1R-01

議事録からの課題抽出と能動学習による精度向上

森田 尚也† 大平 茂輝† 長尾 確†
名古屋大学大学院情報科学研究科†

1 はじめに

議論活動は，人間社会において集団の様々な思考を整理し，問題解決や方針決定をするための重要な人間活動である．そうした会議の議事録を見直すことは今後の活動を円滑に進めるための有効な手段である．活動を進める上で特に注目すべきなのは今後の課題となるような発言（以下，課題発言）であるが，その存在は議事録においてその他の多くの発言の中に埋もれてしまうことがある．

そこで本研究では，この課題発言を機械学習によって自動抽出し，長期に渡る運用に耐え得る抽出精度を維持していく手法を提案する．具体的には，課題発言の判別モデルを最大エントロピー法によって作成しそれを初期モデルとした後，新たに得られる議事録に対して能動学習を行い抽出精度の向上を図る．能動学習では機械学習に用いる教師信号付与のコストを削減することができるのも利点の一つである．本研究では判別精度向上，コスト削減の両面で貢献する手法を提案し，計算機実験を行うことでその有効性を実証する．

2 議論活動の意味構造化記録

我々の研究室では，Discussion Mining (DM)というシステムを用いて，定期的に行われる研究進捗発表のゼミを記録している．DM では，単なるテキストとして議事録を作成するに止まらず，それぞれの発言に対して議論構造における知的情報を付与しながら記録する．知的情報とは例えば，発言者が誰であるかといった属性や，発言間の依存関係等の情報である．

この知的情報の中でも特に，我々の DM 独自で定義しているものについて説明する．1 つ目はマーキング情報である．マーキングとは，会議において発表者が会議後に見直したいと判断した場合に付与する情報である．2 つ目は発言のタイプである．DM では，発言を「導入」と「継続」の 2 タイプに分類し，各発言間にどの発言を受けて行われたかを表すリンク情報を付与することで議論内容の構造化を行っている．導入発言は新しい話題の起点となる発言と定義し，継続発言は直前までの発言内容を受けた発言と定義する．

Extracting Tasks from Minutes and Improving Accuracy by Active Learning
†Naoya Morita †Shigeki Ohira †Katashi Nagao
†Graduate School of Information Science, Nagoya University

3 課題発言抽出のための機械学習モデル

本研究では，課題発言を自動抽出するための機械学習モデルとして，ある発言に関する諸情報を説明変数$\boldsymbol{x}$とし，その発言が課題か否かを人手で付与した 2 値ラベルを目的変数$y \in \{0,1\}$とした最大エントロピーモデルを考える[1]．説明変数について具体的には，発表者名，発言の開始時刻，発言者の属性（教員/学生），発言のタイプ（導入/継続），マーキングの有無，文字数，含まれる文の種類，含まれる形態素 unigram と bigram，発表者の応答があるか否かを用いる．このモデルによって各発言が課題発言である確率を求め，その確率値が 0.5 を超えるものを課題発言と判定する．

4 能動学習

4.1 課題発言抽出に対する能動学習

解析データが増える際，判別モデルは常に新しい方がよい．ただし今回の例のように，増えるデータには正解ラベルが付いておらず，ラベル付けに高いコストがかかる状況では，全てのデータに対してラベル付けをして学習を行うことは困難である．これに対して本研究では，限られたデータの中で最大限のモデル精度向上を図る能動学習[2]と呼ばれる機械学習手法を用いる．

能動学習では，得られた大量のラベル無しデータ$\boldsymbol{U}$の中から，最もモデル更新に寄与する可能性のあるデータをサンプリングする．その選出基準は様々に考案されているが，本研究ではそれらのうち，現モデルに対して判別が最も曖昧なデータを選出する Uncertainty Sampling と呼ばれる手法をもとに考える．課題発言抽出においては最大エントロピーモデルで各発言に対して課題発言である確率が求められるため，その判別が最も曖昧であるもの，つまり課題発言である確率値が 0.5 に最も近いものをサンプリングすることになる(式 1)．

$$\underset{\boldsymbol{x}_i \in \boldsymbol{U}}{\operatorname{argmin}} |P(y=1|\boldsymbol{x}_i) - 0.5| \qquad (1)$$

4.2 能動学習の局所解問題

能動学習は大域的な最適性を保証しない Greedy

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

な手法であるため，外れ値の影響を顕著に受けるという問題がある．これに対して本研究では，能動学習のサンプリング基準の検討と，初期モデルでの学習データ数の検討によって解決を図る．

サンプリング基準について，通常の Uncertainty Sampling では判別が最も曖昧なデータを選出するが，本研究ではこれに加えて次の式 2 によるサンプリングを考える．

$$\underset{x_i \in U}{\operatorname{argmin}} |P(y = 1|x_i) - 1.0| \qquad (2)$$

この式の意味するところは，課題発言である確率が最も高いものをサンプリングするということである．外れ値を含むデータにより学習されたモデルの判別境界は望ましくないものとなる可能性があるが，式 2 による基準では正例の特徴をいち早く捉え，かつ誤判別の場合のモデル更新に大きく寄与できると見込まれる．

初期モデルでの学習データ数について，外れ値の影響を緩和するために大域的なラベル分布を初期のうちに把握するというのも解決策の 1 つである．本研究では，少数のデータから学習を開始する Cold Start と，ある程度のデータを集めてから学習を開始する Warm Start という 2 つの観点で実験を行う．

5 計算機実験

DM により記録された議事録 42 件(発言にして 1,637 件)のデータに対して，Cold Start と Warm Start の 2 つの状況下で，次の 3 通りの手法による 10 分割交差検定の F 値を比較した．

- Active Learning： 1 回のゼミ毎に 10 件の発言を式 2 による能動学習によってサンプリングする
- Random Sampling：1 回のゼミ毎に 10 件の発言をランダムにサンプリングする
- Full Sampling：1 回のゼミ毎に全ての発言をサンプリングする

Cold Start では手法間での公平性を保つため初期モデルに用いる議事録を同一の 1 件とし，Warm Start では初期モデルに用いる議事録を 10 件とした．

Cold Start の状況において，式 2 による実験結果の図 1 では能動学習が外れ値の影響を緩和して他手法よりも高精度を維持できている．一方で，式 1 によるサンプリングでは外れ値の存在により能動学習によるモデルが局所解問題の影響を顕著に受け，他手法よりも低精度となる結果となった．

Warm Start の状況における図 2 では，Cold Start に比べると，学習データ数の増加によって外れ値の影響が緩和されることによる安定した F 値の推移が見られる．また，Cold Start と同様に式 1 よりも式 2 の方が高精度を維持できる結果となった．

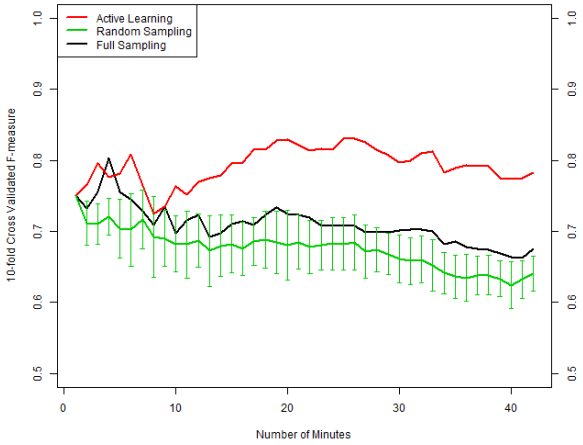

図 1：Cold Start での F 値比較

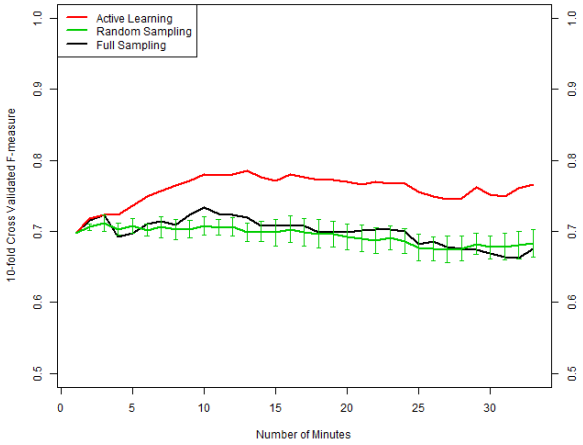

図 2：Warm Start での F 値比較

6 まとめ

議事録から課題を自動抽出する手法を提案し，能動学習によってラベル付けコストを削減しつつ抽出精度を向上させる手法を提案した．今後の課題としては，本稿で得られた知見をもとに，外れ値の含まれたデータに対して柔軟に対応できるサンプリングを適用することが挙げられる．

参考文献

[1] K. Nagao, K. Inoue, N. Morita and S. Matsubara. “Automatic Extraction of Task Statements from Structured Meeting Content.” *Proc. of the 7th International Conference on Knowledge Discovery and Information Retrieval*, 2015.

[2] B. Settles. “Active Learning Literature Survey.” Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.