2Z-6

オンライン講義受講に適したテキスト入力支援システム

小高 正道　井上 亮文　市村 哲

東京工科大学

１．背景

現在、講義形式としてインターネットを介したオンライン講義が注目されている。この形式の講義は、各自のパソコンで、講義資料（パワーポイント（以下 PPT）など）と講師が映った映像を再生させて受講するもので、LAN が繋げられる場所であれば，どこでも受講できるメリットある。また最近、受講者は、講義メモをノート PC にテキスト入力する場合が多い。

以上のような環境を想定した場合，受講者は講師が話した内容を高速にノート PC にキーボード入力する必要がある。そこで、オンライン講義の受講時に適したテキスト入力支援システムを提案する。

２．従来技術

日本語入力システムの従来技術としては、Microsoft IME，ATOK，POBox[2](携帯電話の予測入力システム)等がある。これらは，入力された読み仮名文字列を逐次解析し，日本語変換候補を出力するシステムである。しかし、あらかじめ用意された辞書や利用者の利用履歴のみから予測が行われるため，講義中に説明される新しい語句や専門用語を適切に優先候補として出力することは難しいという問題があった。講師が発した発話を学生が聞き取ってテキスト入力するという用途のためには改善の余地がある。

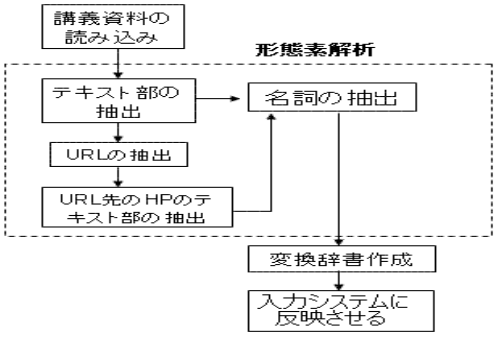

図 1.動作概要

A text input system for online lectures.

Masamichi　Kotaka , Inoue Akifumi ,
Satoshi Ichimura
Tokyo University of Technology

３．提案

講義資料や講義で用いるインターネット情報を参照して適切な日本語変換候補を出力することで，講師が発した発話を学生が高速かつ正確にテキスト入力できるようになるコンテキストアウェア日本語入力システム「CaIME」を提案する。CaIME の機能は以下のようなものである。

１．講師が Web 掲載した１つまたは複数の講義資料から変換候補語句を習得する機能
２．１の講義資料に含まれる URL のリンク先から変換候補語句を習得する機能
３．１または２によって得た変換候補語句を辞書登録し，IME の優先候補語句とする機能

システムの典型的な動作を図 1 に示す。学生（ユーザー）は CaIME を使い、講義 Web ページに掲載されている講義資料（PPT ファイル）のうち，講義内容に関連の深そうな PPT ファイルを１つまたは複数選択する。この時システムは，ユーザによって選択された PPT ファイルをネットワーク経由で取得し，この中に記述されたテキストの抽出を行なう。次にシステムは，テキストを形態素解析（自然言語で書かれた文を意味を持つ最小の文字列に分割し品詞に分ける作業）をして「名詞」を抽出すると共に，正規表現を用いたパターンマッチによりテキストから「URL」を抽出する。

抽出された名詞は，優先候補語句として IME の変換辞書に即座に登録される。一方，URL については，URL のリンク先ホームページのコンテンツをインターネット経由で取得し，そのコンテンツから抽出した名詞を優先候補語句として変換辞書に登録するようになっている。講義資料から URL を抽出する理由は，講師が準備や講義資料作成の際に参考にしたページであることが多く，講義資料に書かれていない専門用語を講師が発話した場合にでもシステムが対応できる可能性が高いためである。

以上のように講義資料から抽出された名詞と URL に基づき変換候補語句を辞書登録することにより，講義で使われる可能性の高い用語が日本語変換候補の上位に現れるようになっている。

４．システムの実装

４．１　変換候補語句を習得

CaIME の実装においては，PPT 文書からテキストを抽出するために，Microsoft オフィスの OLE オートメーション機能を用いた。また，URL のリンク先ホームページの HTML 文書からテキストを抽出するために，HTML タグを消去するというテキスト処理を施した。

抜き出したテキストを形態素解析するために「茶筌」と「termex」[1]を併用する方式を用いている。大学講義において登場する専門用語は，単語を組み合わせて複雑な概念を表す複合語であることが多いが，茶筌は複合語に対応していないという問題や，語句を品詞単位に細かく分割するため，そのまま日本語入力変換候補とするのは適当でないという問題がある。そこで，テキストファイルから専門用語を抽出することを目的に開発されたソフトウェアコンポーネント termex を茶筌と組み合わせることとした。termex は，どの用語が重要であるか判断する機構も備えている。

図 2 は，CaIME の抽出語句表示機能を実行した例である．右端テキスト枠には，茶筌と termex を併用して抽出した変換候補語句のリストが表示されている．変換候補語句のリストは，テキスト形式でログ保存される。

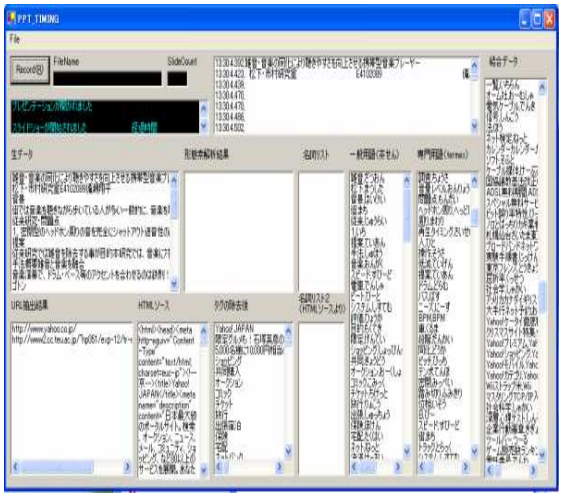

図 2.CalIME の動作例

４．２　ＩＭＥの操作

Windows には，使用中のエディタの IME を，外部アプリケーションから制御することができる IMM API が備わっている。CaIME が抽出した変換候補語句はこの IMM API を介して，変換辞書に登録されるようになっている。専門用語が複合語である場合が多いのを利用し，最初の最小単語の読み仮名で登録するようにした。例えば、「せんもん」の読み仮名で、「専門用語」，「専門用語自動抽出機能」，「専門用語抽出システム」が登録される。

図３は，語句登録用の IMM API「ADD Word」で、ログデータを読み込んだ例である。「せんもん」と入力して，「専門」，「専門用語」，「専門用語自動抽出機能」，「専門用語抽出システム」が，上位変換候補として表示された。図４は，辞書登録した結果が他のアプリでも有効であることを示すため，Windows に添付されている「メモ帳」での動作を示したものである。

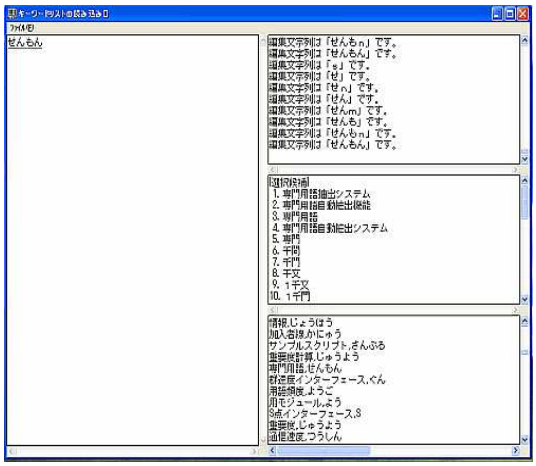

図３.ADD Word の動作例

図４.IME の反映結果（メモ張）

５．おわりに

講師が発した発話を学生が高速かつ正確にテキスト入力できるようになる日本語入力システム CaIME を提案した。語句検出精度を向上させること，PDF 資料に対応すること，実際の講義において評価実験を実施することを今後の課題としている。

参考文献

[1] Termx, 東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム
http://gensen.dl.itc.u-tokyo.ac.jp/win.html.
[2] POBox,
http://pitecan.com/papers/HUC99/HUC99.pdf