がいくつ存在するかに応じた評価値を与える。

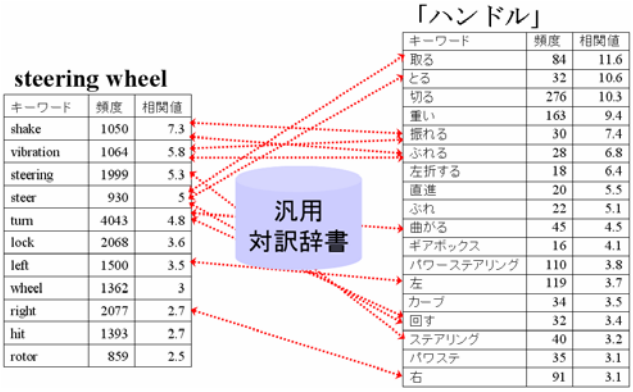

図 4： steering wheel と「ハンドル」の意味的類似性の評価

5. 実験

前述した日米の自動車の不具合データを用いて、前節の翻訳対抽出手法がどの程度有効かを、英語に対する日本語、日本語に対する英語の双方向で調査した。また英語から日本語の方向で、コーパスのサイズと精度の関係を調査した。

5.1. 実験データと実験環境

コーパスとして使用したデータは、日本語側が自動車不具合情報の 20,269 件であり、個々のデータに含まれるテキストの平均長は約 35 文字であった。英語側は Consumer Complaints の 525,055 件であり、個々のデータに含まれるテキストの平均長は約 295 文字（約 51 単語）であった。

翻訳対象表現としては、単言語内でのテキストマイニングにおいて分析上有効な表現として分析者が辞書登録していた表現のうち部位のカテゴリに属する表現を用いた。英語は 100 表現、日本語は 30 表現である。

汎用対訳辞書としては英日機械翻訳用に整備された辞書を利用した。多義性を展開して 1 対 1 にした翻訳対の数は約 16 万件である。但し、表現の微妙な違いから翻訳対がうまくマッチしないケースが予備実験で多く見られたため、ノイズが入るのを覚悟で処理中に E→J→E→J という展開を行った。例え

ば、*“vibration”*の対訳として「振動」が得られると、同じ「振動」を対訳とする*“jar”*や*“joggle”*の対訳である「衝突」や「軽い揺れ」も*“vibration”*の対訳として扱うようにした。

翻訳対抽出の過程で相関の強い表現を抽出する際の閾値は小規模の予備実験で試行錯誤した結果、日本語側を 1.2、英語側を 3.5 に設定した。

5.2. 翻訳対抽出精度

日本語側20,269件と英語側525,055件の全データを使用して、英語→日本語と日本語→英語の双方向での翻訳対抽出を行った際の精度評価結果を表 1 に、出力結果の一部を表 2 に示す。

表1: 翻訳対抽出精度の評価結果

	英語→日本語（100 表現）		日本語→英語（30 表現）	
	件数	精度	件数	精度
1 位の候補が正解	31	31%	10	33%
5 位以内に正解	60	60%	20	66%
10 位以内に正解	72	72%	22	73%
20 位以内に正解	81	81%	27	90%

表 2: 翻訳対抽出の出力例 （下線の表現を正解と判断）

翻訳対象	radiator	headlight	ガソリン	ブレーキ
第 1 候補	冷却	前照灯	gasoline	foot
第 2 候補	冷却水	ヘッドライト	smell	pedal
第 3 候補	ラジエータ	ライト	fuel	parking
第 4 候補	ホース	レンズ	gas	brake pedal
第 5 候補	タンク	方向指示器	tank	lane
第 6 候補	ラジエター	熱	fuel tank	braking
第 7 候補	オーバーヒート	水滴	odor	park
第 8 候補	コア	曇り	gas tank	brake
第 9 候補	ヒーター	制動灯	exhaust	traffic
第 10 候補	水漏れ	雨天	leak	someone

20 位以内に正解が存在する割合が、双方向において 8 割を超えるという結果が得られた。また、表 2 に示される通り、同義表現を抽出してくれるため、テキストマイニングの辞書を作成する上では有効性が高い。その半面、データ内容の異なりにより、翻訳対の抽出が非常に困難と考えられるケースも見られた。例えば、英語のデータでは比較的高い頻度で出現する steering column（4,406 件）や cruise control（3,359 件）という部位に関して日本語のデータでは殆ど言及されておらず、「ステアリングコラム」を含むテキストは 1 件、「ステアリングコラムシャフト」が 2 件、「オートクルーズ」が 7 件、「クルーズコントロール」が 5 件であった。

5.3. コーパスのサイズと精度の関係

日本語のコーパスのサイズを 5 千件から 2 万件まで、5 千件刻みで 4 段階、英語のコーパスを 10 万件から 50 万件まで 10 万件刻みで 5 段階変化させて英語から日本語への方向での翻訳対抽出精度を調べたところ、正解率は、日本語コーパスのサイズに応じて変化するものの、英語コーパスのサイズの変化には影響されなかった。そこで、日本語コーパスのサイズを 2 万件に固定し、英語コーパスのサイズを 5 百件、千件、1 万件、2 万件、4 万件、6 万件、8 万件と変化させたところ、全体的には、英語コーパスのサイズが 2 万件付近から精度が安定する傾向が見られた。従って、日本語コーパスはまだ足りない可能性があるものの、英語コーパスは 2 万件（単語数にして約百万語）程度で充分という可能性が認められた。

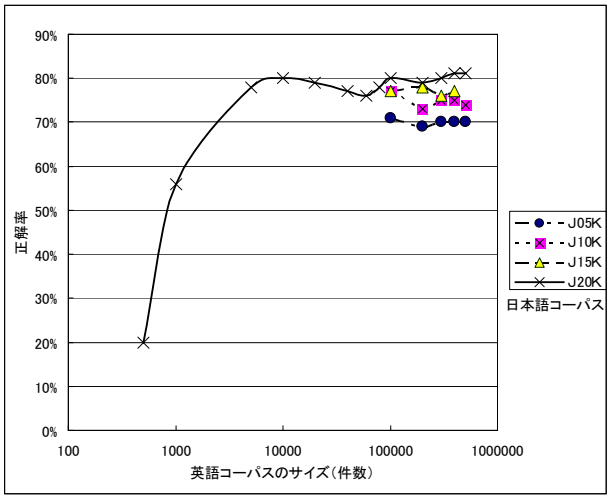

図 5 :正解が 20 位以内にある正解率とコーパスサイズの関係

6. おわりに

需要が高まりつつある言語横断テキストマイニングとその実現に必要となる翻訳対抽出についての特徴を示した上で、そのための翻訳対抽出手法を提案し、その実験結果を示した。

実装実験では、まず英語の表現を日本語の表現へ対応付ける方向でシステムを構築し実験した上で逆方向を試した。その結果、殆ど手間をかけずに逆方向で同等以上の精度を出すことができ、本手法の汎用性の高さを確認できた。20 位以内に正解の含まれる割合が 8 割を超えるうえ、出力に同義表現が多数含まれるため、言語横断テキストマイニングの辞書整備に有益である。但し、異なる言語圏における生活環境の違いから、データ中に頻出している概念が必ずしも一致せず、適切な翻訳表現が対象言語のコーパス中に見つからないケースが散見された。そのため、人手による確認無しに出力をそのままテキストマイニングに利用するためには、例えば正解に確信度を付けるといった、何らかの工夫を加える必要がある。

参考文献

[1] 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

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

[3] H Kaji, Y Kida, Y Morimoto. Learning translation templates from bilingual text. Proc. of 14th COLING, pp. 672-678. 1992

[4] T. Utsuro, T. Horiuchi, T. Hamamoto, K. Hino, and T. Nakayama. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. Proc. of 10th EACL, pp.355-362. 2003.

[5] R. Rapp, Automatic identification of word translations from unrelated English and German corpora, Proc. of 37th ACL, pp.519-526. 1999