大会プログラム自動生成に向けての一考察

小作 浩美　内山 将夫　村田 真樹　内元 清貴　井佐原 均
郵政省 通信総合研究所 関西先端研究センター
{romi,mutiyama,murata,uchimoto,isahara}@crl.go.jp

1 はじめに

インターネットの普及に伴い，電子化テキストの入手が容易になってきている．それらのテキストをより効率的に利用するため，たくさんの言語処理技術が提案されて来ている．しかしながら，機械翻訳や仮名漢字変換を除き，それらの技術を具体的に応用し広く利用する場面が多いとは言えない．

そのような中で，今回大会プログラムを作成する機会を得た．その作成においていくつかの言語処理技術を利用することができないかと考え，言語処理技術を利用した大会プログラムの自動作成を試みた．

本稿では，その一連の実験について報告するとともに，作成中に気づいた問題点を上げ，今後のプログラム作成の自動化に向けた考察を行なう．

2 大会プログラム作成の流れ

2000年言語処理学会第6回年次大会のプログラムを作成するにあたり，申込〆切後に，メイル以外での申込書も含めて，講演申込のものをすべて図1に示すようにデータベース化し大会プログラム作成を行なった．各申込のセッションへの分類方法については詳しくは3.3章で述べるが，まずセッション名として適当なキーワードを抽出し，次に各申込をセッション名たるキーワードの類似度に基づき分類した．

ここまでの処理では，大会プログラムとして想定すべき条件（例えば会場数や時間の制約など）については考慮しなかった．そのため，この結果に対し，後から時間的な制約により1つのセッションでの発表件数（文書数）の調整を行ない，今回の大会プログラムの原案とした．さらに，申込の際に講演の日付希望や，順番を出来ればこうしてほしいとの要望もあり，それらを考慮した修正も行なった．加えて会場のスケジュールや同じ所属の発表が重ならないように考慮して，変更を加え，今回の大会プログラムとし，採否連絡と共に講演番号を連絡した．なお，ポスター発表については，ここでの方法を用いずに，申込者からの講演分野の申請を利用し，会場

```
<APPLICATION>
<MAIL-ID>130</MAIL-ID>
<TYPE>講演発表</TYPE>
<TITLE> 大会プログラム自動生成に向けての一考察
</TITLE>
<AUTHOR id=1>
<KANJI>○小作 浩美</KANJI>
<KANA>オザク ヒロミ</KANA>
<AFFILIATION> 郵政省通信総合研究所（通信総研）
</AFFILIATION>
<NUMBER>552-633-2044</NUMBER>
</AUTHOR>
<AUTHOR id=2>
...
</AUTHOR>
<CATEGORY>d,e</CATEGORY>
<APPLIANCE>OHP</APPLIANCE>
<ABSTRACT>
  いくかの言語処理技術を利用して言語処理学会の
大会プログラムを自動作成することを試みた。
その結果と自動生成するにあたり明らかになった
問題点、改良点について報告する。
</ABSTRACT>
<ADDRESS>
住所:    〒651-2492 神戸市西区岩岡町岩岡588-2
所属:    郵政省通信総合研究所
氏名:    小作浩美
...
</ADDRESS>
</APPLICATION>
```

図 1: 申込書から作成したデータ

数に合わせて講演分野 a，b の組と c，d，e の組の2つに分け，番号をつけた．

次にこのプログラム作成をするために行なった予備実験について報告する．

3 大会プログラム作成実験

大会プログラムを作成する課題を，文書の自動分類課題と捉え，既存の文書分類技術がどれほど利用できるか実験した．なお，この実験や学習には1999年言語処理学会第5回年次大会のメイルでの申込データと第5回年次大会のセッション名（例：「形態素解析」「意味文脈解析」など）とその大会プログラムを正解として利用した．

今回実験した分類方法は，以下の3つの方法である．

1. 最大エントロピー法を用いた分類
2. クラスタリング手法を用いた分類
3. キーワード抽出・検索法を用いた分類

まず，申込フォーマットの発表内容分野は昨年度通りであるので，昨年度の情報をうまく学習し，さらに今年も前回と同様な話題の講演傾向であるならば，学習アルゴリズム，最大エントロピー法を利用したもので良いプログラムを作成できるはずである．

あるいは，クラスタリング手法で，ある程度のグルーピングができ，各グループに含まれる文書数がそれほど大きく違わなければ，そのグループに基づいて大会プログラム原案を簡単に作成できるのではないかと考えた．

しかし，以上2つの方法はこの課題に合わない部分があり，最終的にはキーワード抽出・検索法を利用することになった．次にそれぞれの手法について行なった実験を報告する．

3.1 最大エントロピー法を用いた分類

最大エントロピー法とは，分類先の推定において，解析に用いられる情報の細かい単位（素性）を定義しておくと，学習データから素性の各出現パターンに対して各分類になる確率を求めるものである．この確率を求める際に，エントロピーを最大にする操作を行なうため，この方法は最大エントロピー法と呼ばれている．この方法を用いた文書の自動分類の研究には文献 [1, 2] がある．

ここでは，文献 [3] のシステムを用いた[1]．実際の分類では，そのシステムの出力から各分類になる確率を計算し，その確率が最大の分類を解とする．

最大エントロピー法の利用では学習に用いる素性が必要となる．学習に用いる素性としては，分類すべき論文のタイトルと要旨を JUMAN[5] で形態素解析して形態素列に分解し，その形態素のうち名詞のみを取り出して用いた．また，タイトルに現れたキーワードは特に重要と考え，上記の素性とは別にタイトルから得られた形態素は別個の素性として用いた．さらに，申込書に記述された講演分野も素性の一つとして学習した．

この方法により，昨年度のデータで学習し今年度のデータを分類して得た結果の一部を表1に示す．確率の上位のものは，良い結果となっている．しかしながら，確率が低いものについては，確率の最大のものとそれに続く確率のものとの間に目立った差が存在せず，どのセッションに含まれるべきか判断に困るような結果であったり（表2），分類自体全く見当外れであったりした．

これは，要旨には文字制限があり，内容をすべて網羅するように幅広く記述されている場合や新たな分野への研究である場合に低い確率になるようである．また，講演分野の記述の無い申込や複数の分野を指定されたものなど，申込の講演分野を学習素性とする場合には信頼性のない素性となる．利用できる情報が少ない状態では，利用できる情報はできる限り信頼性の高いものである必要がある．

さらに，この手法では，申込の分類先を学習データとした年の各セッション名とする必要があり，新たな分野の研究が数多く投稿された場合その新たな分野に分類することができない．つまり，分類先が固定化される問題がある．この問題は，最大エントロピー法に限ったものではなく，教師あり機械学習手法で分類を行なうときには必ず生じる問題である．

上記の理由により，ここで得た結果を採用しなかった．

3.2 クラスタリング手法を用いた分類

クラスタリング手法としては，トップダウンの手法 [6] とボトムアップの手法 [7] を用いた．どちらも昨年度のデータを利用し分類実験を行なった．

トップダウンクラスタリングでは，論文集合を分割していくということを，論文集合の大きさが1になるまで再帰的に繰り返す．このとき，分割の際に，出現頻度の分散が最大の単語に着目し，その単語の出現頻度がある一定以上のものと一定以下のものに集合を分割する [6]．

一方，ボトムアップクラスタリングにおいては，各クラスタ間の距離を比べて，最も近いもの同士をくっつけていくということをクラスタ数が1になるまで繰り返す．このときの距離として，ダイバージェンス (K-L 情報量) を利用する [7]．

どちらの手法についても，何らかの意味概念を持った分類が得られたが，各分類に含まれる文書数に大きなバラツキがあり，それらを一つのセッションにまとめるような意味概念を明確にすることができなかった．そのため，これらのクラスタリング手法は採用しなかった．

3.3 キーワード抽出・検索法を用いた分類

まず，この手法の有効性を確認するために，昨年度のデータよりセッション名の候補を選択する時にキーワードの出現数を利用したものと，スコアリングを行なう手法 [8] の2つで実験した．

出現数の高いキーワードには「システム」「本稿」「我々」「利用」「提案」「手法」などのものが並び，セッション名とするには問題のあるものが多かった．一方，スコアリング手法では「対話」「情報検索」「統語」などが得られた．スコアの上位10位のキーワード（システム，対話，統語，情報検索，翻訳，モデル，解析，構文解析，

[1] 今は Web 上に存在していない．文献としては [4] を参照のこと．

表 1: 最大エントロピー法を用いた分類結果

セッション名	確率	申込 ID	タイトル
機械翻訳	0.989	ID90	語彙化されたツリーオートマトンに...
機械翻訳	0.972	ID37	英日・日英機械翻訳の実力
機械翻訳	0.887	ID50	日英機械翻訳における名詞の訳語選択
検索	0.839	ID38	情報検索における絞り込み語提示による...
検索	0.830	ID97	科学論文における要旨—本文間のハイパー...
検索	0.735	ID9	用例利用型翻訳のための類似用例検索手法

表 2: 確率の低い結果

申込 ID	タイトル	セッション名 (確率)
ID25	辞書定義文を用いた..	分類・他 (0.163) タグ付け (0.152) 辞書 (0.126)
ID18	GDAタグを利用した..	分類・他 (0.237) 言語モデ (0.179) 抽出 (0.156)

抽出，辞書，生成，分類，手法）からセッション名として妥当を思われるキーワード 10 個（翻訳，モデル，解析，対話，検索，抽出，辞書，生成，分類）を人手で選択した．これらのキーワードを基に我々の研究 [9] で採用されているツールを利用し，類似している文書を抽出し，抽出結果を分類結果とした．タイトルに重みをつけずに類似度を求めた場合，同じ得点で複数のクラスに現れてしまう文書が存在した．そこでセッション名がタイトルに現れた場合はスコアを高くするようにして類似度を求め分類したところ，昨年度の分類と完全に一致するものではないが，良い分類結果が得られた．この結果を受けて実際の大会プログラムを作成した．(2章参照) なお，どのキーワードとも類似しない文書も存在する．その文書についてはその他として取り扱う．

この手法では，まず，プログラム作成に重要と思われるキーワードを選出し，それに大きなスコアを割当て，つぎに，そのキーワードに基づいて文書を分類する．このときのキーワードは，それ自体がセッション名として，成立するようなものである．(今回のセッション名はそれらのキーワードから決定している.) そして，我々は，これらの単語が非常に重要であることを前もって知っており，これらのキーワードに大きな価値を与える．つまり，セッション作成という観点からは，ある種のキーワードには大きなスコアを与えるべきである．このことは，論文題目の作成過程を考えてみれば明らかである．すなわち，著者は自分の論文を的確に示すような題名を，なるべく，一般性を持つ形で提示しようとする．このときには，自分の論文が所属する分野を示すキーワードと共に，自分の論文の特徴を示すキーワードをいれるというのが普通の手段である．そのため，論文タイトルには分野を示すキーワードが含まれやすい．したがって，そのような分野を示すキーワードは特に重要視する理由がある．この手法は，その点に着目した手法である．

4 自動作成に向けて

今回の大会プログラムはキーワード抽出・検索法の出力を原案とすることとした．それは，セッション名を今回の大会への申込書の中から抽出できる利点，セッションにあった発表を集められる利点が存在するからである．もちろん，抽出した分類に入らないものや時間的な制約で 1 つのセッションにすることが難しいものなどの問題もある．その点は今回は手作業で一部変更を行なった．また，1 つのキーワードに明らかに複数のセッション数分の申込がまとまっている場合は，そのクラス中の文書から新たなキーワードを抽出して再度分類することをくりかえした．しかしながら,「解析」のセッションでは分類に利用するのに良いキーワードが抽出できず，3 つのセッションに跨ることとなってしまった．これは「解析」の次のクラスとなるようなスコアの高い一般的なキーワードはなく，概念的には近い意味を持つ別のキーワードでそれぞれ記述されていたり，あるいは「解析」には変わりがないが研究対象の違いにより，出てくるキーワードが若干異なるなど，表層的な情報からだけでは分類できないことによる．

この実験を行なうにあたり，第一の問題は申込書から実験データを作成することであった．学会では申込書のフォーマットを会報，学会のメイリングリスト，ホームページなどで周知している．また，最近のインターネットの普及により，申込はほとんどメイルでなされるであろうと考えていたのでメイルから申込フォーマット部分

を抽出し，実験データに変換することを考え，自動抽出プログラムを作成していた．

しかしながら，メイルで申込をする場合，申込者側で必要と思われるデータだけ（例えばタイトル，著者名リストのみ）を羅列し申し込む方が多く見うけられた．特に，締め切り当日の夜中に出されたメイルほどそのような傾向が強く，中には講演かポスターの申込なのかも記入されていない場合も存在した．さらに，事務処理をする際に必要な連絡先や会員番号，所属の記述すらない申込メイルもあった．学会運用上では，会員であるかどうかを調べる必要があるように考えられるが，今回は会員であるかどうかのチェックはデータ不備のため，ユーザからの申請を信じる形となっている．また，1つのメイルに複数の申込書が含まれいるものもあり，1メイル1申込と勝手に想定していたためにデータチェックに手間取ってしまった．

さらに，メイル等で周知するための大会プログラムを申込データから自動で作成したが，上記のような入力不備のため所属のない講演者や所属の記述が申込者により異なるため所属の記述が一貫していない．また，大会プログラム作成には関係のない部分であったが，申込書からデータベース化する際に見落としたバグがあり，メイルで周知した大会プログラムに不備が存在した．

これらについては，申込受け付け方法をWeb上で統一するなど，事務局側での申込受け付け体制を検討する必要があると思われる．また，講演の分野の指定や会員番号の記述など申込フォーマットにあっても無記入になりやすい項目については申込フォーマットの項目を検討し直すことや講演分野は申請者に申請してもらうなど申込書の形式から検討する必要もあると考える．

第二の問題としては，申込締め切り期日後の申請者の存在があげられる．大会プログラムを自動作成する場合には，すべての申込をデータベース化してから作業する必要があり，運用上締め切りの期日を確実に守る必要がある．当初，大会プログラム作成の実験をはじめた後に申し込まれたものは，申込を受け付けないか，特別セッションを設けるかなどと考えていたが，最終的には締め切り後の申込についても受け付けることとした．そのため大会プログラムの作成が遅れ，その結果として採録の連絡も遅くなり，多数の方に迷惑をかける結果となった．この場をかりて陳謝したい．

我々は大会プログラム作成を文書分類課題とみなして，実験を行なった．しかし，実際のプログラムにおいては，時間的な制約，場所的な制約も考慮する必要がある．また，所属が同じ講演者の講演が時間的に重なることを避けるような制約も設けた．これは，同じ所属の人同士は，お互いの発表を聞きたいであろうと判断したことによる．そのあたりの条件については，今回は我々がプログラム作成の中で思いつく条件を独断で設けたがもっと別の重要な観点がある可能性もある．

5 おわりに

本研究の実験は，現在の様々な言語処理技術が何か役に立つのでは考え，行なったものである．この実験のおかげで大会プログラム作成はそれほど苦労することなくできた．しかしながら，時間的な制約もあり，多数の技術をじっくり吟味する時間を取ることが出来なかった．今後，もし機会があれば，他の技術についても実験を行なってみたい．また，今回のプログラムについて，発表者や大会参加者より，アンケートにより評価をしてもらおうと考えている．アンケート結果については，別途報告する予定である．発表者および大会参加者にはアンケートの協力をお願いする (http://www.crl.go.jp/nlp にもアンケート情報を掲載予定)．さらに，大会プログラム自動作成に向けて考慮すべき条件をより明確にし，今後の事務処理の簡素化に貢献できればと考えている．

謝辞

北陸先端大学の島津明氏，望月源氏に昨年度のデータの利用に際しご協力いただいた．また，通信総合研究所知的機能研究室の太田公子氏，志水美江子氏にもご協力いただいた．ここに感謝する．

参考文献

[1] 乾裕子, 内元清貴, 村田真樹, 井佐原均, 文末表現に着目した自由回答アンケートの分類, 情報処理学会 自然言語処理研究会 NL128-25, (1998).

[2] Kamal Nigam, John Lafferty, and Andrew McCallum, Using maximum entropy for text classification, *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, (1999).

[3] Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release 1.6 beta, (http://www.mnemonic.com/software/memt, 1998).

[4] Eric Sven Ristad, Maximum Entropy Modeling for Natural Language, *ACL/EACL Tutorial Program, Madrid*, (1997).

[5] 黒橋 禎夫, 長尾 眞, 日本語形態素解析システム JUMAN 使用説明書 Version 3.61,1999.

[6] 田中英輝, 大規模文書集合の高速クラスタリング, 言語処理学会, 第3回年次大会, (1997).

[7] L. Douglas Baker, Andrew Kachites McCallum, Distributional Clustering of Words for Text Classification, SIGIR'98, (1998).

[8] 宮本義男, 加藤茂樹, 後藤康雄, 林博之, キーワード自動抽出システム, 第37回システム制御情報学会,(1993).

[9] Hiromi Ozaku, Kiyotaka Uchimoto, Hitoshi Isahara, Improvement of Intelligent Network News Reader HISHO, *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Languages IRAL'98*, (1998).