自然言語処理 91－6
（1992. 9. 18）

日本語キーワードの自動抽出手法

水野 聡 、島田静雄 、中牟田純 、近藤邦雄 、佐藤 尚

埼玉大学工学部情報工学科

本稿は、法律文や論文中で重要となるキーワードの自動抽出を行なう方法を提案する。自然言語の処理を行なう場合、構文解析を行なうのが一般的である。しかし、構文解析を行なうためには種々の複雑な処理が必要になる。そこで、本研究では構文解析を行なう場合に比べ比較的実現が容易であるという特徴を持った、字種切り法などの字面処理を用いてキーワードの自動抽出を試みている。この方法により、字面処理を用いた場合においても、キーワードの自動抽出においてある程度有効であることがわかった。

Extraction of Meaningful Keywords from Japanese Documents

Satoshi MIZUNO, Shizuo SHIMADA, Jun NAKAMUTA, Kunio KONDO , Hisashi SATO

Department of Information and Computer Sciences
SAITAMA University

Japanese documents have another kind of difficulties to separate words and phrases in sentences in the series of letters mixed with chinese symbolic characters and phonetic letters.The study aims syntactic analysis of natural japanese sentences from an approach extracting meaningful terms until remaining simple semantic structures.The procedure was tested on legal statements such as traffic laws as samples of logically tight statements.Practically good results were obtained in speed and accuracy on the analysis of other statements in technical reports.

1　まえがき

現在、テキスト型データをデータベースに登録する際、原文の特徴を登録する方法としてキーワードが一般に使用されている。本研究の目的は、原文の特徴を効率良く表すキーワード（重要キーワード）を日本語文章中から自動的に抽出することである。また、本研究の特徴として形態素解析を行なわず、字面処理に基づいていることが挙げられる。一般的に、形態素解析を行なわない自然言語処理は、性能的に劣ると考えられている。しかし、形態素解析を行なう場合には大規模な辞書が必要であり、処理も複雑になる。それと比較して字面処理による場合は、処理速度が速い、比較的小規模な処理系で実現できる、などの特徴を持つため本研究では字面処理を採用している。
また、今回報告するキーワードの抽出方法のルールとして使用する日本語文章の特徴と併せて、抽出されたキーワードを日本語文章中から取り除き、残った文章の骨組みを解析することによりコンピュータ上で取り扱いやすい日本語文章の書式を提案していくことが本研究の最終的な目標である。

2　辞書の作成

本研究では、助詞、副詞、接続詞、指示語、接尾語を辞書に登録し使用している。これらの辞書は対象とするテキストとのマッチッングをとるものであり、字面処理用の辞書であるので見出し情報のみを登録し、文法情報や語義情報は登録されていない。

・副詞、接続詞辞書

副詞、接続詞の辞書はＷｎｎの辞書をテキスト形式に変換したものの中から、それぞれの品詞のものを抜き出したテキストファイルである。
副詞辞書に登録されている単語は９３語、接続詞辞書は５４５語である。

・接尾語辞書

接尾語の辞書も、Ｗｎｎの辞書をテキスト形式に変換したものの中から、「前」「中」「内」「外」「等」「上」のそれぞれの文字を含む単語を抜き出す。抜き出された単語から前述の接尾語となる部分を取り除いた文字列をそれぞれの接尾語ごとの辞書に登録したものである。
例：「恒等」の場合は、「恒」を登録する。
接尾語辞書に登録されている語数を以下に示す。
「前」２６語、「中」４４語、「内」４４語
「外」４１語、「等」１８語、「上」５１語

・助詞、指示語辞書

助詞、指示語辞書には以下の単語が登録されている。

・助詞

「を」「が」「は」「の」「と」「で」「に」「や」「から」の９単語

・指示語

「この」「ここ」「その」「それ」「そこ」「あの」「どの」「どこ」の８単語

3　サンプルテキスト

現在、サンプルテキストとして使用している日本語文章は、道路交通法の条文と論文のアブストラクトである。

・論文のアブストラクト

情報処理学会の論文誌、学会誌に掲載された論文のアブストラクトから抜粋し、テキストファイルの形式にしたものである。現在のサンプル数は８個、文字数にすると３０８１文字である。

・道路交通法

ここで法律文を例題としてとりあげた理由は法律文は独特な文体を持っているが、文章構造が論理的な整合性を持つと考えたからである。
今回サンプルテキストとして使用したものは、模範六法[4]の道路交通法の本則部分から図及び表を削除したものをテキストファイルの形式にしたものである。本則は、第一条から第一三二条までの２２２個の条文からなっている。文字数にすると１０万５千文字程度である。

4　キーワードの抽出方法

この章では、キーワードの自動抽出法の手順について説明する。

処理１ 括弧の処理
一般に論文中での括弧の用法は、ただし書きや、注釈、略語などであるため重要キーワードが含まれている可能性が低いと考えられる。したがって、文章内に括弧に括られた部分がある場合はその部分を削除する。

処理２ 副詞、接続詞等の処理
文章を「句読点」、さらに、辞書に登録されている「助詞」、「副詞」、「接続詞」、「指示語」の部分で切る。

処理３ 字種切り法による処理 １
処理２によって分割された文字列の中で「行なえるアルゴリズム」、「文字コードだけ」のように、ひらがなからカタカナに字種が変化するところ、カタカナからひらがなに字種が変化するところで文字列を分割する。

処理４ 字種切り法による処理 ２
処理２・３によって分割された文字列の中で「おく必要」、「よって誘導」のように文字列がひらがなで始まり、なおかつ文字列内に漢字を含むものは漢字部分以降「必要」、「誘導」をキーワードの候補として採用する。

処理５ 形容動詞の処理
処理１～４によって、抽出されたキーワードの候補のなかで「な」を含むものに対して次の処理を行なう。
「な」の前後の一文字がそれぞれ漢字である場合は、この「な」を形容動詞の活用語尾連体形であると断定して「な」の直前にある漢字列を形容動詞の語幹として文字列より削除し、残った文字列部分をあらためて新しいキーワードの候補とする。

処理６ 接尾語の処理
キーワード候補文字列の最後の文字が、「前」「中」「内」「外」「等」「上」のいずれかである場合は、次の処理を行なう。
前述の接尾語の直前の文字がそれぞれの接尾語辞書に登録されていなければ、文字列から接尾語を削除し、登録されていたならばそのままキーワード候補とする。

処理７ 不要キーワードの削除
処理１～６によって抽出されたキーワードのうち１文字になったものはキーワードとして成立しないので削除する。また、文末の表現に良く見られるひらがな列もキーワードとして適さないので、ひらがなのみのキーワード候補も削除する。

処理７’不要キーワードの削除（道路交通法のみ）
この処理は、道路交通法に処理を行なう場合にのみ実行する。道路交通法は法律文という性質上道路交通法の他の条文や他の法律文を参照することが多い。このため「第九条第三項」のような漢字列もキーワードの候補として挙げられてしまう。そこで、このような漢字列をキーワードの候補から外すため、表１に示される漢字のみで構成されているキーワード候補を削除する。

処理８ キーワードの重みつけ
キーワード候補群の中に同じ文字列が含まれる場合は、より文字数の多いキーワードのみを採用し、それ以外のキーワード候補を削除する。さらに、採用されたキーワードの発生頻度に削除されたキーワードの発生頻度を加算する。
（例） キーワード「テキスト」「テキスト断片」があった場合は、「テキスト断片」を採用し、「テキスト」を削除する。さらに、「テキスト」の発生頻度が４、「テキスト断片」の発生頻度が２の場合は「テキスト断片」の頻度を６とする。

※ただし、道路交通法には「車」、「車両」、「自

動車」のように同じ文字列が含まれていても、全く異なった意味を持つ単語が数多く存在するため、道路交通法に処理を行なう場合は処理８は実行しない。

処理９ 重要キーワードの認定
以上の処理により、重みつけをされたキーワードの重みの大きいものを重要キーワードとして採用する。

算用数字、漢数字、刑、民、明、治、大、正、昭、和、平、成、元、本、文、法、律、年、章、節、又、中、前、後、次、同、第、条、項、号、段、各、全

表１：不要キーワードの構成漢字

図１～４に、以上の処理をサンプルテキストに行なった過程を示す。サンプルテキストには、パソコンサーバのフォルトトレラント性実現の一手法. 情報処理学会春季全国大会 ２Ｓ－１ ｐ１９を使用させて頂いた。

5 抽出例

今回報告した手法により処理を行ない抽出されたキーワードの例を以下に示す。（ただし、論文のアブストラクトの場合は重要度の重みが２以上のもの、道路交通法の場合は重み４以上のものに限る）

・論文のアブストラクト
10 誤り検出能力
8 検出誤り
5 辞書
4 単語単位
4 照合し
4 字面処理
4 形態素解析
4 基づく手法
3 評価しなければ
3 程度検出
3 調べた結果
2 明らか
2 必要
2 日本語文章
2 実現性
2 実験
2 我々
2 英語

・道路交通法
19 道路
15 車両
9 運転
8 部分
7 道路標示
6 規定
5 歩行者
5 自転車
5 レール
4 道路標識
4 通行
4 交通
4 区画
4 路面電車

※１ 先頭の数字は、キーワードの重みである。
※２ 論文のアブストラクトの方は、字面処理による日本語文誤り検出の一方式. 情報処理学会全国大会平成４年前期.
のアブストラクトをサンプルテキストとして使用させて頂いたものである。

6 今後の課題

今回、報告した方法では、重要キーワードとしての重みつけを、キーワードの発生頻度によって行なっている。この方法は重要なキーワードほど文章中での発生頻度が高いということを前提としている。しかし、一般的な基本単語（例えば「必

要」、「報告」など）の方が実際には発生頻度が高いこともある。また、発生頻度が低いキーワードの中にも、重要なキーワードが存在する場合もありうる。そこで、今後は発生頻度による重要キーワードとしての重みつけのほかに、文章構造の側面から重要キーワードを認識するルールの調査を行ないたいと考えている。また、現在使用しているルールにはまだ不完全なものもあるため、今後さらにサンプル数を増やしより多くの対象について処理を行い検討することによりさらに強力な処理ルールに変更する必要がある。また、抽出されたキーワードの正解率の定量的な値や他の手法との比較も今後調査する必要がある。

参考文献

[1] 下村、酒井他．字面処理による日本語誤り検出の一方式．情報処理学会春季全国大会 5C－3 p307 （1992）

[2] 岩淵、須田、中田．日本語全文情報の自動索引．情報処理学会春季全国大会 3G－5 p95 （1992）

[3] 榊原、三末．決定木の学習による文書データの分類と日本語キーワードの抽出．情報処理学会人工知能学会研究報告書、82－1 （1992）

[4] 模範六法 平成3年度版、pp．429－459（1990）
三省堂

[5] 小島 和男、やさしい法令用語の解説
公務職員研究協会

[6] 富田 隆行、文法の基礎知識とその教え方
凡人社

キーワード抽出の処理過程

情報処理教育では、プログラミング実習の一環として課題に沿ったプログラムを学生に作成させ、それをレポート形式で提出させることの重要性が指摘されている（1）。このようなレポート受付の媒体・方法は様々あるが、ここでのレポートが通常の実験レポートのように日本語等の自然言語処理で記述されたものでないため、内容を把握することは経験豊かな教員でも困難な場合が少なくない。そこで受付後の処理、すなわちプログラムの動作確認が容易に行なえることを前提とした受付方法をとる必要がある。そのため、筆者等は学生が直接操作するパソコンからＬＡＮを介してレポートサーバにソースプログラムを受付けるシステムを開発し、運用してきた（2）。そのサーバのハードウエアに関するフォルトトレラント性の実現の一手法と、実現したシステムについて報告する。

図１：原文

情報処理教育 プログラミング実習 一環 課題 沿ったプログラム 学生 作成させ レポート形式 提出させる 重要性 指摘されている ようなレポート受付 媒体・方法 様々 レポート 通常 実験レポート よう 日本語等 自然言語処理 記述された ため 内容 把握 経験豊かな教員 困難な場合 少なく 受付後 処理 プログラム 動作確認 容易 行なえる 前提 受付方法 る必要 筆者等 学生 直接操作 パソコン ＬＡＮ 介 レポートサーバ ソースプログラム 受付けるシステム 開発し 運用 きた サーバ ハードウエア 関 フォルトトレラント性 実現 一手法 実現 システム て報告

図２：処理１、２の結果

情報処理教育 プログラミング実習 一環 課題 沿った プログラム 学生 作成させ レポート形式 提出させる 重要性 指摘されている ような レポート受付 媒体・方法 様々 レポート 通常 実験レポート よう 日本語等 自然言語処理 記述された ため 内容 把握 経験豊かな教員 困難な場合 少なく 受付後 処理 プログラム 動作確認 容易 行なえる 前提 受付方法 必要 筆者等 学生 直接操作 パソコン ＬＡＮ 介 レポートサーバ ソースプログラム 受付ける システム 開発し 運用 きた サーバ ハードウエア 関 フォルトトレラント性 実現 一手法 実現 システム 報告

図３：処理３、４（字種切り処理 １、２）の結果

情報処理教育 プログラミング実習 一環 課題 沿った プログラム 学生 作成させ レポート形式 提出させる 重要性 指摘されている ような レポート受付 媒体・方法 様々 レポート 通常 実験レポート よう 日本語 自然言語処理 記述された ため 内容 把握 経験豊かな教員 場合 少なく 受付後 処理 プログラム 動作確認 容易 行なえる 前提 受付方法 必要 筆者 学生 直接操作 パソコン ＬＡＮ 介 レポートサーバ ソースプログラム 受付ける システム 開発し 運用 きた サーバ ハードウエア 関 フォルトトレラント性 実現 一手法 実現 システム 報告

図４：処理４（接尾語処理）、処理５（形容動詞処理）の結果