コミュニティ型機械翻訳サイト「訳してねっと®」

介弘　達哉　　村田　稔樹

インターネットやイントラネットの普及に伴い，Web上に多種多様なジャンルの情報や各種文書が置かれるようになり，そのページ数，文書数も飛躍的に増えている。これらのページや文書は自国語ではない場合も多く，Webページやネット上に置かれた各種文書の翻訳の需要も日々高まっている。このようなさまざまな分野の情報から機械翻訳を使ってある程度意味のわかる訳文を得るためには，翻訳する分野に応じた専門用語辞書の充実が不可欠である。しかし，従来から行われてきたような機械翻訳メーカが専門用語辞書を作成し，ユーザに提供するという方法では，作成できる分野に限界があり，ユーザが満足できるような翻訳結果にはなっていないのが現状だった。

また一方では，Linux*1) に代表されるように，ソースコードを公開し皆で協力してプログラミングを行うというオープンソースの概念でのシステム開発がよく行われるようになった。ソースコードをオープンにしたことと，共同開発を効率的に進めることができるバージョン管理ツールなどの環境を整えたことで，開発スピードが飛躍的に向上した。彼らの開発作業はボランティアベースで行われているが，皆が高いモチベーションを持って開発に当たっている。

同様に翻訳作業においても，同じ分野の技術者同士がインターネット上で仲間を募り，技術マニュアル等の翻訳を協力して行い，翻訳結果をインターネット上に公開しているようなサイトをよく見るようになった。彼らの翻訳作業もボランティアベースで行われていることが多い。

そこで我々は，機械翻訳や翻訳メモリ*2) などが利用でき，彼らの翻訳作業を支援できるようなオープンな環境を提供できれば，彼らの開発も効率的に行えるようになるし，そこに登録された辞書によって翻訳システム自体の訳質も向上すると考え「訳してねっと*3)」というサイトを構築した。

本稿では，現在インターネット上で公開実験を行っている「訳してねっと」および，「訳してねっと」の機能を一部修正し，企業向けにアレンジした「会社で訳してねっと」の概要を説明し，機械翻訳システムを複数のユーザが協調して使うことの有効性を述べる。

サイトの特長

本サイトのシステム構成を**図1**に，トップページを**図2**に示す。本サイトの特長として

- パターン翻訳方式を使った記述能力の高い翻訳機能
- コミュニティごとに辞書データの編集が可能な辞書管理機能
- 専門用語を自動抽出する辞書登録支援機能

などが挙げられる[1)]。

本章ではこれらの特長を順番に説明する。

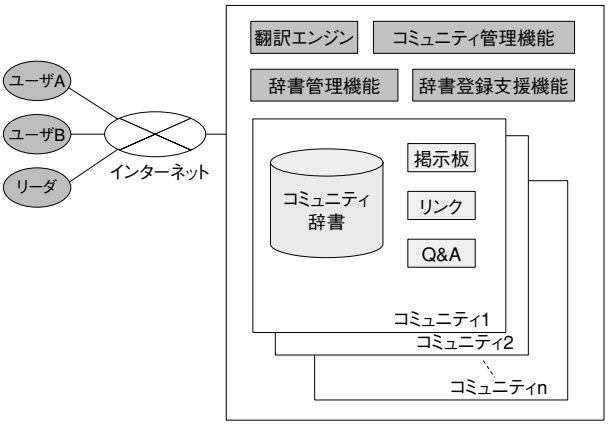

図1　システム構成

(1) 翻訳機能

翻訳は英語から日本語と日本語から英語の2通りが可能で，翻訳エンジンは当社で開発したパターン翻訳方式を用いている[2)]。パターン翻訳では分野に特有な言い回しなど通常のユーザ辞書では記述できないような対訳表現も翻訳パターンとして登録し，それを機械翻訳の翻訳結果に反映させることができる。たとえば

download［NP］at once
⇔［NP］を一括ダウンロードする

*1)Linuxは，Linus Torvaldsの米国およびその他の国における登録商標または商標です。　*2)過去に翻訳した結果やその修正結果を対訳形式でデータベースに蓄積し，再利用できるようにしたもの。翻訳作業の効率化が図れる。　*3)訳してねっと は沖電気工業(株)の登録商標です。

図2　トップページ

といった変数付きの辞書データの登録も可能である。([NP]の部分が変数を意味している。)

本サイトでは，テキスト翻訳，ファイル翻訳，Web翻訳の3種類の翻訳方法を用意している。

① テキスト翻訳

テキスト翻訳は，テキストをWebフォームに入力し，その入力した文章を翻訳する。

② ファイル翻訳

ファイル翻訳はユーザのコンピュータ上に存在するローカルファイルを翻訳する。テキストファイル，HTMLファイル，XMLファイル，Microsoft*4) 社のWord，Excel，PowerPoint*5) などのファイルを指定できる。

図3　Web翻訳結果のページ

③ Web翻訳

図3にWeb翻訳結果の画面を示す。

Web翻訳はURLを指定することによって，指定したWebページの翻訳を行う。図のように対訳で表示する機能もあるので翻訳結果のチェックに便利である。

(2) 辞書管理機能

本サイトのコミュニティ構成を**図4**に示す。

本サイトにはたくさんのコミュニティがあり，それぞ

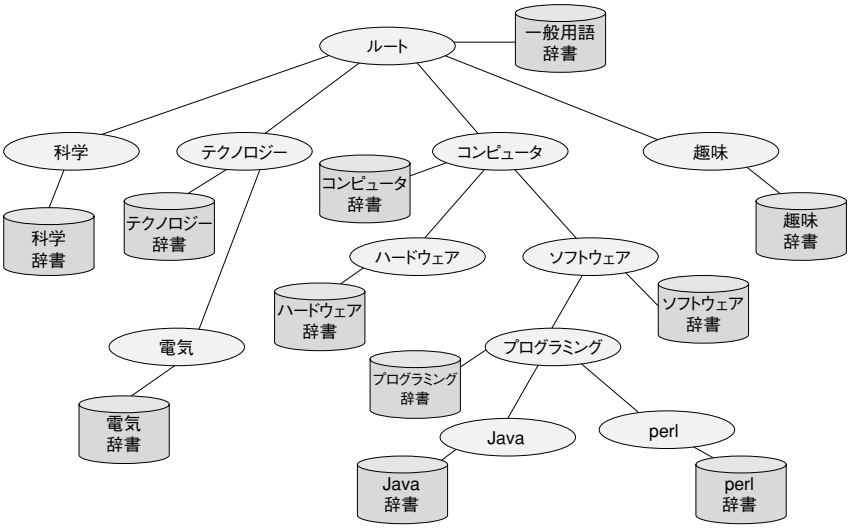

図4　コミュニティ構成

*4) MicrosoftはMicrosoft社の米国およびその他の国における登録商標です。　*5) Microsoft Word, Microsoft Excel, Microsoft PowerPointはMicrosoft社の米国およびその他の国における商標または登録商標です。

れのコミュニティが1つの分野に対応している。コミュニティは図4のようにツリー構造をなしていて，下の層に行くほど細かい分野になっている。それぞれのコミュニティにはコミュニティ辞書と呼ばれる辞書がひとつずつ存在する。

各コミュニティの管理は，ユーザ主導で行われ，そのコミュニティのメンバはそのコミュニティ辞書の追加，修正，削除，検索などを行うことができる。また，コミュニティリーダと呼ばれているコミュニティのまとめ役をおくこともでき，コミュニティリーダは他のメンバが登録した辞書の内容を修正することもできる。

辞書登録ページを図5に示す。

辞書データには，見出し語，訳語，品詞，活用情報，意味情報，その用語の説明文などが登録できる。

辞書登録の方法としては，図5のようにブラウザ上の登録フォームに入力して登録する方法と，Excelなどを用いてCSV形式の辞書データを作成し，サーバにアップロードすることによって一括して登録する方法の2つの方法が用意されている。

コミュニティ内で翻訳処理を行うと，そのコミュニティの辞書および，そのコミュニティの上位の辞書を使って翻訳する。たとえば図4のようなコミュニティ構成においてJava*6) コミュニティで翻訳する場合は，Java辞書，プログラミング辞書，ソフトウェア辞書，コンピュータ辞書，一般用語辞書を使って翻訳する。もし複数の辞書に，同じ辞書データがあった場合は下位のコミュニティの辞書が優先される。

また，翻訳に使用する辞書を自由に選択できるようになっていて，たとえばコンピュータ関係の特許の翻訳を行う場合は，特許特有の言いまわしなどが登録された特許辞書とその特許の分野であるコンピュータ辞書を使うといった指定もできる。辞書データを登録すると，すぐに翻訳用辞書の再構築が行われるので，登録した辞書データが翻訳結果に正しく反映されているかどうか，その場でチェックできるようになっている[3]。

(3) 専門用語自動抽出機能

大量の辞書データを効率よく登録するためには，辞書登録を支援する機能が必要である。我々は，テキスト上の専門用語を自動的に抽出する研究[4] や，英語と日本語の対訳テキストから自動的に語の対応を付ける研究[5] など，文書を解析し，辞書データを半自動的に獲得するようなさまざまな研究を続けてきた。本サイトではユーザが指定したWebページから専門用語を自動的に抽出し，その専門用語を見出し語として提示して辞書登録画面を表示する機能を備えているので，効率的な辞書登録ができる。

(4) その他の機能

図6にコミュニティのメインページを示す。

よく翻訳するサイトや文書は図6のようにあらかじめ登録しておくことができ，一度登録しておくとリンクをクリックするだけで，登録したコミュニティの辞書を使って翻訳処理が行われ，結果が表示される。

また，メンバ間でそのコミュニティに関する意見交換や辞書登録データに関する議論などができるように，コミュニティ毎に掲示板やQ&Aの機能を設けている。

さらに，ユーザの辞書登録に対するモチベーション向上のため，図2のように辞書登録語数をユーザごとやコミュニティごとに順位付けしてトップページに表示するようにしている。

図5　辞書登録ページ

図6　コミュニティメインページ

*6) Java およびすべてのJava関連の商標およびロゴは，米国およびその他の国における米国Sun Microsystems, Inc.の 商標または登録商標です。

訳してねっと の現状

本サイトは現在公開実験中で，インターネットを介してだれでも無料で利用できる[6]。毎日多くのユーザから多数のアクセスがあり，いろいろなコミュニティで翻訳や辞書登録が行われている。活発なコミュニティのひとつにJakartaコミュニティがあり，そこではJa-Jakartaプロジェクト[7] の方々が英文のマニュアルを翻訳するための下訳として本サイトを利用している。彼らは，訳語の統一や工数の削減に本サイトが有効であるとの見解を示している[8]。

会社で訳してねっと

企業のグローバル化に伴い，社内での翻訳の需要も増えてきている。「会社で訳してねっと」は，「訳してねっと」をベースに企業向けに製品化したもので，企業などのイントラネット内に導入して利用する。「会社で訳してねっと」の特長を以下に述べる。

① 自由度の高いコミュニティ構成

コミュニティはそれぞれの企業で自由に作成できるので，部署ごとやプロジェクトごとなどその企業の目的にあった構成にすることできる。

② 充実した辞書登録・管理機能

自社の製品名やその企業特有の言い回しなども辞書データとして簡単に登録できる。文単位の登録や，変数付きの辞書データも登録できるので，改版が頻繁に行われるマニュアルなどを翻訳する場合，一度辞書データを登録しておけば，改版時の翻訳作業量は最小限で済む。

③ 翻訳知識の共有

翻訳の得意な社員あるいはその分野に詳しい社員が辞書データを登録しておけば，英語の苦手な社員や専門知識を持たない社員でも，その辞書データを使った高品質の翻訳が可能となる。また，「訳してねっと」で作られた辞書を利用することもできるので，翻訳に必要な最新の用語などに関する知識も自動的に蓄えられていく。

以上のように「会社で訳してねっと」を利用することにより，企業内の翻訳生産性を向上させることができる。

また，「会社で訳してねっと」はニーズに合わせていろいろとカスタマイズできるように作られているので，企業だけでなく，大学や自治体などにも導入可能である。

あ と が き

ユーザの協調作業により翻訳品質を向上していくことができる機械翻訳サイト「訳してねっと」およびそのイントラネット版である「会社で訳してねっと」について，その思想および機能の概要を述べた。

今後は，もっと沢山の辞書データが集まり，もっと多くの人に利用してもらえるように，ユーザの意見を取り入れながら，より使いやすいシステムに改良していくとともに，集まった辞書データの妥当性の自動チェックや自動分類などの研究[9] も行っていく予定である。

なお本研究は独立行政法人情報通信研究機構平成14年度基盤技術研究促進制度に係る研究開発課題「多言語標準文書処理システムの研究開発」の一環として行われている。 ◆◆

■参考文献

1）Shimohata, S., *et al.*: Collaborative Translation Environment on the Web. , MT Summit VIII, pp.331–334, 2001

2）Kitamura,M., Murata,T.: Practical Machine Translation System allowing Complex Patterns ,MT Summit IX, pp.232–239, 2003

3）Murata,T., *et al.*: Implementation of Collaborative Translation Environment `Yakushite Net’ , MT Summit IX, pp.479–482, 2003

4）Shimohata,S., *et al.*: Retrieving Collocations by Co-occurrences and Word Order Constraints, In Proceedings of ACL-EACL '97（35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics）, pp.476–481, 1997

5）Kitamura,M., Matsumoto,Y.: Practical Translation Pattern Acquisition from Combined Language Resources, First International Joint Conference on Natural Language Processing （IJCNLP-04）, pp.652–659, 2003

6）訳してねっと: http://www.yakushite.net/

7）Ja-Jakartaプロジェクト：http://www.jajakarta.org/

8）小山 博史，他：The Ja-Jakarta Project における分散システム構築と運用，情報処理学会研究会報告 2004-DSM-33/電子情報通信学会技術研究報告 TM2004-4，2004年

9）佐々木美樹，他：コアワードを利用した単語の分野自動判定，FIT（情報科学技術フォーラム）2003論文集，pp.171–172，2003年

●筆者紹介

介弘達哉：Tatsuya Sukehiro. 研究開発本部 ユビキタスシステムラボラトリ

村田稔樹：Toshiki Murata. 研究開発本部 ユビキタスシステムラボラトリ