3M3-OS-07d-2

PrivateCrowdSourcingを用いた言語、音声資源の収集 ～システムの構築と言語収集～

Collecting new vocabularies and sound resources using Private CrowdSourcing System

芦川　将之 *1　　有賀　康顕 *1　　宮村　祐一 *1
Masayuki ASHIKAWA　Michiaki ARIGA　Yuichi MIYAMURA

株式会社東芝 研究開発センター
Corporate Research and Development Center, Toshiba Corporation

In speech recognition and text-to-speech, a large dictionary is indispensable. It is hard to for conventional methods of word extraction, part-of-speech tagging and pronunciation prediction to create a high-quality dictionary. To create a large-scale and high-quality dictionary, this paper presents a method of aquiring unknown words using the private CrowdSourcing system.

1. はじめに

近年、音声でユーザと対話を行うサービスが開始されるなど、音声処理技術の適用先が拡大している。音声を認識・合成するためには、テキストに対して読みやアクセントを推定する必要があるため、形態素解析をはじめとしたテキスト解析技術が用いられる。

形態素解析等のテキスト解析技術では辞書と教師ラベル付きコーパスを用いて統計モデルを学習する手法が一般的であり、解析性能向上のために大規模な辞書やコーパスが求められている。

従来、コーパスから自動的に未知語を獲得する手法が研究されている [Mori 1996][村脇 2010][羽鳥 2011]。これらの研究を組み合わせて用いることで、単語の表記文字列、品詞、読みなどを自動獲得することができる。しかし、例えば読み推定精度は 90 %程度であるなど、自動獲得のみではそのまま辞書として利用できるレベルに至っていない。 そこで我々は高品質な言語資源を獲得するために Private CrowdSourcing System(PCSS) を開発してきた [芦川 2012]。

本報では、音声合成や音声認識に必要な単語表記、品詞、読み、アクセント、からなる「語彙」の獲得を目的として、PCSSを活用することで、大量のウェブテキストコーパスから精度よく語彙を獲得する手法を提案する。

2. Private CrowdSourcing System

2.1 Private CrowdSourcing System の必要性

CrowdSourcing とは単純な作業ではあるが自動化することが困難な作業を、不特定多数の一般人に単純作業 (タスク) として業務を委託することで問題を解決する方法である。

CrowdSourcing には最も有名な「Amazon Mechanical Turk」を始めとして既存のサービスが存在している。しかし既存のサービスはタスク登録者の単位で独立しており、異なるタスク登録者のタスクを参考にすることが出来ない。これは既存のサービスの多くが不特定多数のタスク登録者を対象としたサービスであることからセキュリティ面で不可避である。その為、タスク登録者 A が登録したタスクとタスク登録者 B が登録しようとするタスクが類似していても参考にすることが出来ず、A、B ともに同一の間違いを繰り返し、同一の作業を繰り返してしまうという問題がある。これは精度の面からもコストの面からも望ましくない。

我々の PCSS では Private なグループ内ですべての情報を共有し、グループに所属するタスク登録者はそれらすべてのタスク情報を参照することが可能である。その為初めてのタスクであっても従来の類似したタスクから、発生しうる問題点を推測し、作業適性を持つと推測される作業者を割り振ることで効率よく作業を進めることが出来る。 また、我々が行ったタスク「音声収集」[中田 2013] では既存のサービスでは実装しにくいサーバ側での音声処理の組み込みや、問題が発生した際に他のタスクと比較することでタスク内容依存の問題か、作業者の問題かの判定などの精度向上施策を行なっている。

2.2 PCSS の構築

PCSS を構築するにあたって一番の問題は作業者の募集である。「Amazon Mechanical Turk」の様に既に周知のサービスであれば作業者の募集は容易だが、無名の状態から必要な人数を集めるには多大なコストがかかる。また、「Amazon Mechanical Turk」の様に誰でも作業ができる環境では作業者の質が安定せず、作業結果の質が低下してしまうという問題もある。

本稿の PCSS においては、作業者の募集に関してネットワークリサーチを行なっている外部業者へと委託した。外部業者は既にリサーチ対象となるユーザを数百万規模で管理しており、これらのユーザを PCSS の作業者候補とすることが可能であり、また我々が望む条件に合致するユーザを抽出することが可能である。

これにより我々は外部業者のユーザを作業者として作業を提供し、Web 経由で作業を行なってもらい、さらに外部業者を経由して作業者に報酬を支払うという図 1 の構成を構築することができた。

2.3 精度向上

2.2 節で作業者を予め絞り込むことで作業者の精度向上を行ったが、CrowdSourcing の作業内容は多岐にわたり事前の調査だけでは十分でないことがわかっている。その為、本稿では事前の調査に加えて作業者が PCSS にて作業を開始してからの行動履歴をベースにした精度向上の試みを行なっている。

正解率と経験値

正解率は「正解数／総作業数」で算出し、一定値以下の

連絡先: 芦川将之，(株) 東芝研究開発センター知識メディアラボラトリー，〒 212-8582 川崎市幸区小向東芝町 1，044-549-2243，masayuki.ashikawa@toshiba.co.jp

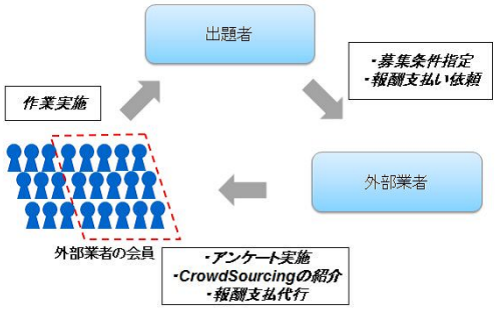

図 1: アンケート業者を利用した CrowdSourcing

作業者は作業不可にすることで作業結果の精度を上げる。また同様に「正解数—不正解数」で算出される経験値を設定している。一定の経験値を持つユーザに対して高報酬、高難易度の作業を提供することで、好成績のユーザのモチベーションを高める目的がある。なお作業において作業結果がどのような条件を満たした場合に「正解」とするかは各作業においてそれぞれ定義されている。

スキル

作業者には「文法に詳しい」「音感が良い」などの作業者自身が自覚していない特徴が多く存在する。そしてそれらの特徴は前述のアンケートでは判定することは出来ず、実際に作業を行いその正解率から判断していくしかない。我々のシステムでは作業者の作業結果から判明した作業者の特徴を作業者ごとに「スキル」として管理し、スキルにあった作業を出題することで精度向上を行なっている。例えば「読み仮名付け」の作業の正解率が高い作業者には「読み仮名付け」のスキルを付与し、「読み仮名付け」の作業は「読み仮名付け」スキルを持つ作業者に優先して出題することで精度向上を行う。

2.4 PCSS における作業プロセス

本稿における PCSS は以下の 3 つのフェーズを作業者が行うことで成り立っている。

1. 複数の作業から選択する作業選択フェーズ
2. 作業内容の説明、練習を行うトレーニングフェーズ
3. 実際に作業を行う作業フェーズ

1. 作業選択フェーズ
   作業者のスキルや正解率に応じて作業可能な作業のリストが変わる。作業者は好みに応じて作業を選択して (2) トレーニングフェーズに移る。

2. トレーニングフェーズ
   作業者は各作業において、初回の作業を行う前にはトレーニングとして説明画面で作業内容を確認する必要がある。ここでは作業の概要や注意点など作業提供者が作業者に注意して欲しいことを表示し、正しく作業ができるかどうか簡易なチェックを行うことができる。トレーニングの終了後 (3) 作業フェーズへ移る。またトレーニングは後から繰り返し行うことも可能である。

3. 作業フェーズ
   作業内容によって内容は変化する。作業者は自分のステータスを確認しながら作業を進めることが出来る。

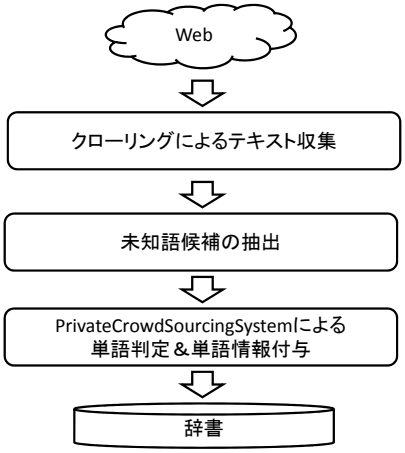

図 2: 語彙抽出フロー

表 1: 獲得したウェブテキスト

獲得ページ数	517,239,154
日本語ページ数	319,570,805
文数	12,504,868,218

3. PCSS を用いた語彙収集

本節では前節までで構築した PCSS を用いて知識処理研究に必要な語彙を収集する方法について述べる。

3.1 概要

語彙収集のフローを図 2 に示す。まず始めに、ウェブクローラを用いて大規模テキストを収集する。続いて、収集したテキストから未知語の候補を自動抽出する。そして最後に、PCSS を用いて未知語候補から単語として適当なものだけを絞り込み、品詞や読み仮名等の単語情報を付与する。

3.2 クローリングによるテキスト収集

ウェブテキストには、固有名詞や新語などの未知語が頻繁に出現する。こうした未知語を獲得するコーパスとして、ウェブテキストを収集する。本稿では、OpenDirectory[*1] の URL をシードとして、Apache Nutch[*2] を用いて収集した。
獲得したテキストの情報を表 1 に示す。5.2 億ページから日本語文 125 億文を得ることが出来た。

3.3 未知語候補の抽出

3.2 節で収集したテキストから未知語の候補を抽出する。抽出処理は以下のステップで行う。

1. テキストに対して点予測手法 [Mori2011] による単語分割を実施
2. 単語分割結果から辞書未登録文字列を取得
3. 単語分割結果を用いて単語 Ngram を作成
4. 単語 Ngram を用いて辞書未登録文字列の中から未知語候補を選出

点予測による単語分割では、入力テキストの各文字間が単語境界かどうか SVM を用いて判定する。単語境界が判定出来

*1 http://www.dmoz.org/World/Japanese/
*2 http://nutch.apache.org/

れば入力テキストを単語列に変換することができる（ステップ 1)。続いて、単語列の中で辞書に未登録な単語を辞書未登録文字列としてリストアップする（ステップ 2)。最後に、ステップ 1 の単語分割結果から単語 Ngram を作成し（ステップ 3)、Ngram を用いて辞書未登録文字列の中から未知語候補を選び出す（ステップ 4)。選出には、辞書未登録文字列の出現頻度や前後の文字種等を用いる。

単語分割を点予測手法で行うことで、複数文字種で構成される未知語の獲得が期待できる。多くの形態素解析器では、形態素ラティスからの最適パス選択には統計モデルを用いるが、未知語の生成は人手で作成したルールを用いる。人手によるルール作成は容易でないため、文字種に強く依存した未知語生成ルールが用いられる。そのため、生成される未知語は主に単一文字種で構成される。しかし、世の中には「写メ」などのように複数文字種で構成される単語も多く、それらに対応できないという問題がある。一方、点予測による単語分割では、人手で作成したルール用いないため、複数文字種で構成される未知語も作成できる。

単語 Ngram を用いて辞書未登録文字列の中から未知語候補を選出する例を紹介する。「レコメンデーション」という入力文が与えられ、「レコメン」が辞書に存在した場合に、「レコメン」「デーション」と単語分割したとする。この場合、「デーション」は辞書未登録文字列となるが、「デーション」が単独で用いられることはなく、単語としては不適切である。このとき、「デーション」の前方の文字種は極めて高い確率でカタカナとなる（「レコメンデーション」や「コンソリデーション」など)。一方、辞書未登録文字列が適切な単語である場合には、その前後には多様な文字種がつながる。例えば、「タブレット」が未登録文字列の場合は「そのタブレット」「東芝タブレット」などが考えられる。そこで、大規模テキストから作成した Ngram を基に辞書未登録文字列の前後の文字種のバリエーションを調査し、著しく偏っているケースを候補から自動除外した。除外されずに残ったものが、未知語候補となる。

3.4 単語判定と単語情報付与

3.3 節の方法で作成された未知語候補には、単語として適当でないものが残っている可能性が高い。また抽出した単語に対して音声処理に必要な情報を付与しなくてはならない。これらの情報収集を PCSS の以下の 4 タスクとして行った。

1. 単語判定タスク
   このタスクでは作業者に対して 3.3 節の方法で作成された未知語候補を「それは (未知語候補) です」という問題文に加工して表示し、「問題文は日本語して自然か否か」という選択をさせた。「日本語として自然である」と回答された場合、その文章に含まれる未知語候補を未知語として扱う。

2. 品詞付与タスク
   画面例を図 3 に示す。このタスクでは名詞とそれ以外の品詞に分ける作業を行なっている。名詞に関しては「人名」「地名」「組織名」「その他の名詞」に再分類している。1) で単語として適切であると判定された未知語に単語抽出元の前後の文章を付与して問題文に加工して表示し、「人名」「地名」「組織名」「その他の名詞」「名詞以外」を選択させた。

3. 読み付与タスク
   このタスクでは 2) で名詞と判定された未知語を問題として表示し、その読みを入力させ、その結果を未知語に対する読みと判定した。

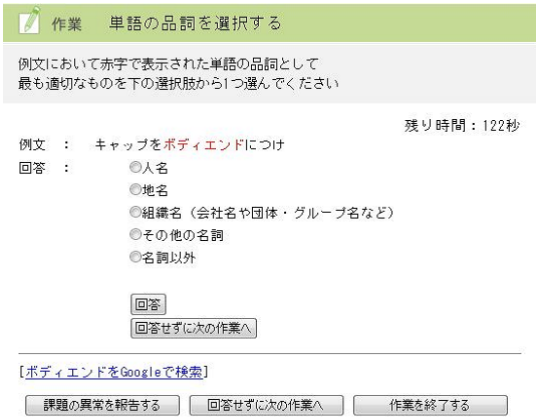

図 3: 品詞判定タスク作業画面

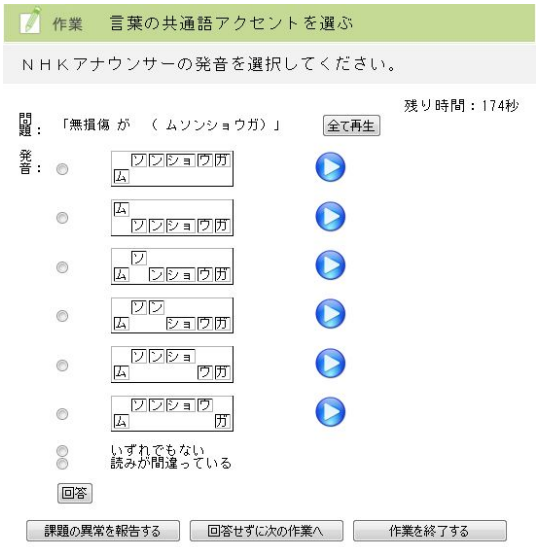

図 4: アクセント付与タスク作業画面

4. アクセント付与タスク
   画面例を図 4 に示す。このタスクでは 3) で付けられた読みから推定されるアクセント候補から合成した音声を用い、どれが自然かを選択させた。その結果を未知語に対するアクセントと判定した。

各タスクは 3 人に出題され、2 人以上一致した回答を有効なデータとして扱う。ただし、1) の単語判定タスクは高精度であることを求められるため、3 人が一致した回答のみを有効なデータとして扱った。

3.5 結果と課題

3.1 節から 3.4 節までの方法を用いて未知語の獲得実験を行った。なお、今回の実験では、3.3 節のステップ 4 で設定する頻度閾値は 100 回以上とした。獲得語彙数を表 2 に示す。125 億文のウェブテキストから 14 万語の語彙を獲得することが出来た。

表 2: 未知語獲得数

未知語候補抽出数	227,367
未知語獲得数	138,546

表 3: 各タスクの作業結果における一致率

	3 人一致	2 人一致	不一致
品詞	71.3 %	27.7 %	0.9 %
読み	82.8 %	11.2 %	1.5 %
アクセント	76.2 %	22.5 %	1.3 %

獲得できた未知語の例としては「Siri」「あっちゃん」「先っちょ」「スンゲー」「ドm」「花立山」「えらそう」「やべえええええ」などが挙げられる。

課題としては、アダルト用語、差別用語などへの対応が挙げられる。今回、未知語抽出元テキストを不特定多数のウェブテキストとしたため、大量のアダルト用語や差別用語が未知語候補として抽出された。こうした語の判定を作業者に行わせ続けることは作業者の精神的負担が大きい。今回の実験では、アダルト用語等を含んでいるケースは作業をスキップするように指示したが、前処理でフィルタリングする等の対策が必要である。

各タスクの作業結果における一致率を表 3 に示す。不一致になった例として「鹿島」があげられる。これは会社名と判断された場合は「かじま」、スポーツチームと判断された場合は「かしま」が正解となる。このような文脈により読み異なる語彙に対しては抽出元の前後の文を作業者に提示して揺れを回避する必要がある。しかし問題文が長文になると作業者の処理速度が低下する。これらの問題に対し適切な出題方法の検討を行なっていく必要がある。

4. まとめ

本稿では、音声認識や音声合成に必要な単語表記、読み仮名、アクセント、品詞からなる「語彙」の獲得を目的として、PCSS を活用することで、大量のウェブテキストコーパスから語彙を獲得する手法を提案した。日本語 125 億文のウェブテキストから未知語候補を抽出し、14 万語を獲得することができた。

3.5 節で述べたように今後の課題としてはアダルト用語や差別用語の排除、読み異なり問題の回避がある。

また、専門的な技術を必要とする作業へはさらなるクオリティコントロールの適用が必要となる。主に CrowdSourcing では内容を簡易化することで精度を上げることが可能であるが、簡易化出来ないような専門的な作業を CrowdSourcing で行う場合は作業者のクオリティをあげなくてはならない。その為には作業者の特性を解析し、適切なスキル付与を行わなければならないが現在は手動でスキル付与を行なっており効率が悪い。このスキル付与を自動かつ高精度で行うことで、大量の専門的な作業を高精度に行うことが可能となる。

さらに本稿では、形態素解析性能を改善させるためには、辞書だけでなく、統計モデル学習用コーパスの大規模化も不可欠である。CrowdSourcing によって、助詞や助動詞等の品詞判定を行わせることは難しいが、名詞などの単語境界や品詞情報などであれば付与できると予想される。そこで、部分アノテーションコーパスからでもモデル学習が行える点予測手法と CrowdSourcing を組み合わせることにより低コストで学習コーパスを増やしていくことが考えられる。

参考文献

[Mori 1996] Shinsuke Mori and Makoto Nagao : Word extraction from corpora and its part-of-speech estimation using distributional analysis, in Proceedings of COLING(1996)

[村脇 2010] 村脇 有吾 , 黒橋 禎夫 : 形態論的制約を用いたオンライン未知語獲得, 自然言語処理, vol. 17(2010), no. 1, pp. 55-75.

[羽鳥 2011] 羽鳥 潤, 鈴木 久美 : 機械翻訳手法に基づいた日本語の読み推定, 言語処理学会第 17 回年次大会 (2011)

[芦川 2012] 芦川 将之, 西山 修, 下郡 信宏 : CrowdSourcing を用いた単語への読み付け、アクセント付け手法の提案, 電子情報通信学会技術研究報告,111(447)(2012), pp. 11-16

[中田 2013] 中田 康太, 芦川 将之 : Private CrowdSourcing を用いた言語、音声資源の収集〜音声収集と品質評価〜, 人工知能学会全国大会,(第 27 回)(2013)

[Mori2011] 森 信介 , 中田 陽介 , Neubig Graham , 河原 達也: 点予測による形態素解析 , 自然言語処理, Vol.18(2011), no. 4, pp. 367-381,