(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3682529号
(P3682529)
(45)発行日　平成17年8月10日(2005.8.10)　(24)登録日　平成17年6月3日(2005.6.3)

(51)Int.Cl.[7]　F I
G06F 17/30　G06F 17/30 220A
G06N 3/00　G06F 17/30 170A
G06F 17/30 180A
G06N 3/00 550Z

請求項の数 15　(全 14 頁)

(21)出願番号　特願2002-23493(P2002-23493)
(22)出願日　平成14年1月31日(2002.1.31)
(65)公開番号　特開2003-223456(P2003-223456A)
(43)公開日　平成15年8月8日(2003.8.8)
審査請求日　平成14年1月31日(2002.1.31)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100119161
弁理士　重久　啓子
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

審査官　辻本　泰隆

最終頁に続く

(54)【発明の名称】要約自動評価処理装置、要約自動評価処理プログラム、および要約自動評価処理方法

(57)【特許請求の範囲】

【請求項１】

コンピュータを用いた要約自動評価処理装置であって、

文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段と、

前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する解－素性対抽出手段と、

前記解と前記素性の集合との組を学習結果として学習結果記憶手段に記憶する機械学習手段と、

前記解－素性対抽出手段により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する素性抽出手段と、

前記学習結果である前記解と前記素性の集合との組をもとに、ベイズの定理にもとづいて前記素性抽出手段から得た前記テキストの素性の集合の場合の各分類になる確率を求め、前記確率の値が最も大きい分類を、求める推定解とする評価推定手段とを備える

ことを特徴とする要約自動評価処理装置。

【請求項２】

コンピュータを用いた要約自動評価処理装置であって、

文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先で

ある解との組からなる解データを記憶する解データ記憶手段と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する解－素性対抽出手段と、
前記解と前記素性の集合との組とを規則とし、前記規則を所定の優先順位によりリストに格納し、前記リストを学習結果として学習結果記憶手段に記憶する機械学習手段と、
前記解－素性対抽出手段により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する素性抽出手段と、
前記学習結果である前記リストに格納された前記規則を優先順序の高い順に前記素性抽出手段から得た前記テキストの素性の集合と比較し、素性が一致した規則の分類先を、求める推定解とする評価推定手段とを備える
ことを特徴とする要約自動評価処理装置。
【請求項３】
コンピュータを用いた要約自動評価処理装置であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する解－素性対抽出手段と、
前記解と前記素性の集合との組から、前記素性の集合が所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布を求め、前記確率分布を学習結果として学習結果記憶手段に記憶する機械学習手段と、
前記解－素性対抽出手段により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する素性抽出手段と、
前記学習結果である前記確率分布にもとづいて、前記素性抽出手段から得た前記テキストの素性の集合の場合の各分類の確率を求め、前記確率の値が最も大きい分類を、求める推定解とする評価推定手段とを備える
ことを特徴とする要約自動評価処理装置。
【請求項４】
コンピュータを用いた要約自動評価処理装置であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する解－素性対抽出手段と、
前記解と前記素性の集合との組を用いて、所定のサポートベクトルマシンモデルの方法により超平面を求め、前記超平面および前記超平面により分割された二つの空間の分類を学習結果として学習結果記憶手段に記憶する機械学習手段と、
前記解－素性対抽出手段により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する素性抽出手段と、
前記学習結果である前記超平面をもとに、前記素性抽出手段から得た前記テキストの素性の集合が前記超平面で分割された空間のいずれに属するかを求め、前記素性の集合が属する空間の分類を、求める推定解とする評価推定手段とを備える
ことを特徴とする要約自動評価処理装置。
【請求項５】
請求項１ないし請求項４のいずれか一項に記載の要約自動評価処理装置において、
前記問題の要約結果に対する解は、機械処理によりなされたものと人手によりなされた

ものをそれぞれ示す二つの分類先からなるものである
ことを特徴とする要約自動評価処理装置。
【請求項６】
要約を自動評価する処理をコンピュータに実行させるためのプログラムであって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理と、
前記解と前記素性の集合との組を学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理と、
前記学習結果である前記解と前記素性の集合との組をもとに、ベイズの定理にもとづいて前記入力されたテキストの素性の集合の場合の各分類になる確率を求め、前記確率の値が最も大きい分類を、求める推定解とする処理とを、
コンピュータに実行させるための要約自動評価処理プログラム。
【請求項７】
要約を自動評価する処理をコンピュータに実行させるためのプログラムであって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理と、
前記解と前記素性の集合との組とを規則とし、前記規則を所定の優先順位によりリストに格納し、前記リストを学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理と、
前記学習結果である前記リストに格納された前記規則を優先順序の高い順に前記入力されたテキストの素性の集合と比較し、素性が一致した規則の分類先を、求める推定解とする処理とを、
コンピュータに実行させるための要約自動評価処理プログラム。
【請求項８】
要約を自動評価する処理をコンピュータに実行させるためのプログラムであって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理と、
前記解と前記素性の集合との組から、前記素性の集合が所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布を求め、前記確率分布を学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理と、
前記学習結果である前記確率分布にもとづいて、前記入力されたテキストの素性の集合の場合の各分類の確率を求め、前記確率の値が最も大きい分類を、求める推定解とする処理とを、
コンピュータに実行させるための要約自動評価処理プログラム。

【請求項９】
要約を自動評価する処理をコンピュータに実行させるためのプログラムであって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理と、
前記解と前記素性の集合との組を用いて、所定のサポートベクトルマシンモデルの方法により超平面を求め、前記超平面および前記超平面により分割された二つの空間の分類を学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理と、
前記学習結果である前記超平面をもとに、前記入力されたテキストの素性の集合が前記超平面で分割された空間のいずれに属するかを求め、前記素性の集合が属する空間の分類を、求める推定解とする処理とを、
コンピュータに実行させるための要約自動評価処理プログラム。
【請求項１０】
請求項６ないし請求項９のいずれか一項に記載の要約自動評価処理プログラムにおいて、
前記問題の要約結果に対する解は、機械処理によりなされたものと人手によりなされたものをそれぞれ示す二つの分類先からなるものである
ことを特徴とする要約自動評価処理プログラム。
【請求項１１】
コンピュータを用いた要約自動評価処理方法であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理過程と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理過程と、
前記解と前記素性の集合との組を学習結果として学習結果記憶手段に記憶する処理過程と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理過程と、
前記学習結果である前記解と前記素性の集合との組をもとに、ベイズの定理にもとづいて前記入力されたテキストの素性の集合の場合の各分類になる確率を求め、前記確率の値が最も大きい分類を、求める推定解とする処理過程とを備える
ことを特徴とする要約自動評価処理方法。
【請求項１２】
コンピュータを用いた要約自動評価処理方法であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理過程と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理過程と、
前記解と前記素性の集合との組とを規則とし、前記規則を所定の優先順位によりリストに格納し、前記リストを学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力

されたテキストから前記素性の集合を抽出する処理過程と、
前記学習結果である前記リストに格納された前記規則を優先順序の高い順に前記入力されたテキストの素性の集合と比較し、素性が一致した規則の分類先を、求める推定解とする処理過程とを備える
ことを特徴とする要約自動評価処理方法。
【請求項１３】
コンピュータを用いた要約自動評価処理方法であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理過程と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理過程と、
前記解と前記素性の集合との組から、前記素性の集合が所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布を求め、前記確率分布を学習結果として学習結果記憶手段に記憶する処理と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理過程と、
前記学習結果である前記確率分布にもとづいて、前記入力されたテキストの素性の集合の場合の各分類の確率を求め、前記確率の値が最も大きい分類を、求める推定解とする処理とを備える
ことを特徴とする要約自動評価処理方法。
【請求項１４】
コンピュータを用いた要約自動評価処理方法であって、
文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段にアクセスする処理過程と、
前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する処理過程と、
前記解と前記素性の集合との組を用いて、所定のサポートベクトルマシンモデルの方法により超平面を求め、前記超平面および前記超平面により分割された二つの空間の分類を学習結果として学習結果記憶手段に記憶する処理過程と、
前記解と前記素性の集合との組を生成する処理により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する処理過程と、
前記学習結果である前記超平面をもとに、前記入力されたテキストの素性の集合が前記超平面で分割された空間のいずれに属するかを求め、前記素性の集合が属する空間の分類を、求める推定解とする処理過程とを備える
ことを特徴とする要約自動評価処理方法。
【請求項１５】
請求項１１ないし請求項１４のいずれか一項に記載の要約自動評価処理方法において、
前記問題の要約結果に対する解は、機械処理によりなされたものと人手によりなされたものをそれぞれ示す二つの分類先からなるものである
ことを特徴とする要約自動評価処理方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、コンピュータを用いて要約を評価する処理に関し、特に教師あり機械学習法を用いて要約の自動評価処理を行う処理装置と、要約の自動評価処理をコンピュータに実行させるためのプログラムと、要約の自動評価処理方法とに関する。

【０００２】
【従来の技術】
近年、情報技術の発展に伴ってコンピュータを用いた文章の自動要約処理が盛んになってきている。そして、様々な自動要約処理手法で作成された要約結果に対する公正な評価の重要性が増してきている。
【０００３】
要約処理は、主に、重要文抽出要約と自由作成要約との２つの類型がある。重要文抽出要約は、要約率に応じて評価対象の文章中に存在する文を抽出して要約する処理である。自由作成要約は、評価対象の文章中の内容を削除したり変更したりして自由に文生成などして要約する処理である。
【０００４】
重要文抽出要約については、文章中のどの文を抽出すると良いかという情報を用いて評価を自動処理することが可能である。例えば、文章中の文に対して、要約結果として抽出されるべき程度を示す重要度を予め付与し、抽出された文の重要度を集計して要約の評価とする。
【０００５】
一方、自由作成要約においては、良い要約は複数あり得るため、あらゆる良い要約すなわち正解の情報を用意しておくことは困難であり、評価を自動処理することは困難である。そのため、従来、自由作成要約の評価は人が知識や経験にもとづいて行っていることが多い。
【０００６】
要約の自動評価処理として、以下の参考文献１に示す従来手法がある。参考文献１では、重要文抽出要約について、コンピュータで抽出処理した文と予め人間が選択しておいた重要文との間の一致度をもとに再現率・適合率・Ｆ値により要約の評価を行っている。また、自由作成要約についても、作成された要約と、予め人間が作成した正解となる要約との類似度を単語の頻度ベクトルを用いて判断して行っている。［参考文献１：野畑周　他、複数の評価尺度を統合的に用いた重要文抽出システム、言語処理学会第７回年次大会発表論文集、pp301-304 , 2001］
【０００７】
【発明が解決しようとする課題】
従来、自由作成要約の評価は、通常では専門家の手により行なわれていた。しかし、人手による評価は、評価者の経験則や価値にもとづくものであるため、同じ要約結果に対しても、異なる評価者では評価が異なってしまう場合や、同じ評価者であっても評価の時期が異なれば評価が異なってしまう場合があった。このように、要約結果に対する評価に再現性がなく、また、公正な評価が困難であるという問題があった。
【０００８】
上記の参考文献１に示された自由作成要約に対する評価処理では、予め用意しておいた正解要約との類似度を単語の頻度ベクトルを用いて判断しているため、内容を示すキーワードの分布さえ類似していれば評価値が高くなってしまう傾向がある。例えば、その要約結果が、正解要約に含まれる単語さえ含んでいれば、文章としての体裁をなさずに非常に読みにくいものであっても一定の良い評価を得ることになってしまう点が問題であった。
【０００９】
したがって、自由作成要約を含めた要約について、評価者の主観に左右されない、再現性のある客観的な評価を行なえるような自動処理が必要である。
【００１０】
ここで、コンピュータで処理された要約結果の評価を、人手によりなされた要約との比較で行うことを考える。一般的にコンピュータでなされた要約結果は、人手によりなされた要約に比べて、要約内容の適切さや文のなめらかさなどの点で要約の精度が低いため、人手による要約と区別ができる程度の自然さしか備えていない場合が多い。「要約結果が良い」とは、その要約結果が専門家の手によりなされた要約と判別が困難な程度に自然なも

のであるということを前提とすれば、コンピュータによりなされた要約結果は、文構造や要約内容などにおいて、人手によりなされた要約に似ているほど良い要約結果であると考えることができる。このことから、「機械による要約」と「人手による要約」という分類先を要約結果の評価として用いることができる。

【００１１】

本発明の目的は、上記の点に鑑み、重要文抽出要約だけでなく自動生成要約であっても、人手によらずに自動的に要約評価処理を行える処理装置と、その処理プログラムと、その処理方法とを提供することである。

【００１２】

さらに、本発明の目的は、要約結果の処理手段の種類を評価の分類先として自動的に要約評価処理を行なえる処理装置と、その処理プログラムと、その処理方法とを提供することである。

【００１３】

【課題を解決するための手段】

上記の目的を達成するため、本発明は、コンピュータを用いた要約自動評価処理装置であって、文章およびその要約結果である問題と前記要約結果に対する評価を示す複数の分類先である解との組からなる解データを記憶する解データ記憶手段と、前記解データの問題である前記文章および前記要約結果から、少なくとも、前記要約結果の文のなめらかさを示す情報および前記要約結果が前記文章の内容を表示しているかどうかを示す情報を含む所定の情報を素性として抽出し、前記解と前記素性の集合との組を生成する解－素性対抽出手段と、前記解と前記素性の集合との組を学習結果として学習結果記憶手段に記憶する機械学習手段と、前記解－素性対抽出手段により抽出される情報を素性とし、入力されたテキストから前記素性の集合を抽出する素性抽出手段と、前記学習結果である前記解と前記素性の集合との組をもとに、ベイズの定理にもとづいて、前記素性抽出手段から得た前記テキストの素性の集合の場合の各分類になる確率を求め、前記確率の値が最も大きい分類を、求める推定解とする評価推定手段とを備える。

【００１４】

また、本発明の要約自動評価処理装置が前記構成である場合に、機械学習手段が、前記解と前記素性の集合との組とを規則とし、前記規則を所定の優先順位によりリストに格納し、前記リストを学習結果として学習結果記憶手段に記憶するものであり、評価推定手段が、前記学習結果である前記リストに格納された前記規則を優先順序の高い順に前記素性抽出手段から得た前記テキストの素性の集合と比較し、素性が一致した規則の分類先を、求める推定解とするものであるように構成されてもよい。

または、機械学習手段が、前記解と前記素性の集合との組から、前記素性の集合が所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布を求め、前記確率分布を学習結果として学習結果記憶手段に記憶するものであり、評価推定手段が、前記学習結果である前記確率分布にもとづいて、前記素性抽出手段から得た前記テキストの素性の集合の場合の各分類の確率を求め、前記確率の値が最も大きい分類を、求める推定解とするものであるように構成されてもよい。

または、機械学習手段が、前記解と前記素性の集合との組を用いて、所定のサポートベクトルマシンモデルの方法により超平面を求め、前記超平面および前記超平面により分割された二つの空間の分類を学習結果として学習結果記憶手段に記憶するものであり、評価推定手段が、前記学習結果である前記超平面をもとに、前記素性抽出手段から得た前記テキストの素性の集合が前記超平面で分割された空間のいずれに属するかを求め、前記素性の集合が属する空間の分類を、求める推定解とするものであるように構成されてもよい。

【００１５】

また、本発明は、前記要約自動評価処理装置で実行する処理をコンピュータに実行させるためのプログラム、または、前記要約自動評価処理装置で実行する処理方法である。

【００１６】

本発明では、機械処理による要約結果および人手による要約結果に解（評価）を付与した

事例を解データとして予め大量に用意しておく。そして、これらの解データの事例ごとに、解と素性の集合との組を抽出し、解と素性の集合との組から、どのような素性のときにどのような解（評価）になりやすいかを機械学習手法により学習する。その後、対象となる要約結果が入力されると、入力した要約結果から素性の集合を取り出して、機械学習の結果を参照して、どのような素性の集合の場合にどのような解（評価）になりやすいかを推定することで、要約結果の評価を行なう。

【００１７】

これにより、評価者の影響を受けない、再現性のある公平な評価を提供することが可能となる。

【００１８】

また、解データの解（評価）として、「機械による要約」と「人手による要約」の２つの分類先を用いることができる。この場合には、本発明では、入力された要約結果が「機械による要約」であるか「人手による要約」であるかを判別する。この２つの分類先は、解データとして用意される要約結果のそのものから自動的に獲得される処理コンピュータにより機械処理により付与されてもよく、また、人手により付与されてもよい。分類先が機械処理で付与されるときは、解を付与する処理負担を軽減することができる。解データの精度を考慮する場合には、専門家により解が付与された解データを用いることもできる。その場合には、３段階や５段階など多段階の評価を行なうために、評価に応じて３つもしくは５つの分類先などを付与することもできる。

【００１９】

なお、本発明の各手段または機能または要素をコンピュータに実行させるためのプログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハードディスクなどの適当な記録媒体に格納することができ、これらの記録媒体に記録して提供され、または、通信インタフェースを介して種々の通信網を利用した送受信により提供されるものである。

【００２０】

【発明の実施の形態】

以下に、本発明の実施の形態を説明する。

【００２１】

図１に、本発明にかかる処理装置の構成例を示す。要約自動評価処理装置１は、解データ記憶部１１と、解－素性対抽出部１２と、機械学習部１３と、学習結果データ記憶部１４と、素性抽出部１５と、評価推定部１６とを持つ。

【００２２】

解データ記憶部１１は、機械学習法を実施する際の教師信号となるデータ（解データ）を記憶する手段である。解データ記憶部１１には、解データとして、問題と解との組である事例が記憶される。

【００２３】

問題は、要約前の文章（テキスト）と要約結果とからなる。要約結果は、機械によるもの、もしくは、人手によりなされたものなどである。

【００２４】

要約結果に対する評価である解は、「機械による要約」または「人手による要約」の２つの分類先とする。２つの分類先は、要約結果から自動的に付与されるようにしてもよいし人手により付与されるようにしてもよい。「機械による要約」または「人手による要約」の２つの分類先を解として用いるのは、要約結果が生成された手段にもとづいて機械的に分類先を付与できるようにするためである。すなわち、コンピュータで自動要約処理された要約結果については自動的に「機械による要約」という解（分類先）が与えられ、人手により生成された要約結果については「人手による要約」という解が与えられた解データを用いる。これにより、解を付与する処理負担が軽減できる。また、解の精度を重視する場合には、専門家の手により解を付与するようにしてもよい。

【００２５】

解－素性対抽出部１２は、解データ記憶部１１に記憶されている事例ごとに、事例の解と素性の集合との組を抽出する手段である。

【００２６】

素性として、１）文のなめらかさを示す情報、２）内容をよく表しているかどうかを示す情報、および、３）自動要約で用いられる特徴的な情報などを抽出する。

【００２７】

１）文のなめらかさを示す情報としては、ｋ－ｇｒａｍ 形態素列のコーパスでの存在、かかりうけ文節間の意味的整合度などを、また、２）内容をよく表しているかどうかを示す情報としては、要約前のテキストにあったキーフレーズの包含率などを、また、３）自動要約で用いられる特徴的な情報としては、その文の位置やリード文かどうか、ＴＦ／ＩＤＦ、文の長さ、固有表現・接続詞・機能語などの手がかり表現の存在などを抽出する。

【００２８】

機械学習部１３は、解－素性対抽出部１２により抽出された解と素性の集合との組から、どのような素性のときにどのような解になりやすいかを教師あり機械学習法により学習する手段である。その学習結果は、学習結果データ記憶部１４に保存される。機械学習部１３は、教師あり機械学習法であればどのような手法で処理を行ってもよい。手法として、例えば、決定木法、サポートベクトル法、パラメータチューニング法、シンプルベイズ法、最大エントロピー法、決定リスト法などがある。

【００２９】

素性抽出部１５は、評価対象の要約２から素性の集合を抽出し、抽出した素性の集合を評価推定部１６へ渡す手段である。

【００３０】

評価推定部１６は、学習結果データ記憶部１４の学習結果データを参照して、素性抽出部１５から渡された素性の集合の場合に、どのような解（評価）になりやすいかを推定し、推定結果である評価３を出力する手段である。

【００３１】

図２に、本発明の処理の流れを示す。なお、要約自動評価処理装置１の解データ記憶部１１には、解データとして、複数の言語のデータに「解」の情報が付与された大量の事例を記憶しておく。

【００３２】

まず、解－素性対抽出部１２は、解データ記憶部１１から、各事例ごとに、解と素性の集合との組を抽出する（ステップＳ１）。例えば、以下のものを素性として抽出する。

【００３３】

素性１：ｋ－ｇｒａｍ の形態素列のコーパスでの存在、
素性２：かかりうけ文節間の意味的整合度、
素性３：ＴＦ／ＩＤＦの値が大きかった「自然言語」の要約後での包含率、
素性４：入力の文章の第一文が用いられているかどうか、
素性５ :出力された要約結果の長さ、
素性６ :接続詞「つまり」が要約抽出箇所の直前にあるかどうか。

【００３４】

解－素性対抽出部１２は、事例ごとに、素性１として、ｋ－ｇｒａｍ 形態素列、例えば「動詞を＜｜＞省略＜｜＞する（＜｜＞は区切りを示す）」で「省略」を省略した「動詞を＜｜＞する」という形態素３ｇｒａｍがコーパスに出現するか否かを調べて、その存在を抽出する。形態素列「動詞を＜｜＞する」がコーパスに出現しないならば、この表現は文としてなめらかでないと推測できるからである。このように、ｋ－ｇｒａｍ 形態素列のコーパスでの存在を素性として利用することで要約のなめらかさを判断できる。

【００３５】

また、解－素性対抽出部１２は、素性２として、例えば「動詞を＜｜＞省略＜｜＞する」で「省略」を省略した「動詞を＜｜＞する」について、「動詞を」の文節が「する」の文節にかかっているものがコーパスにあるか否かを調べ、かかりうけ文節間の意味的整合度

を素性として抽出する。例えば、「動詞を」と「する」にかかりうけがなかった場合には、この表現は文としてなめらかでないと推測できるからである。

【００３６】

また、解－素性対抽出部１２は、素性３として、例えば、要約前のテキストにあったキーフレーズ（自然言語）が要約結果にも含まれるかどうかという、キーフレーズ包含率を抽出する。要約結果にこれらのキーフレーズがなるべく多数含まれている場合には、要約結果がそのテキストの内容をよく表している良い要約と判断できるからである。

【００３７】

キーフレーズの自動抽出処理として、主としてＴＦ／ＩＤＦ法を用いることができる。ＴＦは、文章中でのその語の出現回数もしくは頻度を示す値である。ＩＤＦは、あらかじめ持っている多数の文書群のうち、その語が出現する文書数の逆数である。一般にＴＦとＩＤＦとの積が大きい語ほどキーフレーズとして妥当なものとなる。例えば入力として図３に示すテキストの例があり、キーフレーズが「自然言語」「動詞」「省略」「復元」「表層の表現」「用例」であるとする。これらの語は、このテキストの内容を表現する際に重要な語であるので、要約結果にも出現することが望ましい。解－素性対抽出部１２は、例えばＴＦ／ＩＤＦ法を用いて、上記のようなキーフレーズとなる語を取り出し、ＴＦもしくはＩＤＦの値が高いこれらの語が要約後にも含まれているかどうかを調べ、その包含率を素性として抽出する。

【００３８】

また、解－素性対抽出部１２は、素性４として、入力されたテキストの第一文が用いられているかどうかを素性として抽出する。一般的に重要な文は、文章の初めの方にあることが多いため、文章の初めの方にある文が要約として用いられている場合に良い要約であると判断できるからである。

【００３９】

また、解－素性対抽出部１２は、素性５として、要約結果の長さを調べて素性として抽出する。要約は一般に文を短くすることが目的であるので、要約結果が短いほど良い要約であると判断できるからである。

【００４０】

また、解－素性対抽出部１２は、素性６として、要約結果として抽出した箇所の直前に接続詞「つまり」があるかどうかを素性として抽出する。要約として抽出すると良い文や箇所などを示す接続詞や機能語などの手がかり表現というものがある。例えば、接続詞「つまり」などがあるとき、その接続詞以降は内容をまとめた表現があり、その部分を抽出している場合には良い要約であると判断できるからである。

【００４１】

そして、機械学習部１３は、解－素性対抽出部１２により抽出された、解と上記の素性の集合との組から、どのような素性のときにどのような解（すなわち、「機械による要約」もしくは「人手による要約」）になりやすいかを機械学習法により学習する（ステップＳ２）。機械学習部１３は、教師あり機械学習法として、例えば、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などを用いる。

【００４２】

シンプルベイズ法は、ベイズの定理にもとづいて各分類になる確率を推定し、その確率値が最も大きい分類を求める分類とする方法である。

【００４３】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【００４４】

最大エントロピー法は、あらかじめ設定しておいた素性$f_j$（$1 \leqq j \leqq k$）の集合をＦとするとき、所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確

率分布を求め、その確率分布にしたがって求まる各分類の確率のうち、もっとも大きい確率値を持つ分類を求める分類とする方法である。

【００４５】

サポートベクトルマシン法は、空間を超平面で分割することにより、２つの分類からなるデータを分類する手法である。

【００４６】

決定リスト法および最大エントロピー法については、以下の参考文献２に、サポートベクトルマシン法については、以下の参考文献３および参考文献４に説明されている。

［参考文献２：村田真樹、内山将夫、内元清貴、馬青、井佐原均、種々の機械学習法を用いた多義解消実験、電子情報通信学会言語理解とコミュニケーション研究会, NCL2001-2, (2001)］

［参考文献３：Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)］

［参考文献４：Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/TinySVM/index.html,2000)］その後、素性抽出部１５は、評価を求めたい要約２を入力し（ステップＳ３）、解－素性対抽出部１２での処理とほぼ同様の処理により、入力した要約２から素性の集合を取り出し、それらを評価推定部１６へ渡す（ステップＳ４）。

【００４７】

評価推定部１６は、渡された素性の集合の場合にどのような解になりやすいかを学習結果データ記憶部１４の学習結果データをもとに推定し、推定した解すなわち評価３を出力する（ステップＳ５）。例えば、要約２から抽出した素性の集合にもとづく機械学習法による処理により、要約２の解が「人手による要約」であると判断された場合には、「人手による要約」もしくは「良い要約」などの評価３を出力する。また、要約２の解が「機械による要約」であると判断された場合には、「機械による要約」もしくは「良くない要約」などの評価３を出力する。

【００４８】

以上、本発明をその実施の形態により説明したが、本発明はその主旨の範囲において種々の変形が可能である。例えば、本発明の実施の形態では、解データ記憶部１１で記憶する解データとして「人手による要約」と「機械による要約」との２つの分類を解とする例を説明したが、３以上の分類を解とすることも可能である。

【００４９】

【発明の効果】

本発明によれば、大量の解データを用意して教師あり機械学習法により要約の評価の推定を行う。これにより、自由作成要約についても評価を自動処理することが可能となり、再現性のある公正な評価をすることができるという効果を奏する。

【００５０】

また、本発明により要約結果に対し同等の評価を何回も繰り返すことができ、システムを少しずつ改良するなどのチューンナップを容易に行うことができるという効果を奏する。

【００５１】

また、本発明により同等の評価を再現することができるため、要約処理方法に対する評価の共有化が可能となるという効果を奏する。

【図面の簡単な説明】

【図１】本発明にかかる処理装置の構成例を示す図である。

【図２】本発明の処理の流れを示す図である。

【図３】

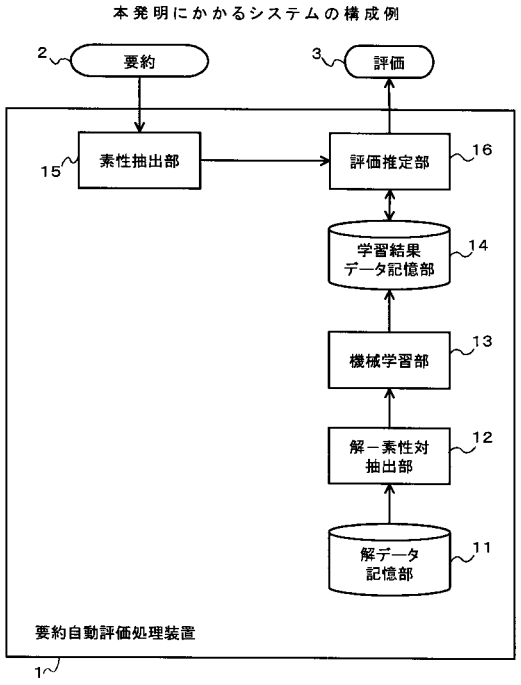
処理対象となるテキストの例を示す図である。

【符号の説明】

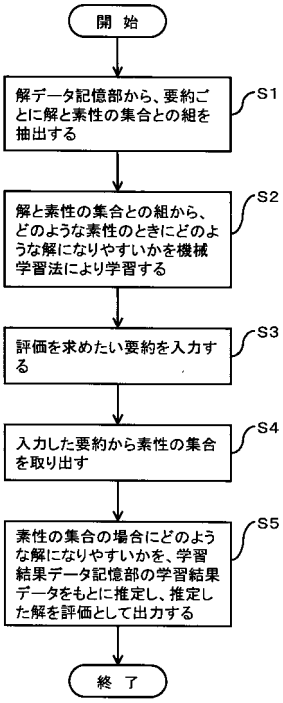
１　要約自動評価処理装置

１１　解データ記憶部
１２　解－素性対抽出部
１３　機械学習部
１４　学習結果データ記憶部
１５　素性抽出部
１６　評価推定部
２　要約
３　評価

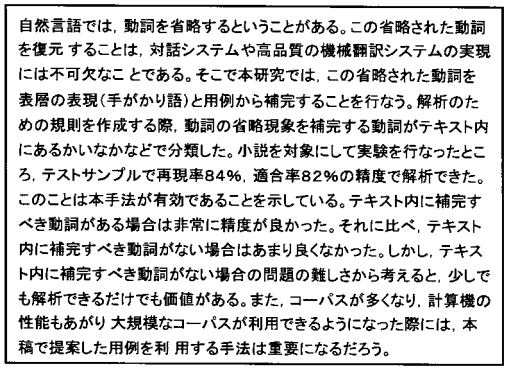
【図１】




【図２】



【図３】


自然言語では，動詞を省略するということがある。この省略された動詞を復元 することは，対話システムや高品質の機械翻訳システムの実現には不可欠なこ とである。そこで本研究では，この省略された動詞を表層の表現（手がかり語）と用例から補完することを行なう。解析のための規則を作成する際，動詞の省略現象を補完する動詞がテキスト内にあるかいなかなどで分類した。小説を対象にして実験を行なったところ，テストサンプルで再現率84%，適合率82%の精度で解析できた。このことは本手法が有効であることを示している。テキスト内に補完すべき動詞がある場合は非常に精度が良かった。それに比べ，テキスト内に補完すべき動詞がない場合はあまり良くなかった。しかし，テキスト内に補完すべき動詞がない場合の問題の難しさから考えると，少しでも解析できるだけでも価値がある。また，コーパスが多くなり，計算機の性能もあがり 大規模なコーパスが利用できるようになった際には，本稿で提案した用例を利 用する手法は重要になるだろう。

フロントページの続き

(56)参考文献 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, 日本, 言語処理学会, １９９９年 7月１０日
難波 英嗣, 奥村 学, 第2回ＮＴＣＩＲワークショップ自動要約タスク（ＴＳＣ）の結果および評価法の分析, 情報処理学会研究報告 Ｖｏｌ．２００１ Ｎｏ．６９, 日本, 社団法人情報処理学会, ２００１年 7月１７日

(58)調査した分野(Int.Cl.[7], ＤＢ名)
G06F 17/30
G06N 3/00
JICSTファイル(JOIS)