報道発表

- **NICTと沖電気、Webページから新語を獲得する技術を共同開発**
  ～1億文字、Web 4万ページ分を1日で処理し、最新用語を継続的に獲得～

- **平成17年7月21日**

独立行政法人情報通信研究機構（理事長：長尾 真。以下、NICT）と沖電気工業株式会社（社長：篠塚 勝正。以下、沖電気）は、このたびWebページから新語を獲得して属性を判別する技術を共同開発しました。インターネット検索エンジン等に代表される情報検索・抽出システムの精度向上に効果を発揮します。今後、本技術は一般ユーザ向けインターネットサービスへの導入推進が予定されています。

＜背景＞
インターネット検索エンジンの普及により情報検索や情報抽出※1の技術は身近なものとなりました。検索や抽出を行うに際しては対象となる文書中のテキストの解析が必要になりますが、辞書に登録されていない新語が含まれると解析がうまくいかないことが多くあります。インターネット上では日々新語が生み出されており、これが検索や抽出の精度を下げる原因となっています。そのため、新聞記事などを用いて新しい用語を獲得する研究が行われてきています。ところが、新聞とは違ってインターネットのWebページは内容・用語・書式などが様々であり、新語の自動獲得は容易ではありませんでした。また、新しく獲得された用語を、実際のシステム、特に情報抽出のシステムで利用するためには、その用語が人名や組織名なのか専門分野に関する語なのかといった属性を判別することが必要となります。これに関しても自動で判別することは困難でした。

＜成果＞
今回開発した技術は、収集した大量のWebページに対して形態素解析※2を行い、文中の形態素列の頻度と、その前後の形態素の異なり数とを指標とした関数を用いて用語を獲得します。この技術を用いることにより、名詞だけから構成される用語だけでなく、助詞などを含む用語をも獲得することができます。さらに固有表現抽出※3や既存辞書とのマッチングを行って用語を構成する形態素に素性を割り当て、その情報を利用して用語全体の属性を判別します。形態素に素性を割り当てることができない場合にも、その形態素の用語全体に対する影響を考慮することにより属性を推定することができます。また、実システムへの導入を考慮して処理を高速化しています。テキストで200メガバイトの収集済みのWebページ（約1億文字、2年分の新聞記事に相当）を平均1日で処理し、用語の獲得を行うことができます。従来は容易に追加することができなかった最新用語を、高速でWebページから獲得・判別することが可能となります。それによりインターネットユーザが新語をリアルタイムで検索できるようになります。例えば、大学や企業のWebページから獲得した技術用語を継続的にシステムに反映することにより、ユーザが探している最新技術の名称を常に正しく検索・提示することができるため、最新用語による情報収集、および技術探索への効果が期待されます。

＜今後＞
本技術は、NICTけいはんな情報通信オープンラボにおける沖電気とNICTの共同研究の成果によるものです。今後もNICTと沖電気は当技術の向上を目指して共同研究を継続します。尚、沖電気は、メールで受け取る情報収集支援サービス「MAILPIA(R)」※4、および産学連携支援ツール「Bluesilk(R)」※5への本技術の導入を進める予定です。

7月22日（金）、23日（土）に沖縄県宜野湾市(健康文化村 カルチャーリゾート フェストーネ)で開催される情報処理学会自然言語処理研究会（電子情報通信学会言語理解とコミュニケーション研究会と合同開催）にて本件に関する技術報告をいたします。
尚詳細は第168回自然言語処理研究会のホームページ
(http://www.jaist.ac.jp/nlp/SIGNL/NL168program.html)をご覧ください。

＜問い合わせ先＞
情報通信研究機構 総務部 広報室
奥山 利幸、大野 由樹子
Tel: 042－327－6923、Fax: 042－327－7587

沖電気工業株式会社
広報部　原
Tel: 03-3580-8950

＜内容に関するお問い合わせ＞
独立行政法人情報通信研究機構
情報通信部門
けいはんな情報通信融合研究センター
自然言語グループ
井佐原　均
Tel:0774-98-6830

沖電気工業株式会社
研究開発本部

【用語解説】

※1　情報抽出

一般に、文章から有効な情報を取り出すことをいう。文章から特定の属性を持つ語を抽出して整理・再構成することにより、短時間で大量の情報を概観できる。最近では、情報検索の結果得られた文書からさらに必要な情報だけをユーザに提示するなどの目的で「Q&Aシステム」などにも適用されている。語の抽出は固有表現抽出などの手法で行われる。

※2　形態素解析

文を形態素という文法的に意味のある最小の単位に分割して品詞情報を付与することをいう。テキストを処理する技術の中で最も基本的なものであり、情報検索におけるキーワードの抽出やかな漢字変換に始まり、文書要約や機械翻訳などほとんどすべての応用ソフトウェアに使われている。

※3　固有表現抽出

文章から、固有表現(Named Entity)と呼ばれる、人名・地名・組織名（会社名、団体名、他）・数値表現（時間、金額、他）などを表す部分を取り出し、それがどのような固有表現であるか（人名か、地名か、組織名か、など）を決定することをいう。

※4　MAILPIA

沖電気で開発し、プロバイダー向けに提供中のWebページなどからの情報収集を支援するサービス。指定したWebページの指定したキーワードに関する内容が更新された際に、更新された部分だけを切り出してユーザの携帯電話やPCにメールで送付することができる。
詳細はhttp://www.mailpia.jp/

※5　Bluesilk

沖電気と株式会社三菱総合研究所とで共同開発中の次世代型の検索エンジン。産学連携支援や発想支援のツールとして利用されることを想定している。任意の文章を入力として概念検索を行うことができる。また、出力には、関連人名・関連組織名・関連技術用語をリストアップすることができる。
詳細はhttp://www.bluesilk.biz/