帰納的学習を用いた電子メイルにおける Subject の自動生成手法の提案

康　錦紅[†]　荒木　健治[†]　栃内　香次[‡]
[†]北海道大学大学院工学研究科電子情報工学専攻
〒060-8628 札幌市北区北 13 条西 8 丁目
[‡]北海学院大学大学院経営学研究科
〒062-8605 札幌市豊平区旭町 4-1-4

１．はじめに

最近の電子メイルの普及は目覚めしく，非常に多くのメイルが日々送られてくるという環境にいる人が急激に増大している．電子メイルにおいて Subject は相手にまず自分が何を伝えたいのかを表現する部分なので非常に重要ではあるが，時間がなくなってくると十分に考えずに Subject を書いてしまうことも多く，そのことがメイルを受け取った人に誤解を招たり，後ほどメイルを保存しておいてメイルを見ながら内容を確認する際に問題になる．このような状況を考え，本稿では過去の Subject とメイル本文の組より帰納的学習[1]により Subject の生成ルールを獲得し，それらのルールを用いて Subject をメイル本文より自動的に生成する手法を提案する．

２．予備実験

まず，メーリングリストから Subject が入ってるメイル 100 通を出現順に選択して予備実験を行った．予備実験の結果を表１に示す．ここで，Subject を含む文を重要文と定義し，重要文中５回以上出現した語をキーワードと定義する．予備実験の結果から重要文は文の位置，日時表現と重要文中で高い確率で使用された語との関係が深いことが分かったので，重要文の決定に使用される要素を表２に示すように決定した．

表　１　重要文決定の予備実験結果

段落	第一段落	第二段落	第三段落	他の段落
頻度	8	84	6	2
段落中文の位置	一番目	二番目	三番目	その他
頻度	94	6	0	0
日時表現	ある	なし		
頻度	34	６６		
キーワード	ある	なし		
頻度	63	37		

３．　提案手法

本手法の流れを図１に示す．

本手法ではあるユーザの過去のメイル本文と Subject の組より学習を行うことにより Subject を生成することができる．また，システムがユーザに動的に適応することができる．

A Study on Generation Method of Subject of E-mail Using Inductive Learning
†Jinhong Kang,Kenji Araki (Hokkaido University)
‡Koji Tochinai (Hokkai-Gakuen University)

具体的なアルゴリズムとして，以下の２段階で Subject を生成する．
(１)Subject を含むと推測される文（重要文）を抽出する．
(２)重要文より品詞列パターンを用いて Subject を生成する．

アルゴリズムとしては，まず入力文に対して形態素解析ツール「茶筅」[2]を用いて形態素解析を行う．次に，予備実験の結果より，第何段落目か，段落中文の位置，日時表現の有無，重要文中で高頻度で出現する語の有無などの情報を用いて重要文を決定する．これらの要素の値は評価式を用いて算出し，評価値が最大のものを重要文と決定する．ここで利用される重みは予備実験中の各要素の出現頻度により決定したものである．このようにして決定された重要文より Subject を決定する．Subject の決定は品詞列のパターンを用いて決定する．パターンは以下のようなもので構成されている．

[名詞列]＋[助動詞]
[名詞列]＋[助詞]
[名詞列]＋[助詞]＋[名詞-サ変接続]
[名詞列]＋[助詞]＋[動詞-自立]

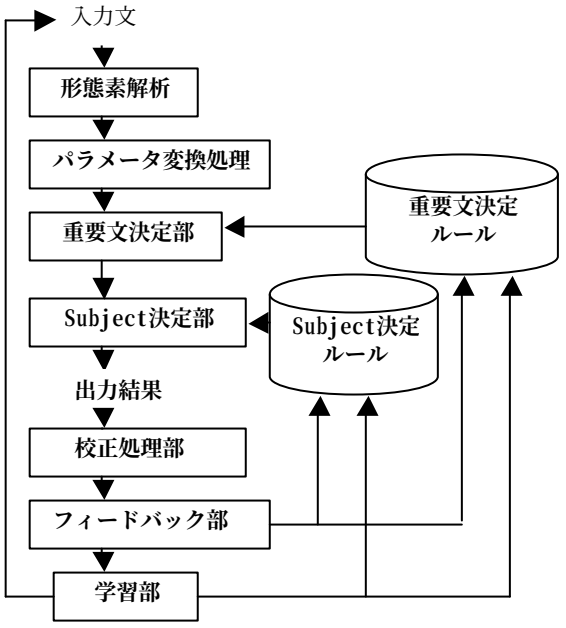

図　１

生成した Subject は校正処理部で正誤の判定後，

誤った結果はユーザにより校正され，学習部で新たなルールとして追加される．また，生成結果の正誤により，フィードバック部で尤度が変化する．尤度は適用したルール正誤の変換度数により変化する．

4．　概要

4.1　パラメータ変換処理

まず，入力文に対して形態素解析ツール茶筅[2]を用いて形態素解析を行う．次に，表２に示す要素で入力文の一文一文を４桁の数字に変換する．例えば，第一段落の一番目の文で，日時表現があってキーワードがある文のパラメータは１１１１のようになっている．

表　2　パラメータ決定要素

１．　段落
１：第一段落　２：第二段落　３：第三段落 4：その他の段落
２．　段落中文の位置
１：一番目の文　２：二番目の文　３：三番目の文
４：その他
３．　日時表現
１：ある　２：なし
４．キーワード
１：ある　２:なし

4.2　重要文の抽出

重要文は，変換されたパラメータを用いて決定する．重要度は式（１）により計算する．

重要度＝ルールの重み×∑要素の重み　…（1）

ここで、ルールの重みとは，ルールと一致した文が重要文になった確率である．入力メイルの一文ごとの重要度で重要度が一番大きい文を重要文として抽出する．

4.3　Subject 決定部

Subject はパターンマッチにより決定する．パターンは以下のようなもので構成される．

[名詞列]＋[助動詞]
[名詞列]＋[助詞]
[名詞列]＋[助詞]＋[名詞-サ変接続]
[名詞列]＋[助詞]＋[動詞-自立]

ここで名詞列とは，名詞または接頭詞，名詞，未知語の品詞が連続に並んだ単語列あるいは接頭詞，名詞，未知語が「と」,「の」,「や」で 連結されたものである．マッチした文字列は、尤度の高さにより，一番高い尤度を用いる文字列からSubject が生成される．Subject は文字列の名詞列である．

例を以下に示す．

重要文：

[記号-空白]第[接頭詞-名詞接続]３[未知語]回[名詞-一般]国際[名詞-一般]バラ[名詞-一　般]と[助詞-格助詞-一般]ガーデンニングショウ[未知語]が[助詞-格助詞-一般]5/18[未知語]（[記号-括弧開]金[名詞-一般]）[記号-括弧閉]～[記号-一般]23[未知語]（[記号-括弧開]水[名詞-一般]）[記号-括弧閉]まで[助詞-副助詞]西武[名詞-固有名詞-組織]ドーム[名詞-一般]で[助詞-格助詞-一般]開催[名詞-サ変接続]する[動詞-自立]れる[動詞-接尾]ます[助動詞]。

マッチした文：

（１）第[接頭詞-名詞接続]3[未知語]回[名詞-一般]国際[名詞-一般]バラ[名詞-一般]と[助詞-格助詞-一般]ガーデンニングショウ[未知語]が[助詞-格助詞-一般]」

（２）「西武[名詞-固有名詞-組織]ドーム[名詞-一般]で[助詞-格助詞-一般]開催[名詞-サ変接続]する[動詞-自立]

ここで，マッチした文が 2 つなので，尤度を計算し，計算結果により尤度が高い(1)から Subject が決定される．Subject は,「第３回国際バラとガーデンニングショウ」となる．

5．評価実験と考察

5.1 実験結果

今回はメーリングリスト[3]のメイルを使って実験を行った．実験結果を表３に示す．

表　3　　実験結果

メイル数	正しい結果数	誤った結果数
100	52	48

5.2 考察

今回，結果の正誤の判定は実例と比較して評価した．システムが出力した結果が実例の Subject と同じ場合正しいと評価，違う場合誤った結果として評価した．実例の Subject の中に必ずしも正しいとはいえないものが８通あったので，実験結果に影響を与えた．また，重要文の抽出が正しくないものが１９通あつたので, Subject の決定に影響を与えた．

６．おわりに

本稿では，帰納的学習により電子メイルの本文から Subject を自動的に決定する手法を提案し，本手法による評価実験を行った．今回は実験で用いたメイルの数が少なかったので，今後，より多いのメイルを用いて実験を行う予定である．また，重要文の決定が誤った場合 Subject の出力も必ず間違うので，重要文の決定ルールの有効性についても検討する必要がある．

参考文献

（１）　荒木健治,“帰納的学習を用いた自然言語処理の有効性について”，信学技報、TL99－41、pp.33－40、2000.

（２）　松本裕治，北内啓，山下達雄，平野善隆,“日本語形態素解析システム『茶筅』version 2.0 使用説明書”，NAIST Technical Report，NAIST-IS-TR99008，April 1999

（３）　http://www.egroups.co.jp/messages/eco-school