日英放送原稿翻訳者のための類似用例提示型翻訳支援システム

熊野 正[†] 田中 英輝[†] 松田 伸洋[‡] 浦谷 則好[†] 江原 暉将[†]

デモ１６ [†]NHK 放送技術研究所 [‡](株) 漢字情報サービス

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を開発している [4]．これは，日英のニュース原稿から構築した日英対訳記事データベース [1]（1995 年 3 月～1997 年 2 月分，14,701 対）を用い，ユーザの入力に類似した表現を含む記事対，あるいはユーザが入力した記事と類似した内容の記事対を検索・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．

本システムは，以下のような特徴を持っている．

- ユーザが表現を入力したときに，入力に類似した表現を含む日英の「記事」全体を提示する．これによって，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を判断することができる．
- ユーザが記事全体を入力したときに，入力と単語の出現傾向の類似した記事を提示する．あらかじめ単語の出現頻度を用いてデータベースの記事集合をクラスタリングしておき，高速な検索を実現する．また，クラスタリング結果の階層木をグラフィカルに表示し，記事ブラウザとして利用することができる [5, 6]．
- 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 [2, 3]．

本システムの設計方針や構成については既に文献 [4] で紹介した．本稿では，本システムがユーザに提供する機能について概説する．

2. システムの提供する機能

2.1. 表現検索

単純な完全一致検索と類似表現検索を提供する．現在，システムは日本語からの検索のみを提供する．ユーザは必要に応じて明示的に検索手法を使い分けることができる．

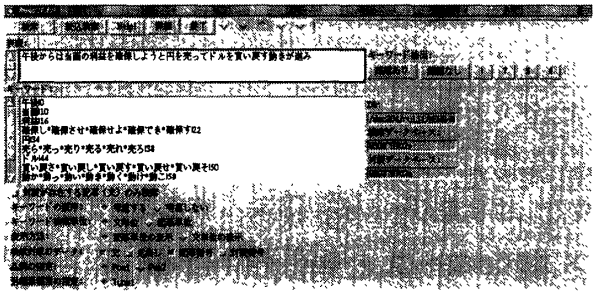

図 1: 表現検索入力ウィンドウ

類似表現検索では，以下の手順で検索を行う [7]（図 1）．

1. 入力表現を形態素解析して内容語をキーワード集合として抽出
2. 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などを付加して展開
3. キーワードの出現順序と出現間の距離を考慮して完全一致検索を行い，1 文中に全てのキーワードが出現するような記事リストを出力

また，ユーザの要求に応じて，キーワードの数を減らして検索を行うことができる．これによって，より広範に類似表現を検索することができる．

記事リストの各記事は，2.3 節で述べる対訳記事表示ウィンドウで見ることができる．

2.2. 類似記事検索

日本語記事を入力として与えて，単語の出現傾向が入力に類似した記事リストを検索する機能を提供する．

この機能の実現のために，あらかじめデータベース中の全日本語記事を内容語の出現頻度によってクラスタリングを行って階層木を作成しておき [5]，検索時には入力から取り出した内容語出現頻度リストを用いて階層木をたどり，リーフに割り当てられている記事リストを出力する [6]．

クラスタリング結果の階層木や，検索の結果リーフにたどり着くまでの階層木上での経路は，グラフィカルに表示することができる（図 2）．階層木の任意のノードやリーフをマウスクリックすることで，その下の折り畳まれた階層木を展開して表示し，またその下に含まれる記事リストを得ることができる．記事リストの各記事は，2.3 節で述べる対訳記事表

"Translation Example Browsing Environment for News Articles"

KUMANO Tadashi(kumano@strl.nhk.or.jp),
TANAKA Hideki, MATSUDA Shin'yo,
URATANI Noriyoshi, EHARA Terumasa
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, JAPAN 157

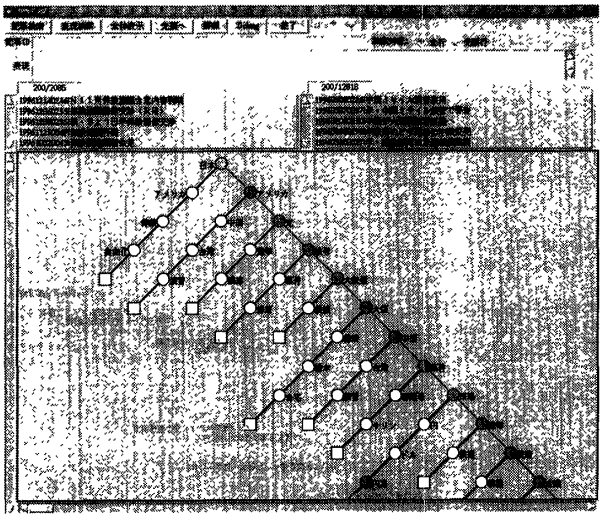

図 2:　類似記事検索ウィンドウ

示ウィンドウで見ることができる．

2.3. 対訳記事表示

2.1, 2.2 節で述べた検索の結果得られた日英記事対を，左右に対比させて表示する（図 3）．表現検索の結果の表示の場合には，日本語側の相当する表現の部分を色を変えて表示する．

日英の表現間の対応を理解する手掛かりとして，日英文間の対応を推定し表示することができる．日英の文の上にマウスカーソルを置くと，相手側の対応する文の色が変わる．

この機能の実現のために，あらかじめ以下の処理を行ってデータを作成しておく．

1. 対訳記事データベース全体から日英単語共起情報（共起度）を抽出 [2]
   - 記事を共起回数計算の単位とし，1 記事中に同じ単語の出現する回数と記事長で共起回数を正規化
   - 共起度の尺度として t-score を使用
2. 1. の共起度を尺度にして，日英文間の関連度（日英単語間の相互情報量の積）を計算 [3]

文の上にマウスカーソルを置いたときには，相手側の各文との関連度を調べて上位の文の色を変えて表示する．

3. 今後の方針

今後は，本システムの翻訳現場での実用化に向けて，以下の要素技術の改良を行う．

- 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などの改良を測る．

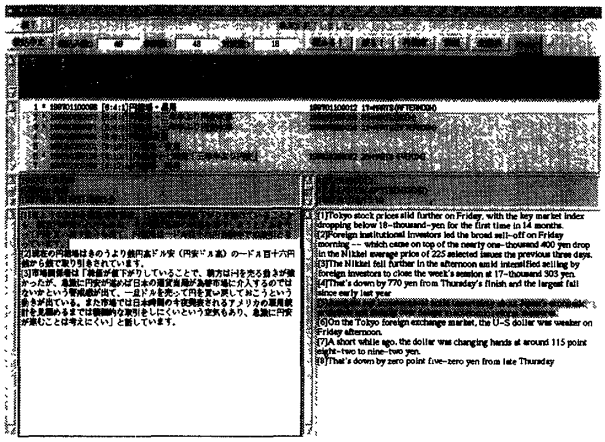

図 3:　対訳記事表示ウィンドウ

- 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する手法を検討する．

また，現在入力言語には日本語しか使用できないが，早期に英語側からの検索を行えるよう拡張を行う予定である．

さらに，Web ブラウザをユーザインタフェースとするサーバ・クライアント型の実用システムを試作し，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．

参考文献

[1] 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.

[2] 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 回情報処理学会全国大会, 第 2 巻, pp. 2:53–54, 1996.

[3] 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.

[4] 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. 第 3 回言語処理学会年次大会, pp. 529–532, 1997.

[5] 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, pp. 249–252, 1997.

[6] 田中英輝, 熊野正, 松田伸洋. 階層分割型クラスタリングを使った文書ブラウザ. 第 55 回情報処理学会回全国大会, 6N–7, 1997.

[7] 田中英輝. 長い日本語表現の高速類似検索手法. 第 121 回自然言語処理研究会, 1997.