周辺文脈を利用した固有名詞のクロスリンガル情報検索

後藤功雄　江原暉将
{igoto, eharate}@strl.nhk.or.jp
NHK放送技術研究所

1 はじめに

NHKでは海外向け放送や音声多重放送のニュースに用いる英語の原稿を、日本語原稿から翻訳して作成している。我々はこれまでに、この翻訳業務を支援するために、翻訳用例提示システム[1]を開発して現場に導入し、翻訳作業の効率化を図った。翻訳用例提示システムとは、過去の対訳記事をデータベースとして保持し、ユーザが検索したい表現を入力すると、その表現に似た表現を含む記事とその対訳の記事を提示するものであり、翻訳メモリ型の翻訳支援システムである。このシステムの運用開始後にユーザへのアンケートを実施したところ、固有名詞の調査のための利用が多いことが分かった。

ニュースは最新の出来事を取り上げるため、新しい外国の人名等の固有名詞が翻訳対象の日本語原稿中に含まれることが多い。それらの固有名詞の訳の調査は、翻訳作業の中では大きな負担となっている。

そのため、外来語の固有名詞の効率的な調査方法が必要とされている。しかし、日本語の問い合わせで英語の文書群の検索を行うクロスリンガル情報検索では、一般的には検索対象の単語が辞書に登録されていない場合に検索することができず、新しい固有名詞の検索を行うことは困難であった。

本稿では、日本語原稿中のカタカナで表現された外国の人名等の固有名詞をインターネット上のWebページからクロスリンガル検索を行う手法について述べる。この手法の特徴は、辞書に登録されていない固有名詞であっても、周辺文脈の単語を利用することによって検索を可能にしていることである。また、この手法による実験の結果についても報告する。

2 周辺文脈を利用したクロスリンガル情報検索手法

2.1 周辺文脈の利用

日本語ニュース記事中の、カタカナで表現された新しい外来語の固有名詞の英訳語の調査は、辞書を引いて調べるというわけにはいかない。そのため、インターネット上の膨大な情報から検索しようとしても、検索対象の英訳をキーワードとしてクロスリンガル検索を行うことはできない。

その問題を解決するために、周辺文脈を利用して、情報検索を行うことを考える。

翻訳対象の日本語記事に含まれるカタカナで表現された外国の人名等の固有名詞については、記事の文脈中にその固有名詞の所属や肩書き等の社会的位置付けの説明が含まれている。

また、Webページなどの大きなデータベース中から、外来語の固有名詞を検索する際にも、同じ様な綴りで全く別のものも存在する可能性があり、検索された固有名詞についても、その単語に関連している周辺文脈を無視することはできない。つまり、検索結果についても、検索された単語の社会的位置付けが明らかにされている情報でなければ役に立たない。

そこで、日本語記事中の外来語の固有名詞に関係の深い社会的位置付けを説明している周辺文脈を英語に翻訳し、それらを検索キーワードとして検索を行い、得られた情報の中から、キーワード周辺の固有名詞を推定して、日本語の固有名詞と発音を基準とした比較を行い、類似度の高い単語を検索結果とすることによって、クロスリンガル情報検索を行う方法を提案する。

2.2 周辺文脈の抽出

翻訳者は、対訳エディタを中心とした統合翻訳支援環境-Translaters' Workbench-[2]を利用して翻訳作業を行うことにより、図1のように日本語記事の検索対象部分をマウスで選択して検索ボタンを押すだけで、システムは検索対象の固有名詞だけでなく、周辺文脈を取得することができる。

この検索手法では、周辺文脈で検索対象の固有名詞に関係の深い単語を適切に選択することが必要である。

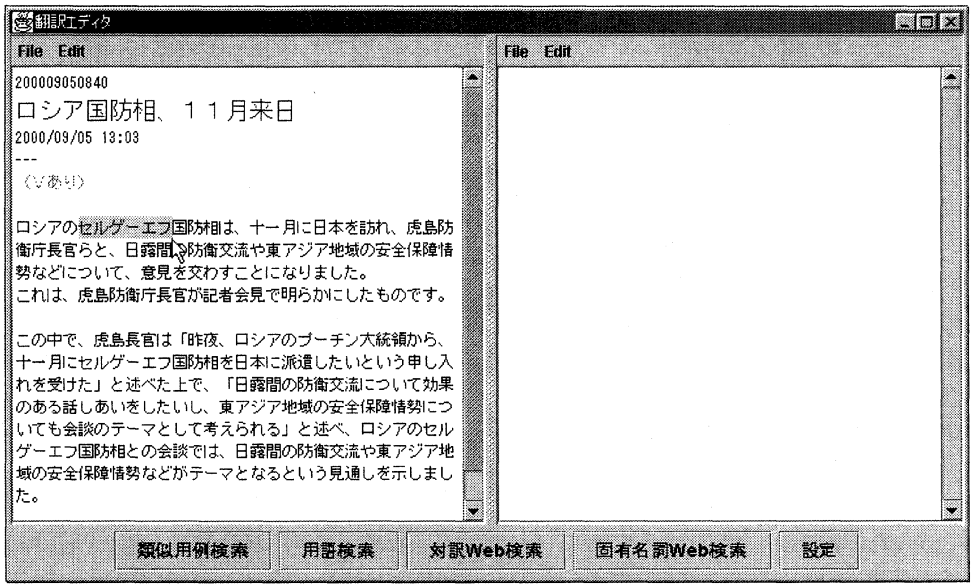

図1 Translaters' Workbench

固有名詞が人名であれば、社会的位置付けは、図2のように、係り受けの関係から所属する組織名が人名の前に表現され、肩書きが人名の後に表現される場合が多い。

そこで、人名を検索対象とした場合、形態素解析結果の人名部分より前に位置している「組織名」または「地名」を表す単語が「所属」となり、人名部分の後に位置する名詞類が「肩書き」になる場合が多いと考えられる。

2.3 クロスリンガル検索

周辺文脈より抽出したキーワードによりクロスリンガル検索を行うために、辞書を用いて英語に変換する。人名の場合は、「所属」と「肩書き」それぞれを英語に変換し、訳語が何通りもある場合は、それぞれ「所属」と「肩書き」の組み合わせを作成して、キーワードとする。

このようにして作成したキーワードを用いて検索を行った結果については、「所属」と「肩書き」の2つのキーワードが位置的に近くに存在するものほど、検索対象の人名が含まれている可能性が高いと考えられるので、近いものを優先する。さらに、検索結果内のキーワード近辺の固有名詞を推定して抽出する。

この流れを図3に示す。

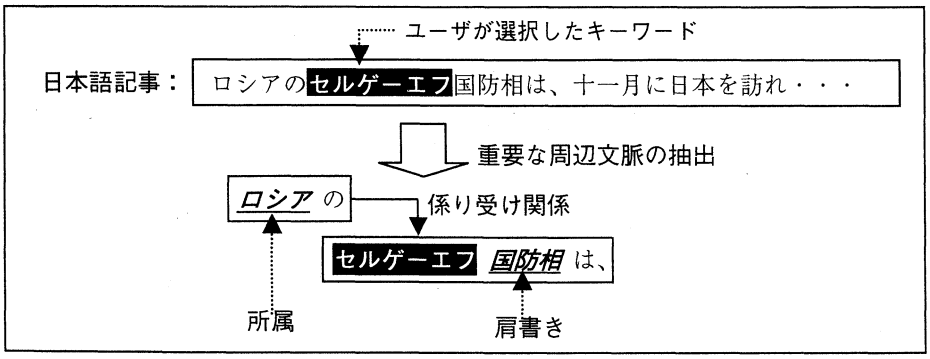

図2 典型的な人名の周辺文脈

図3 クロスリンガル検索の流れ

2.4 発音を元にしたカタカナと英語表現の比較

外来語は、元の言語の発音に近いように表記されているので、日本語のカタカナで表現された固有名詞と、抽出した英語の固有名詞とを、発音を評価の基準として比較を行う。

比較は、どちらの単語も発音記号へ変換し、それらの文字列の DP マッチングによって行う。ここでは、発音を示す記号として、ローマ字のアルファベット表記とする。

英単語の文字列ではなく、発音を基準とするのは、カタカナ文字から元の英単語を推測するよりも英単語から発音を推測するほうが、精度が高く実現できると考えたためである。

ローマ字からアルファベットへの変換は一意に決まるが、英単語からローマ字のアルファベットへの変換は、それほど単純ではない。

しかし、英単語をローマ字に変換する研究はすでに行われており[3]、簡易な処理で比較的精度の高い変換を行うことができる。そこで、基本的な変換処理は、この手法を元にして行う。ただし、変換結果は、アルファベットとし、また、変換にあいまい性のある部分については、全ての組み合わせとした。例を図4に示す。

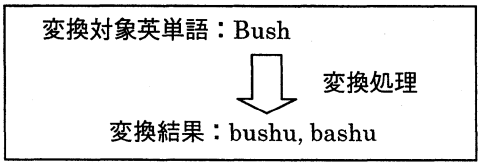

図4 あいまい性を残した英単語からローマ字のアルファベットへの変換処理例

英単語の発音を推定して得られたローマ字のアルファベットの表記と、日本語の固有名詞のカタカナをローマ字のアルファベットに変換したものとの類似度をDPマッチングによって計算する。類似度を計算する際には、英単語のローマ字のアルファベット表記では、複数の候補があるが、一番類似しているものを採用する。

このようにあいまい性を含んだ変換によって、英語以外の言語圏の英語表記など、同じ綴りでも読み方に差が出るような単語に対しても、比較の精度を高めることができる。

3 実験

NHK 国際放送局で利用している翻訳用例提示システムに対する検索クエリーの2000年10月のログに含まれる人名の中から、4つを選択して小規模な実験を行った。テストセットとしての日本語記事は、2000年8月～10月20日までの実際に英語記事に翻訳された実績のある記事の中から、新しいものから最大5つを選択した。また、記事中に複数回、同じ人名が含まれている場合は、記事中の最初に出現した人名部分を対象とした。

実験は、以下の仕様のシステムを作成して行った。

周辺文脈として、検索対象の人名を含む1文とした。

「所属」の推定は、人名より前に位置する文字列の中で、形態素解析結果で「組織」または「地名」となったもの全てとした。「肩書き」の推定は、人名の後に位置する文字列の中で形態素解析結果が「名詞類」と「接尾語」となった一番近い単語の集まりとした。ただし、一部の接尾語は対象外とした。

キーワードを日本語から英語へ変換する辞書は、NHKの日英対訳記事データベースよりパターンを抽出した独自のものを使用した。

検索結果のWebページから、ページの内容をHTMLのタグ情報や文を認定することにより分割し、それらの

検索対象人名	社会的位置づけ	テスト記事数	検索成功記事数	検索失敗記事数	検索成功の際に利用した周辺文脈の単語	検索結果の英単語
ブルグ	イスラエルの国会の議長	1	1	0	イスラエル、議長	Brug
シャロン	イスラエルの右派政党リクードの党首	4	4	0	リクード、党首	Sharon
クレバノフ	ロシアの副首相	3	2	1	ロシア、副首相	Klebanov
セルゲーエフ	ロシアの国防相	5	3	2	ロシア、国防相	Sergeyev

表1 実験結果

中に「所属」と「肩書き」の2つのキーワードが含まれているものを抽出し、固有名詞の推定は、それらの中で、大文字で始まる全ての単語とした。

Webページの検索エンジンには、「AltaVista[1]」を利用し、「所属」と「肩書き」が近い位置に存在するページを優先的に検索する代わりに、それぞれのキーワードを"NEAR"演算子を用いてクエリーを作成した。"NEAR"演算子は、演算子の前後のキーワードが10単語以内にあるものを検索する。

実験は、2001年1月に行い、周辺文脈の各キーワードの組に対してAltaVistaの検索結果を最大100件取得し、それらのWebページの内容を取得して、その中から正しい英訳語を自動で抽出することができれば「検索成功」、できなければ「検索失敗」とした。

検索対象の人名と社会的位置付け、検索結果を表1に示す。

実験結果では、周辺文脈の適切な単語をキーワードとして選択できた場合には検索に成功している。

よって、周辺文脈を利用する検索手法が有効であることが分かる。

テスト記事中の3記事については検索に失敗しているが、その原因は、検索対象の人名を含む文中に「所属」を示す「ロシア」が含まれていなかったためである。

検索に失敗した記事は、いずれも検索対象の人名が2文目以降に出現しており、記事の冒頭で「ロシア」に関する記事であることが明記されていて、繰り返しになるため人名を含む文に「所属」の説明がなかったのである。

記事の最初の文に検索対象の人名が含まれている場合には、その文の中に「所属」が含まれているが、記事の2文目以降の場合には、含まれていないことがある。そこで、検索対象の人名を含む1文だけでなく、記事の最初からの情報も文脈として必要であることが分かる。

[1] http://www.altavista.com/

4 おわりに

本稿では、周辺文脈を利用した固有名詞のクロスリンガル情報検索手法について論じた。

この手法により、次々と新しい固有名詞が出現するニュース翻訳の現場において、クロスリンガル検索により、インターネットから英訳語の情報を検索することができると考えられる。

また、この手法のうち、比較的簡易に実現できる機能のみを用いて実験システムを構築し、インターネットの情報を対象に、実際に翻訳に用いられた人名と記事を用いて情報検索の小規模な実験を行い、基本的動作の有効性を示した。

今後は、より高度な周辺文脈の解析手法の検討と、検索対象の固有名詞を含むWebページの特徴を分析することにより、さらに精度の高い検索が可能になるように研究を進める予定である。

参考文献

[1] 熊野正, 田中英輝, 浦谷則好, 江原暉将. 日英放送原稿翻訳支援のための類似用例提示システム. 言語処理学会第3回年次大会, pp.529-532, 1997.

[2] 熊野正, 後藤功雄, 江原暉将. Translators' Workbench: 対訳エディタを中心とした統合翻訳支援環境. 言語処理学会第6回年次大会, pp.143-146, 2000.

[3] 住吉英樹, 相沢輝昭. 英語固有名詞の片カナ変換. 情報処理学会論文誌, Vol.35, No.1, pp.35-45, 1994.