D_038

ファンサイトのチャットデータを用いた番組メタデータ自動生成

Automatic Meta Data Generation from TV Viewers' Live Chat

有安　香子†　　八木　伸行†
Kyoko Ariyasu　Nobuyuki Yagi

1．背景・目的

大容量記録ディスクの普及により、大量の映像コンテンツを蓄積することが容易になってきた。この大量の映像コンテンツを様々な用途に有効活用するために、セグメントメタデータ（番組内容に沿ってシーンに分割し内容を表すメタデータ）の自動生成の研究が多数行われている。対象番組の特徴に応じて画像認識技術や音声認識技術を組み合わせ、メタデータを付与する研究[1]や、スポーツ番組などを対象として、ライブチャットデータを用いて、盛り上がり部分を検出し、評判分析によりその盛り上がりが「肯定的」か「否定的」かを解析する研究[2]など、様々な手法が提案されている。

ライブチャットとは、インターネットの発達に伴い自然派生的に出現した、新しい番組視聴形態のひとつであり、番組放送時に視聴者同士がリアルタイムでチャットを行いながら意見を交換し合うものである。情報を伝えることを主な目的としたニュースや情報番組などとは違い、ドラマ番組は視聴者が一般常識や内容に関する予備知識などを用いて物語を理解する事を前提として作られている。この様な前提の上に作られている番組のメタデータ生成に、視聴者の感想や意見を含めたライブチャットデータを用いることで、「何が映っているか」だけではなく、「何について何故盛り上がっているか」という視聴者の観点から見た番組内容を表す番組メタデータを自動生成することが可能であると考え、ドラマを対象としたライブチャットデータからのメタデータの自動生成を試みた。本報告では、メタデータ生成の初期段階として、番組映像のセグメント化（内容的な区切りに沿ったシーンに分割）と各シーン内での登場人物の推定を行い、システム出力を実際のデータと比較し、検討を行った。

2．メタデータ生成手法概要

2.1　ライブチャットデータ整形

システムの入力は実際にネット上で行われているライブチャットデータと、実際に放映された番組映像を用いる。各ファンサイト特有の表現など、データの偏りを出来るだけ少なくするため、複数のコミュニティのライブチャットデータを時間軸（各発言毎に付加されているタイムスタンプ）毎に並べ替え、一元的に扱えるようにした。その後、番組内容と関係ないデータを除去するため以下の作業を行った。

- 参加者同士のコミュニケーションを目的とした会話や不適切な発言の除去
  →他者のIDや発言を引用している発言を除去
  不適切な発言をNGワードとして指定し、それを含む発言を除去
- 複数行のアスキーアートの除去
  →複数行に渡り半角記号が使われている発言を除去
- コピー&ペースト発言の除去
  →チャットに相応しくない長い発言（200文字以上）を除去

上記の発言の削除を行った後、データを一元的に解析するために、名称の統一を目的として名前の置換えを行った。対象番組の公式サイトから、劇中役名と俳優名の対応データを入手し、主要出演者149人について「劇中役名（苗字）」「劇中役名（名前）」「俳優名（苗字）」「俳優名（名前）」を劇中役名（フルネーム）に置き換え統一した。また、実際のライブチャットデータの一部を参考に、名寄せ（登場人物の別呼称の置換え）、表記ゆれの吸収（例：かとり・蚊取り・か鳥→香取）の置き換えも行った。

2.2　番組映像のセグメント化アルゴリズム

以上の整形を行ったデータを用いて番組映像のセグメント化を行う。セグメント化の手法を、図1に記す。まず初めに番組映像を[3]の手法を用いて映像の特徴量の似た区間（カット）毎に区切る。これにより、短いもので3秒程度、平均18.9秒程度の映像を元にした区切りを番組の時間軸につける。次にライブチャットデータから得られた各発言をchasen[4]を用いて形態素解析を行い「登場人物名」と「名詞」「動詞」「形容詞」のセットとして保持する。その後、ある発言のセット（人物A・「行け」）に着目し、その発言と時間的に近い発言の他のセットの中から、同じ内容が保持されているセットを探す（図1の例では人物Bと「行け」のセット）。これらの発言を一つの塊として考え、全ての発言に同様の処理を行い、話題が変化するポイントを検出し、変化するポイントから一番近いカット点を起点として番組の内容的な区切り（シーン）としてセグメント化を行った。

出演者数が少ない回などは、うまくセグメント化できない場合がある。このようなケースの際は、下に記すキー表現を伴う発言に重みを持たせ、セグメント化を行った。

1)「登場人物名」＋「アスキーアート」
2)　複数IDによる同一文章の書き込み
3)文末の語を複数回繰り返すことによる強調

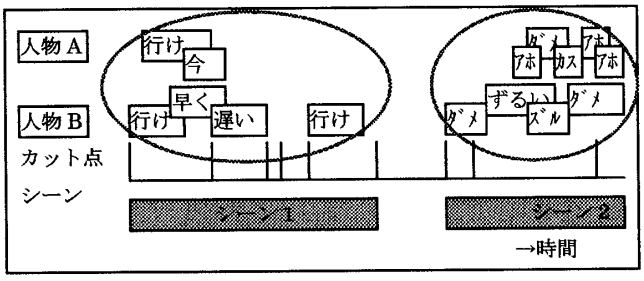

図1：シーン分割方法

†NHK放送技術研究所　人間・情報　知能処理

予備実験の結果から、劇中に新しい登場人物が出現したシーンに 1）の表記が、映像的・内容的にインパクトの強いシーンに 2)の表記が、ストーリー上感情的に盛り上ったシーンにおいて 1)、3)の表記が多いことが得られた。そこで、上に述べた手法でセグメント化が出来なかった場合に、これらの表現を含む発言に重みを付け、その発言部分を強制的に一塊とみなし、セグメント化を行った。

尚、実際のデータのタイムスタンプは、チャット参加者の書き込み時間の分だけデータに遅延が生じる。この遅延分を補正するため、分割されたシーンの冒頭のタイムスタンプと実際の放送時間の差分の平均を遅延時間として算出し、出力結果の補正を行うこととする。

2.3 登場人物推定アルゴリズム

登場人物の推定は、下記に示した 3 つの手法を用いた。

A)セグメント化された各シーン内の発言に登場する登場人物のうち、予備実験により最適とされる回数を算出し、その回数（閾値）以上の頻度で登場した人物名について登場人物と推定する。

B)2.2 のキー表現を伴う発言にマッチした場合に、出現頻度に一定値を乗算し、総合的に出現頻度の値が上位のものを登場人物と推定する。

C)出現頻度の少ない登場人物について、シーン内での言及回数／その回の全ての言及回数を計算し、この値がある値を超えたら登場人物とみなす。

3. 実験・考察

NHK で 2004 年 1 月から 12 月まで放送された、大河ドラマ「新撰組!!」38 話分（1710 分）についてデータ解析を行った。ライブチャットデータは各回 4～6 サイトから収集を行い、2.1 で述べた方法でデータ整形を行った。その後、セグメント化アルゴリズム・登場人物推定アルゴリズムを適用し、シーン毎の推定登場人物の適合率と再現率を算出した（図 2）。適合率・再現率を計算する際の正解データ数「実際の登場人物数」は、各シーンにおいて 2 文以上の台詞を話す、番組公式サイトに役名が記されている登場人物の出演シーンを数えたものとした。

上記セグメント化アルゴリズムにより、各話 45 分を 6～8 のセグメントに分割できることが確認された。また、セグメント化がうまくなされなかった回については、キー表現を含む発言の重み付けを行いうことで、6～8 のセグメントに分割できることが確認された。

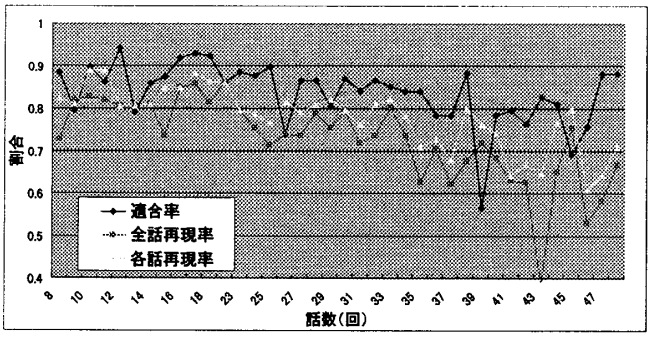

図 2：各回における登場人物の適合率と再現率

セグメント化された各シーン内における登場人物の適合率は平均 82%に達した。適合率を下げる原因として考えられるのは、人気がない又は有名ではない役者が演じている役については視聴者が全く言及しないこと。同一シーン内に登場する役者の苗字が同じ（例：山本耕史・山本太郎）場合や、役者苗字と劇中名苗字が同じ（例：伊東四郎・伊東甲子太郎）場合振り分けられず、データを無効として扱っている事などがあげられる。

各シーン内の登場人物の再現率は平均 69%であった。番組内容的に、途中の回で死亡する主要登場人物が多く、後の回において台詞に名前が出てくるものの、出演しない登場人物が多かったため、話数が進むにつれ、出演していない人に関する言及が多くなり、再現率が下がったと考えられる。そこで、各話の出演者を事前情報を与えると、話の後半の再現率が上がり（図 2 中の各話再現率）平均 74%となる事がわかった。

2.1 に記した名寄せ・表記ゆれの吸収に関しては、これを行うことで、登場人物推定の基準となる発言が多くなるものの、無駄な雑談なども拾ってしまうい、再現率の低い回には有効であるが、適合率には有用性がみられないことがわかった。

登場人物の推定については、A)の手法において閾値を 2 回とした場合が、最も再現率・適合率がよくなることがわかった。B）の手法において、元々キー表現を含む発言で言及される登場人物に関しては、言及数が多いため、重み付けを行っても登場人物の推定結果にはほとんど影響がないことがわかった。C）の登場人物ごとの言及割合については、ストーリーに関係のない登場人物に関する突発的な言及をこの手法で回避することができず、かえって余計な人物名を拾ってしまい、結果が悪くなることがわかった。

尚、データの平均遅延時間は 27.1 秒であったので、この数値を用いて出力結果のタイムスタンプの補正を行った。

4. まとめ

以上の実験により、各話 45 分の番組をその内容に沿って大まかに 6～8 シーンに分け、適合率 82%、再現率 74%の割合で登場人物を推定できることがわかった。精度を上げるための改良点として、現在、暫定的に行っている重要度による重み付けの精査、役者名や役名が同じデータを時間軸を加味して推定し有効データに振り分けを行う、名寄せ・表記ゆれの効率的な吸収方法などが挙げられる。今後はこれらの課題を解決し、本来目的である「何について何故もりあがっているか」の検出を行う。

参考文献

[1] M.Sano, et al.," Generating Metadata from Acoustic and Speech Data in Live Broadcasting", ICASSP2005, MSP-P2.4, March 2005

[2] H.Miyamori, et al., " Generation of Views of TV Content Using TV Viewers' Perspectives Expressed in Live Chats on the Web", ACM Multimedia2005, pp.853-861, November 2005

[3] T.Mochizuki, et al., "Baseball Video Indexing using Patternization of Scenes and Hidden Markov Model", ICIP2005, pp 1212-1215, September 2005

[4]chasen, http://chasen.naist.jp/hiki/ChaSen/