

複数ドメインの意見分析コーパスを用いた アンサンブル学習による意見分析システムの提案

高村 慎太郎

吉岡 真治

関 洋平

北海道大学大学院 情報学研究科 筑波大学大学院図書館情報メディア研究科

takamura,yoshioka@kb.ist.hokudai.ac.jp,yohei@slis.tsukuba.ac.jp

1 はじめに

インターネットの発展に代表されるような社会の情報化により、大規模な情報が容易に手に入るようになった。その反面、入手可能な情報の量がユーザの情報処理能力を超えてしまい、逆にその効率が落ちてしまうということが起こる。そのような状況を改善するため、多くの情報から有益な情報を発見するテキストマイニング技術が1つの重要な方法となっている。テキストデータから意見や評判などの情報を抽出する意見文抽出分野の研究もその中の一つであり、商品や映画の評判分析、トレンド分析、政策や選挙のための情報分析、世論調査などについて実用化が進められている。

利用用途によって対象となる文書ジャンルは様々である。報道機関が発信するニュースに始まり、Web上の商品や企業に対するレビューサイト、近年では、個人の意思や自己体験を記述するBlogやTwitterを対象とした研究も盛んに行われており、それに伴って、意見分析に用いられる多くの学習用コーパスが入手可能な状況となっている[1]。これらは意見表現の特徴においても様々であり、新聞・ブログ・議事録などの異なるタイプのドメインに応じて、文書ドメインの性質に応じた特徴的な意見の記述が存在する。

ただし、これらの特徴的な意見の記述方法は、必ずしも、異なるタイプのドメインにだけ存在するわけではなく、各々のドメインで得た意見情報の抽出に関する知識は、他のドメインにおいても有用であることが考えられる。

このような背景のもと、本研究では、「内容よりも表現による影響を考慮する重要性」に着目し、複数ドメインの文書に関する様々な意見分析システムを構築し、その結果を集約することにより、全体のシステムの向上をはかるアンサンブル学習による意見分析システムを提案する。これにより、どんなドメインに対しても、多様なドメイン特徴を持つ意見の記述を抽出できることを目指す。

2 関連研究

2.1 アンサンブル学習

アンサンブル学習[2]は、汎化能力向上の為の学習法であり、単一の学習機械の性能向上ではなく、個々に学習した複数の学習機械を何らかの形で融合させて汎化性能を向上させる。個々に学習した識別器を複数個用意し、それらを、単純には出力の平均を用いる等して、まとめ合わせて一つの識別器を構築する学習法である。これは、一人で考え込むよりも複数人で議論したほうが、より良い結論が得られるという諺「3人寄れば文殊の知恵」を地で行く学習法であり、実際、識別器の個数を増加することにより識別能力が向上して行く事は、理論的にも実験的にも確認されている。アンサンブル学習は、その高い識別能力に加え単純性（識別器を複数個用意するだけ）と汎用性（任意の識別器に適用可能）という利点からもさまざまな分野での応用が行われている。

我々は、このアンサンブル学習の内、特に、「異なる学習タスクに対し、同じ学習機械で学習した結果をアンサンブルするという形態」に焦点を当てる。

2.2 特徴集合と分類アルゴリズムを用いたアンサンブル学習による意見分析

Rui Xiaら[3]は、Amazon.comから取得した複数の製品レビュー記事（映画、書籍、DVD、エレクトロニクス、キッチン）を対象に、ドメイン毎に有効な特徴や機械学習アルゴリズムは異なることから、個々の欠点を克服し、お互いのメリットの恩恵を受ける為にはアンサンブル学習が有効であると考え、特徴集合と分類アルゴリズムの変化によって複数の意見抽出システムを構築し、それらをアンサンブル学習することを提案した。

この研究は、単体システムの性能向上ではなく、複数システムを活かすという点で意義があり、意見抽出に対してアンサンブル学習は有効であると考えられる。

しかし、意見分析は、内容よりも表現による影響を考慮することが重要であると考えられ、単語の品詞情報や N-gram 法に焦点をあてることで、それぞれのドメインに特徴的な要素を獲得出来ているかは、議論の余地がある。

3 複数ドメインの意見分析コーパスを用いたアンサンブル学習

本研究では、NICT 意見 (評価表現) 抽出ツールと複数のドメインコーパスとして KO:国会議事録:5807 文、YC:yahoo!知恵袋:1907 文、YB:yahoo!Blog:6924 文、BO:書籍批評:406 文 (以上、特定領域研究「日本語コーパス」研究成果 [1])、NP:新聞記事:6670 文 (提供元:NTCIR-MOAT [4])、KB:京都観光ブログ:1267 文 (提供元:ALGIN¹) を用い、複数の意見抽出システムを構築する。最終的に、アンサンブル学習を行う事で、単独の場合との性能比較を行う。

3.1 NICT 意見 (評価表現) 抽出ツール

NICT 意見 (評価表現) 抽出ツール²は、与えられた文からの意見抽出を行うツールである。本ツールは、文からの評価表現抽出モジュール、抽出した結果に対する評価タイプ分類モジュール、極性判定モジュール、評価保持者抽出用モジュールから構成されている。

本研究では、文を単位とした意見性の有無 (2 値判定) について、意見判定と呼ばれるサブタスクを対象としているため、文からの評価表現抽出モジュールについてのみ利用していると考えて良い。

なお、本来であれば、全てのコーパスについて、意見部分を抽出して学習させるのが適切であるが、作業量的に困難であるため、本研究では、意見文と判断されたものは、文全体が意見の記述と設定して学習を行った。

4 評価実験

4.1 実験の方針

本研究では、NTCIR 意見分析タスク [4] における、文を単位とした意見性の有無 (2 値判定) について、意見判定と呼ばれるサブタスクを対象とした実験を行う。

まず、異なる学習コーパスを用いて複数の単一識別器を構築し、その性能を検証する。次に、本研究の提案手法である「アンサンブル学習を用いた識別器」の評価を行い、その有効性を示す。

4.2 単一識別器による評価

本研究では、特定のドメインのコーパスから以下の異なる学習コーパスを用いて作成した 3 種類の識別器 (7 識別器) により意見抽出を行った。

LOCAL(1) 対象ドメインのみを利用した識別器

本実験では、自身所属ドメインのデータセットを均等な文書数に 5 分割し、5 分割交差検定用の識別器を作成した。

OTHER(5) 他ドメインでの学習による識別器

残りの 5 つのドメインの各々について、全ての文書を利用して識別器を作成した。

ALL(1) 全てのドメインを利用した識別器

対象ドメインについては、LOCAL と同様に、5 分割交差検定用の文書群を利用し、それに、残りのドメインの全文書を用いて識別器を作成した。

これらの識別器を用いた評価を各ドメインに対して行った。その結果を表 1 に示す。表の中で、評価対象と識別器が同じものについては、LOCAL を用いた 5 分割交差検定の平均値であり、識別器 ALL についても、自身所属ドメインの 5 分割交差検定の情報とその他のドメインを組み合わせた識別器による 5 つの識別器による平均値を示している。また、各評価対象で、最も性能の良いものを太字で示し、ALL については、LOCAL での性能よりも優れている場合には下線を引いて示す。

太字で示される最大値を見ると、必ずしも、LOCAL による学習結果が、必ずしも、最善となっていないことが確認された。また、ALL の下線に注目すると、コーパスの量を増やすことが必ずしも、性能改善に寄与しないことも確認された。

次に、ALL を用いた場合の影響について分析する。LOCAL の性能の良い、YB,NP では、ALL を使った場合に、性能が低下する可能性があるが、それ以外のドメインでは、コーパスを増やすことが性能向上につながることを確認した。また、OTHER についてであるが、YB ドメインでの学習による識別器は、他のどのドメインを評価しても再現率の向上が見られることから、YB ドメインが豊富な意見表現を含んでいる為であると考えられる。これは、文書ドメインの性質に応じた意見表現が、自身の所属ドメインのみではなく、他のドメインにおいても有効な表現である場合があると考えられる。

また、各々の識別器の精度は、十分に高く、アンサンブル学習に利用するのに十分な精度を持っていると考える。

次節において、アンサンブル学習による識別器の評価結果を示す。

¹<http://www.alagin.jp>

²<http://alaginrc.nict.go.jp/opinion/index.html>

表 1: 各ドメイン毎の 5 分割交差検定と他のドメインによる学習の影響比較

評価対象									
識別器	KO			YC			YB		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
KO	0.713	0.855	0.777	0.704	0.651	0.676	0.855	0.526	0.651
YC	0.757	0.918	0.830	0.748	0.760	0.751	0.822	0.753	0.786
YB	0.682	0.911	0.780	0.663	0.904	0.765	0.835	0.856	0.845
BO	0.688	0.910	0.784	0.672	0.781	0.723	0.835	0.670	0.743
NP	0.794	0.826	0.809	0.753	0.505	0.604	0.881	0.439	0.586
KB	0.659	0.440	0.528	0.652	0.687	0.669	0.756	0.649	0.698
ALL	0.818	<u>0.857</u>	0.833	0.730	<u>0.812</u>	0.769	<u>0.844</u>	0.777	0.808
識別器	BO			NP			KB		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
KO	0.801	0.882	0.840	0.638	0.690	0.663	0.676	0.753	0.713
YC	0.793	0.867	0.829	0.620	0.850	0.717	0.673	0.837	0.746
YB	0.763	0.957	0.849	0.606	0.840	0.704	0.642	0.957	0.769
BO	0.776	0.913	0.828	0.651	0.764	0.703	0.666	0.883	0.759
NP	0.868	0.753	0.806	0.778	0.764	0.770	0.708	0.557	0.623
KB	0.758	0.652	0.701	0.614	0.489	0.544	0.740	0.787	0.757
ALL	<u>0.828</u>	0.864	<u>0.840</u>	0.731	<u>0.798</u>	0.763	0.696	<u>0.862</u>	0.769

4.3 アンサンブル学習による評価

本研究の提案手法であるアンサンブル学習手法としては、単純多数決 (7 つの識別器の多数決 (X=4) もしくは、一定数 (X) 以上が意見と判断) の手法 (表 2) と、自身所属ドメインで学習した識別器に対する重みづけを行った手法 (表 3) を行った。重みづけ手法では、自身所属ドメインが意見と判定した場合には意見と判断し、意見と判断しなかったものの内、残りの X 件 (OTHER もしくは、OTHER+ALL) 以上が意見と判断した場合に意見と判断する。

表 2,3 では、各々の行の最大値を太字で、ALL 単独の学習結果よりも良いものには、下線を引いている。この結果、単純多数決では、精度重視のシステムができることを確認した。また、重みつき多数決で、ALL も一つのシステムとして加え、X=5 という設定の場合に、単純に ALL で学習した場合に比べ、最悪の場合でも、あまり、性能を落さずに、全体の性能が向上できることを確認した。

また、単純な多数決では、精度の向上ははかれるものの、再現率の低下が起こることを確認した。これに対し、重みつき多数決では、自身所属ドメインで作成した識別器が有用な場合に、その性能改善が期待できることを確認した。

4.4 考察

本論文で対象としてきた各ドメインのコーパスは、そのドメインに依存して、その意見表現の傾向が異なる。

ゆえに、各々のドメインによって学習した識別器の精度がある程度高い場合には、アンサンブル学習を用いる事で、自身所属ドメインに加えて他のドメインに特徴的な意見表現も捉えられるのではないかと考えた。

このことは、実験結果において単一ドメインでの学習と比較することで、自身所属ドメインの学習データからは意見と判断出来なかったものの中から、精度をあまり落とすことなく意見を発見でき、全体として F 値の向上が見られたと推測できる。

また、多数決の設定によって、精度重視のシステムを構築できる事が明らかになった。

この分析によって、複数ドメインコーパスを用いたアンサンブル学習による意見分析は、多岐に渡る意見表現を獲得しやすくなるという点で再現率が向上する一方で、複数の識別器の組合わせにより、精度の最小限を低く抑えることが出来ることを確認した。また、必要に応じて、精度重視の設定などが行えることも確認した。

ドメインに強く依存する意見表現の特徴を特定するという試み、本試みの正当性をさらに裏付ける試みは、本研究の今後の課題とする。

表 2: 単純多数決に基づくアンサンブル学習の意見判定実験結果

評価対象	X=4			X=5			X=6			X=7		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
KO	0.763	0.921	0.835	0.812	0.852	0.831	<u>0.846</u>	0.702	0.767	0.875	0.352	0.502
YC	0.727	0.794	0.759	<u>0.762</u>	0.686	0.722	<u>0.809</u>	0.534	0.643	0.868	0.331	0.479
YB	<u>0.868</u>	0.718	0.786	<u>0.901</u>	0.583	0.708	<u>0.927</u>	0.431	0.588	0.941	0.237	0.379
BO	0.704	0.803	0.750	0.718	0.713	0.716	0.734	0.602	0.661	0.735	0.358	0.482
NP	0.690	0.809	0.745	<u>0.740</u>	0.719	0.729	<u>0.791</u>	0.581	0.670	0.840	0.303	0.445
KB	0.626	0.799	0.702	0.643	0.697	0.669	0.656	0.541	0.593	0.664	0.322	0.434

表 3: 重みつき多数決に基づくアンサンブル学習の意見判定実験結果

評価対象	X=4 (ALL なし)			X=5 (ALL なし)			X=5 (ALL あり)			X=6 (ALL あり)		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
KO	0.727	<u>0.952</u>	0.825	0.746	<u>0.876</u>	0.806	0.727	0.966	0.830	0.746	<u>0.941</u>	0.832
YC	0.711	<u>0.835</u>	0.768	0.737	0.793	0.764	0.703	0.858	0.773	0.727	<u>0.825</u>	0.773
YB	0.834	<u>0.858</u>	<u>0.845</u>	0.834	<u>0.855</u>	<u>0.844</u>	0.832	0.863	0.847	0.834	<u>0.857</u>	<u>0.845</u>
BO	0.696	0.789	0.739	0.694	0.771	0.730	0.695	0.799	0.743	0.695	0.785	0.737
NP	<u>0.748</u>	<u>0.803</u>	<u>0.775</u>	0.776	0.767	<u>0.772</u>	0.731	0.837	0.781	<u>0.761</u>	0.792	<u>0.776</u>
KB	0.625	0.810	0.705	0.627	0.725	0.672	0.619	0.833	0.710	0.26	0.782	0.696

なお、各ドメインのラベル付きコーパスのラベル付けの方針が一貫していないことで、各々の「意見性を持つ文」の定義が異っており、少なからずシステムに何らかの影響を及ぼしていると考えられる。

また、今回の実験において、NICT の意見抽出ツールを利用したが、実際の利用方法とは若干異なる部分があった。このツールは、コーパスでの学習の際に、文中の意見対象部分を正確に特定してシステムに伝える必要があるが、今回は意見文一文全てを意見対象部分とした。この事が要因となり、精度が下がった事も予想される。これらを詳細に定義することによる更なる性能向上も今後の課題である。

5 おわりに

本研究では、複数ドメインの意見分析コーパスを用いたアンサンブル学習に基づき、単一ドメインでの学習よりも高性能な意見分析システムを構築することを目指した。そのために、複数ドメインコーパスとシステムを組み合わせる事で意見分析システムを多数構築し、アンサンブル学習を行う事に着手した。Blog, 新聞記事, レビュー記事等の様々なドメインを学習コーパス, 評価対象とした実験から、アンサンブル学習は、意見分析を行う上で、高い再現率を実現できる十分なカバレッジを得ている事を確認した。今後は、極性や意見発言者等の詳細なドメイン情報を利用する事で性能向上を目指す。

謝辞

本研究の一部は、科研費基盤研究 (B) 21300029 により行われた。また、本実験では、特定領域研究「日本語コーパス」による意見コーパス NTCIR のコーパス、ALAGIN Forum のコーパス、意見抽出ツールを利用した。これらのコーパスの作成者に感謝の意を示す。

参考文献

- [1] 関洋平. Bccwj を利用した意見分析コーパスの構築について. 「現代日本語書き言葉均衡コーパス」完成記念講演会予稿集, pp. 125–130, 2011.
- [2] 上田修功. アンサンブル学習の新展開. 電子情報通信学会技術研究報告. NC, ニューロコンピューティング, pp. 31–36, 2002.
- [3] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Inf. Sci.*, Vol. 181, pp. 1138–1152, March 2011.
- [4] Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. Overview of multilingual opinion analysis task at ntcir-8. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, And Cross-Lingual Information Access*, pp. 209–220, 2010.