様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年 １０ 月 ２７ 日現在

機関番号：２２６０４
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ１３１７３

研究課題名（和文）アクセント情報付き大規模単語データベースの構築

研究課題名（英文）Construction of a large word database with accent information

研究代表者
岡　照晃（Oka, Teruaki）

東京都立大学・システムデザイン研究科・特任助教

研究者番号：５０７８２９４２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：クラウドソーシングを利用し、形態素解析用電子化辞書UniDicへアクセント情報の付与を行なった。作業者は不特定多数の非専門家であるため、アクセントを特定したい単語だけでなく、それに連接してアクセントを変化させないような後続語を同時に提示し、音声合成による発話の中からなじみのあるアクセントを選択するタスクを設定した。すでにアクセントが既知の単語をgoldとして使用するフィルタリングと、ベイズ推定による作業者と各設問のレベル推定を用いて、重み付き多数決によるアクセント情報付与を実施した。フィルタリングとタスクに対する作業者レベルの予測から居住地の違いに影響されない大規模なアクセント付与を実現した。

研究成果の学術的意義や社会的意義
単語へのアクセント付与作業は、居住地や出身地の影響を受けるため、非専門家には難しく、大規模な実施は困難だった。クラウドソーシングの普及とともに発展した設問や作業者のレベル推定手法を使うことで、専門家を時間的空間的に拘束することのないアクセント付与のフローを実現した。

研究成果の概要（英文）：Crowdsourcing was used to add accent information to UniDic, an electronic dictionary for morphological analysis. Since the participants were unspecified non-specialists, we set them the task of selecting familiar accents from speech synthesized by simultaneously presenting not only the word whose accent they wanted to identify, but also its successor words that would not change the accent of the word. Filtering was performed using words with known accents as gold, and Bayesian level estimation of the worker and each question was used to assign accent information by weighted majority voting. The filtering and the prediction of the worker's level for the task resulted in large-scale accent assignment that was not affected by differences in place of residence.

研究分野：自然言語処理

キーワード：アクセント　形態素解析辞書

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

日本語において「何を『単語』と見なすか？」は自明でなく、日本語母語話者の間でも、皆が納得するような明確な『単語』の定義はない。しかし語数調査や語彙表などを作成する際には何かしら『単語』の定義（言語単位の設定）を行わなければ、作業自体を開始できない。国立国語研究所では図１の３種類の斉一で階層的な言語単位を定義し、形態論情報付きコーパスや、語彙表、語数表の作成を行なっている。短単位は用例収集を目的とした言語単位である。３単位中最も斉一に規定され、１単位当たりの字数も短い（少ない）ため、用例を広く多く集めることに向いている。一方、長単位は文節を自立部と付属部に分け、自立部内の短単位をまとめ上げた言語単位である。短単位では捉え難い複合語を扱うことができる。中単位は音声の研究に向けて設計された言語単位であり、短単位と長単位の中間に位置する。

国語研では、生テキストデータに対して短単位のリッチな形態論情報（e.g.，単位境界、品詞、活用、語種、仮名表記…）を自動付与する手段として、短単位自動解析用辞書 UniDic を配布している。形態論情報を自動付与する日本語の電子化辞書は複数あるが[3]、アクセント情報を持つのは UniDic のみであり、UniDic の大きな特徴といえる。そのため近年ではスマートスピーカーの発する音声の機械生成（音声合成）にも利用されている。しかしこうした社会からの需要に対し、UniDic へのアクセント情報付与作業は現在休止状態であり、再開の目途もなく、アクセント情報の付与されていない約２万の短単位を抱えたまま（図２）見出し語の追加だけが現在も継続して行われている。

図１　斉一で階層的な言語単位

長単位	固有名詞仮名表記				を	調査し		た
中単位	固有名詞		仮名表記		を	調査し		た
短単位	固有	名詞	仮名	表記	を	調査	し	た

図２　UniDic中でアクセント情報が付与されていない短単位数

名詞	4,637
接尾辞	1,356
接頭辞	384
動詞	10,038
助動詞	1,297
形容詞	932
副詞	282
計	18,926

２．研究の目的

日本語のコーパス言語学研究の一環として、アクセント情報付き短単位辞書の構築を行う。各短単位のアクセント情報の決定には、通常ならば専門家を集め、時間をかけた協議が必要になる。そのため短単位辞書 UniDic では、この作業が現在休止状態にある。本研究の目的は、UniDic 短単位へアクセント情報を網羅的に付与することである。

３．研究の方法

そこで１つの短単位に対し、取りうるアクセントの情報を正誤問わず網羅的に付与、それぞれの発音を音声合成で機械生成する。これによりアクセント情報付与は「並べられた発音から最も自然に聴こえるものを１つ選ぶ」作業に簡単化できる。この作業を、クラウドソーシングを使って

大規模かつ効率的に実施する。【活用】アクセント情報付き短単位辞書を作ることで、短単位の上位単位で、音声の研究に向けて設計された中単位の自動解析が可能になる。

４．研究成果

クラウドソーシングを利用し、形態素解析用電子化辞書 UniDic へアクセント情報の付与を行なった。作業者は不特定多数の非専門家であるため、アクセントを特定したい単語だけでなく、それに連接してアクセントを変化させないような後続語を同時に提示し、音声合成による発話の中からなじみのあるアクセントを選択するタスクを設定した。すでにアクセントが既知の単語を gold として使用するフィルタリングと、ベイズ推定による作業者と各設問のレベル推定を用いて、重み付き多数決によるアクセント情報付与を実施した。フィルタリングとタスクに対する作業者レベルの予測から居住地の違いに影響されない大規模なアクセント付与を実現した。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Teruaki Oka, Yuichi Ishimoto, Yutaka Yagi, Takenori Nakamura, Masayuki Asahara, Kikuo Maekawa, Toshinobu Ogiso, Hanae Koiso, Kumiko Sakoda and Nobuko Kibe
２．発表標題 KOTONOHA: A Corpus Concordance System for Skewer-Searching NINJAL Corpora
３．学会等名 12th Edition of its Language Resources and Evaluation Conference (LREC2020)（国際学会）
４．発表年 2020年

１．発表者名 久本 空海，山村 崇，勝田 哲弘，竹林佑斗，高岡 一馬，内田 佳孝，岡 照晃，浅原 正幸
２．発表標題 chiVe: 製品利用可能な日本語単語ベクトル資源の実現へ向けて ～ 形態素解析器Sudachiと超大規模ウェブコーパスNWJCによる分散表現の獲得と改良 ～
３．学会等名 第16回テキストアナリティクス・シンポジウム
４．発表年 2020年

１．発表者名 岡 照晃
２．発表標題 クラウドソーシングによる形態論情報付与付き辞書整備
３．学会等名 日本言語学会第158回大会
４．発表年 2019年

１．発表者名 岡 照晃
２．発表標題 UniDic非コアデータ : 解析用UniDicのID情報にひも付く追加情報の公開について
３．学会等名 言語資源活用ワークショップ2019（LRW2019）
４．発表年 2019年

１．発表者名 Teruaki Oka
２．発表標題 New words in Japanese and the design of UniDic electronic dictionary
３．学会等名 Globalex Workshop on Lexicography and Neologism 2019 (GWLN 2019)（国際学会）
４．発表年 2019年

１．発表者名 河村宗一郎，久本空海，真鍋陽俊，高岡一馬，内田佳孝，岡照晃，浅原正幸
２．発表標題 chiVe 2.0: SudachiとNWJCを用いた実用的な日本語単語ベクトルの実現へ向けて
３．学会等名 言語処理学会第26回年次大会(NLP2020)
４．発表年 2020年

〔図書〕　計2件

１．著者名 沖森 卓也	４．発行年 2021年
２．出版社 朝倉書店	５．総ページ数 560
３．書名 日本語文法百科	

１．著者名 村上征勝、金明哲（同志社大学教授）、小木曽智信（国立国語研究所教授）、中園聡（鹿児島国際大学教授）、矢野桂司（立命館大学教授）、赤間亮（立命館大学教授）、阪田真己子（同志社大学教授）、宝珍輝尚（京都工芸繊維大学教授）、芳沢光雄（桜美林大学教授）、渡辺美智子（慶應義塾大学教授）、足立浩平（大阪大学教授）	４．発行年 2019年
２．出版社 勉誠出版	５．総ページ数 850
３．書名 文化情報学事典	

〔産業財産権〕

〔その他〕

東京都立大学　自然言語処理研究室　研究発表
https://cl.sd.tmu.ac.jp/research/publications
「UniDic」国語研短単位自動解析用辞書
https://unidic.ninjal.ac.jp/
UniDic非コアデータ
https://teru-oka-1933.github.io/unidic_non_core/
短単位自動解析用辞書『suwad』
https://teruaki-oka.com/madic/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関