BBS 対話における発話間の応答関係の判定

荒牧英治 †　　阿辺川武 †　　村上陽平 ‡　　灘本明代 ‡
† 東京大学 ‡NICT
aramaki@hcc.h.u-tokyo.ac.jp abekawa@p.u-tokyo.ac.jp
{nadamoto, yohei}@nict.go.jp

1　はじめに

インターネットの急速な発展とともに，利用できるテキストの量は日々増加している．なかでも特に爆発的な増加をみせるテキストはブログや掲示板（以降，BBS）上のテキストであり，これらを有効活用できれば重要な知識源とみなせるる可能性がある．

表 1 に BBS の書き込み（以降，**コメント**）の例を示す．コメント (1) による質問に，コメント (3) が返信しており，さらに (5) が補足している．これらのコメント群を合わせると,「小さくて軽い *MP3* プレイヤー」としては「*iriver* の *N12*」があるが，それは「生産中止」であることが分かる．このように，BBS 上には複数の人間によって，集合知とも言える知識が蓄積されているが，表中に見られるように省略が多く，さらにその省略を推定しようにも対応するコメント間にギャップ（別のコメントの割り込み）がみられるため，その扱いが困難である．このギャップは BBS において頻繁に見られる現象であり，図 1 が示すように，ほぼ半数の対応するコメント間にギャップが存在している．この問題に対応するため，本研究では 2 つのコメント間が対応しているかどうかを識別する問題に取り組む．この問題は BBS を解析するために必須のものであると同時に，新しい会話分析タスクとしても意義のあるものだと考える．

この問題に対して，3 つの異なる手がかりが存在する，と我々は仮定する．まず，両コメントの内容的な関連性である．例えば，コメント (1) と (3) では「小さくて」「相当小さい」といった関連のある語が含まれており，本稿ではこれを**内容的関連性**と呼ぶ．内容的関連性は，文同士の類似度と近い概念であり，我々は web 上での単語の共起にもとづく類似度 [2] で，これを計算する．

もう一つの手がかりは,「教えてください：はどうでしょうか」といった両コメントで呼応*した表現である．このようなコメント間で呼応する表現による手がかりを本稿では**機能的関連性**と呼ぶ．機能的関連性を捉えるために，我々は大量の BBS を収集し（17,300,000 コメント），そのうち高い確信度で対応しているとなっているものを用いて，呼応表現を抽出する．

表 1: BBS 書き込みの例.

(1) 小さくて軽い MP3 プレイヤーを教えてください．やっぱりシャッフルが一番なんでしょうか？
(2) バッテリーがまだ残っているのに，ipod が止まってしまいます．
(3) iriver の N12 はどうでしょうか．相当小さい上，カラーですよ．
(4) バッテリー表示は近似なので，バッテリー切れでもちょっと残っていたりしますよ．
(5) N12 はもう生産中止ですよ．

最後の手がかりは，注目しているコメントの外部（コンテキスト）からくる情報である．例えば，コメント間の距離や時間差，過去に会話していたかどうか，など様々な情報がこれにあたる．これらを用いることで，精度の向上は見込めるが，本研究では会話を分析することを研究の目的においているため，コンテキストから得られる情報を用いなかった．

提案手法のポイントは次の 2 点である: (1) 対応するコメントの関連性を**内容的関連性**と**機能的関連性**という 2 つの指標と仮定し，それらを定式化する点，(2) 機能的関連性を計算するために，大量のコーパス（コメント対）を自動構築する点．

2　提案手法

まず，本研究の問題を定式化する：

*一般に**呼応**表現は「決して〜ない」といった文内で共起する語ペアを指すが，本稿ではコメント間での呼応に対して用いる．

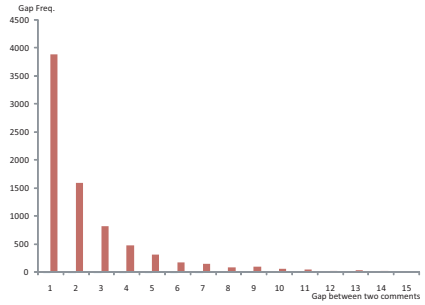

図 1: 対応するコメント間のギャップの長さとその頻度．ギャップの距離は 5 章で述べる実験のテストセットから得た．

入力: ある BBS 内の 2 つのコメント（i 番目のコメントと j 番目のコメント $(j > i)$.

出力: True または False（もし 2 つのコメントが対応しているならば True，そうでないなら False）.

以降は，簡便のため i 番目のコメントを P，j 番目のコメントを Q と表記する．この問題を解くために，内容的関連性 (2.1 節)，機能的関連性 (2.2 節) を素性として，SVM で学習させる (2.3 節).

2.1 内容的関連性

本研究では web 上での単語の共起頻度にもとづいた単語類似度 ($WEBPMI$) を利用し，内容的関連性 ($sim_r(P,Q)$) を定義する．

$$sim_r(P,Q) = \sum_{p \in W_p} \max_{q \in W_q} WEBPMI(p,q), \qquad (1)$$

ここで，P は W_p に含まれる語の集合，W_q は Q に含まれる語の集合，$WEBPMI$ は次の式によって定義される：

$$\mathrm{WEBPMI}(p,q) = \begin{cases} 0 & \text{if } H(p \cap q) \le c, \\ \log \frac{\frac{H(p \cap q)}{N}}{\frac{H(p)}{N} \frac{H(q)}{N}} & \text{otherwise}, \end{cases} \qquad (2)$$

ここで，$H(p)$ はクエリ「p」によって検索エンジンが返す文書数．$H(q)$ はクエリ「q」によって検索エンジンが返す文書数．$H(p \cap q)$ は「$p + q$」によって検索エンジンが返す文書数．N は検索エンジンが持つ文書数である．小さな値によるノイズを避けるため，閾値 c よりも小さいものは棄却した†.

2.2 機能的関連性

機能的関連性 (sim_d) を計算するために，我々は Corresponding-PMI ($CPMI$) を提案する．これは $WEBPMI$ と似ているが，以下の 2 点で異なる：

†先行研究 [2] にもとづいて $c = 5$ とした.

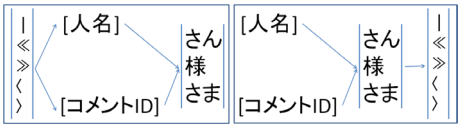

図 2: 抽出パターン.

(1) $WEBPMI$ は web での共起頻度を用いるが，$CPMI$ は対応するコメント (P,Q) 間での共起頻度を用いる.

(2) $WEBPMI$ は一語しか扱わないが，$CPMI$ は語群（n-gram）を扱う $(n = 1..3)$.

$$sim_d(P,Q) = \sum_{p \in N_P} \max \sum_{q \in N_Q} CPMI(p,q), \qquad (3)$$

ここで，N_P は，P に含まれる n-gram の集合，N_Q は Q に含まれる n-gram の集合，$CPMI$ は次式によって定義される：

$$CPMI(p,q) = \begin{cases} 0 & \text{if } H_c(p \cap q) \le c, \\ \log \frac{\frac{H_c(p \cap q)}{M}}{\frac{H_a(p)}{M} \frac{H_b(q)}{M}} & \text{otherwise}, \end{cases} \qquad (4)$$

ここで，$H_a(p)$ は n-gram$p$ の P における出現数．$H_b(q)$ は n-gram$q$ の Q における出現数．$H_c(p \cap q)$ は n-gram 対 $(p:q)$ の共起頻度数である.

2.3 SVM による学習

内容的関連性と機能的関連性により 2 つの値を得ることができる，これらの値に加えて，P と Q に含まれる語彙を素性として，SVM に学習を行う.

学習に必要なトレーニングデータは，正例は次章に述べる手法で得るコメントペア $(P:Q)$ を用いた．負例は $(P:Q)$ の応答コメント (Q) をそれより前のコメントペアの応答コメント (Q') と無作為に入れ替えることによって作成した.

3 Web からのコメントペアの自動抽出

本章では機能的関連性と SVM 学習に必要とされるコメントペアを Web から自動抽出する手法を述べる.

まず，130,000 の BBS をクロールし，17,300,000 コメントを収集した．これらのコメント間の対応関係は多くの場合分からないが,「＞ 30」や「太郎さん＞」以下のように，発言者 ID や発言者名が示されている場合は対応関係を得ることができる．そこで図 2 に示されるようなパターンを用いて，対応しているコメントを抽出した.

ここで，長い（文字数が多い）コメントは，複数のコメントへのレスポンスや，長い引用など，複雑な現

象を含んでいる場合が多く，本研究の問題設定（対応する/しないの二値を出力）に沿わない場合がある．そこで 100 文字以上の長いコメントは棄却した．この結果，121,699 コメントペア（全コメント量の 1.4%）を得ることができた．

4 関連研究

本研究のように応答かどうかを判別するという問題設定は会話研究では珍しく，同様の研究は，我々の知るかぎりでは，徳永ら [10] によるチャットの発話の応答関係判別のみである．徳永らは人手による辞書を用いて発話のタイプ（アクト）を決定し，それを素性の一部としていた．一方，本研究はタイプといった恣意的な区別を導入しないかわりに，呼応表現を学習するというアプローチをとっている点で新規性を持つ．

他の多くの会話／談話の先行研究は，DAMSL[4] や RST-DT[3] や discourse graph-bank[9] といった少量ではあるが注意深くアノテートされたコーパスにもとづいて研究されてきた．本研究で扱うデータは，それらの先行研究で用いられたコーパスと比較して，より粗い単位（コメント単位）で構成され，さらに 1 種類の関係（対応しているかどうか）しか扱っていない．以上のような欠点があるものの，本研究はかつてない大きな対話データを扱っており，これが統計的手法（PMI）の導入を可能としている．

もう一つの関連分野は 同じトピックを持つ文章を特定するタスクである Topic Detection and Tracking (TDT) [1] である．多くの TDT の手法はクラスタリング手法 [5, 6, 8] をベースとしているが，コメント毎にトピックが変わるような BBS に対しては効果を期待できない．

5 実験設定

テストセットはトレーニングセットからコメントペアを無作為抽出することによって得た．実験では次の 2 つのテストセットを用いた．

SMALL-SET: 140 コメントペア．人間も参加する小規模なデータ．

LARGE-SET: 8400 コメントペア．

また，次の手法を比較した．

human-A, B, and C: 人間（ 3 人）による判定結果．

表 2: SMALL-SET の結果．

	Accuracy	Precision	Recall	$F_{\beta=1}$
human-A	79.28	83.33	75.34	79.13
human-B	75.71	78.26	73.97	76.05
human-C	70.71	71.62	72.6	72.10
$Overlap$	61.42	58.71	87.67	70.32
sim_r	61.42	72.09	42.46	53.44
sim_d	65.71	66.23	69.86	67.99
SVM	63.28	64.44	79.45	72.10

表 3: 人間とシステム間の一致率と Kappa 値．

	Human-B	Human-C	$Overlap$	sim_r	sim_d
Human-A	0.78 (0.56)⊕	0.74 (0.49)⊕	0.52 (0.08)⊖	0.60 (0.20)	0.65 (0.28)
Human-B		0.73 (0.47)⊕	0.54 (0.09)⊖	0.60 (0.21)	0.62 (0.25)
Human-C			0.59 (0.15)⊖	0.52 (0.05)⊖	0.62 (0.25)
$Overlap$				0.63 (0.21)	0.45 (0.13)⊖
sim_r					0.56 ((0.16)⊖

* 括弧内の数字は κ 値を示す．⊖ は κ 値の解釈が「slight」であることを示す．⊕ は κ 値の解釈が「moderate」であることを示す．

Overlap: 語の一致率による精度（ベースライン）．語の一致率が閾値より高ければ TRUE を出力し; そうでなければ FALSE を出力する（以降の sim_d と sim_r も同様に閾値によって判定する）．

sim_r: sim_r のみを使う．

sim_d: sim_d のみを使う．

SVM: 提案手法．

また，$WEBPMI$ の計算にあたっては正確なドキュメント数を得るために TSUBAKI[7] を用いた．

5.1 結果

表 2 に SMALL-SET での各手法の精度を示す．*Overlap*,sim_r と sim_d の精度は閾値に依存するため，様々な閾値で実験し，もっとも高い Accuracy を示した値を用いた．

人間の精度の上限

人間の精度はたかだか 70–79%しかなく，本タスクの難しさを示している．これは短い返答（「そう思います」や「ありがとうございます」など）による False Positive，専門的すぎて評価者では判断できない返答による false negative が原因である．

表 4: 高い $CPHI$ を持つ呼応表現の例.

n-gram in P	n-gram in Q	$CPHI$
行き ます	お 待ち して	8.43
どこ に ある	あり ます	8.37
はじめ まして	はじめ まして	7.86
教えて ください	と 思い ますよ	7.62
いかがでしょう	早速	7.47
でき ます	やって み	7.38
と 思い ます	あり が とう	7.12
か な ？	多分	6.93
あり が とう	いえいえ	6.80
私 は	私 も	6.73
か ？	と 思い ます	6.72

このような限界はあるものの人間同士の一致度は表 3 に示されるように高く（κ value = moderate），これらの限界は評価者間で一致しているものと考えられる．以上から，本タスクは難しいいものの不合理ではないと言える．

2 つの関連性の独立性

表 2 に示されるように，sim_d は sim_r や $Overlap$ よりも高い精度を示した．より重要なことは，$Overlap$ と sim_r はわずかに相関しているが（fair agreement;$(0.2 < \kappa < 0.4)$），これらの両方とも sim_d に対してわずかな相関しかみせていないことである (slight agreement; $\kappa < 0.2$)．これらの結果から，sim_r (or $Overlap$) と sim_r は互いに独立であることを示し，2 つの指標の妥当性を示している．表 4 に高い $CPHI$ を持つ呼応表現の例を示す．表にみられるように，これらの関連性を内容的関連性でとらえることは困難だと想像される．

SVM の精度

表 2 に示されるように,SVM は，sim_d や sim_r の中間の accuracy を示し，両指標をうまく使い分けることができなかった．効果的な素性をデザインし，両指標のよい部分を組み合わせることが今後の課題である．

コーパスサイズと精度

表 3 の対応ペアを計算するために用いるコメントの数と sim_d の関係を示す．表において精度はまだ飽和しておらず，今後，より大規模なデータを用いることで，さらに高い精度が期待される．

6 まとめ

本論文は，内容的関連性と機能的関連性という 2 つの指標を用いて，発言が対応しているかどうかを判定する手法を提案した．実験の結果，2 つの関連性はそれぞれ独立に作用し，提案する指標の妥当性を示すことができた．本研究により，今後，かつてない大規模な統計的な対話研究が可能となることを信じている．

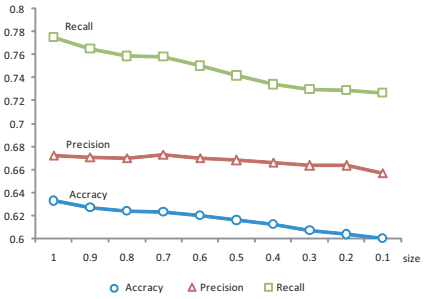

図 3: トレーニングセットのサイズと sim_d の精度（LARGE-SET）.

参考文献

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[2] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of 16th International World Wide Web Conference (WWW 2007)*, pp. 757–766, 2007.

[3] Carlson, D. Marcu, and M. E. Okurowski. Rst discourse treebank, 2002.

[4] Mark G. Core and James F. Allen. Coding dialogues with the DAMSL annotation scheme. In *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pp. 28–35. American Association for Artificial Intelligence, 1997.

[5] K. Rajaraman and A. Tan. Topic detection, tracking and trend analysis using self-organizing neural networks. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001)*, pp. 102–107, 2001.

[6] J. M. Schultz and M. Liberman. Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of DARPA Broadcast News Workshop*, pp. 189–192, 1999.

[7] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology (to appear). In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.

[8] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In *Proceedings of DARPA Broadcast News Workshop*, pp. 193–198, 1999.

[9] Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, Vol. 31, No. 2, pp. 249–287, 2005.

[10] 徳永泰浩, 乾健太郎, 松本裕治. チャット対話における発話間の継続関係と応答関係の同定. 自然言語処理, Vol. 12, No. 1, pp. 79–105, 2005.