決定木学習を応用した日本語要約文の自動作成

原口 良胤

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： 自動要約, 決定木学習, 機械学習.

近年，電子化されたテキストが氾濫し，必要な情報を得るために多大な労力が要求されるようになってきている．そこで，必要な情報のみを効率良く抽出することが可能な，自動要約技術の必要性がますます高まってきている．

決定木学習を要約作成に利用するには，要約をある文書中の重要文の集合ととらえ，人手によってあらかじめ重要文を決定する．この文集合から，各文についてその特徴を属性値として計算し，それらの属性値を用いて訓練を行い決定木を構成する．次に要約の対象となる文書の各文について属性値を計算し訓練によって得られた決定木を用いて重要文を抽出し要約を作成する．

決定木学習は，あらかじめ用意された複数のクラスに分類された訓練事例をもとに決定木を構成し，それによって新たな事例を複数のクラスに分類する．ここで，あるクラスの事例数が他のクラスの事例数と比較して極めて少ないと，それらの事例はノイズとして扱われ，うまく分類できない．要約作成を行う場合，重要文は非重要文と比べて事例数が極端に少ないので，決定木学習をそのまま用いると重要文の事例がうまく分類されない．

本研究では，2 段階の決定木を構成する手法 (TLDT:Two Level Dicision Tree construction) を適用することで，決定木学習を要約作成に適用する際の問題を解消することを試みる．TLDT では，訓練事例中の重要文の属性値と類似の属性値を持つ非重要文 (ニアミス) を抽出し，重要文と合わせて新たなクラス mix を構成する．一方，重要文の属性値と類似していない属性値を持つ非重要文でも新たなクラス far を構成する．1 段階目の決定木学習では，これら 2 つのクラスをよく分離する決定木を構成する．2 段階目の決定木学習では，クラス mix を重要文クラスと非重要文クラスに分離するような決定木を構成する．

以下に示す実験を行った．

実験では，1995 年の日経新聞の記事，社説，コラム合計 75 テキストを用いた．重要文は人手によって決定されたものを用いた．まずそれぞれの文の属性を計算した．属性を以下に示す．

- 文の位置
  テキスト中の先行する文数を正規化した値
- 文の長さ
  文の文字数を正規化した値
- 段落内位置
  段落内で先行する文数を正規化した値
- $tf * idf$
  $tf * idf$ を正規化した値
- 態度表現
  態度の種類を示す離散値
- 語彙的連鎖
  テキスト中の語彙的連鎖に含まれる語の数を正規化した値．
- 助詞
  助詞の種類を示す離散値
- 接続詞
  接続詞の種類を示す離散値
- 前方照応
  前方照応の種類を示す離散値

これらの属性データを訓練セットとテストセットに分割する．訓練セットは，mix クラスと far クラスに分割する．ニアミスデータの類似度を計算する際，以下に示す 3 種類の方法を利用した．

- 文中のすべての属性を利用した類似度
- 文中の一部の属性を利用した類似度
- 文中の離散属性のみを用いた類似度

また，決定木の構築に当っては 2 通りの方法を用いる．ひとつは，ニアミス文を抽出し決定木を構築する際，連続値をもつ属性を離散値に変換する．もうひとつの方法は，連続値をそのまま利用する方法である．

表 1: C4.5 と TLDT の比較 (mix:far=1:2)

	recall	precision	f-measure
C4.5	0.222	0.461	0.300
TLDT	0.280	0.454	0.347

同時に，mix クラスと far クラスの事例数の偏りを補正するために 2 通りの方法を用いる．ひとつは far クラスから mix クラスと同数の事例を取り出して訓練データを構築する．もうひとつは，mix クラスと far クラスの事例数を 1:2 にする方法である．far クラスから事例を選択する際には，類似度を低いものから選ぶ．

1 段目の決定木を訓練セットを用いて構成し，2 段目の決定木はクラス分けし直したデータから構成する．その後，テストデータによって評価を行う．これを 10 回行い交差検定を行った．

結果は，C4.5 を単独で利用した場合よりも recall,precision,f-measure ともに比較的よい結果を得た．表に示す．

ここで f-measure は，$f-measure = \frac{2 \times recall \times precision}{recall + precision}$ である．

結果より，以下の傾向が見られる．つまり，$mix : far$ の比を増加させると recall が低下し，precision が向上する．その結果，f-measure が若干向上する．

今後の課題として，まず，ニアミスを抽出する際をもう少し工夫することで全体的な recall の向上する可能性がある．次に，要約率を利用することも考えられる．要約率はテキストから抽出される重要文の割合である．訓練データを要約率に応じて構成することで，抽出される重要文数を調整できる可能性がある．