

係り受け構造を利用した発言の賛否の分類

Classifying Pros and Cons of Utterances Using their Dependency Structures

森田 一*¹ 池田 大介*¹ 奥村 学*²
MORITA Hajime IKEDA Daisuke OKUMURA Manabu

*¹東京工業大学 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

*²東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

Conversations and discussions can be easily and efficiently referred in the form of their summaries. Since their discourse has rather complex structure, simply extracting important utterances would not yield useful summaries. Therefore, we think it is inevitable to take into account the relationship between utterances and the distinction of utterances whether they are an opinion or not in utterance extraction. In this work, we present a method to judge whether a meeting utterance is pros or cons of the previous utterance. Although Support Vector Machines with bag-of-words and/or N-gram features are commonly used in text classification task, we transform utterances into tree structures such as dependency trees and use syntactic patterns for learning, since it is difficult to obtain enough information from short meeting utterances for the classification.

1. はじめに

ネットワークを通じたコミュニケーション手段としてIM(Instant Messenger)が広く普及するようになった結果、大量のテキストデータ化された会話のログが残るようになった。会話のログは個人の興味や意見が色濃く現れ、かつ非常にリアルタイム性にも優れている。しかし、日々会話ログが蓄積される一方、その膨大さに埋もれてしまい、後々会話中の有用な情報を活用することが難しくなっている。この膨大なログを有効に使うためには、会話で扱われる情報を再利用可能な形に自動的にまとめる必要がある。いずれ携帯電話による会話内容もテキストデータ化されログとして残るようになれば、膨大なログを扱えるようにする必要性は一層増すことになると考えられる。

会話に含まれる情報を効率的に参照するには、会話で話された内容の要約を作り一覧性を高める方法が考えられる。しかし、会話の内容は一般の文章に比べ複雑な構造を持ち、主語が省略されるため、発言を抜き出すだけでは何について述べている発言なのか分からなくなってしまうことが予想される。そのため、自然で意味の通る要約を作るためには、各発言同士の繋がり、参照関係を分析し、その関係を考慮した上で要約に組み込むべき発言を選び、要約文を生成するべきである。

このような理由から、発言同士の参照関係を同定する前段階として、発言同士の関係のなかでも重要な位置を占め、また会話の要約を作る際には重要なポイントとなりやすい、賛成と反対の発言の抽出を行った。

賛成・反対を表す発言の推定は、賛否を表す発言とそうで無い発言を分類する問題として考えられるため、分類問題の手法を利用して推定を行った。

2. 関連研究

Galley[2] は会議中の発言の応答ペアの同定を、CRF(Conditional Random Field)を用いて行ってお

り、また、Thomas et al.[9] は会議中の人物単位での賛成・反対の同定を行っている。

発言単位での賛成反対を分類する先行研究として、Hillard et al.[4] や Galley et al.[3] などが上げられる。Hillard et al.[4] は ICSI による議事録のサブセット [6] の発言の賛否の同定を行っている。分類器には決定木を用いており、素性としては、発言の語数、発言に含まれるポジティブなキーワード数、発言に含まれるネガティブなキーワード数、最初の語のクラス (ポジティブ、ネガティブ、バックワード、ディスコース、マーカー、その他)、それらに加えて、ポーズや間隔、基本周波数といった音韻の素性を利用している。この結果 78% の正解率を得ている。なお、キーワードは 5 回以上共起し、60% 以上の確率を閾値に選択している。また、書き起こしテキストだけでなく、ASR によるテキストについても 71% の正解率を得ている。

また、Galley et al.[3] も同様に MRDA(The ICSI meeting recorder dialog act corpus)[8] の発言の賛否の同定を行っている。Bayesian Network を用いて分類しており、素性としては、前後の発言の話者と同一話者かどうか、前後の発言が同一の話者に対して発言されているか、発言の時間的長さ、他の話者との時間的重なり、発言間の空白、発話率、発言の語数、発言中の内容語の数、Agree, Disagree, backword, other の 4 つのクラスへの所属確率、発言の最初と最後の語、ポジティブな形容詞、ネガティブな形容詞、手がかりとなる句や賛成反対の手がかりとなる語句の数。また、文脈素性として、クラスラベル間の依存関係を加えている。結果、86.9% の正解率を得ている。

これらの先行研究と本研究とは、会議中の発言からその賛否を分類するという点では一致しているが、発言そのものの情報を、単語もしくはその集合としてではなく順序付きラベル木として扱う点が異なる。本研究では発言を木として扱う事で、発言の構文的な情報を考慮し分類する手法を提案する。

3. 提案手法

3.1 手法の概要

本手法では図 1 に示す手順で発言を分類する。

テキストの分類は bag-of-words や N-gram を素性とした、SVM(Support Vector Machines) をはじめとする分類器

連絡先: 森田 一, 東京工業大学 総合理工学研究科,
morita@lr.pi.titech.ac.jp

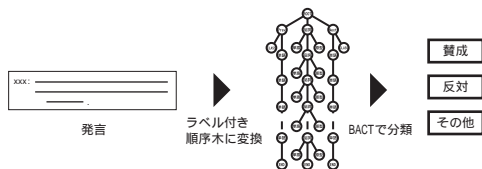


図 1: 手法の概要

を用いる手法が一般的である, それに対し本手法ではまず, 個々の発言を発言中の単語をノードとしたラベル付き順序木に変換する. ラベル付き順序木とは木の各ノードにラベルが付き, 兄弟ノード間の順序関係が保たれるような木構造である.

発言の木への変換をおこなった後で, 各ラベル木をラベル付き順序木の分類器である BACT(Boosting Algorithm for classification of Trees)*1を用いて賛成, 反対, その他の3クラスへと分類する.

本手法が発言をベクトルとしてではなく, 順序付きラベル木として分類するのは, 分類の対象とする個々の発言が短く省略も多いため, BOW のような構造を考慮しない素性では分類に十分な情報が得られないためである.

短い発言から分類に十分な情報を得て, 構文的な情報を分類に取り入れる目的で, 順序付きラベル木として発言の分類を行う.

3.2 BACT

BACT は工藤ら [5] による Boosting を用いたラベル付き順序木の分類器である. Boosting は弱学習器と呼ばれる学習器を多数組み合わせ, 精度の良い学習器を構成する手法で, 学習器に対して, 逐次的に新しい弱学習器を追加して改良を行う. その際, 現在の学習器が誤った事例の重みを大きく, 正しく分類出来ている事例の重みを小さく設定する事で, その誤った事例を正しく分類可能な弱分類器を追加するようにする.

また, 各弱学習器を学習時の分類精度で重み付けし, 分類時には各弱学習器の重みを足しあわせる事により, 分類スコアを決定する. BACT では, Boosting のための弱学習器として DecisionStumps を用いる. Decision Stumps は, 一つの素性の有無だけによって分類を行う単純な分類アルゴリズムであり, ここでは, 部分木の有無に基づいて分類を行う Decision Stumps を考える. つまりこの場合には, ある部分木の有無で分岐する, 深さ 1 の決定木とみなすこともできる. 例えば, ある単語の有無や特定の係り受けのペアが存在するか, などが分類の素性として用いられることになる.

工藤ら [5] による BACT を用いた文のモダリティの分類実験で, 係り受けと N-gram についてそれぞれ F 値の平均で 0.866, 0.859 という高い値を出し, 構造を考慮して分類を行う有効性を示す結果が出ている.

3.3 順序木の作り方

発言を木構造に変換する方法として, 以下の 2 通りの方法を提案する. どの木構造も, 木の根である事を示す ROOT ノードを親として, 発言からそれぞれの木の構成法で作られた木を木構造に加える.

- N-gram 木

発言を単語単位で分割し, 語順通りに一列に並べた形の木構造を作る. この木の各部分木が単語の N-gram となっている. 発言の末尾には発言の末尾である事を示す END

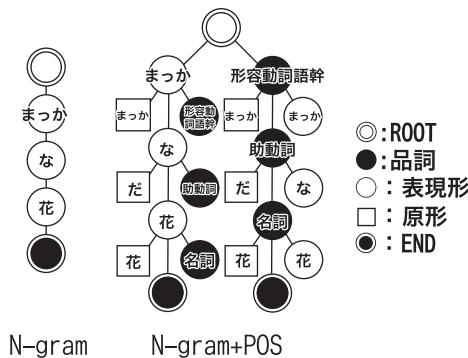


図 2: N-gram 木イメージ図

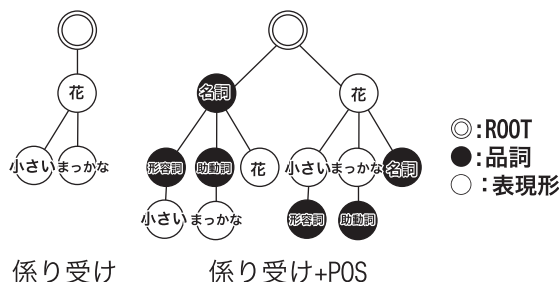


図 3: 係り受け木イメージ図

ノードを追加する. N-gram 木のイメージ図を図 2 (左) に示す.

- 係り受け木

発言に対して構文解析を行い, その結果を係り受け木に変換する. この木の各部分木は発言中の係り受け関係を示している. 構文解析には Charniak Parser[1], 係り受け木への変換には ptbconv[10] を使用した. 係り受け木のイメージ図を図 3 (左) に示す.

以上二つの木構造を元に, 発言中の単語以外の情報を木に含める拡張を行う.

- 共通のバイナリ素性

発言内容のみではなく, 分類に利用可能な情報を与えるため, 前発言, 前々発言と話者が同じかどうか, 指示語の有無, 発言長, 前発言との時間的距離をそれぞれ素性として加えた. これらの素性は, ROOT ノードの子ノードとして木構造に含める.

- 品詞

各単語を表すノードにその子ノードとして品詞をラベルとしたノードを加えた木と, 各ノードのラベルを品詞に置き換え, その子ノードとして単語をラベルとしたノードを加えた木を並べ, 共通の ROOT ノードを親にもつ一つの木にまとめる. 品詞拡張を加えた N-gram と係り受けそれぞれの木の概念図を図 2 (右) と図 3 (右) に示す. 品詞情報は発言を TreeTagger[7] にかける事によって得た.

- 前後の発言と動的素性

分類の対象とする発言だけでなく, 前後の発言も順序付きラベル木に組み込む. 前後の発言を木として組み入れ, 先

*1 <http://www.chasen.org/taku/software/bact/>

図 4: ICSI meeting recorder dialog act (MRDA) corpus

```
13377,134.16,Btr002-c6_0133770_0134160,A,
133.77+133.87+i|133.87+134.02+don't|134.02+134.16+know,s|no,
Btr002-c6_fe069,s`no,,,,,
133.79,135.12,Btr002-c7_0133790_0135120,A,
133.79+134.05+maybe|134.05+134.4+very|134.4+134.66+very|134.66+135.12+soft,
s`ba,Btr002-c7_fe008,s`ba,,,,,
135.12,137,Btr002-c7_0135120_0137000,A,
135.12+135.41+maybe|135.41+135.58+with|135.7+135.86+the|
135.96+136.5+microphone|136.5+137+noise,
s`cs,Btr002-c7_fe008,s\symbol{50E}cs,,,,,
```

頭に以前の発言であれば NEXT, 以後であれば PREV それぞれのラベルをつけ, PREV にはそのクラスラベルを動的素性として加える. クラスラベルは訓練時には正解データを, 実際分類時にはその分類結果を用いる.

4. 実験

4.1 データ

実験のためのデータとして, ICSI Meeting Corpus[8] の書き起こしテキストを用いた. ICSI Meeting Corpus は議事録の書き起こしテキストと音声データを含むコーパスで, 全発言に対して発話開始時間と終了時間, 発言 ID, 各語の発話開始時刻と終了時刻を含む発話内容, 意図タグ, 会議名, 発話者, 応答ペア, ホットスポットがラベル付けされている. コーパス中の発言の例は図 4 を参照して欲しい. 意図タグは, 発話の意図を全 4 種類のタグとして発話に対して付与したもので, タグの例としては, Accept や Reject, TopicChange などがある. このうち, 意図タグとして, 賛成・反対と見なせるタグを定め, それぞれの意図タグが付けられている発言に賛成・反対のタグをつけた. 賛成と見なしたタグは, Acknowledge-answer(“bk”), Accept(“aa”), Accept-part(“aap”), Maybe(“am”) の 4 つ, 反対と見なしたタグは, Reject(“ar”), Reject-part(“arp”) の 2 つである. この定義で発言に対してタグを付けた所, 全 112027 発言中, 17255 発言に賛成, 1192 発言に反対のタグが付与された.

4.2 設定

コーパスをトレーニングデータとテストデータに 9:1 の割合で分割し, トレーニングデータで学習, テストデータ中の発言の賛否あるいはそれ以外を分類する実験を行った. 比較対象のベースラインとしては, 一般的にテキスト分類に用いられる SVM の多値分類への拡張として Yamcha*2 を用いて分類を行った.

N-gram 木を用いた提案手法とベースラインの Yamcha では, “前後を含む発言の表現形と品詞と動的素性+共通バイナリ素性”について, 賛成・反対・その他の 3 値での分類実験を行った. 係り受け木を用いた手法では, “前後を含む発言の表現形と品詞と動的素性+共通バイナリ素性”に加えて, 品詞情報を除いた“前後を含む発言の表現型と動的素性+共通バイナリ素性”についても分類実験を行った.

N-gram 木と係り受け木ではそれぞれ各拡張を組み合わせ, Yamcha では前後を含む発言の先頭と末尾の 3 語分の表現形と品詞の対と動的素性, 共通のバイナリ素性を素性として用い, 分類を行った.

4.3 実験結果

ラベルごとの実験結果を表 1 に示す. 実験の評価は各ラベルごとの F 値とその平均で行った.

N-gram 木を拡張した場合が最も平均した F 値が高く, ベースラインと比較して F 値が 0.034 向上した. 係り受け木を拡

*2 <http://www.chasen.org/taku/software/yamcha/>

張した場合がその次に F 値が高く, ベースラインを 0.032 向上した. 各ラベルごとに F 値を比べると, 賛成は品詞を用いない係り受け木が, 反対では N-gram 木が最も高い F 値を示した.

5. 考察

N-gram 木, 係り受け木ともに, 木構造を考慮しないベースラインを上回り, 会議中の発言の分類に関して, 構造を考慮する有効性を示す事ができた.

構造を考慮して学習が行われている例として, N-gram に品詞を加えた実験では,

(RB(VVN(trained)))

(PP(we)(MD(will)(VV)))

などの部分木を素性として学習出来ている. これは, RB trained や, we will VV のような単語と品詞の組み合わせを学習出来ているということで, 構文的パターンが学習されていることを示す. 係り受け木での実験では,

((not)anyway)

などの係り受けを素性として学習出来ている. 係り受けを学習する事で, 単語同士の文章内での位置が離れていてもその単語間に関係があれば, その単語のペアを学習し, 分類のための素性として利用できる.

一方, N-gram 木と係り受け木を比較すると, 係り受け木が N-gram よりもやや精度で劣る. これは, 口語表現が多く構文解析が難しい, 発言に対して構文解析を行い係り受け木を作ったために, 係り受け木を作る段階での精度が低いという可能性が高い. また, 係り受け木に対して品詞を加えることで, 逆に精度が下がる結果になったが, これは係り受け木の精度の問題とともに, 係り受け木では品詞を考慮せずとも語同士の関係は構造として考慮に入れられているため, かえって品詞を入れる事で分類の対象となる木の肥大化を招き, 学習の妨げとなってしまったのではないかと考えられる.

6. まとめ

本研究では, 会話文の要約を作成する事を目標とし, 発言の賛成と反対の同定を行った.

情報量の乏しい短い発言を分類するために, 発言からラベル付き順序木を作成し, その木を分類する手法を提案し, 実験によりその有効性を確認した.

本研究では, N-gram 木と係り受け木それぞれに対して品詞と共通のバイナリ素性, 動的素性の 3 つの拡張を加え実験を行い, さらに係り受け木に対しては共通のバイナリ素性と動的素性のみを加えた実験を行った.

実験では, Yamcha と BACT を比較し, 発言のもつ構造を考慮した木構造として扱うことの優位性と, 本研究の手法により構文パターンが学習されている事, BACT においても従来用いられていた各種バイナリ素性が有効である事を確認した. 動的素性を加えて前後の発言とその分類結果を用いて分類を行い, 木構造の形で前後の素性を扱うことが有効である事を確認した. 評価実験の結果, 3 値の分類タスクにおいて F 値 0.621 を得, 本手法を用いない場合に比べて 0.034 ポイントの向上が見られ, 本手法の有効性を確認する事ができた.

7. 今後の課題

ある発言が何らかの賛意や否定を示している場合, それは会話の文脈と無関係に現れるわけではなく, 会話中の他の発言に対しての意思の表明である可能性が非常に高い. このように,

表 1: 実験結果: ラベルごとの F 値

	ベースライン (SVM)	N-gram 木	係り受け木	係り受け木 (品詞無し)
賛成	0.587	0.588	0.562	0.591
反対	0.587	0.653	0.580	0.636
平均	0.587	0.621	0.571	0.619

発言同士の関係を推定することにおいて、ある発言が賛否のいずれかを表しているかどうかは有用であると考えられる。また逆に、発言同士の関連も発言が賛否を示しているかどうかを推定する役に立つと考えられる。発言同士の関係と発言の意図の推定を合わせて解く事により、互いの精度の向上を計りたいと考えている。

参考文献

- [1] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the first conference on NAACL*, pp. 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [2] Michel Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on EMNLP*, pp. 364–372, Sydney, Australia, July 2006. Association for Computational Linguistics.
- [3] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 669–676, Barcelona, Spain, July 2004.
- [4] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *NAACL '03: Proceedings of the 2003 Conference of the NAACL on Human Language Technology*, pp. 34–36, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
- [5] Taku KUDO and Yuji MATSUMOTO. A boosting algorithm for classification of semi-structured text. *IEICE technical report. Artificial intelligence and knowledge-based processing*, Vol. 103, No. 726, pp. 27–32, 20040310.
- [6] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proceedings of the First International Conference on Human Language Technology Research*, pp. 00–00, 2001.
- [7] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- [8] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The icsi meeting recorder dialog act (mrda) corpus. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIG-dial Workshop on Discourse and Dialogue*, pp. 97–100, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.
- [9] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pp. 327–335, 2006.
- [10] 山田寛康, 松本裕治. Support vector machine を用いた決定性上昇型依存構造解析 (自然言語). *情報処理学会論文誌*, Vol. 45, No. 10, pp. 2416–2427, 20041015.