5N－2

決定木学習を用いたテキスト自動要約手法に関するいくつかの考察

奥村 学*，原口良胤**，望月 源*

* 北陸先端科学技術大学院大学 情報科学研究科

** NTT インテリジェントテクノロジ

1　はじめに

電子化されたテキストが世の中に満ち溢れ，情報洪水という言葉が使われるようになってからかなりの歳月を経ている．そのため，自動要約技術などにより，読み手が読むテキストの量を制御できることが求められている．

テキスト自動要約の一つの手法である重要文抽出では，文に関するさまざまな情報が用いられてきているが，一般に複数の情報を同時に用いた方が精度を改善できるとの考え方に基づき，複数の情報を統合して用いる研究が近年数多く見られる [5]．

その中でも，重要文集合を訓練データとして，機械学習手法などを用いることにより，複数の情報の統合方法を最適化する研究が盛んになってきている．決定木学習アルゴリズム C4.5[4] を用いて訓練データから決定木を学習し，学習した決定木により，テキスト中の文を重要文/非重要文に分類することで重要文抽出を行なう手法もその一つである [6]．決定木学習を用いた重要文抽出では，あらかじめ重要文/非重要文の 2 つのクラスに分類済みの訓練データに，文に関するさまざまな情報を付加しておき，そのデータを正しく分類できるような，(文に関する情報の組み合わせとしての) ルール集合を決定木の形で学習することになる．決定木学習は，学習結果が人間にも見やすいルールの形で得られ，また，そのルールを人手で修正できることから，機械学習手法として優れているとされる．

しかし，決定木学習では，訓練データ中の事例が特定のクラスに偏って分布する場合，事例数が少ないクラスの事例はノイズとして扱われてしまい，その結果，精度の良い決定木が学習できないという問題が近年指摘されるようになっている [1, 3]．重要文抽出に決定木学習を用いる場合，重要文数は非重要文数に比べ極端に少なく，したがって，決定木学習のこの問題の影響を受け，精度の良い重要文抽出が実現できない可能性がある．

本研究では，この問題に対して，

1. 非重要文事例を重要文事例と同数サンプリングして，決定木学習を行なうことが精度向上につながる，
2. しかし，サンプリングして学習した決定木は，過剰に重要文クラスに文を分類する傾向があり，設定した要約率以上に重要文を出力する，
3. したがって，学習した決定木が重要文と分類した文のうち，信頼度が高いものを，要約率まで出力することで，さらに精度を向上させることができる

ことを，考察から指摘し，実験により示す．

2　決定木学習を用いた重要文抽出

前節で述べたように，決定木学習を用いた重要文抽出では，訓練データ中で重要文/非重要文間に事例数の大きな偏りがあり，重要文クラスに属する事例数が統計的に誤差と考えられてしまう可能性があり，そのため，大多数の文を非重要文と分類する決定木が学習され，精度のうち，特に，再現率 (後述) が良くない傾向がある．

事例数の偏りによる分類精度の低下を回避する一つの単純な手法としては，重要文 (事例数が少ないクラス) とほぼ同数の非重要文をランダムに抽出し，それらのみで (それ以外の非重要文は用いないで) 決定木学習を行なうものが考えられる [2]．これにより，事例数が偏った訓練データで学習した場合に比べ，高い再現率が得られることが予想される．

しかし，高い再現率は，過剰に重要文を出力することによっても達成できる．重要文と非重要文の数が同数の訓練データからは，テキストの総文数中の重要文数の割合 (要約率) を誤って大きく見積もった決定木が学習される可能性があり，その場合，設定した要約率以上に重要文を出力することが考えられる．

一方，決定木学習で学習された分類のための知識には，信頼度が付与されている．したがって，信頼度の高い (確からしい) 知識に基づいて分類された重要文のみを上位から抽出することが可能である．これにより，より精度の高い分類が可能となる．

我々は，これらの知見から，サンプリングした非重要文事例を利用した訓練データで決定木学習し，学習した決定木を利用して，信頼度の高いものから，要約

Some Observations on Automatic Text Summarization Based on Decision Tree Learning
OKUMURA Manabu, HARAGUCHI Yoshitsugu, MOCHIZUKI Hajime
School of Information Science, Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292, Japan
oku@jaist.ac.jp

実験条件	要約率 (%)	Recall(%)	Precision(%)	F-measure
単純な決定木	6.6	23.1	52.1	0.320
サンプリング	24.2	50.0	30.7	0.381
サンプリング + 信頼度	9.7	53.9	37.9	0.439

表 1: 重要文抽出精度の比較

率に達するまで，重要文を抽出することにより，サンプリングしない元の訓練データで学習した決定木，サンプリングした訓練データで学習した決定木をそのまま (信頼度を考慮せず) 用いた場合と比較し，より精度の高い重要文抽出が可能になるとの仮説を立てる．次節では，実験を行ない，この仮説の通り，大幅に精度を向上できることを報告する．

3 実験

実験には，1995 年の日本経済新聞のコラム，社説，記事それぞれ 25 テキスト，計 75 テキストからなる，野本ら [6] のデータを利用する．このデータでは，総文数の 10%程度が重要文として大学 (院) 生により，あらかじめ選択されている (総文数 1424 のうち，重要文数は 202 である)．

決定木学習に利用する，文に関する情報としては，以下の 9 つを用いる．

1. 文のテキスト中での位置
2. 文の長さ (文字数)
3. 文の段落中での位置
4. 文中の単語の tf.idf 値の総和による文の重要度
5. 文末の態度表現の型 (13 種類)
6. 文を通過する語彙的連鎖に基づいて計算した文の重要度
7. 文中の主題，主語の有無
8. 文頭の接続詞の種類
9. 文頭の指示詞の種類

評価尺度としては，以下に示す再現率 (recall;R)，適合率 (precision;P)，F-measure を用いる．

$$R = \frac{\text{決定木が抽出した正解重要文数}}{\text{正解重要文数}} \quad (1)$$

$$P = \frac{\text{決定木が抽出した正解重要文数}}{\text{決定木が抽出した重要文数}} \quad (2)$$

$$F - measure = \frac{2 \times P \times R}{P + R} \quad (3)$$

表 1 に，10 回の交差検定により行なった実験の結果の平均を示す．実験から，

1. 単純な決定木学習では，設定した要約率よりも少ない文数しか重要文と分類しておらず，再現率が低い，
2. 非重要文事例をサンプリングした訓練データで学習した場合，要約率以上に重要文を出力しており，再現率が向上している，
3. 信頼度の高い，上位 10%の重要文を抽出することで，サンプリングのみの場合と比較して，さらに精度を向上できる

ことを確認できた．この結果から，提案した，

> サンプリングした非重要文事例を利用した訓練データで決定木学習し，学習した決定木を利用して，信頼度の高いものから，要約率の長さだけ，重要文を抽出する

手法が有効であることを示せたと言える．

参考文献

[1] T. Honda, H. Mochizuki, T.B. Ho, and M. Okumura. Generating decision trees from an unbalanced data set. In *Proc. of the 9th European Conference on Machine Learning*, 1997.

[2] I. Mani and E. Bloedorn. Machine learning of generic and user-focused summarization. In *Proc. of the 15th National Conference on Artificial Intelligence*, pp. 821–826, 1998.

[3] R.J. Passonneau and D.J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, Vol. 23, No. 1, pp. 103–139, 1997.

[4] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.

[5] 奥村 学，難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理,「テキスト要約のための言語処理」特集号, Vol. 6, No. 6, pp. 1–26, 1999.

[6] 野本忠司，松本裕治. 人間の重要度判定に基づいた自動要約の試み. 情報処理学会自然言語処理研究会報告, pp. 71–76, 1997. 120-11.