対訳文書の文・単語対応付け技術を利用した対訳例検索システム

2H－1

北村 美穂子　　山本 秀樹

沖電気工業株式会社 研究開発本部 関西総合研究所

1　はじめに

日本人がビジネスや学会などの専門的な分野で英語文書を使用する機会が増えている．これらの文書は，市販の辞書には存在しない専門用語に代表される文書固有の表現を多く含む[1]．英語文書を作成する際，市販の辞書だけでなく，手本とする対訳文書から検索された訳語や表現を利用できれば，正確かつ微妙なニュアンスを伝える英語文書を作成することができる．

対訳文書は，翻訳者の訳語の知識だけでなく，対象言語や対象分野などの種々の特徴を反映した総合的な知識を持つ文書である．既存の対訳文書をデータベース化し，それを直接利用するインターフェイスを用意することによって，データベースの使用者は，分野や書式などの文書情報や，対象言語や対象分野などの総合的な情報を英語文書の作成時に直接利用することができる．

本稿は，対訳文書の文対応付けの技術と専門用語や文書固有の表現の対訳ペアを自動的に抽出するという技術を利用して，翻訳支援の資源として既存の対訳テキストを有効活用できる対訳例検索システムを提案する．

2　文・単語の対応付け

通常，人間が翻訳した対訳文書は，英文と日本語文が必ずしも1対1に対応せず，2文が1文にまとめられて翻訳されたり，時には翻訳されなかったりする．したがって，対訳文書を単に電子化しただけのものは，ある表現が対訳文書中のどの文でどう翻訳されているかを検索する対訳例検索には利用できない．

また，仮に対訳文書データベースの英文と日本語文とが1文単位で対応がついていて，利用者が欲する特定の表現を含む対訳文を検索できるようになったとしても，特定の表現が対応する対訳文中のどこに対応するかを見つけるのは利用者にとって簡単ではない．以上により，対訳例検索システムには，電子化された対訳文書の文対応を自動的にとる技術および，対訳文の中で表現間の対応をとる技術が必要となる．

日本語と英語の対訳文書を文や単語単位で対応付ける試みは，対訳文書に出現する単語の統計情報や対訳辞書を用いる方法が知られている[2]．機械翻訳用の十分な対訳辞書を保持するため，我々は文の対応付けの方法として対訳辞書を用いた方法[3]を用いる．これは，各文書を1文ごとに分割し，形態素解析を行ない，対訳辞書を用いて日本語文と英語文の類似度を計算し，この類似度の和が文書全体で最大になるような組合せを選択することによって文の対応付けを行なう方法である．

一方，単語の対応付けは，単語単位だけでなく，複数の単語からなる対訳表現も抽出できる方法[4]を用い，さらに対訳辞書の情報も利用できるように拡張する．この方法は，対訳文書の統計情報を用いて，共起性の高い日本語と英語の単語列ペアを抽出する．抽出された対訳表現を対訳辞書に登録し，その対訳辞書と対訳文書の統計情報を用いて新たな対訳表現を順に抽出していく方法である．

以上の対応付けの方法を用いて，OCRで読み取った約1万文の科学技術論文の対訳文書で実験した結果，文の対応付けの精度は91%となった．さらに，その対応付けされた文書を用いて対訳表現の抽出を行なった結果，市販の辞書に登録済みの2,676ペアの対訳表現と，登録されていない1,325ペア(精度82%)の対訳表現を抽出することができた．

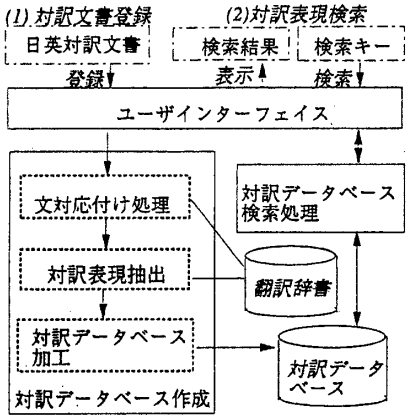

図 1: 対訳例検索システムの概略

3　対訳例検索システム

3.1　システム構成

Translation Retrieval System using Alignment Data from Parallel Texts
Mihoko Kitamura and Hideki Yamamoto
Oki Electric Industry Co., Ltd.
Kansai Laboratory, Reseach & Developement Group
{kita,hyama}@kansai.oki.co.jp

```
{対訳DB}          ::=  {対訳文書の並び}
{対訳文書の並び}  ::=  {対訳文書} | {対訳文書}{対訳文書の並び}
{対訳文書}        ::=  {英語文書}{日本語文書}
{英語文書}        ::=  {英語文書タグ}{文の並び}
{日本語文書}      ::=  {日本語文書タグ}{文の並び}
{文の並び}        ::=  {文} | {文}{文の並び}
{文}              ::=  {文タグ}{単語の並び}
{単語の並び}      ::=  {単語} | {単語}{単語の並び}
{単語}            ::=  {単語タグ}{見出し} | {見出し}
{見出し}          ::=  {出現形}\\{標準形}
{英語文書タグ}    ::=  < text lang="eng" id= 文書識別番号>
{日本語文書タグ}  ::=  < text lang="jap" id= 文書識別番号>
{文タグ}          ::=  < sent id= 文識別番号>
{単語タグ}        ::=  < r id= 単語識別番号>
```

図 2: 対訳データベースの定義 (BNF 記法)

対訳例検索システムは，事前に登録した日本語と英語の対訳文書を対訳データベースとして持ち，検索キーとして与えられる原言語の表現に対応する目的言語の表現を容易に検索できることを目的とする．日本語から英語，英語から日本語の双方の検索が可能であり，原文に出現する形(出現形)と辞書に登録されている形(標準形)の両方で検索することができる．検索キーに対応する訳語とその検索キーを含む原文および訳文の全文を検索結果として表示する．

本システムの構成図を図 1に示し，対訳文書の登録，対訳表現の検索の 2 つの処理について説明する．

3.2 対訳文書の登録

対訳例として使用したい対訳文書を指定する．指定された対訳文書は，文献 [3] の方法によって文単位で対応付けられ，次に文献 [4] の方法によって対訳表現が抽出される．抽出された文対応付け結果と対訳表現は，図 2の定義にしたがって加工され，対訳データベースに格納される．

対応付けの情報は，文書，文，単語の単位で，SGML 方式 [5][6] に準拠したタグを用いて付与される．各タグには識別番号が与えられ，同じ識別番号を持つ文書，文および単語同士が対応関係にある．また，単語は出現形と標準形の両方で格納される．

3.3 対訳例の検索

本システムは，使用者が入力した検索キーの文字種により，検索キーの言語を認識し，英語文書または日本語文書から検索を行う．検索キーは，単語だけでなく，「事実上困難になる」のような句や節レベルの検索もできるようにしている．実際の検索は，英語文書または日本語文書を構成する各単語の出現形をつなぎ合わせて原文を復元し，復元した文と検索キーと比較する．両者が一致しない場合は，標準形で同様の処理を行なう．

両者が一致すれば，一致した単語が持つ単語識別番号，文識別番号，文書識別番号を取り出し，それと同じ番号を持つもう一方の言語の単語，文，文書を抽出し，検索キーを含む原文，それに対応する訳文，検索キーに対応する訳語を表示する．抽出結果が複数存在する場合は，全ての結果を抽出し，表示する．

4 実装

本システムを WWW の CGI コマンドとして実装した．検索結果の表示例を図 3に示す．

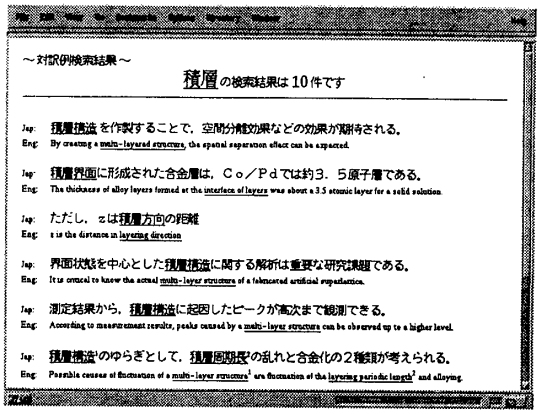

図 3: 検索結果表示例

5 おわりに

対訳文書の文対応付けの技術と，専門用語や文書固有の表現の対訳ペアを自動的に抽出する技術を用いて，既存の対訳文書から対訳例を検索するシステムを提案した．

本システムを使用する翻訳作業者は，対訳例を伴った訳語の提示によって，翻訳対象となる文書の分野や文中における使用方法を考慮に入れた訳語選択が可能となる．

現段階では，検索結果の表示は文書の出現順序であり，結果の分類など特別な処理は行なっていないが，文や単語の対応付け時に得られる対応度を利用したり，検索キーを含む対訳文の出現単語情報を利用することにより，検索結果の分類も可能である．

今後，本システムを評価し，使用者が効率良く対訳例を検索できるための改良を行なう予定である．

参考文献

[1] 北村美穂子，松本裕治：対訳コーパスを利用した翻訳規則の自動獲得，情報処理学会論文誌，Vol. 37, No. 6, pp. 1030–1040 (1996).
[2] 宇津呂武仁，松本裕治：コーパスを用いた言語知識の獲得，人工知能学会誌，Vol. 10, No. 2, pp. 197–204 (1995).
[3] 介弘達哉，下畑さより，松下久明：差分翻訳システムにおける対訳文書の文の対応付け，電子情報通信学会 1995 年総合大会講演論文集 (D-116), p. 122 (1995).
[4] 北村美穂子，松本裕治：対訳コーパス中の共起頻度に基づく対訳表現の自動抽出，情報処理学会研究報告 96-NL-1，第 96 巻 (1996).
[5] Bond, F., Takahashi, Y., Yamada, S. and Nishigaki, M.: Still Tagging an Aligned Japanese/English Corpus, 言語処理学会 第 2 回 年次大会発表論文集, pp. 205–208 (1996).
[6] 田中洋一：文書記述言語 SGML とその動向，情報処理学会誌，Vol. 32, No. 10, pp. 1118–1125 (1991).