評価文書分類における異言語翻訳データの利用法

乾 孝司　山本 幹雄
筑波大学大学院 システム情報工学研究科
{inui,myama}@cs.tsukuba.ac.jp

1　はじめに

近年，Web のコモディティ化に伴い，Web 上には Web ユーザによって生成された多数の製品やサービスに関する意見や評判が記述されたテキストが存在している．そして，これらの評判情報を自動的に整理・要約する評判分析 [6, 12] に関する技術開発が活発に進められている．評価文書分類は，評判分析の中核を担う要素技術のひとつであり，評判が記述された文書に対して，その評判の内容が肯定的か否定的かを判定，分類する技術である．

評価文書分類は，トピック文書分類と同様，主に教師あり学習に基づく手法が主流であり [7]，分類精度の向上には，新しい分類手法の開発と共に，十分な教師ありデータの確保が重要な課題となる．しかしながら，一般に教師ありデータは様々な面で高価であり，その確保には多大なコストを費やす必要がある．そのため，低コストで入手でき，教師ありデータの代替として利用可能なデータを準備し，これらのデータを用いて分類器を学習するアプローチが検討されている．例えば，Aue ら [2] は，十分な量の教師ありデータの確保が困難な新規ドメインの評価文書分類に対して，ナイーブベイズ法に EM アルゴリズムを組み合わせることで教師ありデータと教師なしデータを混合して分類器の学習に利用する Nigam らの半教師あり学習手法 [5] を適用している．

上記の半教師あり学習のように，教師ありデータと教師なしデータを混合するスタイルの他に，言語の異なる複数の教師ありデータを混合するスタイルの手法も幾つか検討されている [3, 10]．複数言語のデータを混合するには言語間のギャップを取り除く必要があるが，通常は機械翻訳を介することで，このギャップの解消にあたる．しかしながら，副作用としてデータに翻訳誤りが混入するという別の問題が生じる結果となり，このことが最終的な分類精度に影響を及ぼすことになる．

本稿では，上記の問題に対して，翻訳誤りの影響を低減する手法を検討する．ここで，翻訳誤りの影響を取り除くもっとも直接的な方法として，翻訳された文の信頼性を評価することで，翻訳の誤り箇所を検知し，この部分の翻訳結果を修正する，あるいは，学習データから除外する等の適切な処置を施すことが考えられる．しかしながら，翻訳文の信頼性を評価することは翻訳処理自体と同程度に困難な課題である．また，信頼性評価のためには豊富な参照訳を必要とするが [8]，豊富な参照訳の確保にはやはり多大なコストを要することになる．

以上を踏まえ，本稿では参照訳を利用しない簡便な手法を検討する．ここで，教師データとなる文書内の文を表 1 のように，翻訳誤りの有無と評価文書分類への貢献度の観点から 4 種類に分割して考えると，A に該当する文のみからなる要約文書が最適な教師データであり，逆に D からなる要約文書は教師データとして適さないと言える．理想的には A に該当する文で構成される要約文書を作成したいが，先に述べた通り，AB と CD を区別することは参照訳を必要とするため，ここでは放棄する．そこで，AC と BD を区別することを考え，A に該当する文で構成される要約文書の近似として，AC に該当する文で構成される要約文書を作成する手法を提案する．このような要約文書を作成することで，少なくとも D に該当する文が与える悪影響を回避できると考えられる．

表 1: 翻訳誤りと分類への貢献度

		翻訳誤り	
		無	有
分類への貢献度	高	A	C
	低	B	D

なお本稿では，最終的に評価文書分類の精度評価を実施する対象言語として日本語を想定し，また，日本語データに混合する異言語データとして英語を想定して議論を進める．異言語データとして英語を選択した理由は，英語は日本語を含めたその他の言語

と比較して，使用可能なデータ量が豊富に蓄積されているためである．以下本稿では，日本語を英語に翻訳することや翻訳されたデータのことを「日本語→英語」,「日→英」のように参照することがある．同様に，英語から日本語への翻訳は「英語→日本語」,「英→日」とする．

2 提案手法

前節で述べた要約文書を作成するために，まず，翻訳された文書中の文に対する文選択の基準に関する基本的な考え方を述べる．その後，具体的な文選択処理（要約処理）の手続きについて述べる．

2.1 基本的な考え方

文書中の各文に対して文書分類への貢献度を見積もることを考える．そして，貢献度の低い文を文書から除外し，貢献度の高い文のみ文書内に保持することで文選択を実現する．ここで，学習データを利用して貢献度を自動推定することも考えられる．しかしながら，今回は 1 節でも述べたように，代替データではなく純粋な意味での学習データが十分に確保できない設定下を想定していることから，学習データに基づく自動推定は取り扱わない．代わりに，次のような関係を仮定し，文内に含まれる評価表現の情報に基づいて貢献度を見積もる．

仮定：
評価表現を含む文は分類への貢献度が高い．

2.2 評価表現に基づく文選択

上記の仮定に従えば，評価表現に基づく文選択処理は次のようになる．

評価表現に基づく文選択処理：
文書中のある文が評価表現を含んでいれば，その文をそのまま保持し，逆に，評価表現を含んでいなければ，その文を文書から除外する．

文選択処理は翻訳後の文書に対して施すが，翻訳過程で評価表現に翻訳誤りが生じる可能性もある．この場合，これまでの議論は全て成り立たなくなる．そこで，文選択処理で参照している評価表現の翻訳の信頼性を担保するために，文選択処理において保持と決定された文に対して以下に示す追加条件を課すことにする．もし，条件を満たさない場合は，保持ではなく除外とする．

- 条件 1：対訳関係にある翻訳元の文が評価表現を含んでいる．
- 条件 2：翻訳前後で参照している評価表現の評価極性が一致している[1]．

3 異言語翻訳データに基づく分類手法

異言語翻訳データを文書分類に利用する手法が既に幾つか提案されている [3, 10]．前節の手続きによって文選択処理が施された文書も，通常の文書と同様にこれらの手法が適用できる．

本研究では，次に示す既存の 3 つの手法によって文書分類をおこなう．なお，いずれの手法にもデータの翻訳過程があるが，本研究では翻訳の後処理として文選択処理を組み入れる．

Training Translation Model 学習データを評価データ側の言語に翻訳し，翻訳後のデータを利用して分類器を学習する．評価データはそのまま使用する．

Test Translation Model 学習データはそのまま使用して分類器を学習する．評価データを学習データ側の言語に翻訳し，翻訳後のデータを利用して分類をおこなう．

Co-training Model Blum ら [4] によって提案された Co-training の枠組みに従って，上記の両モデルを組み合わせて利用する手法．

なお，先行研究では，教師あり学習データをもつ言語を単一言語に限定しているが，一般には，規模の程度差があるものの，任意言語で教師あり学習データを用意できる．本研究でも日英両言語で教師あり学習データを利用できる環境にあるため，各モデルにおける分類器の学習には両言語のデータを利用する．すなわち，Training Translation Model では，英語を日本語に翻訳したデータと，日本語のデータを混合して，分類器を学習する．また，Test Translation Model では，英語のデータと日本語を英語に翻訳したデータを混合させる．

4 評価実験

実験を通して提案手法の有効性を評価した．

4.1 実験の設定

実験用データとして，Amazon.co.jp [2] 内の日本語レビューと Amazon.com [3] 内の英語レビューを使用した．このレビューには 5 段階の製品評価が付与されており，評価 4 および 5（つまり，良い評価）が与えられたレビューを肯定的なレビュー，評価 1 および 2（つ

[1] 評価表現間に対訳関係があるかどうかは評価しない．
[2] http://www.amazon.co.jp/
[3] http://www.amazon.com/

表 2: 評価表現辞書の統計情報

		内訳	
	全体	肯定	否定
日本語エントリ数	724	340	384
英語エントリ数	1,392	609	783

表 3: 辞書照合に成功する文書／文の割合

言語情報	文書レベル	文レベル
日本語	96%	47%
日本語→英語	99%	67%
英語	97%	63%
英語→日本語	83%	32%

まり，悪い評価）が与えられたレビューを否定的なレビューとみなした．使用した各レビューの総数・内訳は以下の通りである．

- 日本語レビュー 1,000 件（肯定 500 ／否定 500）
- 英語レビュー 10,000 件（肯定 5,000 ／否定 5,000）

両言語とも全て MP3 プレーヤーに関するレビューである．また，同一製品についてのレビューが日英どちらのデータセットにも含まれることがあるが，これらの間に対訳関係はない．

翻訳処理には，「日→英」，「英→日」の両方向ともエキサイト翻訳サービス[4] を用いた．

異言語翻訳データを利用した文書分類の手法として，3 節で示した 3 つの手法を用いた．これらの内部で使用する分類アルゴリズムには SVM[9] を採用し，素性情報には日英両言語とも単語ユニグラムを用いた．ただし，頻度情報は利用しなかった．日本語データの単語分割は MeCab[5] でおこなった．また，英語データに小文字化処理を施した．

実験では 10 分割の交差検定をおこない，分類性能を正解率で評価した．ただし，Co-training Model では，開発データが必要となるため，学習 6：開発 3：評価 1 の割合で分割した．

文選択処理では文内の評価表現を認定する必要がある．本研究では，高村ら [13] の手法によって自動生成された評価表現辞書に人手による修正を加えたものを準備し，この辞書との照合操作によって評価表現の認定をおこなった．利用した評価表現辞書の統計情報を表 2 に示す[6]．また，各実験データにおいて辞書照合に成功する文書／文の割合を表 3 に示す．この表から，文書レベルでみれば，ほぼ全ての文書中にひとつ以上の評価表現が含まれていることがわかる．文レベルでみると，32%〜67%までの幅があるが，大雑把に言えば，およそ半数の文に評価表現が含まれており，文選択処理によって，文書サイズが半分程度になっていたことがわかる．

4.2 実験結果

実験結果を表 4 に示す．表の各列に 3 つの分類手法の結果を示す．各行は文選択処理の違いを表しており，PNDIC が本稿で提案した文選択処理である．RANDOM は，性能比較用に，PNDIC と同数の文をランダムに選択した場合の結果であり，WITHOUT は文選択処理をおこなわい場合の結果である．

まず，WITHOUT の行に注目し，分類手法の比較をおこなう．表 4 から，Training Translation Model および Test Translation Model のどちらよりも，それら両者を同時に考慮する Co-training Model の性能が良いことがわかる．このことは，文選択処理を実施した他の行の結果からも読み取れる．

表の結果とは別に，学習に日本語データのみを利用した場合（翻訳なし，WITHOUT）の正解率を算出したところ，77.9 であった．つまり，我々の実験設定では，Training /Test Translation Model 単独手法では一度性能が悪化し，Co-training Model によって改善に転じていることがわかる．

次に，文選択処理の有効性を検証する．WITHOUT と RANDOM の行の比較から，ランダムに文選択を実現するだけでは性能改善に繋がらないことがわかる．WITHOUT と PNDIC を比較すると，どの分類手法においても，PNDIC の性能が WITHOUT のそれを上回っている．このことから，本稿で提案した文選択処理は，評価表現辞書を用いた簡便な手法であるものの，ある程度有効に働くことが確認できる．

5 関連研究

Agarwal ら [1] は，文書分類のベンチマークとして利用される Reuters-21578 [7] と 20-newsgroups [8] をデータとして，文書分類におけるノイズ混入の影響を調査した．具体的には，各データに人工的に単語の綴り

[4] http://www.excite.co.jp/world/

[5] http://mecab.sourceforge.net/

[6] 本研究では利用していないが，この辞書には人手によって対訳関係情報も含まれている．日本語単語エントリあたりの平均対訳英単語数は 3.15，英単語エントリあたりの平均対訳日本語単語数は 1.64 である．

[7] http://www.daviddlewis.com/resources/

[8] http://people.csail.mit.edu/jrennie/20Newsgroups/

表 4: 文選択処理の効果（正解率）

		分類手法		
		Training Translation	Test Translation	Co-training
文選択手法	WITHOUT	73.6	73.7	78.4
	RANDOM	73.0	69.0	77.5
	PNDIC	77.0	78.1	81.7

誤り（spelling error）を混入させ，綴り誤りの割合と分類性能の関係を評価している．彼らは，調査結果から 50%〜70%の割合で誤りを混入させても，それほど分類性能が低下しないことを報告している．しかし，翻訳誤りとして単語の綴り誤りは現れないことから，翻訳誤りが混入したデータに対する文書分類が，上記の結果と同様になるわけではないだろう．

6 おわりに

本稿では，評価文書分類において異言語翻訳データを利用する際，評価表現に基づく文選択処理を施すことで，翻訳誤りに起因する分類誤りを改善することを試みた．

本研究では文レベルで文書内情報を処理したが，一般的な素性選択の観点から見れば，単語レベルの対処法についても検討の余地がある．また，転移学習 [14] の枠組みで，多言語データを扱うアプローチもあり [11]，これらと本研究との関連性についても検討していきたい．

参考文献

[1] Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. How much noise in text is too much: A study in automatic document classification. In *Proceedings of the 7th IEEE International Conference on Data Mining*, pp. 3–12, 2007.

[2] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing*, 2005.

[3] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 127–135, 2008.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

[5] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.

[6] Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, 2008.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 76–86, 2002.

[8] Sylvain Raybaud, Caroline Lavecchia, David Langlois, and Kamel Smalili. Word- and sentence-level confidence measures for machine translation. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pp. 104–111, 2009.

[9] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[10] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pp. 235–243, 2009.

[11] Qiang Yang, Yuqiang Chen, Gui rong Xue, Wenyuan Dai, and Yong Yu. Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of the ACL-IJCNLP*, pp. 1–9, 2009.

[12] 乾孝司, 奥村学. テキスト評価分析の技術とその応用. 情報処理, Vol. 48, No. 9, pp. 995–1000, 2007.

[13] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌, Vol. 47, No. 2, 2006.

[14] 神嶌敏弘. 転移学習. 人工知能学会誌, Vol. 25, No. 4, pp. 572–580, 2007.