日英放送原稿翻訳支援のための類似用例提示システム

熊野 正　　田中 英輝　　浦谷 則好　　江原 暉将

NHK 放送技術研究所

1. はじめに

我々は現在，日英ニュース翻訳現場での利用を目指して，類似用例提示型翻訳支援システムの開発を進めている [2–4]．これは，日英の放送原稿から構築した日英対訳記事データベースを用い，ユーザの入力に類似した表現を含む記事対を検索・提示することでユーザの翻訳作業を支援するものである．

ユーザの入力に対して類似した表現を検索し，その対訳をユーザに示すことによって翻訳作業の支援を行う仕組みはすでに提案されており（[7,8] 他），実際に IBM Translation Manager/2 [5] や TRADOS Translator's Workbench [10] などの商用システムが稼働している．これに対して我々のシステムは，以下のような特徴を持っている．

1. 該当表現を含む,「記事」というある大きさの文章を提示することによって，ユーザが前後の文脈とその表現の関わりを理解し，用例が適切であるかどうかの判断材料とすることができる．
2. 日英記事対を提示する際には，日英表現の対応情報を提示する．提示に必要な対応情報をあらかじめデータベースの記事対に対して付与しておく．

本稿では，今回作成したプロトタイプシステムについて，その設計方針について述べ，構成と必要な要素技術について概説する．

2. システムの設計方針

本システムの開発に先立ち，我々はユーザである翻訳現場の翻訳者に対して，どのようなサービスを提供することが現場の支援につながるか，について意見を求めた．

翻訳者のスキルの程度は個人差が大きく，支援システムに求める機能も人によってかなり異なる．熟練した翻訳者は，訳すべき日本語原稿を前にして何らかの支援を要することは少ないが，

- 過去に固有名詞がどのように訳されたかなどを効率的に参照できる

ことを望んでいる．これに対してあまり熟練していない翻訳者にとっては，過去に訳された記事を参照して学習できることは大切であり，

- これから訳そうとしている記事と同じような記事が過去にどのように訳されているか，日英での記事構成はどのように変化しているかを，日英記事対を並べて比較できる
- ある表現がある文脈の中でどのように訳され，別の文脈の中ではどのように訳されるかを，多くの翻訳例を参照して把握できる

といった機能があることが望ましい．

我々はシステム構築にあたって，このようなさまざまなレベルの情報を柔軟に参照できるシステムを目指している．そして，ユーザのどのレベルの要求に対しても，常に記事対全体を提示することで応えるシステムが望ましいと考えている．理由は以下の通りである．

- ある表現の訳され方は前後の文脈に左右されるので，表現を文脈の中で提示することは重要である．
  記事は平均 5 文程度，たかだか 10 文程度なので，提示される情報量としてそれほど大量ではない．また，検索結果の表現とそれに対応する相手言語側の表現（表現対）を記事対全体の中で強調して提示することで，ユーザはまず強調された表現対を見て，必要に応じてその前後の文脈を参照することができる．
- 記事単位の対応は自明であるのに対して，文より細かい表現単位での対応づけの判断基準は明確でなく，対応づけは人手でもなかなか難しい．そのため，計算機で推定した表現対を切り出して提示してしまうと，対応づけが誤っていた場合にユーザは情報を全く得られない．
  記事対全体を提示する場合，表現対応情報は補

"Translation Example Browser for News Articles"
KUMANO Tadashi (kumano@strl.nhk.or.jp),
TANAKA Hideki, URATANI Noriyoshi,
EHARA Terumasa
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, JAPAN 157

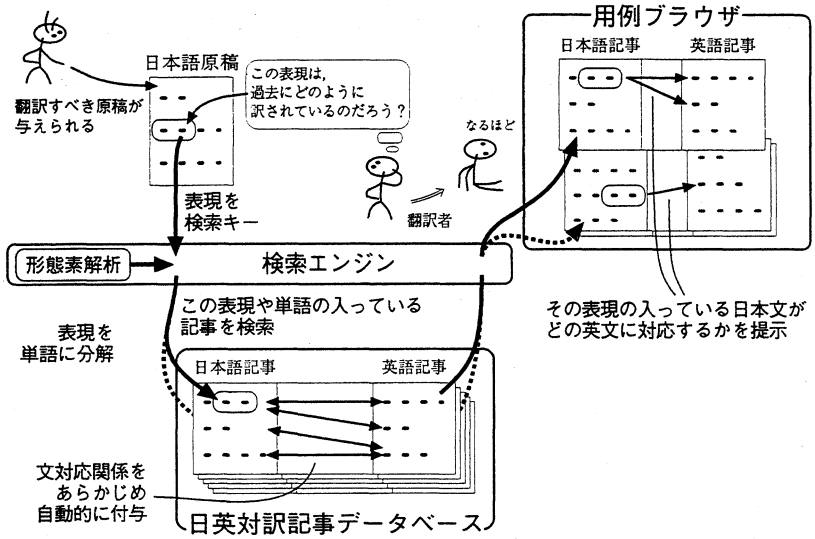

図 1: システムの概要・処理の流れ

助的な情報であり，対応づけを参考に表現の対応を読む，という使い方ができる程度に対応づけの精度が達していれば有用である．

このようなシステムは，ユーザの情報の取捨選択に多くを委ねることになる．従って，システムの使い勝手は，ユーザがいかに高速に多くの用例を「斜め読み」できるかにかかっている．システムの設計にあたってはこの点に留意し，

- 単語から記事全体まで入力できる種々の検索手段
- 多くの用例を高速に閲覧できる高速な検索と軽快な用例ブラウザ

を提供すること目標にする．

3. システムの構成

本システムは，大きく分けて以下の 3 つの構成要素からなる．

1. 日英対訳記事データベース
2. 検索エンジン
3. 用例ブラウザ

本システムの概要と処理の流れを 図 1 に示す．

3.1. 日英対訳記事データベース

日英対訳記事データベースは，NHK が日々作成している日英それぞれのニュース原稿からなっている．英語原稿は基本的には日本語原稿を人手で翻訳して作られているが，翻訳結果は自然な英語のニュース記事になるように表現や内容が変更されており，元の日本語記事の構造が保たれているとは限らない [2]．

日英の原稿は別個に作成されているが，日英の原稿には記事を一意に特定する記事番号がつけられており，これを基に日英の記事対を作成することができる．現在は，1995 年 3 月～1996 年 6 月の日英記事対 8,491 対からデータベースを構築している．

記事対中の日英表現間の対応情報として，現在は日英文間の対応情報を推定・付与している．文対応情報は，対訳データベース全体から抽出した単語共起情報に基づいて推定する．以下ではその手法を概説する（詳細は [4] を参照）．

3.1.1. 日英単語の共起度 (t-score) の計算

データベース中に出現する日英の内容語（形態素解析で付与した品詞で判断）の共起度を，t-score [1] を尺度として計算した．

単語の共起の計算においては，あらかじめ分かっている対応単位である「記事」を共起単位とした．従来共起の計算には「文」が共起単位として用いられることが多かったが，ニュース記事では十分な精度で文対応をとっておくことが難しいため，「記事」を共起単位とした．記事は文より大きな単位であるが，平均 5 文程度であり，文に近い扱いをすることができると考える．ただ，

- 記事は文より単語の出現数が大きく，同じ記事の中に同じ単語が繰り返し出現することを無視することは危険である

- 1文より1記事の方が大きさの変動が大きいため，小さな記事間の単語共起と大きな記事間の単語共起を同列に扱うことは問題である

という点を考慮して，日英単語の共起回数を計算する際に，記事中での単語の出現回数と記事長（記事の単語数）とで正規化を行っている．

3.1.2. 日英文間の関連度の計算

あらかじめデータベース全体から求められた日英単語共起度を基に，日英記事対中の日英文間の結び付きの強さ（関連度）を求め，対応関係の推定を行う．ある日英文間の関連度の計算は，以下の手順で行う．

1. 日本語文側の各単語に対して，英語文側の単語の中で最も共起度の大きな単語を選び，相互情報量[1]を計算する．
2. 1. で求めた相互情報量を全ての日本語文側の単語について積算し，日本語文の単語数によって正規化する．

求められた関連度を用いて対応関係を検定する手法については，現在いくつかの方法を検討中である．

3.2. 検索エンジン

現在，システムは日本語からの検索のみを提供する．現在実装している日本語の検索機能は下記の2つがあり，ユーザは必要に応じて各機能を使い分けることができる．

- 字面検索部
- 類似検索部

いずれも全文検索エンジンを使った文字列照合を使って実現している．全文検索には長尾・森の手法 [6] を採用し，各文字に行番号情報を付与するという拡張を行って使用している．

字面検索部は，任意の入力文字列をそのまま検索キーとして全文検索を行い，これを含む文（とその記事）を検索する．

類似検索部の動作は以下の通りである．

1. キーワードの抽出
   入力表現を形態素解析して，名詞と動詞相当の自立語を抽出する．
2. 活用語キーワードの展開
   活用するキーワードはキーワードの語幹に可能な活用語尾や助数詞類を付加して展開する．
3. 検索
   以上のキーワード集合の各キーワードが出現する文を最初に検索する．この結果を利用して，キーワード全てが（活用語を展開したキーワードは展開形のどれか1つ）出現する文（とその記事）を出力する．また，後に述べるユーザの要求によって，キーワードの数を1つずつ減らしながらその数のキーワードを含む文（とその記事）を出力する．これらの検索はキーワードの順番を考慮している．

3.3. 用例ブラウザ

用例ブラウザは以下の2種類のウィンドウからなるインタフェースを持つ（**図 2**）．

1. 検索表現入力ウィンドウ
   字面検索のキーを入力する部分と，類似検索のための表現を入力する部分を持つ．類似検索のための表現を入力後，ボタンを押すことによって，形態素解析を行った結果を字面検索のキーとすることができる．
   検索開始ボタンを押すと記事対提示ウィンドウが開かれ，検索結果を参照できる．
2. 記事対提示ウィンドウ
   入力表現を含んだ記事対のリストを表示し，リストをクリックすることで記事対全体を表示することができる．記事対の表示の際には，
   - 入力表現に該当する部分を強調表示
   - 日英の文にマウスカーソルを重ねると，相手言語側の対応する文を反転表示

   することができる．また，3.2 節類似検索部の説明で述べた，出現するキーワードの数を，ユーザの指示で増減することができる．

GUI は Tcl/Tk[2] を利用して作成した．

4. 今後の方針

今後は，本システムの翻訳現場での実用を目指して，以下の要素技術の改良を行い，データベースの構築を継続するとともに，現場の翻訳者との議論を重ねて，より使いやすいシステムを構築していく予定である．

- 日英文間の対応推定の精度向上
  対応づけの結果に対して客観的な評価を行う方法について検討し，文間の関連度の計算方法や対応判定手法，さらに現在の手法とは基本的に異なる手法まで含めてより高精度な手法の開発を行いたい．

[1] 正確には相互情報量の対数をとる前の値

[2] Tcl/Tk のバージョンはそれぞれ 7.5jp/4.1jp を使用．日本語化は（株）SRA の西中芳幸，酒匂寛，石曽根信の各氏の手による．

図 2: 用例ブラウザ

- 節や句といったより細かな表現間の対応関係の推定
  現在文単位で行っている対応推定を，より小さな単位に対して適用したい．
- 類似記事の検索の実用化
  現在，1 記事全体を入力として内容が類似した記事を検索する機能を実装していない．どのような手法によってこの機能を実現するかは未定だが，そのための研究の一環として，記事のクラスタリングの研究を行っている [9].

参考文献

[1] K.W. Church and R.L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19–1, pp. 1–24, 1993.

[2] 熊野正, 田中英輝, 金淵培, 浦谷則好. 日英ニュース原稿の対訳コーパス化に関する基礎調査. 第 2 回言語処理学会年次大会, pp. 41–44, 1996.

[3] 熊野正, 田中英輝, 江原暉将. 統計的手法を用いた日英放送原稿の単語対応づけ. 第 52 回情報処理学会年次大会, pp. 2:53–54, 1996.

[4] 熊野正, 田中英輝, 浦谷則好, 江原暉将. 日英放送原稿の文間の対応関係の推定. 自然言語処理シンポジウム「大規模資源と自然言語処理」, 1996. `http://www.etl.go.jp/etl/nl/nlsympo/96/kumano.ps.gz`.

[5] J.-M. Langé. MT at IBM France: Research and Products. In *Premières journées franco-japonaises sur la traduction assistée par ordinateur*, pp. 205–207, 1993.

[6] M. Nagao and S. Mori. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In *14th COLING*, pp. 611–615, 1992.

[7] 中村直人. 用例検索翻訳支援システム. 第 38 回情報処理学会年次大会, pp. 357–358, 1989.

[8] 武田明子, 古郡廷治. 例文をもとにした英文書作成支援システム. 情報処理学会論文誌, 35–1, pp. 53–61, 1994.

[9] 田中英輝. 大規模文書集合の高速クラスタリング. 第 3 回言語処理学会年次大会, A3–3, 1997.

[10] TRADOS GmbH. *TRADOS Translator's Workbench.* `http://www.trados.com/`.