

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6269953号
(P6269953)

(45) 発行日 平成30年1月31日(2018.1.31)

(24) 登録日 平成30年1月12日(2018.1.12)

(51) Int.Cl. F 1
G 0 6 F 17/27 (2006.01) G 0 6 F 17/27 6 7 0

請求項の数 5 (全 21 頁)

(21) 出願番号	特願2014-142404 (P2014-142404)	(73) 特許権者	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
(22) 出願日	平成26年7月10日(2014.7.10)	(73) 特許権者	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
(65) 公開番号	特開2016-18489 (P2016-18489A)	(74) 代理人	110001519 特許業務法人太陽国際特許事務所
(43) 公開日	平成28年2月1日(2016.2.1)	(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
審査請求日	平成28年7月15日(2016.7.15)	(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) 【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に基づいて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異なり数又は曖昧性に関する統計量に基づいて、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】

前記分割位置推定部は、前記生コーパスに含まれる、特定の文字種で構成さ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、及び単語分割しない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。

10

【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、単語分割する位置を推定し、推定された単語分割する位置のうち、

前記元分野の単語分割コーパスに基づいて学習された、単語分割する位置を判定するための単語分割プログラムを用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た推定結果と一致する位置に、単語分割する位置を示すラベルを付与し、前記推定結果と一致しない位置に、分割有無不明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。

20

【請求項 4】

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に基づいて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異なり数又は曖昧性に関する統計量に基づいて、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、

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、

30

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、

40

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 5】

コンピュータを、請求項 1 ~ 請求項 3 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】

50

【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。何をもちて単語とするか、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と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 Conference 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】

10

20

30

40

50

【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。

10

【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。

20

【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。

30

40

【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

50

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。

10

【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。

20

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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。

50

【 0 0 2 1 】

< 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。

【 0 0 2 2 】

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。

10

【 0 0 2 3 】

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。

20

【 0 0 2 4 】

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_1^n に対して、下記(1)式で表される $AV(x_1^n)$ の値となる。

30

【 0 0 2 5 】

【数1】

$$AV(x_1^n) = \min \{ AV_L(x_1^n), AV_R(x_1^n) \}$$

.....(1)

40

【 0 0 2 6 】

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_1^n に対して下記(2)式で表される $H(X | X_1^n = x_1^n)$ の値である。

【 0 0 2 7 】

【数2】

$$H\left(X\left|X_1^n = x_1^n\right.\right) = -\sum_{x \in V_x} P\left(x\left|x_1^n\right.\right) \log P\left(x\left|x_1^n\right.\right)$$

.....(2)

10

【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が出来る。これは、部分文字列 x_1^n の次の文字のエントロピーと、 x_1^{n-1} の次の文字(x_1^n 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_1^n 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】

20

【数3】

$$\Delta H\left(X\left|X_1^n = x_1^n\right.\right) = H\left(X\left|X_1^n = x_1^n\right.\right) - H\left(X\left|X_1^{n-1} = x_1^{n-1}\right.\right)$$

.....(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。

30

40

【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

50

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くと、「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。

10

【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。学習にあ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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

30

【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。

40

【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

50

単語分割コーパスと、単語分割コーパス記憶部 4 2 に記憶されている単語分割コーパスと、単語辞書記憶部 4 4 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 5 0 は、学習素性抽出部 5 2 と、モデル学習部 5 4 とを備えている。

【 0 0 4 7 】

学習素性抽出部 5 2 は、単語辞書記憶部 4 4 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 4 0 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 4 2 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。

10

【 0 0 4 8 】

(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 $x_{i-m+1} \dots x_{i-1} x_i x_{i+1} \dots x_{i+m}$ 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 $x_{i-m+1} \dots x_{i-1} x_i x_{i+1} \dots x_{i+m}$ 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。

20

(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 $x_{i-k+1} x_{i-k+2} \dots x_i$ が単語として単語辞書記憶部 4 4 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 $x_{i+1} x_{i+2} \dots x_{i+k}$ が単語として単語辞書記憶部 4 4 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 $x_{i-j+1} x_{i-j+2} \dots x_{i-j+k}$

【 0 0 4 9 】

【 数 4 】

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

30

【 0 0 5 0 】

が単語として単語辞書記憶部 4 4 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。

【 0 0 5 1 】

モデル学習部 5 4 は、部分的単語分割コーパス記憶部 4 0 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 4 2 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 9 0 に出力すると共に、モデル記憶部 6 0 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 $Li b L i n e a r$ や、サポートベクタマシンを用いた分類器の学習を行う $S V M l i g h t$ などを利用する。

40

【 0 0 5 2 】

モデル記憶部 6 0 には、モデル学習部 5 4 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。

【 0 0 5 3 】

50

< 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。

10

20

30

40

50

【 0 0 6 4 】

ステップ S 1 0 2 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。

【 0 0 6 5 】

次に、ステップ S 1 0 4 では、ステップ S 1 0 2 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。

【 0 0 6 6 】

次に、ステップ S 1 0 6 では、ステップ S 1 0 2 において取得した部分文字列の各々について、ステップ S 1 0 4 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。

10

【 0 0 6 7 】

次に、ステップ S 1 0 8 では、ステップ S 1 0 2 において取得した部分文字列の各々について、ステップ S 1 0 4 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。

【 0 0 6 8 】

次に、ステップ S 1 1 0 では、ステップ S 1 0 8 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 4 0 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。

20

【 0 0 6 9 】

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。

【 0 0 7 0 】

まず、ステップ S 2 0 0 では、単語分割コーパス記憶部 4 2 に記憶されている単語分割コーパスを読み込む。

【 0 0 7 1 】

次に、ステップ S 2 0 2 では、部分的単語分割コーパス記憶部 4 0 に記憶されている部分的単語分割コーパスを読み込む。

【 0 0 7 2 】

次に、ステップ S 2 0 4 では、単語辞書記憶部 4 4 に記憶されている単語辞書を読み込む。

30

【 0 0 7 3 】

次に、ステップ S 2 0 6 では、ステップ S 2 0 4 において取得した単語辞書に基づいて、ステップ S 2 0 2 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 S 2 0 0 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。

40

【 0 0 7 4 】

次に、ステップ S 2 0 8 では、ステップ S 2 0 0 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 S 2 0 2 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 S 2 0 6 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。

【 0 0 7 5 】

次に、ステップ S 2 1 0 では、ステップ S 2 0 8 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 6 0 に記憶す

50

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。

10

【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。

20

【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。

30

【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】

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

・・・(4)

A : 適合率 = 正解数 / システムが出した単語数

10

B : 再現率 = 正解数 / 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。

20

【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变化はなく、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。

30

【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。

40

【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

50

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。

【 0 0 9 1 】

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 4 0 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 4 0 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 4 0 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。

10

【 0 0 9 2 】

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 3 0 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 5 0 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。

20

【 0 0 9 3 】

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。

【 0 0 9 4 】

第 2 の実施の形態においては、モデル学習装置 1 0 0 の、追加学習コーパス部 3 0 における分割位置推定部 3 4 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。

【 0 0 9 5 】

分割位置推定部 3 4 は、統計量計算部 3 2 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ある場合に、分割有無不明位置を示すラベルを付与する。

30

【 0 0 9 6 】

コーパス出力部 3 6 は、分割位置推定部 3 4 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 4 0 に記憶する。

40

【 0 0 9 7 】

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

50

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。

【 0 0 9 8 】

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。

【 0 0 9 9 】

第 3 の実施の形態においては、モデル学習装置 1 0 0 の、追加学習コーパス部 3 0 における分割位置推定部 3 4 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。

【 0 1 0 0 】

分割位置推定部 3 4 は、統計量計算部 3 2 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。

【 0 1 0 1 】

コーパス出力部 3 6 は、分割位置推定部 3 4 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 4 0 に記憶する。

【 0 1 0 2 】

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。

【 0 1 0 3 】

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。

【 0 1 0 4 】

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。

【 0 1 0 5 】

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 】

【 0 1 0 6 】

- 1 単語分割装置
- 1 0 入力部
- 2 0 演算部
- 2 2 生コーパス記憶部
- 2 4 統計量記憶部
- 3 0 追加学習コーパス部
- 3 2 統計量計算部

10

20

30

40

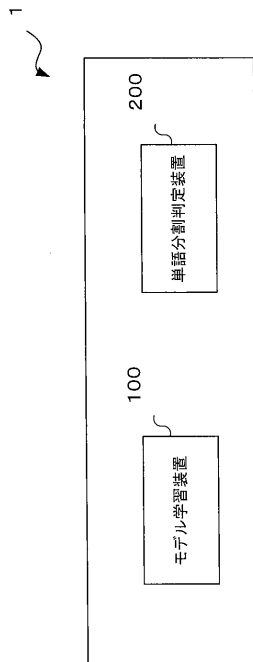
50

- 3 4 分割位置推定部
- 3 6 コーパス出力部
- 4 0 部分的単語分割コーパス記憶部
- 4 2 単語分割コーパス記憶部
- 4 4 単語辞書記憶部
- 5 0 統計モデル学習部
- 5 2 学習素性抽出部
- 5 4 モデル学習部
- 6 0 モデル記憶部
- 9 0 出力部
- 1 0 0 モデル学習装置
- 2 0 0 単語分割判定装置
- 2 1 0 入力部
- 2 2 0 演算部
- 2 3 0 単語辞書記憶部
- 2 3 2 モデル記憶部
- 2 4 0 単語分割処理部
- 2 4 2 素性抽出部
- 2 4 4 二値分類部
- 2 4 6 データ変換部
- 2 5 0 単語列記憶部
- 2 9 0 出力部

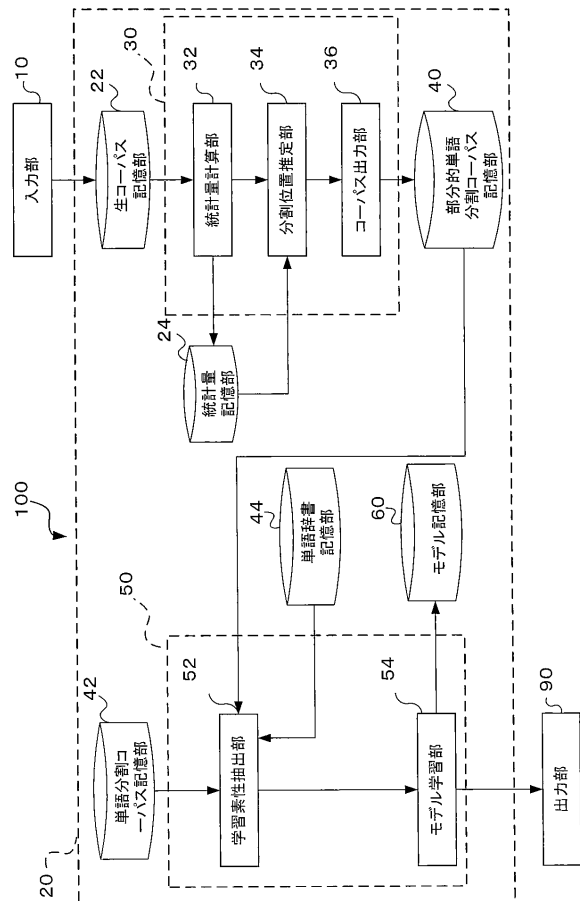
10

20

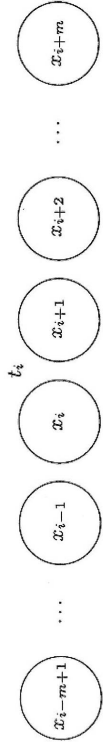
【図1】



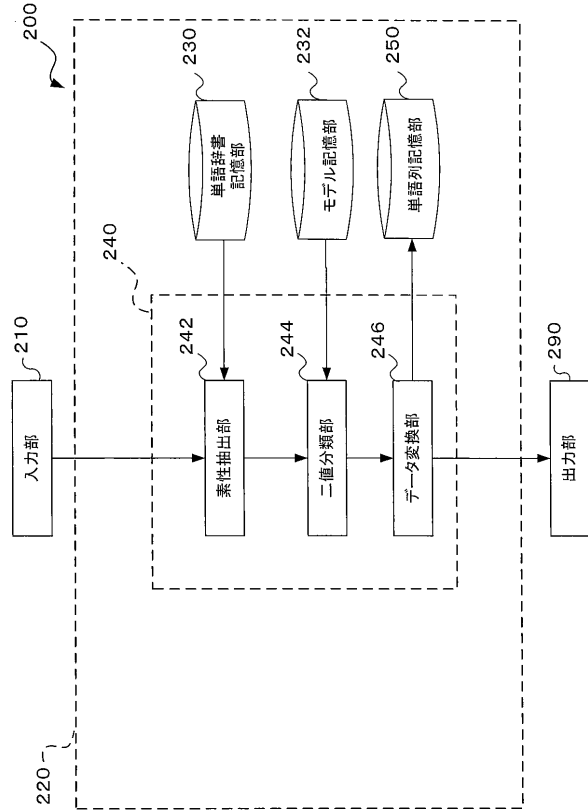
【図2】



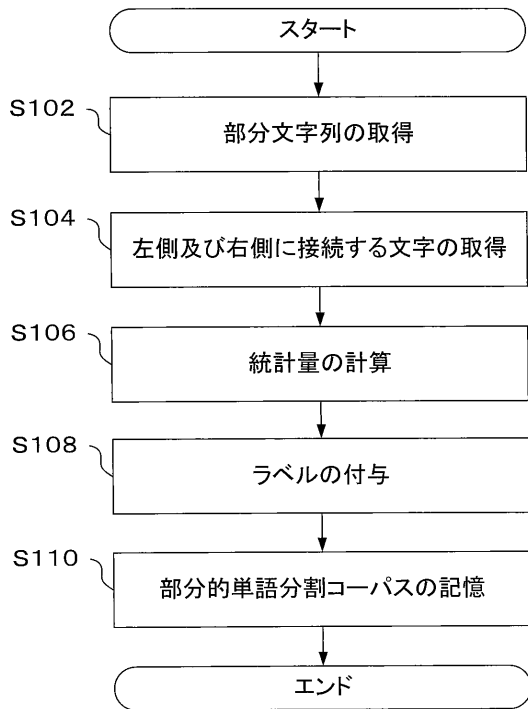
【図3】



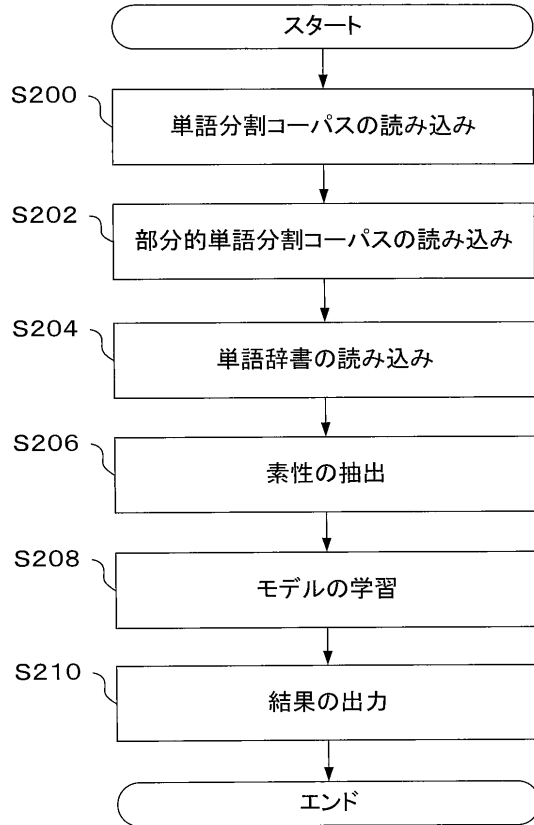
【図4】



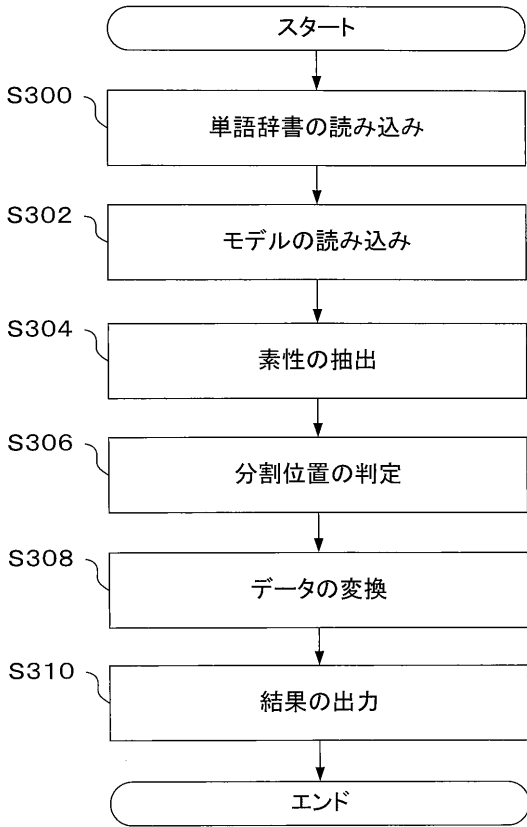
【図5】



【図6】



【図7】



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3番地1 国立大学法人京都大学内

審査官 長 由紀子

(56)参考文献 特開2001-249922(JP,A)

特開平09-270917(JP,A)

特開2006-031295(JP,A)

森 信介 外3名, 点予測による形態素解析, 自然言語処理, 日本, 言語処理学会, 2011年9月30日, 第18巻第4号, p.367-381

森 信介 外2名, 点予測による自動単語分割, 情報処理学会論文誌 論文誌ジャーナル Vol.52 No.10 [CD-ROM], 日本, 一般社団法人情報処理学会, 2011年10月15日, 第52巻第10号, p.2944-2952

森 信介 外1名, 3種類の辞書による自動単語分割の精度向上, 情報処理学会研究報告 平成21年度3 [CD-ROM], 日本, 社団法人情報処理学会, 2009年10月15日, p.1-6

森 信介 外1名, 擬似確率的単語分割コーパスによる言語モデルの改良, 自然言語処理, 日本, 言語処理学会, 2009年10月10日, 第16巻第5号, p.7-21

坪井 祐太 外4名, 日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習, 情報処理学会論文誌 論文誌ジャーナル Vol.50 No.6 [CD-ROM], 日本, 社団法人情報処理学会, 2009年6月15日, 第50巻第6号, p.1622-1635

坪井 祐太 外4名, 部分的かつ曖昧なラベル付き構造データからのマルコフ条件付確率場の学習, 情報処理学会研究報告, 日本, 社団法人情報処理学会 Information Processing Society of Japan, 2007年11月19日, 第2007巻第113号, p.67-74

(58)調査した分野(Int.Cl., DB名)

G06F 17/20 - 28