集合知を利用した語彙情報収集・共有・管理システム

佐々木 浩† 中野 鐵兵† 藤江 真也† 小林 哲則†
†早稲田大学

1 はじめに

音声認識アプリケーション開発における最も重要かつ困難な作業の一つとして，継続的な語彙情報のメンテナンスが挙げられる．アプリケーション用に設計した語彙は一度設計したら完成というものではなく，新規語彙の追加や既存語彙の修正などの持続的なメンテナンスが必要である．例えばPodcastle[1]のような，新規性・話題性の高い単語が多く現れる音声を対象としたディクテーションシステムでは，関連する話題のコーパスをウェブから収集したテキストを用いて逐次更新し，言語モデルと語彙を頻繁に更新することで音声認識の精度向上を図っている．さらに，カーナビゲーションシステムのような音声認識アプリケーションでは，住所名の変更や新規の施設名への対応が求められる．このような日々増殖・変化する語彙を逐次更新するため，ウェブ上の情報資源を基に語彙リストを構築する手法がとられることがある．しかし，ウェブ上で提供されるサービスは語彙情報取得に特化しているわけではなく，そこから必要な情報の抽出が必要である．さらにこれらの枠組みは，アプリケーション開発毎に個別に設計・実装される必要があり，作業負荷や精度・効率の面で問題がある．また，更新された語彙情報を実利用環境に配信するための枠組みも求められる．

そこで本研究では，アプリケーションに用いる語彙情報を開発者で共有する枠組みを提案する．あらゆる分野の語彙情報を一元化されたオンラインデータベース上に蓄積し，共有のウェブサービスとして提供する．音声認識アプリケーションの語彙情報の管理にウェブシステムを利用した例として，w3voice[2]やMusicNavi[3]がある．w3voiceでは音声認識アプリケーションをWebサービスとして共有し，ユーザの作成したWebサイトから利用できるようにする枠組みを提案している．また，音声認識用辞書を共有することにより，認識用語彙情報のユーザ間での更新を実現している．MusicNaviでは楽曲に関する語彙の辞書をオンラインデータベースの形で共有し，音声認識システムに用いている．しかし，これらのシステムで蓄積された語彙情報は指定されたアプリケーションのみでしか扱えず，ユーザの作成したアプリケーションに組み込むための枠組みは提供されていない．また，音声認識・言語処理アプリケーション全般を対象とした語彙情報の共同管理の枠組みは存在しない．それに対して本研究では，自由に利用可能な形式で語彙情報を集約し，簡単な要求を投げるだけで必要な語彙情報を得ることができるようなウェブベースの枠組みを実現する．また，クローラによるWeb資源からの自動収集の枠組みと，利用者の集合知を利用した半自動的な語彙情報作成の枠組みを構築し，データベースの増強を図る．さらに，データベース上の語彙の情報が更新されたり，新規の語彙が追加された際に利用者やアプリケーションへ反映されるような機構を用意する．こうした枠組みを利用することで，語彙情報の動的な更新が可能な音声認識アプリケーション開発の新しい枠組みの実現を目指す．

A Collaborative Lexical Data Design System for Speech Recognition Application Developers
†Hiroshi SASAKI, Teppei NAKANO, Shinya FUJIE, Tetsunori KOBAYASHI Waseda University

2 基本アプローチ

本研究での提案システムの基本的なアプローチを述べる．

Data Intensive Systems 音声・言語アプリケーションで必要な語彙情報が集約され，単一の語の読み情報の取得から，アプリケーション用語彙の作成・管理まで，語彙情報に関連する全ての作業を提案システムで完結できるようにする．またシステムを広く公開し，自然と語彙に関連する情報が集約されるような枠組みを提供する．

Lexicon Lifecycle アプリケーション用の語彙の定義から，その継続的な更新まで包括的な解法を提供する．

Cooperative Framework 語彙情報を必要とするアプリケーション同士のゆるやかな連携を可能にする．すなわち，アプリケーションで使用する語彙の定義と追加・修正された語彙情報の共有を可能にする．

以降，これらの特徴を満たしたシステムを構築するための，本研究における具体的な実装のアプローチを述べる．

2.1 WWW上の語彙資源の利用

利用者が様々な語彙を取得できるようにするため，データベースは初期の段階で十分な量の基本語彙とその語彙に含める情報が用意されている必要がある．データベース上の語彙に含める情報としては語の綴りの情報，読みの情報，収集元を用いる．語彙の整備には例えばipadic[4]などのWWW上で入手可能な既存の辞書を活用する．また，ユーザのアップロードによる語彙追加の枠組みも設け，語彙の増強を図る．加えて，クローラを用いてWikipedia[5]などのWWW上の語彙資源から随時語彙情報を収集するようにする．こうすることにより，データベース内の語彙の新規性の維持が実現できる．さらに，他所から情報を引用する際には語彙情報の収集元の情報も保持，明記し，権利上の問題に配慮する．

2.2 タグを用いた目的語彙の選別

例えばシステムがデータベースに蓄積した様々な語彙から，施設名の一覧のみを取得するなど，特定の語彙の集合のみを取り出せる必要がある．また，データベースに新しく登録された語（新着語と呼ぶ）についても同様で，新着語の中で目的語彙に沿った語のみが利用者の持つ語彙に追加される必要がある．

そこで語彙に対してのメタ情報として自由なタグ付けを許し，そのタグの付与条件（タグ条件と呼ぶ）によって語彙を選別する方法をとる．例えば，“早稲田大学”などの大学名を表す語にはタグとして“名詞”や‘大学”などを付与し，“名詞 and 大学”などの条件で選別できるようにする．タグの情報は既存の語彙辞書やユーザ定義，Web上の分類の情報，解説文の解析などから幅広く回収し，多くの要求への対応を図る．

2.3 アプリケーション連携の枠組みの提供

利用者間での語彙の定義・修正の情報を共有し，利用者の開発負荷を低減させる必要がある．例えば

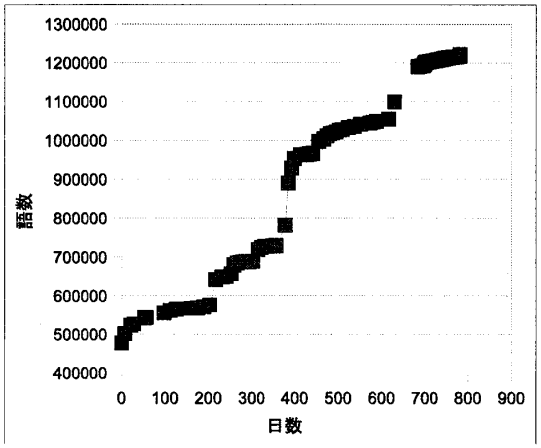

図 1: 総語数の遷移（急激に増加している点は新たな情報源を追加した時点．最後に情報源を追加した後で毎日 200 語以上増加している．）

利用者がホテル名の一覧を作成したいとき，他の利用者がホテル名の一覧を作成していた場合には，その情報を活用できるような機能などが求められる．ここで，他の利用者が作成した語彙をそのまま共有するのではなく，その語彙を表すタグ条件を共有する方法をとる．タグ条件により語彙の選択基準を明確にし，アプリケーションを越えた語彙の利用の促進を図る．

3 システムの開発

2 節の要件を踏まえつつ，実際にシステムの開発を行った．データベースには 2010 年 1 月 4 日時点で 1,221,016 語が登録されており，毎日 200 語以上の語が新規に登録され続けている（図 1）．本サービスは Web アプリケーションとして動作し，Web ブラウザ上または WEB API 経由で利用される．ユーザは本サービスに対して利用したい語の集合のタグ条件を送信すると，データベース内でのそのタグ条件を満たす語の集合を得ることができる．その結果を基に必要な語の追加や不必要な語の削除を行うと，よりユーザの希望に沿ったタグ条件が推奨される．もし適切なタグ条件が得られなければ，編集した結果を表すようなタグ条件の登録を促す．ユーザはこのタグ条件をクエリとして，語彙リストのファイルの形式でのダウンロードや WEB API 経由での参照を行うことができる．

ユーザの定義した語の集合は本サービスに蓄積され，いつでも再利用することができる上に、Web ブラウザ上で自由に修正・編集ができる．さらに，語の集合を定義するタグ条件を基に，その条件を満たした新着語が追加語彙候補としてユーザに通知される仕組みを持つ．ユーザは追加語彙候補を確認することで語の集合の新規性の維持が可能となる．また，WEB API を用いて追加語彙候補を含めた語彙リストを利用することもできる．

4 語彙情報サービスの応用

本サービスの適用例を紹介する．音声による項目選択を利用した Web ブラウザ [6] において，Web サイトの項目に用いられる語彙を音声認識させるため，その読みの情報を本サービスを用いて取得し，音声認識に用いている．新規性が高く，分野も特定されない語彙の読みを適切に取得するために本サービス活用している．また，音声コマンドの操作を用いた車載情報端末 [7] において，地名や施設名などの音声認識辞書を構築する際にも本サービスを用いている．都道府県名＞市区町村名＞地域名などの階層構造を持つ項目を構築するために本サービスのタグ情報も用いている．さらに，音声コンテンツのメタ情報のトピックを推定し，そのトピックに沿ったコーパスを選択肢し，音声認識用の言語モデルの適応を行う手法 [8] を用いる際にも本サービスのタグ情報を用いている．音声コンテンツのメタ情報やコーパスのテキスト情報を増やすため，それらに含まれる語彙のタグ情報を本サービスを用いて抽出し，活用した．これらの利用法に関し，本サービスの語彙が効果的に活用されていることを確認している．

5 まとめと今後の予定

音声認識アプリケーション開発の問題として，システムが認識可能な語彙の適切な設計と，実際に利用されている語彙のメンテナンスを挙げ，これらの問題を解決するために，集合知を利用した語彙情報の収集・共有・管理システムを提案した．具体的には，語彙情報を集中管理するためのオンラインデータベースシステムを構築し，それをウェブシステムとして利用者・アプリケーションに公開する．提案システムでは，Web 資源からの自動収集の枠組みを備え，アプリケーション用の語彙の新規作成から，その継続的な更新まで包括的な解法を提供する．また，実際にシステムの開発を行い，2010 年 1 月 4 日時点で合計 1,221,016 語のデータを保有するデータベースを構築した．今後も引き続き公開を行い，継続的な改良を行っていく．

参考文献

[1] M. Goto, J. Ogata, and K. Eto. Podcastle: A web 2.0 approach to speech recognition research. *In Proc. Interspeech 2007*, pages 2397–2400, August 2007.

[2] R. Nisimura, J. Miyake, H. Kawahara, and T. Irino. *Development of Speech Input Method for Interactive VoiceWeb Systems*, volume 5611. Springer Berlin / Heidelberg, Nara, Japan, July 2009.

[3] S. Hara, C. Miyajima, K. Itou, and K. Takeda. An online customizable music retrieval system with a spoken dialogue interface. *The Journal of the Acoustical Society of America*, pages 3378–3379, November 2006.

[4] ipadic version 2.7.0,`http://chasen.naist.jp/stable/ipadic/ipadic-2.7.0.tar.gz`

[5] Wikipedia,`http://ja.wikipedia.org/wiki/`

[6] 秋元啓孝，中野鐵兵，小林哲則，音声による Web リンク選択インタフェースの検討，情報処理学会全国大会講演論文集，2009.

[7] Teppei Nakano, Tomoyuki Kumai, Tetsunori Kobayashi, Yasushi Ishikawa, Design and Formulation for Speech Interface Based on Flexible Shortcuts, *Proc. Interspeech 2008*, pp.2474-2477, Sept. 2008.

[8] 佐々木 浩，中野 鐵兵，緒方 淳，後藤 真孝，小林 哲則，集合知に基づく語彙情報を用いたトピック依存言語モデリング，情報処理学会研究報告，SIG-SLP-075，pp.57-62，Feb. 2009.