

先進的学習支援のための連合ポータルシステム

A United Portal System for Advanced Learning

三輪 譲二

Jouji MIWA

岩手大学工学部

Faculty of Engineering, Iwate University

<あらまし> e-Learning による専門教育や日本語教育の先進的な学習を支援するため、Web 上の TEXT、HTML、PDF ファイルの日本語字句解析結果から、単語の直接リンクを自動生成し、複数の検索サイトに簡単に直接行き、専門や言語の先進的学習のための連合ポータルシステム (uPal: United Portal for Advanced Learning) を構築した。なお、複数の単語辞書サイトを連合して簡単に検索できる機能と共に、外国人の日本語学習を支援するため、テキスト音声合成 (TTS:Text-To-Speech) 機能を用いて、任意の単語の音声を聞くことができる機能も加えた。

<キーワード> e-Learning、辞書検索、言語インタフェース、ポータル、日本語教育、音声合成

1. まえがき

日本語や英語の辞書や Wikipedia などの専門用語の言語情報が Web 上に多数存在するようになってきたが、これらの情報を検索するには、検索フォームに入力する手間がかかる問題点があった。このため、Web 上の TEXT、HTML、PDF ファイルの日本語字句解析結果から、単語の直接リンクを自動生成し、複数の検索サイトに簡単に直接行き、専門や言語の先進的学習のための連合ポータルシステム (uPal: United Portal for Advanced Learning) を構築した。なお、検索と同時に、テキスト音声合成 (TTS:Text-To-Speech) 機能を用いて、任意の単語の音声を聞くことができる機能も加えた。本報告では、uPal システムの構成や利用例などを述べる。

2. システム構成

Java による Web アプリケーションのシステムの構成を、図 1 に示す。システムは、JSP (JavaServer Page) でプログラミングされ、PDF テキストは、PDFBox の API により、テキストが抽出され、HTML などのテキストは、Sen の API により、字句解析され、名詞、動詞、形容詞、形容動詞の単語のみ、リンク付の HTML を自動生成する。また、リンクをクリックすることにより、単語辞書の検索や音声合成 (TTS) の処理に移動することができる。

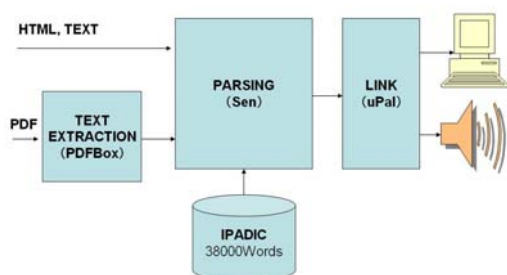


図1 uPal のシステム構成

3. 専門教育や日本語教育への利用

図 2 に読売新聞社のページの表示例を示す。この例において、学習者が検索したい単語をクリックすると右上のポップアップウィンドウが表示され、複数の辞書や、Wikipedia の検索ができる。このように、Web ページや PDF ファイルから、カットアンドペーストすることなしに、効率的に複数の辞書を検索することができる。

なお、この図の例では、マウスを「強盗」の文字の上に置くと、ステータスバーに「強盗」の文字が表示され、検索語を誤り無く選択できるようにしている。

また、本システムでは、単語ごとにリンクが付与されるため、元のリンクは、図 2 の例のようにプレゼントボックスの画像に変換される。さらに、リンクや画像などの相対 URL は、絶対 URL に自動変換され、元の Web ページと同一のページ構成の表示となる。このように、リンクページも自動的に uPal で絶対 URL に変換され、元のリンクを辿ることができる。

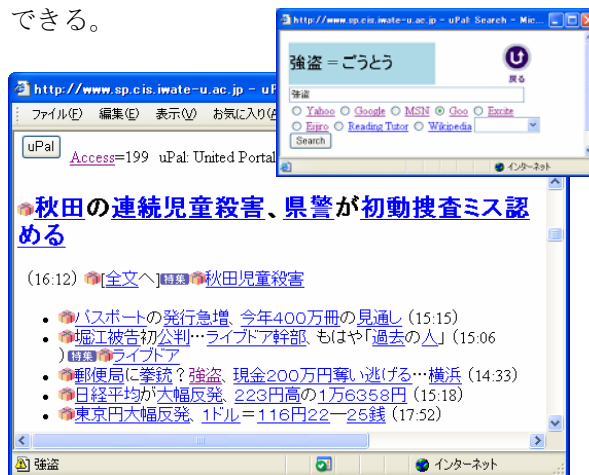


図2 読売新聞社のページの uPal 利用例

図 3 に HTML ソースコードの例を示す。このように、元のリンクは `img` タグにリンクされ、動詞などの活用のある単語は、ひらがなの読みと共に、辞書検索に容易になるように、原形も付与されており、活用に不慣れな日本語学習者の辞書検索のサポート機能としている。また、マウスが検索単語の上に置かれたときに表示される単語の指定や、検索単語外の際に無表示になる指定を加えている。

図 4 の複数辞書の検索例が示すように、Yahoo, Google, MSN, Goo, Wikipedia などの辞書サイトを選択指定して検索でき、複数の辞書を連合して学習でき、上級者の専門科目などの利用にも、利用できる特色を持っている。なお、Wikipedia の検索時には、日本語と英語の自動判別の他に、Chinese, Dutch, Finnish, French, German, Italian, Korean, Portuguese, Russian, Spanish, Turkish の言語を `select` タグにより指定することができる。

```
<a href="parsej.jsp?url=http://iwate-u.ac.jp/my.html">
</a>
<a href="javascript:search(0,'行き','いき','行き')
onMouseOver="window.status=' 行き '; return true;"
onMouseOut="window.status=";return true">
行き</a>
```

図 3 uPal で変換した HTML ソースの一部の例

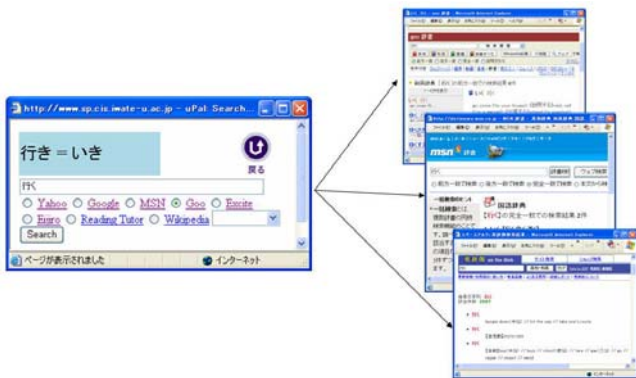


図 4 複数辞書の検索利用例

4. システムの考察

本システムでは、国際化のため UTF-8 の文字コードを使用しており、また、検索情報の繰り返し利用や表示の見易さなどのため、JavaScript、クッキー、ポップアップウィンドウの機能を用いている。このため、本システムを利用するには、これらの機能を受け付ける設定にブラウザをする必要がある。

また、本システムでは、負荷を低減するため、HTML については 1000 行毎に、また、PDF ファイルは 1 ページ毎表示するようにしている。

さらに、文字コードは、Content-type から自動判定しているが、meta タグしてある場合は、後者を優先しているが、設定が誤っていると文字化けが生じる場合がある。

構文解析とリンクづけ処理は、Sen を使用することにより、複雑な新聞社のページでも、数秒で処理が終わり、高速処理可能となった。また、生成された HTML ファイルは、元のファイルの約 6 倍程度

に増大するため、転送の負荷が増大する問題がある。

図 5 に、本文を例として、PDF ファイルのテキスト抽出とリンク表示例を示す。この例から、1 つのテキストボックスが、1 つの行で表現され、本文とレイアウトがよく一致していることが分かる。

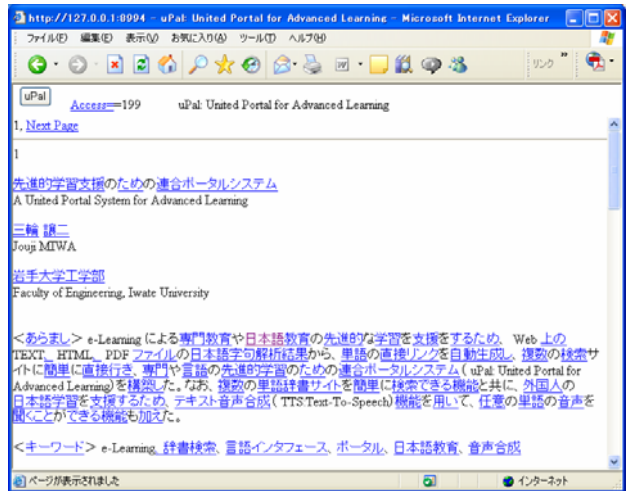


図 5 PDF ファイルのテキスト抽出と表示例



図 6 PDF の表示と Wikipedia 検索例

5. むすび

本報告では、先進的学習支援のための連合ポータルシステムについて述べた。今後、未知語の処理など、さらに使いやすく改善していく必要がある。

なお、本システムの一部は、2006 年 9 月 1 日より、<http://www.sp.cis.iwate-u.ac.jp/icampus/u/> の URL で公開中である。

謝辞

本研究の一部は、平成 16-18 年度科学研究費補助金・基盤研究 (B)「インターネット・文字音声言語統合型日本語教育システムの地球規模の運用実験と評価」による。

参考文献

1)阿部雅嗣, 三輪譲二:Web 対応型音声合成システムとその応用, 電子情報通信学会、音声研究会, 技術報告, SP2004-138, pp.7-12, Tokyo (Jan. 27, 2005).