(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2016-18489
(P2016-18489A)
(43)公開日　平成28年2月1日(2016.2.1)

| (51)Int.Cl. | | F I | | テーマコード（参考） |
|---|---|---|---|---|
| G06F 17/27 | (2006.01) | G06F 17/27 | E | 5B091 |

審査請求 未請求　請求項の数 6　ＯＬ　（全 22 頁）

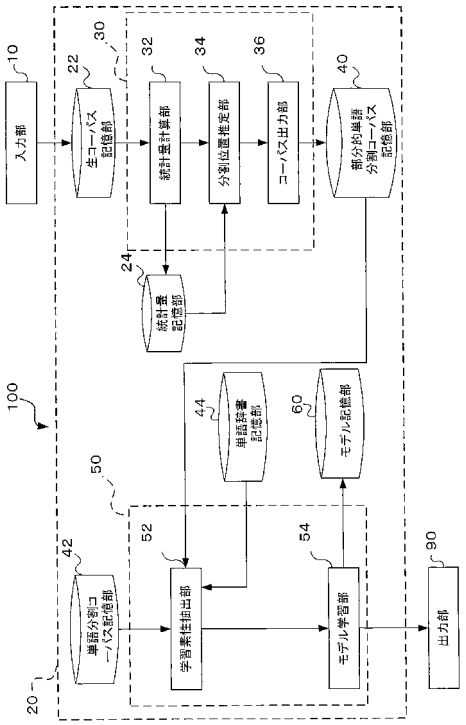
(21)出願番号　特願2014-142404 (P2014-142404)
(22)出願日　平成26年7月10日 (2014.7.10)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(71)出願人　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　須藤　克仁
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　永田　昌明
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】単語分割装置、方法、及びプログラム

(57)【要約】
【課題】対象分野の文字列について精度良く単語分割をすることができる。
【解決手段】分割位置推定部３４により、生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベルを付与し、学習素性抽出部５２により、単語分割コーパスに含まれる文字列の各々に対して、ラベルが付与された文字間の各々についての素性を抽出し、生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出し、モデル学習部５４により、ラベルが付与された文字間の各々についての素性に基づいて、単語分割モデルを学習し、素性抽出部２４２により、入力された文字間の各々についての素性を抽出し、二値分類部２４４により、文字間の各々についての素性と、単語分割モデルとに基づいて、入力された文字列の単語分割する位置を判定する。
【選択図】図２



【特許請求の範囲】
【請求項１】
対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与する分割位置推定部と、
前記対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出する学習素性抽出部と、
前記学習素性抽出部により抽出した、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、前記対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習するモデル学習部と、
入力された前記対象分野の文字列に含まれる文字間の各々についての素性を抽出する素性抽出部と、
前記素性抽出部により抽出した前記文字間の各々についての素性と、前記モデル学習部により学習された前記単語分割モデルとに基づいて、前記入力された前記対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定する二値分類部と、
を含む、単語分割装置。
【請求項２】
前記生コーパスに含まれる文字列に基づいて、部分文字列毎に、前記部分文字列の前後に接続される文字の統計量を計算する統計量計算部を更に含み、
前記分割位置推定部は、前記統計量計算部において前記部分文字列毎に計算された前記部分文字列の前後に接続される文字の統計量に基づいて、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与する請求項１記載の単語分割装置。
【請求項３】
前記分割位置推定部は、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置及び単語分割しない位置を推定し、文字間の各々に、単語分割する位置を示すラベル、単語分割しない位置を示すラベル、及び分割有無不明位置を示すラベルの何れか一つを付与し、
前記学習素性抽出部は、前記単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出する請求項１記載の単語分割装置。
【請求項４】
前記分割位置推定部は、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に、単語分割する位置を示すラベル、及び分割有無不明位置を示すラベルの何れか一方を付与する請求項１記載の単語分割装置。
【請求項５】
分割位置推定部と、学習素性抽出部と、モデル学習部と、素性抽出部と、二値分類部とを含む単語分割装置における、単語分割方法であって、
前記分割位置推定部は、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与し、

前記学習素性抽出部は、前記対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出し、

前記モデル学習部は、前記学習素性抽出部により抽出した、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、前記対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、

前記素性抽出部は、入力された前記対象分野の文字列に含まれる文字間の各々についての素性を抽出し、

前記二値分類部は、前記素性抽出部により抽出した前記文字間の各々についての素性と、前記モデル学習部により学習された前記単語分割モデルとに基づいて、前記入力された前記対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定する

単語分割方法。

【請求項６】

コンピュータを、請求項１～請求項４の何れか１項記載の単語分割装置を構成する各部として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、単語分割装置、方法、及びプログラムに係り、特に、入力された文字列について単語分割するための単語分割装置、方法、及びプログラムに関する。

【背景技術】

【０００２】

日本語や中国語など正書法において単語区切りを明示しない言語を対象とする言語処理システムでは、通常単語分割処理を初期の段階で行い、入力文書あるいは入力文を構成する文字列を単語列に変換する。何をもって単語とするか、という厳格な定義は通常容易でなく、ある種の品詞体系に基づいて単語の単位を定めて利用することが一般的である。近年の言語処理システムではIPAdic、UniDicと呼ばれる辞書で用いられている品詞体系を利用して単語の単位を定めている。そうした単語の定義に基づいて行われる単語分割処理として、近年主流となっているのは、単語分割情報が付与された言語データ（以後、単語分割コーパスとする。）を利用して単語分割のための統計モデルを学習し、その統計モデルに基づいて入力文の単語分割処理を行う方法である（非特許文献１、非特許文献２）。

【０００３】

また、分野適応と呼ばれる技術が知られている。単語分割に対する分野適応の方法としては大きく２種類の方法がある。１つは対象分野の単語分割コーパスを用意し、元の単語分割コーパスと結合して統計モデルを学習する、もしくは学習済みのモデルを追加学習する方法である（非特許文献３）。非特許文献３の技術は単語分割を各文字間が単語の分割位置になるか否かの二値分類の問題として扱い、対象分野の文に対して学習済みの統計モデルを利用して単語分割を行い、分割の確信度が小さい箇所に対して人手で正解を与えることで部分的な単語分割の正解を作成して統計モデルの追加学習を漸進的に行う方法を記載している。もう１つは対象分野の単語分割されていないコーパス（以後、生コーパスとする。）から得られる文字列の統計量を単語分割時の特徴量（以後、素性とする。）として利用する方法である（非特許文献４、非特許文献５）。非特許文献５では、Accessor Variety（非特許文献６）と呼ばれる、ある部分文字列両端に接続する文字の異なり数がその部分文字列が独立した単語らしさを表すことを利用して、Accessor Varietyの値を素性として用い、Accessor Varietyの値が単語分割に貢献する度合いを元分野の単語分割コー

パスから学習する。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】Taku Kudo他，Applying Conditional Random Fields to Japanese Morphological Analysis，Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing，2004
【非特許文献２】丸山宏他，確率的形態素解析，日本ソフトウェア科学会第8回大会論文集，pp.177-180，1991
【非特許文献３】森信介他，点予測による自動単語分割，情報処理学会論文誌Vol.52，No.10，pp. 2944-2952，2011
【非特許文献４】Yiou Wang他，Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data，Proceedings of 5th International Joint Conference on Natural Language Processing，pp 309-317,2011
【非特許文献５】Zhen Guo他，Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation. Proceedings of the 1st CCF Confer- ence on Natural Language Processing & Chinese Computing，pp 121-131，2012.
【非特許文献６】Haodi Feng他，Accessor Variety Criteria for Chinese Word Extraction.Computational Linguistics，volume 30，pp 75-93，2004
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかし、非特許文献１及び非特許文献２の方法においては、単語分割コーパスと類似した文に対しては高い精度で単語分割を行うことができる一方で、異なる対象分野や記述様式（以下、対象分野とする）の文に対しては十分に対応できず単語分割の精度が相対的に低くなる傾向にあるという問題がある。
【０００６】
また、日本語の単語分割においては、単語分割コーパスもしくは外部の辞書から得られる語彙の情報、文字列における漢字、ひらがな、カタカナ、数字といった文字の情報が素性として有効であることが知られているが、カタカナや漢字で構成される長い複合語については、語彙の情報が不足していると単語分割の有効な手がかりが得られず正しく分割することが難しいという問題がある。
【０００７】
また、上記の従来技術において、対象分野の正解データを用いた追加学習による方法では正解データを少量なりとも作成する必要があるし、文字列の統計量を素性として使う方法では生コーパスから得られる文字列の統計量の貢献度合いを元分野の単語分割コーパスから学習するため、生コーパスと元分野の単語分割コーパスとの間の共通部分が少なくなると、貢献度合いの学習が容易でないという問題がある。
【０００８】
また、素性として利用する方式では、新たに単語分割しようとする文中の文字列に対して毎回大規模な生コーパスに基づく統計量の素性を付与する必要があり、単語分割処理の計算時間が増加するという問題がある。
【０００９】
本発明では、上記問題を解決するために成されたものであり、対象分野の文字列について精度良く単語分割することができる単語分割装置、方法、及びプログラムを提供することを目的とする。
【課題を解決するための手段】
【００１０】
上記目的を達成するために、第１の発明に係る単語分割装置は、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字

間の各々に単語分割する位置を示すラベルを付与する分割位置推定部と、前記対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出する学習素性抽出部と、前記学習素性抽出部により抽出した、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、前記対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習するモデル学習部と、入力された前記対象分野の文字列に含まれる文字間の各々についての素性を抽出する素性抽出部と、前記素性抽出部により抽出した前記文字間の各々についての素性と、前記モデル学習部により学習された前記単語分割モデルとに基づいて、前記入力された前記対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定する二値分類部と、を含んで構成されている。

【００１１】

第２の発明に係る単語分割方法は、分割位置推定部と、学習素性抽出部と、モデル学習部と、素性抽出部と、二値分類部とを含む単語分割装置における、単語分割方法であって、前記分割位置推定部は、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与し、前記学習素性抽出部は、前記対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記モデル学習部は、前記学習素性抽出部により抽出した、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、前記対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、前記素性抽出部は、入力された前記対象分野の文字列に含まれる文字間の各々についての素性を抽出し、前記二値分類部は、前記素性抽出部により抽出した前記文字間の各々についての素性と、前記モデル学習部により学習された前記単語分割モデルとに基づいて、前記入力された前記対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定する。

【００１２】

第１及び第２の発明によれば、分割位置推定部により、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与し、学習素性抽出部により、対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出し、モデル学習部により、抽出した、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、素性抽出部により、入力された対象分野の文字列に含まれる文字間の各々についての素性を抽出し、二値分類部は、抽出した文字間の各々についての素性と、学習された前記単語分割モデルとに

基づいて、入力された対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定する。

【００１３】

このように、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベルを付与し、対象分野とは異なる元分野の文字列の集合であって、かつ、文字間の各々に単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性を抽出し、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性に基づいて、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、入力された対象分野の文字列に含まれる文字間の各々についての素性を抽出し、抽出した文字間の各々についての素性と、学習された単語分割モデルとに基づいて、入力された対象分野の文字列に含まれる文字間の各々から、単語分割する位置を判定することにより、対象分野の文字列について精度良く単語分割をすることができる。

【００１４】

また、第１の発明において、前記生コーパスに含まれる文字列に基づいて、部分文字列毎に、前記部分文字列の前後に接続される文字の統計量を計算する統計量計算部を更に含み、前記分割位置推定部は、前記統計量計算部において前記部分文字列毎に計算された前記部分文字列の前後に接続される文字の統計量に基づいて、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に単語分割する位置を示すラベルを付与してもよい。

【００１５】

また、第１の発明において、前記分割位置推定部は、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置及び単語分割しない位置を推定し、文字間の各々に、単語分割する位置を示すラベル、単語分割しない位置を示すラベル、及び分割有無不明位置を示すラベルの何れか一つを付与し、前記学習素性抽出部は、前記単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出してもよい。

【００１６】

また、第１の発明において、前記分割位置推定部は、前記生コーパスに含まれる文字列の各々に対して、単語分割する位置を推定し、文字間の各々に、単語分割する位置を示すラベル、及び分割有無不明位置を示すラベルの何れか一方を付与してもよい。

【００１７】

また、本発明のプログラムは、コンピュータを、上記の単語分割装置を構成する各部として機能させるためのプログラムである。

【発明の効果】

【００１８】

以上説明したように、本発明の単語分割装置、方法、及びプログラムによれば、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベルを付与し、対象分野とは異なる元分野の単語分割コーパスと、生コーパスとに対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、対象分野の文字列について単語分割する位置を判定するための単語分割モ

デルを学習し、学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルに基づいて、対象分野の文字列の単語分割する位置を判定することにより、対象分野の文字列について精度良く単語分割をすることができる。

【図面の簡単な説明】

【００１９】

【図１】本発明の第１の実施の形態に係る単語分割装置の機能的構成を示すブロック図である。

【図２】本発明の第１の実施の形態に係るモデル学習装置の機能的構成を示すブロック図である。

【図３】ある文字間$t_i$の分類に参照する文字の例を示す。

【図４】本発明の第１の実施の形態に係る単語分割判定装置の機能的構成を示すブロック図である。

【図５】本発明の第１の実施の形態に係るモデル学習装置における部分的単語分割コーパス処理ルーチンを示すフローチャート図である。

【図６】本発明の第１の実施の形態に係るモデル学習装置におけるモデル学習処理ルーチンを示すフローチャート図である。

【図７】本発明の第１の実施の形態に係る単語分割判定装置における単語分割判定処理ルーチンを示すフローチャート図である。

【発明を実施するための形態】

【００２０】

以下、図面を参照して本発明の実施の形態を詳細に説明する。

【００２１】

＜本実施の形態の原理＞

まず、本実施の形態における原理について説明する。本実施の形態は、特に日本語の単語分割においてカタカナや漢字で構成される複合語における誤りが多いことに注目してなされるものである。本実施の形態は、大規模な対象分野の生コーパスにおいて単語分割の手がかりとなる文字列の統計量を計算し、生コーパス中で単語境界であると期待できる箇所を自動的に判定し、その箇所を単語分割のための追加学習用データとして利用することを特徴とする。

【００２２】

本実施の形態で利用する単語分割プログラムは、非特許文献３記載のものと同様の、ある文字間が単語境界であるか否かを二値分類で判定するものであるとする。本実施の形態は、単語分割コーパスと、単語分割コーパスとは別の分野である対象分野の生コーパスを利用して統計モデルを学習し、対象分野の入力文字列に対して単語分割処理を施す機能を有する。

【００２３】

次に、本実施の形態において用いる単語分割プログラムについて説明する。単語分割プログラムは、文字間が単語分割する位置であるか否かを、周辺の文字の情報等を素性として二値分類で判定し、その結果に基づいて単語分割した文字列を出力する。素性には非特許文献３記載の、周辺の文字や文字種の情報、周辺の部分文字列が外部の辞書に登録された単語であるか否か、などが利用できる。また、二値分類の方法としては、ロジスティック回帰やサポートベクタマシン(SVM)など既知の統計的手法が適用可能である。二値分類の統計モデルの学習にあたっては、単語分割コーパス、および部分的単語分割コーパスが利用可能である。ここで、単語分割コーパスは文書もしくは文全体に渡って単語分割位置の情報が付与されているもの、部分的単語分割コーパスは単語分割位置の情報が部分的に付与されているものと区別する。二値分類による単語分割方法は、部分的単語分割コーパスによって統計モデルが学習できることが、ごく局所的な情報のみでも統計モデルを強化することができるという面で大きな利点がある。

【００２４】

次に、本実施の形態における生コーパスの利用について説明する。生コーパスから得ら

れる特徴量として、非特許文献６記載のAccessor Varietyや、非特許文献７（Zhihui Jin他. Unsupervised Segmentation of Chinese Text by Use of Branching Entropy, Proceedings of the COLING/ACL 2006, pp 428-435, 2006.）記載のBranching Entropyなどがある。Accessor Varietyはある部分文字列両端に接続する文字の異なり数であり、長さｎの部分文字列ｘ$^{n}_{1}$に対して、下記（１）式で表されるＡＶ（ｘ$^{n}_{1}$）の値となる。

【００２５】

【数１】

$$AV\left(x_1^n\right)=\min\left\{AV_L\left(x_1^n\right),AV_R\left(x_1^n\right)\right\} \quad \cdots\cdot(1)$$

【００２６】

ここで、ＡＶ$_{L}$（ｘ$_{n}$）は生コーパス中で部分文字列ｘ$_{n}$の左側に接続する文字の異なり数、ＡＶ$_{R}$（ｘ$_{n}$）は右側に接続する文字の異なり数である。Branching Entropyはある部分文字列の次に接続する文字のエントロピーであり、長さnの部分文字列ｘ$^{n}_{1}$に対して下記（２）式で表されるＨ（Ｘ｜Ｘ$^{n}_{1}$＝ｘ$^{n}_{1}$）の値である。

【００２７】

【数２】

$$H\left(X\middle|X_1^n=x_1^n\right)=-\sum_{x\in V_x}P\left(x\middle|x_1^n\right)\log P\left(x\middle|x_1^n\right) \quad \cdots\cdot(2)$$

【００２８】

なお、Ｘおよびｘを部分文字列の前に接続する文字と考えることで、部分文字列の前に接続する文字のエントロピーを計算することも可能である。また、下記（３）式のようにBranching Entropyの差分値を見ることもできる。これは、部分文字列ｘ$^{n}_{1}$の次の文字のエントロピーと、ｘ$^{n-1}_{1}$の次の文字(ｘ$^{n}_{1}$の最後の文字に相当)のエントロピーとの差分である。直観的には部分文字列が長い方が次の文字の曖昧性が小さく、エントロピーも小さくなることが予想されるが、ｘ$^{n}_{1}$の直後が単語境界である場合は、続いて出現し得る単語の個数分の曖昧性があるためエントロピーが増加する可能性が高いため、単語境界を見つけるための有力な手がかりとなる（非特許文献７）。

【００２９】

【数３】

$$\Delta H\left(X\middle|X_1^n=x_1^n\right)=H\left(X\middle|X_1^n=x_1^n\right)-H\left(X^-\middle|X_1^{n-1}=x_1^{n-1}\right) \quad \cdots\cdot(3)$$

【００３０】

これらの値が大きい箇所は接続する文字の曖昧性が高い、すなわち前後に様々な単語が接続する単語の境界であることが予想されるため、生コーパスのみを使って単語分割位置

を推定することができる。非特許文献６はAccessor Varietyを利用して中国語の単語を発見することを目的としており、非特許文献７はBranching Entropyを利用して、単語分割コーパスを用いずに単語分割を行うことを目的としている。しかし、これらの統計量による単語分割位置の推定のみですべての単語に対して高精度の単語分割をすることは難しいため、本実施の形態のように単語分割コーパスによる統計モデルの学習と合わせて利用することで、単語分割コーパスと合致する分野では高精度だが対象分野の十分な学習が行えない単語分割コーパスのみによる方法と、対象分野の単語分割がある程度可能だが全体的な単語分割精度が十分でない生コーパスのみによる方法を両立させることが重要である。なお、本実施の形態においては、生コーパスからAccessor Varietyの特徴量を取得する場合を例に説明する。

【００３１】

本実施の形態では、部分的単語分割コーパスによって学習が可能な二値分類による単語分割方法の利点を、単語分割位置が推定可能な大量の生コーパスを利用することで活用する。つまり、分野の異なる単語分割コーパスのみでは十分に適応できなかった対象分野特有の語彙に対する単語分割を、対象分野の生コーパスの中で自動的に推定された単語分割情報を利用して改善することが本実施の形態のアプローチである。上述したように、生コーパスから得られる統計量はそれだけでは高精度の単語分割精度を達成し得ないが、特に長い複合語においては、通常複合語を構成する単語は異なる文脈でそれぞれ利用されているため、前記文字列の統計量に顕著な違いが現れることが期待でき、本実施の形態の目的に好適である。

【００３２】

単語分割位置の推定にあたっては、文字列の統計量に対して閾値を設定して、閾値を超える箇所について「単語分割する位置」を意味するラベルを付し、そうでない箇所については「分割有無不明位置」を意味するラベルを付す。ここでさらに、閾値を下回る箇所、もしくは別のより小さな閾値を下回る箇所について「単語分割しない位置」を意味するラベルを付してもよい。また、元分野の単語分割コーパスとの整合性を保つため、元分野の単語分割コーパスで学習された単語分割プログラムで得られる単語分割有無と一致した箇所のみを保持し、一致しない箇所は「分割有無不明位置」としてもよい。さらに、分野適応における効果が大きいと予想される複合語分割の改善に注力するため、「カタカナとカタカナの間」「漢字と漢字の間」などの特定の対象に絞って「単語分割する位置」「単語分割しない位置」のラベルを付すようにしてもよい。これを生コーパス全体に渡って行うと、「単語分割する位置」「分割有無不明位置」「単語分割しない位置」のラベルが付与されたコーパスが得られる。「分割有無不明位置」のラベルが存在することで、このコーパスは部分的単語分割コーパスとなる。なお、本実施の形態においては、文字列の統計量が閾値を超える箇所について「単語分割する位置」を意味するラベルを付し、第１の閾値以下であって、かつ第２の閾値よりも大きい箇所について「分割有無不明位置」を意味するラベルを付し、第２の閾値以下の箇所について「単語分割しない位置」を意味するラベルを付す。

【００３３】

得られた部分的単語分割コーパスは、単語分割の統計モデルの学習に利用できる。学習に当たっては元の単語分割コーパスと組み合わせて統計モデルを始めから学習し直してもよいし、追加された部分的単語分割コーパスから得られるモデルとの混合モデル（混合比は別途定める）を構成してもよい。学習の方法については公知の単語分割プログラムKyTeaで利用している方式等が利用できる。

【００３４】

＜本発明の第１の実施の形態に係る単語分割装置の構成＞

次に、本発明の第１の実施の形態に係る単語分割装置の構成について説明する。図１に示すように、本発明の第１の実施の形態に係る単語分割装置１は、モデル学習装置１００と、単語分割判定装置２００とを含んで構成されている。

【００３５】

＜本発明の第１の実施の形態に係るモデル学習装置の構成＞

次に、本発明の第１の実施の形態に係るモデル学習装置の構成について説明する。図２に示すように、本発明の第１の実施の形態に係るモデル学習装置１００は、ＣＰＵと、ＲＡＭと、後述する部分的単語分割コーパス処理ルーチン及びモデル学習処理ルーチンを実行するためのプログラムや各種データを記憶したＲＯＭと、を含むコンピュータで構成することが出来る。このモデル学習装置１００は、機能的には図２に示すように入力部１０と、演算部２０と、出力部９０とを備えている。

【００３６】

入力部１０は、対象分野の文字列の集合である生コーパスを受け付け生コーパス記憶部２２に記憶する。

【００３７】

演算部２０は、生コーパス記憶部２２と、統計量記憶部２４と、追加学習コーパス部３０と、部分的単語分割コーパス記憶部４０と、単語分割コーパス記憶部４２と、単語辞書記憶部４４と、統計モデル学習部５０と、モデル記憶部６０とを備えている。

【００３８】

生コーパス記憶部２２には、入力部１０において受け付けた生コーパスが記憶されている。

【００３９】

追加学習コーパス部３０は、生コーパス記憶部２２に記憶されている生コーパスに含まれる部分文字列の各々の前後に接続される文字の統計量に基づいて、部分的単語分割コーパスを取得し、部分的単語分割コーパス記憶部４０に記憶する。また、追加学習コーパス部３０は、統計量計算部３２と、分割位置推定部３４と、コーパス出力部３６とを備えている。

【００４０】

統計量計算部３２は、まず、生コーパス記憶部２２に記憶されている生コーパスに含まれる、長さＮの部分文字列の各々を取得する。ここでは、文字列に、長さＮの窓を走査し、１文字ずつずらしながら長さＮの部分文字列を取得する。このとき、同一の文字列から構成される部分文字列は同一の部分文字列として扱う。次に、統計量計算部３２は、取得された部分文字列の各々について、当該部分文字列の左側に接続する文字、及び右側に接続する文字を取得する。そして、部分文字列の各々について、当該部分文字列の各々について取得された左側に接続する文字、及び右側に接続する文字の各々に基づいて、上記（１）式に従って、部分文字列両端に接続する文字の異なり数であるAccessor Varietyの統計量を計算し、文字列統計量データとして統計量記憶部２４に記憶する。なお、部分文字列と接続する文字の抽出は処理の高速化のために生コーパスを適当なサイズに分割して並列化することが可能であり、部分文字列に対するエントロピーの計算等は処理の高速化のために部分文字列毎に並列化することが可能である。また、部分文字列の長さＮは、非特許文献５や非特許文献７で行われているように、複数のものを並行して利用してもよい。

【００４１】

分割位置推定部３４は、統計量記憶部２４に記憶されている部分文字列毎に計算された当該部分文字列の前後に接続される文字の統計量に基づいて、単語分割する位置を推定し、部分文字列毎に、当該部分文字列の前後の文字間に、分割される位置を示すラベル、分割されない位置を示すラベル、又は分割有無不明位置を示すラベルを付与することにより、生コーパスの全ての文字間の各々にラベルを付与する。具体的には、予め定められた第１の閾値及び第２の閾値（第１の閾値＞第２の閾値）を定めておき、部分文字列の各々について、当該部分文字列の前後に接続される文字の統計量が、予め定められた第１の閾値よりも大きい場合に、分割される位置を示すラベルを付与し、当該部分文字列の前後に接続される文字の統計量が予め定められた第１の閾値以下であり、かつ予め定められた第２の閾値よりも大きい場合に、分割有無不明位置を示すラベルを付与し、当該部分文字列の前後に接続される文字の統計量が予め定められた第２の閾値以下である場合に、分割されない位置を示すラベルを付与する。

【００４２】
コーパス出力部３６は、分割位置推定部３４においてラベルが付与された生コーパスを、部分的単語分割コーパスとして、部分的単語分割コーパス記憶部４０に記憶する。部分的単語分割コーパスの形態としては、例えば、公知の単語分割器KyTeaの部分的単語分割コーパスで利用されている、単語分割される位置は文字間に“|”を、単語分割されない位置は文字間に“-”を、単語分割不明な位置は文字間に空白文字もしくは“?”を、それぞれ挿入した文字列とする。
【００４３】
部分的単語分割コーパス記憶部４０には、コーパス出力部３６において取得された部分的単語分割コーパスを記憶している。
【００４４】
単語分割コーパス記憶部４２には、上記対象分野とは異なる分野である元分野の文字列の集合であって、かつ、文字間の各々の単語分割する位置を示すラベル及び単語分割しない位置を示すラベルが予め付与された単語分割コーパスが記憶されている。
【００４５】
単語辞書記憶部４４には、予め定義された複数の単語の各々からなる単語辞書が記憶されている。
【００４６】
統計モデル学習部５０は、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスと、単語分割コーパス記憶部４２に記憶されている単語分割コーパスと、単語辞書記憶部４４に記憶されている単語辞書と、に基づいて、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習する。また、統計モデル学習部５０は、学習素性抽出部５２と、モデル学習部５４とを備えている。
【００４７】
学習素性抽出部５２は、単語辞書記憶部４４に記憶されている単語辞書に基づいて、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、単語分割コーパス記憶部４２に記憶されている単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出する。例えば、素性として、ある文字間$t_i$については、下記（ａ）～（ｃ）を抽出する（非特許文献３）。図３にある文字間$t_i$の分類に参照する文字を示す。
【００４８】
（ａ）文字ｎ－ｇｒａｍ：文字間の位置ｉの前後の部分文字列であり、窓幅ｍと長さｎのパラメータがある場合、長さ２ｍの文字列$x_{i-m+1}\cdots x_{i-1}x_i x_{i+1}\cdots x_{i+m}$の長さｎのすべての部分文字（文字ｎ－ｇｒａｍ）からなる素性である。
（ｂ）文字種ｎ－ｇｒａｍ：文字間の位置ｉの前後の部分文字列であり、窓幅ｍと長さｎのパラメータがある場合、長さ２ｍの文字列$x_{i-m+1}\cdots x_{i-1}x_i x_{i+1}\cdots x_{i+m}$の長さｎのすべての部分文字（文字ｎ－ｇｒａｍ）に含まれる文字種からなる素性である。ここで、文字種は、漢字、片仮名、平仮名、ローマ字、数字、及びその他の６つである。
（ｃ）単語辞書素性：各長さｋに対する、文字間の左の部分文字列$x_{i-k+1}x_{i-k+2}\cdots x_i$が単語として単語辞書記憶部４４に記憶されている単語辞書に含まれているか否か、文字間の右の部分文字列$x_{i+1}x_{i+2}\cdots x_{i+k}$が単語として単語辞書記憶部４４に記憶されている単語辞書に含まれているか否か、及び文字間をまたぐ部分文字列$x_{i-j+1}x_{i-j+2}\cdots x_{i-j+k}$
【００４９】

【数４】

$$(1 \leq j < k)$$

【００５０】
が単語として単語辞書記憶部４４に記憶されている単語辞書に含まれているか否か、とからなる素性である。
【００５１】
モデル学習部５４は、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスに含まれる文字列の各々の文字間毎のラベル及び学習素性抽出部において抽出された素性と、単語分割コーパス記憶部４２に記憶されている単語分割コーパスに含まれる文字列の各々の文字間毎のラベル及び学習素性抽出部において抽出された素性と、に基づいて、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、出力部９０に出力すると共に、モデル記憶部６０に記憶する。モデルの学習には、ロジスティック回帰又はサポートベクタマシンを用いた分類器の学習を行うＬｉｂＬｉｎｅａｒや、サポートベクタマシンを用いた分類器の学習を行うＳＶＭｌｉｇｈｔなどを利用する。
【００５２】
モデル記憶部６０には、モデル学習部５４において学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルが記憶されている。
【００５３】
＜本発明の第１の実施の形態に係る単語分割判定装置の構成＞
次に、本発明の第１の実施の形態に係る単語分割判定装置の構成について説明する。図４に示すように、本発明の第１の実施の形態に係る単語分割判定装置２００は、ＣＰＵと、ＲＡＭと、後述する単語分割判定処理ルーチンを実行するためのプログラムや各種データを記憶したＲＯＭと、を含むコンピュータで構成することが出来る。この単語分割判定装置２００は、機能的には図４に示すように入力部２１０と、演算部２２０と、出力部２９０とを備えている。
【００５４】
入力部２１０は、対象分野の文字列を受け付ける。
【００５５】
演算部２２０は、単語辞書記憶部２３０と、モデル記憶部２３２と、単語分割処理部２４０と、単語列記憶部２５０と、を備えている。
【００５６】
単語辞書記憶部２３０には、モデル学習装置１００の単語辞書記憶部４４と同一の単語辞書が記憶されている。
【００５７】
モデル記憶部２３２には、モデル学習装置１００のモデル記憶部６０と同一の、対象分野の文字列について単語分割する位置を判定するための単語分割モデルが記憶されている。
【００５８】
単語分割処理部２４０は、入力部２１０において受け付けた、対象分野の文字列について、単語列に分割する。また、単語分割処理部２４０は、素性抽出部２４２と、二値分類部２４４と、データ変換部２４６とを備えている。
【００５９】
素性抽出部２４２は、モデル学習装置１００の学習素性抽出部５２と同様に、入力部２１０において受け付けた文字列の文字間の各々について、単語辞書記憶部２３０に記憶されている単語辞書を用いて、素性を抽出する。
【００６０】
二値分類部２４４は、素性抽出部２４２において抽出した入力部２１０において受け付

けた文字列の文字間の各々の素性と、モデル記憶部２３２に記憶されている対象分野の文字列について単語分割する位置を判定するための単語分割モデルとに基づいて、文字間の各々が単語分割する位置か否かを判定する。なお、単語分割する位置か否かの判定には、ロジスティック回帰やサポートベクタマシンを代表とする公知の様々な分類器を用いることが可能である。

【００６１】

データ変換部２４６は、二値分類部２４４において判定された結果に基づいて、入力部２１０において受け付けた文字列を単語列に分割し、単語列データとして単語列記憶部２５０に記憶すると共に、出力部２９０に出力する。なお、出力として端末やファイルに単語列データを出力する際、典型的には分割する位置となる文字間に空白文字を挿入することで分割位置を表すが、データの形式は特に限定しない。

【００６２】

＜本発明の第１の実施の形態に係るモデル学習装置の作用＞

次に、本発明の第１の実施の形態に係るモデル学習装置１００の作用について説明する。まず、対象分野の文字列の集合である生コーパスを受け付け、生コーパス記憶部２２に記憶する。そして、生コーパス記憶部２２から生コーパスを読み出すと、モデル学習装置１００は、図５に示す部分的単語分割コーパス処理ルーチンを実行する。また、部分的単語分割コーパス処理ルーチンが終了すると、モデル学習装置１００は、図６に示すモデル学習処理ルーチンを実行する。

【００６３】

まず、図５に示す部分的単語分割コーパス処理ルーチンについて説明する。

【００６４】

ステップＳ１０２では、読み込んだ生コーパスに含まれる、文字列を長さＮの部分文字列の各々を取得する。

【００６５】

次に、ステップＳ１０４では、ステップＳ１０２において取得した部分文字列の各々について、左側及び右側に接続する文字を取得する。

【００６６】

次に、ステップＳ１０６では、ステップＳ１０２において取得した部分文字列の各々について、ステップＳ１０４において取得した当該左側及び右側に接続する文字に基づいて、上記（１）式に従って、当該部分文字列の前後に接続される文字の統計量を計算する。

【００６７】

次に、ステップＳ１０８では、ステップＳ１０２において取得した部分文字列の各々について、ステップＳ１０４において取得した当該部分文字列の前後に接続される文字の統計量と、予め定められた第１の閾値及び第２の閾値とに基づいて、当該部分文字列の前後の文字間に、分割される位置を示すラベル、分割されない位置を示すラベル、又は分割有無不明位置を示すラベルを付与することにより、生コーパスの全ての文字間の各々にラベルを付与する。

【００６８】

次に、ステップＳ１１０では、ステップＳ１０８において取得した、文字間の各々にラベルが付与された生コーパスを、部分的単語分割コーパスとして、部分的単語分割コーパス記憶部４０に記憶し、部分的単語分割コーパス処理ルーチンを終了する。

【００６９】

次に、図６に示すモデル学習処理ルーチンについて説明する。

【００７０】

まず、ステップＳ２００では、単語分割コーパス記憶部４２に記憶されている単語分割コーパスを読み込む。

【００７１】

次に、ステップＳ２０２では、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスを読み込む。

【００７２】
次に、ステップＳ２０４では、単語辞書記憶部４４に記憶されている単語辞書を読み込む。
【００７３】
次に、ステップＳ２０６では、ステップＳ２０４において取得した単語辞書に基づいて、ステップＳ２０２において取得した部分的単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、ステップＳ２００において取得した単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出する。
【００７４】
次に、ステップＳ２０８では、ステップＳ２００において取得した単語分割コーパスに含まれる文字列の各々の文字間のラベルと、ステップＳ２０２において取得した部分的単語分割コーパスに含まれる文字列の各々の文字間毎のラベルと、ステップＳ２０６において取得した、単語分割コーパス及び部分単語分割コーパスに含まれる文字列の各々の文字間について取得した素性の各々とに基づいて、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習する。
【００７５】
次に、ステップＳ２１０では、ステップＳ２０８において取得した対象分野の文字列について単語分割する位置を判定するための単語分割モデルを、モデル記憶部６０に記憶すると共に、出力部９０に出力してモデル学習処理ルーチンを終了する。
【００７６】
＜本発明の第１の実施の形態に係る単語分割判定装置の作用＞
次に、本発明の第１の実施の形態に係る単語分割判定装置２００の作用について説明する。まず、入力部２１０から、モデル学習装置１００において学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルが入力され、モデル記憶部２３２に記憶される。そして、処理対象となる文字列を受け付けると、単語分割判定装置２００は、図７に示す単語分割判定処理ルーチンを実行する。
【００７７】
まず、ステップＳ３００では、単語辞書記憶部２３０に記憶されている単語辞書を読み込む。
【００７８】
次に、ステップＳ３０２では、モデル記憶部２３２に記憶されている対象分野の文字列について単語分割する位置を判定するための単語分割モデルを読み込む。
【００７９】
次に、ステップＳ３０４では、入力部２１０において受け付けた文字列の文字間の各々について、ステップＳ２０６と同様に、素性の各々を抽出する。
【００８０】
次に、ステップＳ３０６では、ステップＳ３０４において取得した入力部２１０において受け付けた文字列の文字間の各々の素性の各々と、ステップＳ３０２において取得した対象分野の文字列について単語分割する位置を判定するための単語分割モデルとに基づいて、文字間の各々が単語分割される位置か否かを判定する。
【００８１】
次に、ステップＳ３０８では、ステップＳ３０６において取得した文字列の文字間の各々について判定された結果に基づいて、入力部２１０において受け付けた文字列を単語列に分割し、単語列データとする。
【００８２】
次に、ステップＳ３１０では、ステップＳ３０８において取得した単語列データを、単語列記憶部２５０に記憶すると共に、出力部２９０に出力し、単語分割判定処理ルーチン

を終了する。
【００８３】
＜実験例＞
本実施の形態で説明した手法を用いた実験において、一般的な日本語書き言葉の単語分割コーパスと、大量の日本語特許生コーパスを利用した場合、日本語特許文を単語分割した時の分割精度（Ｆ値）が本実施の形態による分野適応によって９６．１４％から９７．４２％に向上した。つまり、エラー率が３．８６％から２．５８％と約３０％軽減されたことになり、この効果は大きい。なお、分割精度（Ｆ値）の定義を下記（４）式に示す。
【００８４】
【数５】

$$\text{分割精度}\quad F = 2/\left(\frac{1}{A}+\frac{1}{B}\right) \qquad \cdots\cdot(4)$$

$$A:\text{適合率}=\text{正解数}/\text{システムが出した単語数}$$

$$B:\text{再現率}=\text{正解数}/\text{正解データ中の単語数}$$

【００８５】
以上説明したように、本発明の第１の実施の形態に係る単語分割装置によれば、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベル、単語分割しない位置を示すラベル、又は分割有無不明位置を示すラベルを付与して、部分的単語分割とし、対象分野とは異なる元分野の単語分割コーパスと、部分的単語分割コーパスとに対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルに基づいて、対象分野の文字列の単語分割する位置を判定することにより、対象分野の文字列について精度良く単語分割をすることができる。
【００８６】
また、生コーパス中の部分文字列の各々の前後に接続される文字の統計量から、ある部分文字列の文字間が十分に単語分割する位置であると期待できる箇所を自動的に判定し、その結果を単語分割の統計モデルの追加学習データとして利用することで単語分割の分野適応を可能にする。単語分割プログラムが利用する素性に変化はなく、学習データが増加するのみであるので、学習時間の増加は見込まれるものの、単語分割処理自体の時間は大きく変化しないことが期待できる。
【００８７】
また、対象分野の大量の生コーパスを利用することで、利用できる単語分割コーパスが対象分野と異なるものであっても、対象分野の単語分割を精度良く行うことができる。
【００８８】
なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。
【００８９】
例えば、第１の実施の形態において、部分文字列の各々について計算したAccessor Varietyの統計量を文字列統計量データとする場合について説明したが、これに限定されるも

のではなく、部分文字列の各々について、Branching Entropyの統計量を上記（２）式に従って、計算したエントロピーを文字列統計量データとしてもよい。また、部分文字列の各々について、Branching Entropyの差分値を文字列統計量データとしてもよい。この場合、長さＮ－１の部分文字列についても同様に統計量を計算する必要がある。

【００９０】

また、第１の実施の形態においては、部分文字列の前後に接続される文字の統計量に基づいて、分割される位置を示すラベル、分割されない位置を示すラベル、又は分割有無不明位置を示すラベルを付与する場合について説明したが、これに限定されるものではない。例えば、元分野の単語分割コーパスとの整合性を保つ目的で、元分野の単語分割コーパスのみで学習した単語分割器で単語分割した結果と整合する単語分割される／されない位置のみにラベルを付与するようにしてもよい。

【００９１】

また、第１の実施の形態においては、学習素性抽出部において、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出する場合について説明したが、これに限定されるものではない。例えば、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスに含まれる文字列の各々に対して、更に、分割有無不明位置を示すラベルが付与された文字間の各々について素性を抽出してもよい。また、部分的単語分割コーパス記憶部４０に記憶されている部分的単語分割コーパスに含まれる文字列の各々に対して、単語分割する位置を示すラベルが付与された文字間の各々についてのみ素性を抽出してもよい。

【００９２】

また、第１の実施の形態においては、単語分割装置が、モデル学習装置と、単語分割判定装置の２つの装置とから構成される場合について説明したが、これに限定されるものではない。例えば、追加学習コーパス部３０の機能を有する追加学習コーパス装置と、統計モデル学習部５０の機能を有する統計モデル学習装置と、単語分割判定装置の３つの装置とから構成されてもよい。また、単語分割装置に、モデル学習装置、及び単語分割判定装置の機能をもたせ、１つの装置として構成してもよい。

【００９３】

次に、第２の実施の形態に係る単語分割装置について説明する。

【００９４】

第２の実施の形態においては、モデル学習装置１００の、追加学習コーパス部３０における分割位置推定部３４において、部分文字列の前後の文字間に、分割される位置を示すラベル、又は分割有無不明位置を示すラベルを付与する点が第１の実施の形態と異なる。なお、第１の実施の形態に係る単語分割装置１と同様の構成及び作用については、同一の符号を付して説明を省略する。

【００９５】

分割位置推定部３４は、統計量計算部３２において部分文字列毎に計算された当該部分文字列の前後に接続される文字の統計量に基づいて、単語分割する位置を推定し、部分文字列毎に、当該部分文字列の前後の文字間に、分割される位置を示すラベル、又は分割有無不明位置を示すラベルを付与することにより、生コーパスの全ての文字間の各々にラベルを付与する。具体的には、予め閾値を定めておき、部分文字列の各々について、当該部分文字列の前後に接続される文字の統計量が、予め定められた閾値よりも大きい場合に、分割される位置を示すラベルを付与し、当該部分文字列の前後に接続される文字の統計量が予め定められた閾値以下でる場合に、分割有無不明位置を示すラベルを付与する。

【００９６】

コーパス出力部３６は、分割位置推定部３４においてラベルが付与された生コーパスを、部分的単語分割コーパスとして、部分的単語分割コーパス記憶部４０に記憶する。

【００９７】

以上説明したように、本発明の第２の実施の形態に係る単語分割装置によれば、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベル、又は分割有無不明位置を示すラベルを付与して、部分的単語分割コーパスとし、対象分野とは異なる元分野の単語分割コーパスと、部分的単語分割コーパスとに対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルに基づいて、対象分野の文字列の単語分割する位置を判定することにより、対象分野の文字列について精度良く単語分割をすることができる。

【００９８】

次に、第３の実施の形態に係る単語分割装置について説明する。

【００９９】

第３の実施の形態においては、モデル学習装置１００の、追加学習コーパス部３０における分割位置推定部３４において、部分文字列の前後の文字間に、分割される位置を示すラベルのみを付与する点が第１の実施の形態と異なる。なお、第１の実施の形態に係る単語分割装置１と同様の構成及び作用については、同一の符号を付して説明を省略する。

【０１００】

分割位置推定部３４は、統計量計算部３２において部分文字列毎に計算された当該部分文字列の前後に接続される文字の統計量に基づいて、単語分割する位置を推定し、部分文字列毎に、当該部分文字列の前後の文字間に、分割される位置を示すラベルを付与することにより、生コーパスの文字間の各々にラベルを付与する。具体的には、予め閾値を定めておき、部分文字列の各々について、当該部分文字列の前後に接続される文字の統計量が、予め定められた閾値よりも大きい場合に、分割される位置を示すラベルを付与する。

【０１０１】

コーパス出力部３６は、分割位置推定部３４においてラベルが付与された生コーパスを、部分的単語分割コーパスとして、部分的単語分割コーパス記憶部４０に記憶する。

【０１０２】

以上説明したように、本発明の第３の実施の形態に係る単語分割装置によれば、対象分野の文字列の集合である生コーパスに含まれる文字列の各々に対して、文字間の各々に単語分割する位置を示すラベルを付与して、部分的単語分割コーパスとし、対象分野とは異なる元分野の単語分割コーパスと、部分的単語分割コーパスとに対して、単語分割する位置を示すラベルが付与された文字間の各々についての素性、及び単語分割しない位置を示すラベルが付与された文字間の各々についての素性を抽出し、対象分野の文字列について単語分割する位置を判定するための単語分割モデルを学習し、学習された対象分野の文字列について単語分割する位置を判定するための単語分割モデルに基づいて、対象分野の文字列の単語分割する位置を判定することにより、対象分野の文字列について精度良く単語分割をすることができる。

【０１０３】

また、第３の実施の形態においては、統計量の値が大きい箇所は単語分割される位置であることが多い反面、特に短い単語の周辺において単語分割される位置であっても統計量の値が比較的小さいことがあるため、単語分割されない位置の推定精度は必ずしも高くないことを鑑みて、単語分割される位置を示すラベルのみを用いている。

【０１０４】

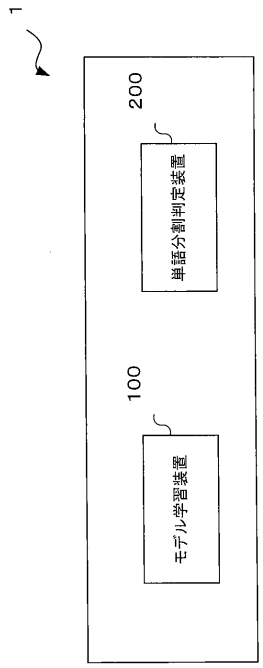
なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

【０１０５】

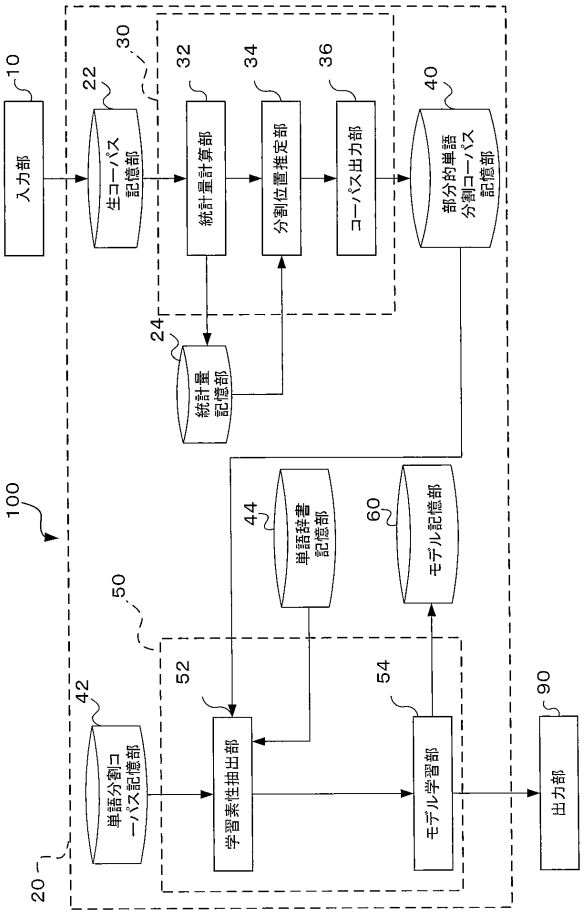
また、本願明細書中において、プログラムが予めインストールされている実施形態として説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能であるし、ネットワークを介して提供することも可能である。

【符号の説明】
【０１０６】
１　　単語分割装置
１０　入力部
２０　演算部
２２　生コーパス記憶部
２４　統計量記憶部
３０　追加学習コーパス部
３２　統計量計算部
３４　分割位置推定部
３６　コーパス出力部
４０　部分的単語分割コーパス記憶部
４２　単語分割コーパス記憶部
４４　単語辞書記憶部
５０　統計モデル学習部
５２　学習素性抽出部
５４　モデル学習部
６０　モデル記憶部
９０　出力部
１００　モデル学習装置
２００　単語分割判定装置
２１０　入力部
２２０　演算部
２３０　単語辞書記憶部
２３２　モデル記憶部
２４０　単語分割処理部
２４２　素性抽出部
２４４　二値分類部
２４６　データ変換部
２５０　単語列記憶部
２９０　出力部

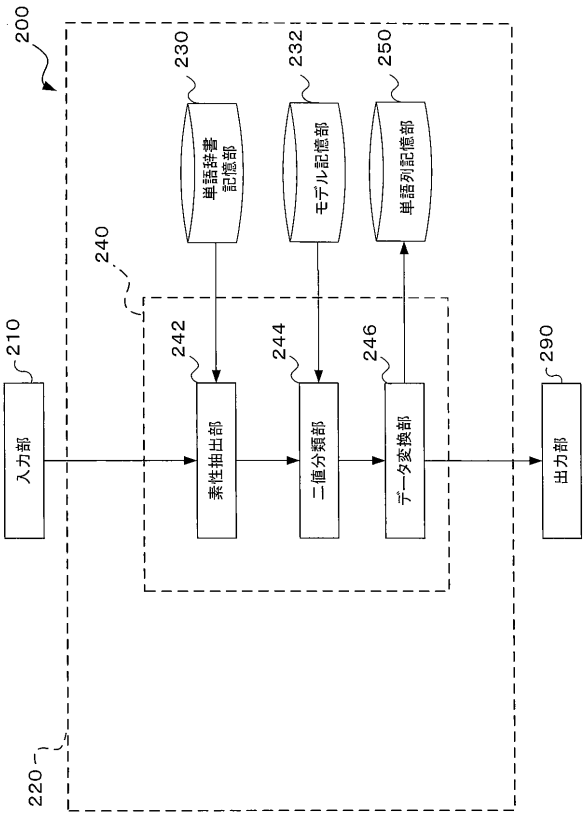
【図１】


1
200
単語分割判定装置
100
モデル学習装置

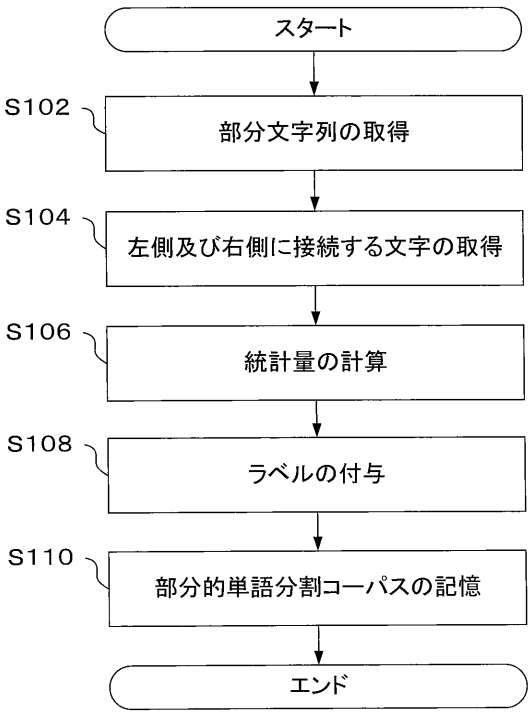
【図２】


10
入力部
22
生コーパス記憶部
30
32
統計量計算部
34
分割位置推定部
36
コーパス出力部
40
部分的単語分割コーパス記憶部
24
統計量記憶部
100
44
単語辞書記憶部
60
モデル記憶部
50
42
単語分割コーパス記憶部
52
学習素性抽出部
54
モデル学習部
90
出力部
20

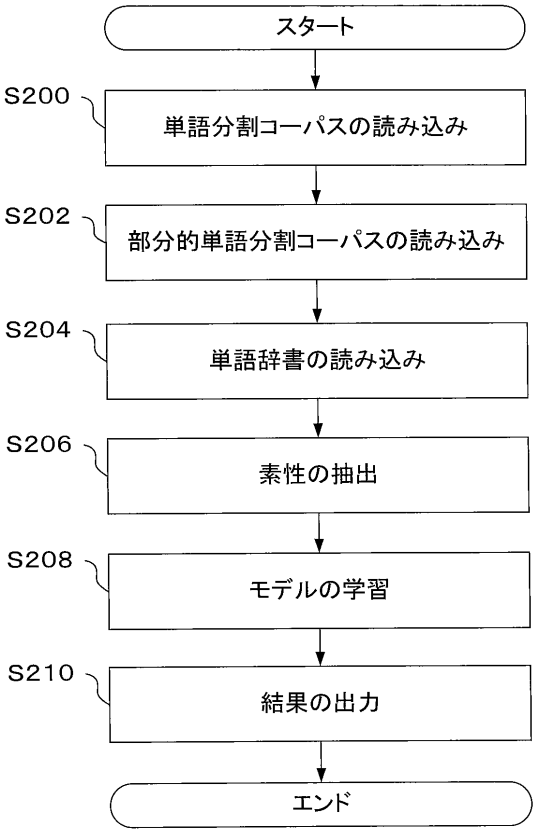
【図４】


200
230
単語辞書記憶部
232
モデル記憶部
250
単語列記憶部
240
210
入力部
242
素性抽出部
244
二値分類部
246
データ変換部
290
出力部
220

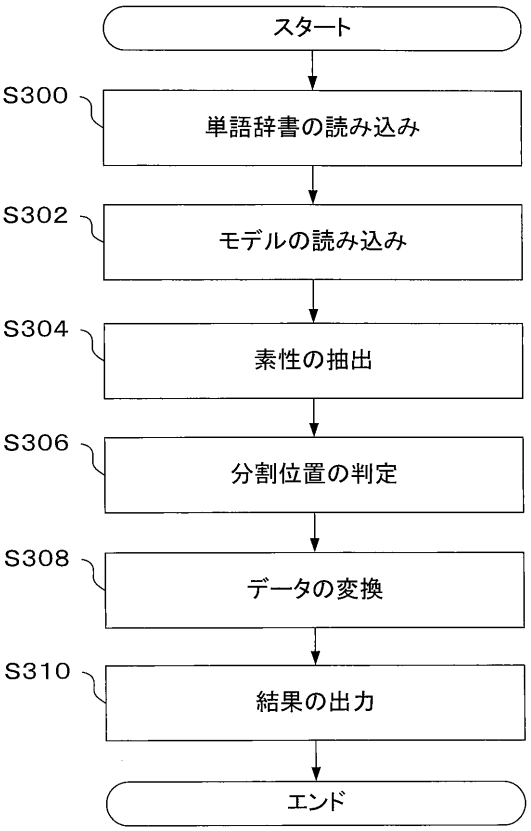
【図５】


スタート
S102
部分文字列の取得
S104
左側及び右側に接続する文字の取得
S106
統計量の計算
S108
ラベルの付与
S110
部分的単語分割コーパスの記憶
エンド

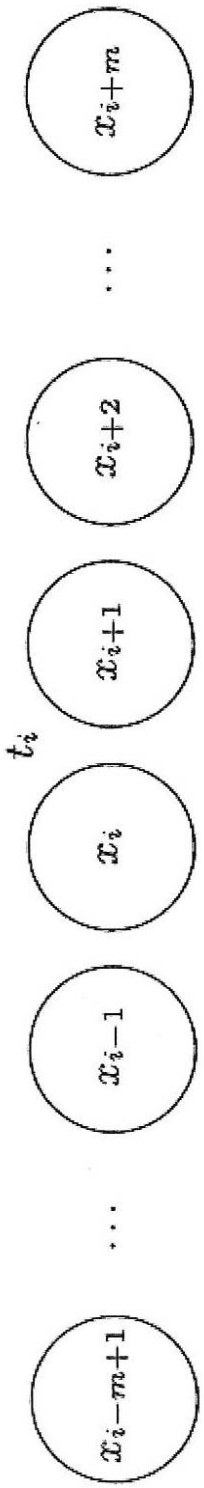
【図６】


スタート
S200
単語分割コーパスの読み込み
S202
部分的単語分割コーパスの読み込み
S204
単語辞書の読み込み
S206
素性の抽出
S208
モデルの学習
S210
結果の出力
エンド

【図７】


スタート
S300
単語辞書の読み込み
S302
モデルの読み込み
S304
素性の抽出
S306
分割位置の判定
S308
データの変換
S310
結果の出力
エンド

【図３】



フロントページの続き

(72)発明者　森　信介
京都府京都市左京区吉田本町３６番地１　国立大学法人京都大学内
Ｆターム(参考) 5B091 CA02　EA01