多言語間翻訳情報共有システムの開発

安藤 真一　佐藤 研治　奥村 明俊
e-mail: {ando, satoh, okumura}@ccm.cl.nec.co.jp
NEC C&C メディア研究所

1. はじめに

マニュアルや仕様書の翻訳では用語や文体の統一が問題となる。特にこれらの文書には頻繁に改訂が加わるものもあり、実際の翻訳対象が文書中の一部分だけであっても、用語や文体の統一のために文書全体を検証する場合が少なくない。これに対して、従来より類似用例検索システム[1,2]が提案されている。これらは翻訳用例として格納された対訳データの中から、ユーザの入力に類似した文を検索することで、「訳し方の見本」を提供するものである。この手法では類似文検索は原言語のみを対象とするため、対訳を用意するだけで多言語化も容易に実現できる。ここではさらにユーザが独自に収集した対訳例文も用例として利用することができる。ただし開発サイクルが早い分野などでは個々のユーザが用例を充実させるには限界がある。そこで本稿ではこれらの翻訳用例を複数のユーザ間で共有し、より効率的に互いの翻訳ノウハウを利用することを考える。特にここではその実現例として、当社のマニュアル翻訳部門向けに我々が開発した多言語間翻訳情報共有システム Web TransFinder について報告する。

2. 翻訳用例の共有

従来の類似用例検索システムでは文単位の翻訳用例を独立に扱っている。しかし個々のユーザが対訳例文を登録することを考えると、ここには以下の3つの問題がある。

1) 対訳例文に分野特有の表現が含まれる場合、これらを未整理のまま用例に登録すると様々な分野に固有な表現が混在してしまい、後の用語や文体を統一する作業に悪影響を及ぼす。このため、これらの表現については何らかの整理が必要となるが、個々の翻訳用例に対して予め分野などを設定することは一般に難しい。
2) 対訳例文が照応表現や省略を含む場合、これを用例データに追加したとしても、類似性の判定に必要な語彙がその例文の外にあるために、類似用例として出力できない場合がある。
3) 出力されたとしても、照応表現や省略を含んだ文だけでは、それが用例として適切かどうかの判断は難しい。

そこで我々は文単位だけでなく、出典元の文書情報を含めて翻訳用例を管理する多言語間翻訳情報共有システムを開発した。各用例の出典元文書や、その中での出現順序を保持することで、出典元文書の持つ分野情報や、また各用例に対する出典元での文脈情報が利用できるようになる。これらの情報を利用することで、本システムでは下記の機能を実現した。

1) 対象分野を指定する検索対象文書選択
2) 文脈を考慮した類似文検索
3) 検索結果からの文脈表示

以下では上記の機能に加えて、本システムの文書単位管理を支える用例管理機構やユーザ文書登録機構を取り上げ、各々について述べる。

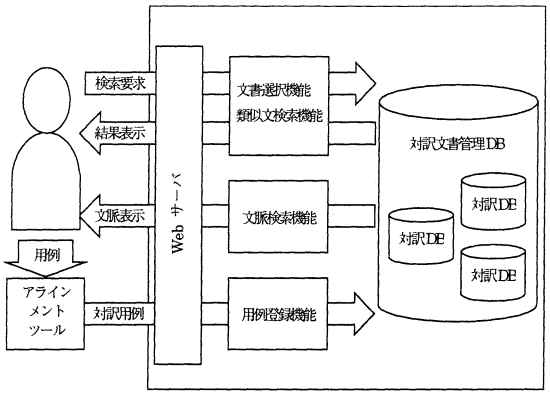

図1．システムの概念図

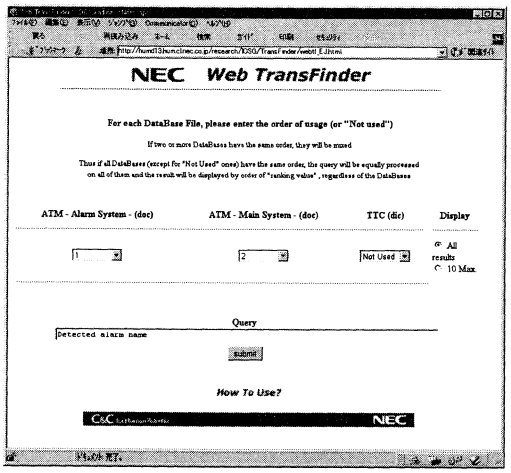

図2．検索インターフェイス画面例

3. システムの概要

本システムは過去の翻訳用例を文書単位で管理し、類似文検索機能によって必要な用例を提供するシステムである。システム全体の概念図を図1に示す。以下では、本システムにおける翻訳用例の管理、アクセス、登録の各々について述べる。

3-1. 翻訳用例の管理

本システムでは翻訳用例を文書単位で管理する。図1中の対訳 DB は原文と訳文の対応関係を管理するデータベースであり、さらに出典元文書での各文の出現順序も保存している。また対訳文書管理 DB はシステム内の全文書に対して、対応する対訳 DB 名と文書名、登録者名、アクセス可能なユーザ名、文書種別を格納する。ここで文書種別としては現在、文から構成されるいわゆる文書を意味する「文書」と、索引などの用語集を表わす「辞書」を用意している。この2つは利用形態や文脈の扱いなどの点で違いがあるためである。

3-2. 翻訳用例へのアクセス

本システムは Web サーバの上に構築されており、類似文検索機能も一般の Web ブラウザから利用することができる。以下では本システムの特徴である、検索対象文書選択、文脈上での類似性を取り入れた類似文検索、検索結果からの文脈表示について述べる。

3-2-1. 検索対象文書の選択

一般に同じ分野の文書では同一の専門用語が用いられ、また使用される表現も酷似している。このため翻訳対象と似た分野の文書を検索対象とすることで、翻訳用例として有用な用例が得られやすいと考えられる。そこで本システムでは、翻訳用例として管理している文書毎に検索対象とするか、検索時の優先順位はどうするかをユーザが指定できるようにした。図2に類似文検索のインターフェイス画面例を示す。この画面中段には本システムが管理している文書の文書名と種別が一覧表示されており、その各々の下には検索順序を指定するリストボックスが用意されている。ユーザは検索順序の値を変更することによって、検索対象とする文書の選択や、複数の検索対象文書に対する検索優先順序を変更することができる。

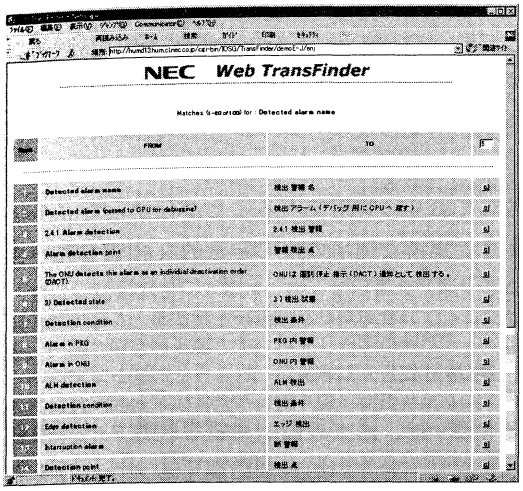

図３．検索結果の表示画面例

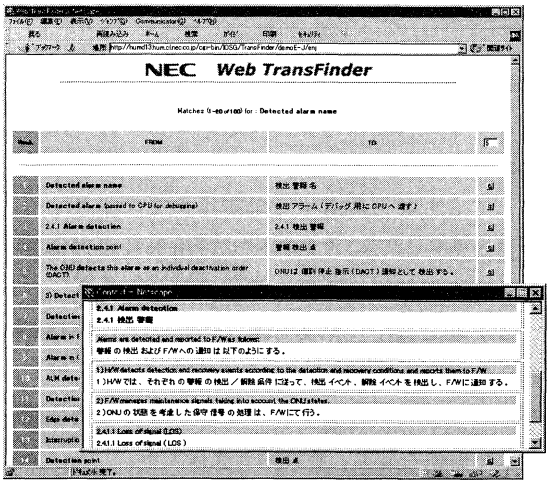

図４．検索結果の文脈表示例

3-2-2. 文脈を考慮した類似文検索

類似文検索では、入力文と各用例との間の類似度を計算し、その大きさに従って類似用例を出力する。ここで単文間の類似度には、ユーザにとっても直感的に分かりやすい尺度として、２文間で一致した形態素数と２文の全形態素数との割合に基づく尺度を用いた。ただし検索対象に含まれる照応表現や省略を含む用例への対処として、実際の類似度の計算では文脈の影響を考慮している。具体的には入力文 Q と検索対象文書の i 番目に出現する用例 Xi との間の類似度 Sim(Q,Xi)として、下式を用いた。

$$Sim(Q,Xi) = 0 \qquad \text{if } Sim'(Q,Xi) = 0$$

$$Sim(Q,Xi) = Sim'(Q,Xi) + \Sigma(j)\{\alpha(i\text{-}j)\,Sim'(Q,Xj)\} \qquad \text{otherwise}$$

ここで Sim'(Q,X)は、Q と X の間で一致した形態素に基づいて計算した類似度である。また上の第２式第２項は文脈の影響を表わす項であり、$\alpha(i\text{-}j)$は Xi からの距離が(i-j)である文脈 Xj の影響を表わすパラメータである。この式を用いることにより、入力文と用例の間で直接一致する形態素が少なくても、その用例の前後に入力と一致する形態素が多く出現すれば類似度が大きくなる。このため照応表現や省略を含んだ用例についても、その先行詞が用例の近くに現れることで高い類似度を持つことになり、これらを検索結果の上位に出力することができるようになる。

3-2-2. 検索結果の文脈表示

検索結果の表示画面例を図３に示す。ここでは検索結果として得られた用例が入力文との間の類似度に従って上から順に表示される。また用例中の入力文と一致した部分はハイライト表示される。図３では一行に一文単位の用例を表示しているが、本システムは出典元文書やその中での位置情報も保持しており、元文書中での使われ方も参照することができる。この文脈表示画面の例を図４に示す。出力された用例が照応表現や省略を含む場合や、文脈の影響で類似用例と判定された用例を参照する場合には、文単位ではその用例の妥当性を判断しにくいため、

文脈表示が有用であると考える。

3-3. 翻訳用例の登録

より多くの対象において翻訳支援を行うためには、なるべく多くのユーザから対訳文書を提供してもらい、共有する翻訳用例を充実させる必要がある。このためユーザの用例登録作業に対する支援機能も重要な機能の一つとなる。本システムでは１）原文と訳文の対応付け作業を含む対訳ＤＢの作成、２）対訳ＤＢの Web サーバへの登録という２つの手続きによる用例登録作業を前提として、各作業を支援する。

まず１）については、原文訳文の対応付け作業を支援することを目的として、アラインメントツールを開発した。通常、アラインメントツールは文章を文単位に区切った後、辞書や統計情報を用いて原文と訳文の実際の対応関係を確認する。しかしここで対象とする文書が仕様書やマニュアルであり、文の対応関係についてもほとんど忠実に翻訳されるため、本ツールでは文切りまでの機能でこれに対応している。逆に今回対象とする文書には、章見出しや個条書き、あるいは表、図といった特殊なフォーマットが多く含まれる。このため、これらのフォーマットの解析機能を導入し、各々の構造の間にも適切に区切りを挿入できるようにした。

また２）については類似文検索と同様に Web インターフェイスの上で提供されている。ユーザはここで自分の環境にある対訳ＤＢに対して文書名、文書種別などを指定することで、このＤＢをサーバ上に登録することができる。

4. 運用状況

本システムは現在、ATM に関する日英マニュアル(3 冊分、7500 文)、PC 周辺機器に関する英仏マニュアル(1 冊分、3000 文)、交換機に関する英西マニュアル（1 冊分、2000 文)を翻訳用例として用いて、約 20 人のユーザを対象に試験的な運用を行っている。現在、各機能に対する定性的な評価を行っているが、特に文脈表示については専門用語の定義などを調べるためにも利用されており、有用であることが分かった。今後より大規模な運用の中で定量的な評価も進めていく。

5. おわりに

本稿では、過去の翻訳用例を複数ユーザの間で共有し、お互いの翻訳に関するノウハウを利用するための環境として、我々が開発した多言語間翻訳情報共有システムについて報告した。ここでは特に出典元文書やそこでの出現順を含めて翻訳用例を管理することにより、自然な形でユーザの分野選択を助け、また文脈までを含めた形で翻訳作業を支援できることを述べた。

今後の展開としては、翻訳支援の観点からは機械翻訳システムとの統合が、あるいは文書単位の管理という観点からは文書管理システムとの統合が考えられる。特に後者については、文書作成時の参照／非参照の関係や引用／非引用の関係から各々の文書を分類し、新たな文書を作成する際の支援に役立てる環境を開発している[3]。今後この環境との統合についても検討していく予定である。

参考文献

[1] 隅田,堤:"翻訳支援のための類似用例の実用的検索法",信学会論文誌 Vol.J74-D-II, No.10, pp.1437-1447,1991.

[2] "Translation Memory", http://crl.nmsu.edu/Research/Projects/oleada/tm.html, 1996.

[3] K.Satoh, A.Okumura: "Documentation Know-how Sharing by Automatic Process Tracking", Proc. of IUI99, pp49-56, 1999.