自然言語処理エンジン
Natural Language Processing Engines

野上　宏康　NOGAMI Hiroyasu
鈴岡　節　SUZUOKA Takashi
梶浦　正浩　KAJIURA Masahiro

電子化文書が日々作成・蓄積され，膨大な情報にインターネットなどを経由してアクセスすることが可能になっているが，これらの情報を利用することはますます困難になっている。これらの情報を有効に利用することを支援するには，テキストの内容を高精度に解析する自然言語処理技術が不可欠である。当社は，業界トップレベルの解析技術をベースに，情報フィルタリング，Web検索，機械翻訳，インターネットロボットの各エンジンによるASP（Application Service Provider）でのサービスを実現した。（株）ニューズウォッチの新聞記事配信サービス，（株）フレッシュアイのリアルタイム検索サービスなどに利用されている。今後，更に，これらのエンジンを応用し新規サービスを実現していく。

An immense volume of digitized documents can now be accessed via the Internet. The more information there is, however, the more difficult it is for users to find the documents they really need. Natural language processing technology is required in order to alleviate this problem.

Using its highly accurate analysis technology, Toshiba has developed four key engines: information filtering, Web-page search, English-Japanese machine translation, and Internet robot. Toshiba provides services using these engines as an application service provider (ASP), including a newspaper article distribution service (NewsWatch Inc.), a Web real-time search service (FreshEye Corp.), and others. We will realize further new services in the future utilizing these technologies.

1 まえがき

インターネットの普及に伴い，膨大な電子化情報へのアクセスが可能になっている。しかし，これらの情報は蓄積される一方であり，この中から利用者が必要な情報を有効に利用することはますます困難となっている。

電子化情報の中でも，自然言語で記述されたテキスト情報は，ものごとを記述するうえで中心的な役割を果たしている。これらの情報を利用者が有効に利用できるように支援するには，このテキスト情報を高精度に解析する技術が不可欠である。当社の自然言語処理技術は，1977年に日本で最初に商品化したワープロの研究からスタートし，その後現在に至るまで，継続して解析技術の研究と膨大な言語データの蓄積を行っている。その結果，これまで常に業界で最高レベルの解析精度を実現している[1]。この技術をベースに，情報フィルタリング，検索，英日／日英相互機械翻訳，インターネットロボットの各エンジンによるサービスを実現した。

各エンジンの関係を**図1**に示す。ロボットにより取得されたWeb情報（テキスト情報）は各エンジンへの入力となる。また，各エンジンの出力は他のエンジンへの入力としても利用される。例えば，フィルタリングの出力を機械翻訳への入力として利用することにより，フィルタリング結果を英語で提供する処理が可能になる。このエンジン間のデータの流れを矢印で示している。これらのエンジンは，（株）ニューズウォッチ（http://www.newswatch.co.jp）で，新聞記事を利用者ごとに選択して毎朝配信するサービスで用いられ，そのフィルタリング精度についてユーザーから高い信頼を得ている。また（株）フレッシュアイ（http://www.fresheye.com）では，Web検索サービスに用いられている。Web上に掲載されたページが作成／更新後，最速10分で検索可能となっており，“新鮮検索No.1”のサービスを提供している。

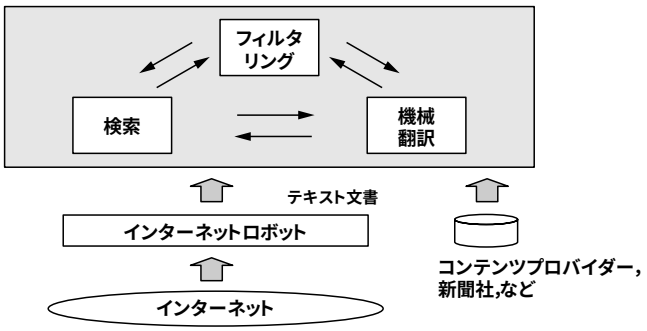

図1．自然言語処理エンジンの関係　各エンジンの出力は，他のエンジンへの入力となる。
Relationships among natural language processing engines

2章以降で各エンジンの概要，技術，サービスについて述べる。検索については4章でロボットと連携して述べる。

2 情報フィルタリングエンジン

2.1 情報フィルタリングの概要

情報フィルタリングとは，時々刻々生成されるテキスト情報（文書）をふるいにかけ，ユーザーの必要な情報だけを抽出する処理を表す。当社は，新聞社などが日々提供する膨大な記事の中から，ユーザーが欲するものだけを抽出し提供する情報フィルタリングエンジン“NEAT（News Extractor with Accurately Tailored profiles）”を開発した[2]。

2.2 技術の概要

NEATは，二つの単語検索方式を融合した検索方式，及び記事の構造・書式に対応した複数の検索条件ベクトルを用いている。これにより，高い再現率（全適合文書に対する出力文書数の割合）及び適合率（出力文書に対する適合文書の割合）でフィルタリングを行うよう設計されている。検索要求と文書との合致度（類似度）の算出方法は，一般的な検索方式であるベクトル空間モデルを基本としており，主に次の2点の拡張を行っている。

2.2.1 単語検索／文字列検索　一般にベクトル空間モデルでは，1要素が1単語の出現頻度（若しくはそれを加工した値）に対応するベクトルを用いる。日本語は膠着語（こうちゃくご）であるため，英語の場合と異なり，形態素解析によって単語切りを行う必要がある（以下，形態素解析結果を用いた検索方式を“単語検索”と呼ぶ）。通常，形態素解析には単語切り誤りによる誤差が含まれる（例えば「東京/都」と「東/京都」）ので，検索要求又は文書の形態素解析での誤差は再現率の低下を招く。

一方，文字列の完全マッチによる検索方式（以下，“文字列検索”と呼ぶ）では，単語検索のようなもれがない代わりに過剰にマッチしてしまう（例えば「スプリング」に「プリン」や「リング」がマッチする）場合がある。この性質は適合率の低下を招く。

NEATでは，それぞれの検索方式の長所短所を補うため，二つの検索方式による類似度を合成して類似度を算出している。実験により，いずれか一方の検索方式を用いるよりも，それぞれの検索方式を対等（五分五分）に合成した場合が，もっとも精度が高くなることを実験により確認している[2]。

2.2.2 文書の各領域に対応した複数ベクトルの利用

通常，文書の内容を表す単語は文書に一様に分布しているわけではない。例えば，文書の内容を表す重要な語は，見出し，最初の一文／一段落，概要などに集中していることが多い。単一のベクトルのモデルでは，文書内の単語出現分布の非一様性の取扱いが難しい。NEATでは，文書の各領域（見出し，本文，文，段落など）に対応したベクトルを設定し，それらを合成することによって類似度を算出している。

2.2.3 サービス　NEATは，電子化された新聞記事を毎日フィルタリングしてユーザーに配信を行う（株）ニューズウォッチ社において，96年から実運用されている。同社におけるNEATの全体の構成とデータの流れは図2のとおりで，大きく分けて，24時間稼働する“受信－前処理部”と一定時刻に起動される“フィルタリング－配信部”の二つから成る。

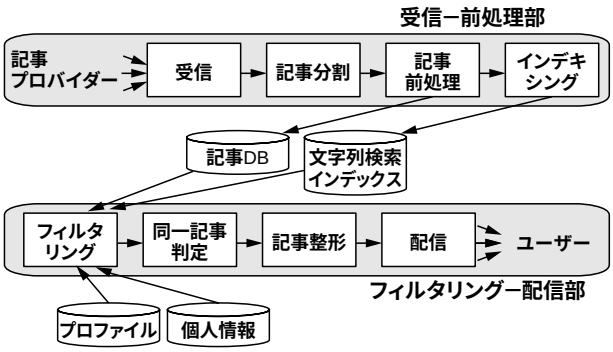

図2．NEATの構成と処理の流れ　NEATは大きく二つの部分から構成される。類似度計算は，フィルタリングで行われる。
Structure and processing flow of NEAT

前者は終日，インターネット，ISDN回線，アナログ回線などを用いて記事プロバイダー（新聞社や雑誌社など）から送られてくる記事を受信する。受信した記事データを1記事ごとに分割後，記事ごとに書式解析や形態素解析などの前処理及び文字列・単語検索用のインデックスの作成を行う。

後者は，作成された記事データベース（DB），インデックス，関係DBに格納されている個人情報，及びユーザーからあらかじめヒアリングを行って作成しておいた検索要求（プロファイル）を基にフィルタリングを行う。その後，同一内容の記事の削除などを行った後，配信媒体に応じた記事の整形を行い，メールやWWWによってユーザーに配信・提供する。実運用開始以降，高精度化や高速化を継続して行っており，フィルタリング精度に対してユーザーから高い評価を得て，業界でNo.1のサービスとなっている。

3 機械翻訳エンジン

3.1 機械翻訳の概要

インターネットを利用して母国語以外で記述されたWebページへも，母国語のページと同様にアクセスしたいという要求が非常に高くなっている。この問題を解決することを目的として，英日／日英相互機械翻訳のASPサービスをインターネット上で実現した。

3.2 技術の概要

機械翻訳の処理はトランスファー方式を採用し，形態素解析，構文解析，意味解析，変換，生成の過程から構成している。特徴は，構文解析に当社独自の語彙（ごい）遷移網文法を用いていることである。この文法は，解析を品詞レベル

と語彙レベルに明確に分けたことが特徴で，これにより膨大な文法・意味情報による高精度な解析が可能となった。

また，翻訳対象分野の自動推定及び文書のフォーマット情報による訳語の訳し分け技術も用いている。これらの技術により，業界で最高レベルの翻訳精度を実現している[1]。

機械翻訳ASPサービスの一般的な構成を**図3**に示す。一般に翻訳実行機能は，翻訳サービスを行うサイトが提供するアプリケーションの画面に埋め込まれる。図3において，①ユーザーがブラウザから翻訳を実行すると，②アプリケーションが翻訳エンジンに翻訳リクエストを出し，⑤翻訳エンジンは要求に対する翻訳処理を行い結果を返す。⑥アプリケーションは翻訳結果を整形しユーザーのブラウザに返す。翻訳対象はWebページ，サイトで提供するコンテンツ，ユーザーの入力する文などである。翻訳対象がWebページの場合は，翻訳エンジンは対象ページへ直接アクセスしページデータを取得する(③，④)。このサービスの特長は，ユーザーはブラウザソフトウェアさえあれば，日々生成される最新用語の辞書，最新の機能でのサービスを受けることが可能な点である。

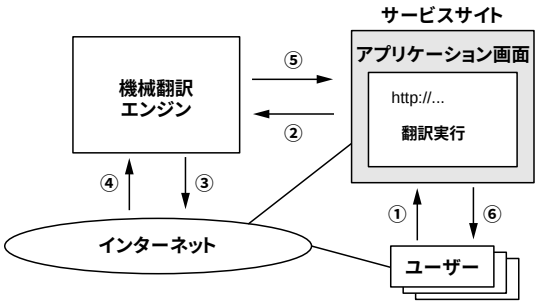

図3．機械翻訳ASPサービスの処理の流れ　ユーザーがブラウザから翻訳を実行すると，サイトで提供するアプリケーションが翻訳エンジンにリクエストを出す。
Processing flow of machine translation service

3.3　サービス

インターネット上のサイトに対して行っている機械翻訳ASPサービスについて述べる。

3.3.1　各サイトで提供する他のサービス機能との連携

連携される代表的なサービスとしては，検索機能，フィルタリング機能，メール・チャット機能，テキスト自由入力機能などがある。現在，Webページの検索機能，ニュース記事のフィルタリング機能，テキスト自由入力機能などと連携して提供されている。

3.3.2　各サイトで提供するコンテンツとの連携

(1)　コンテンツの分野が特定の場合　　サイトが特定分野の情報を提供している場合であり，特定分野のポータルサイトなどがこの例に当たる。事前にその分野に特化した辞書を構築することにより，高精度な翻訳を提供することが可能である。また，翻訳文書作成支援機能を提供することも可能である。現在，(株)セミコンダクタポータル(http://www.semiconductorportal.com)で，半導体関連のコンテンツと連携して提供されている。

(2)　コンテンツの分野が特定でない場合　　種々の分野のコンテンツを提供するサイトなどがこの例に当たる。文書の分野は多岐にわたり，各々の文書量は少ない場合が多い。この場合は，分野の自動推定による訳語の訳し分け技術が，翻訳精度向上のために非常に重要となる。現在，ポータルサイトなどでそのサイトのサービスと連携して提供されている。

4　インターネットロボットエンジン

4.1　インターネットロボットの概要

Webページは指数関数的速度で増大しており，その増加量は半年で倍とも言われている。Webページでの情報発信の良いところは手軽さにある。だれでも簡単に，すぐにページを作成して世界に向けて発信することができる。この簡易性により，厳密さより速報性に重きを置かれたものが多い。このため，ページが作成されたならば速やかに発見することが望ましい。しかし，Webページ数は日増しに増大しており，新鮮な情報はこの中に埋没しつつある。この中から新鮮な情報だけを抽出することが非常に重要な課題となっている。この解決を目的として，インターネット上に散在しているWebページを自動的に収集するインターネットロボット(スパイダーと呼ばれることもある)を開発した。

4.2　技術の概要

このロボットの特長は，新鮮な情報を優先的に収集することができることである。Webページは**図4**のように，過去の情報のほうが多く新規分の比率は低い。新鮮な情報を収集するために必要な技術は，次の二点である。

(1)　あるページが新しいかどうかを判定する。

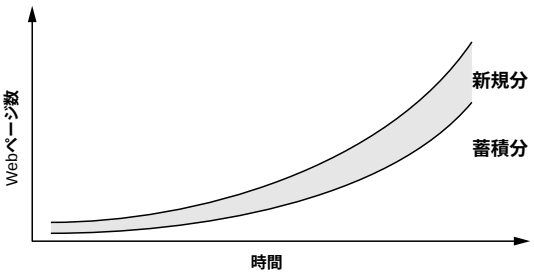

図4．Web ページにおける新鮮な情報の割合　Webページ全体に対して，新鮮なページが占める割合は低い。
Ratio of fresh Web-page information

⑵　どのあたりに新しいページができそうかを予測する。

⑴については，以下の情報を用いて判断する。これらの情報は正しいとは限らないので，個々の情報を総合的に判断する必要がある。

ⓐ　収集先のサーバが返す最終更新時刻情報

ⓑ　URL（Uniform Resource Locator）から推定される日時

ⓒ　ページの文章から推定される日付

ⓓ　過去の収集との差分情報

⑵については，取得箇所の指定と更新頻度学習を用いる。

ⓐ　取得箇所の指定　　特定のURL又はそのページに記述されているリンクに新しい情報があるという指定を外部的手段により与える。この情報に基づいて定期的に指定箇所の収集を行う。プレスリリースやWhat 's newのページを指定することが有効である。

ⓑ　更新頻度の学習　　サイトやページごとに更新頻度を自動的に学習し，その情報に基づいて，新たに作成又は変更されるページを予測して収集を行う。

このロボットの構成を図5に示す。このロボットはプロセスを複数同時に走らせることにより，そのプロセス数に応じたリニアな高速性が得られる。

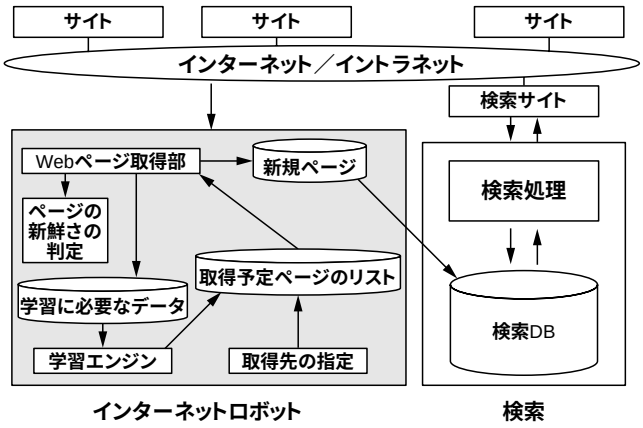

図5．ロボットと検索エンジンの構成　　ロボットにより取得した新規ページは，リアルタイムで検索DBに登録している。

Relationship between Internet robot and Web-page search engines

4.3　検索エンジンと連携したサービス

このロボットは検索エンジンと組み合わせて，（株）フレッシュアイで利用されている[3]（図5）。ロボットは，プレスリリースページなど頻繁にページが生成されるサイトへは高頻度で訪問するよう設定されている。ロボットにより取得したページは，リアルタイムで検索DBに登録している。これにより，Web上に掲載されたページは作成／更新後，最速10分で検索できる。検索結果のURLには新着の表示として，日付ではなく分単位の時刻を用いている。

また，一般にWeb検索サービスにおいて，検索結果ページが既に消滅しており，「そのようなWebページはありません（Not Found）」というメッセージが表示されることがよくある。このロボットはこの問題を軽減する能力も持っている。新鮮な情報を峻別して収集する能力は，裏返せば古い情報を分別することができるということである。これにより，古い情報や消滅したWebページをDBから削除することができる。（株）フレッシュアイでは，検索結果ページが消滅している確率を2%にとどめている。このような技術を使わない場合に比べて，これは一けた小さい値である。

ここで実現している検索サービスは，検索ASPサービスとして提供することも可能である。現在，インターネットの各サイト内のコンテンツに対して，フレッシュアイ経由で実際にサービスを提供している。

5　あとがき

今後も電子化テキストは日々作成され蓄積されていく。利用者のニーズを的確にとらえ，更に有用な新規サービスを迅速に実現していく予定である。各自然言語処理エンジンを様々に組み合わせることで，新規サービスは比較的容易に実現できる。例えば，外国のWebサイトから有用な情報を利用者別に日本語で提供するようなサービスなどである。

各エンジンの機能についても，日々のサービスに対するユーザーからの評価をフィードバックし，機能向上を継続して行っていく。

文　献

⑴　翻訳ソフト 学力テスト．日経ゼロワン2001．2001-04，p.64－69.

⑵　梶浦正浩，ほか．情報フィルタリングシステムNEATの開発．第54回情報処理学会全国大会．1997，分冊3-p.299－300.

⑶　鈴岡　節，ほか．WWW情報フィルタリング・検索システム（FreshEye）－全体システムの構成と動作．第57回情報処理学会全国大会．1998，分冊3-p.149－150.

野上　宏康　NOGAMI Hiroyasu

iバリュー クリエーション社 技術部参事。

インターネットサービスの開発に従事。情報処理学会，言語処理学会，人工知能学会会員。

iValue Creation Co.

鈴岡　節　SUZUOKA Takashi

iバリュー クリエーション社 技術部参事。インターネットサービスの開発に従事。情報処理学会会員。

iValue Creation Co.

梶浦　正浩　KAJIURA Masahiro

iバリュー クリエーション社 技術部主務。

情報フィルタリング／検索技術の開発に従事。情報処理学会，電子情報通信学会会員。

iValue Creation Co.