4B-6 構文の照合による柔軟なテキスト検索機能を備えた翻訳支援システム

隅田 英一郎, 堤 豊
(日本アイ・ビー・エム株式会社, 東京基礎研究所)

1. まえがき

計算機を使った翻訳システムには, 機械翻訳システムと翻訳支援システムの2種類があるといわれている[melby]. 従来, 翻訳支援システムが翻訳に及ぼす効果の大きさは認められながらも, どのような機能が必要かという点の議論にあまり進展がみられなかった.

一方, 最近の計算言語学では, 電子化辞書の高度な検索の研究が盛んである[wachowicz]. これらの従来の電子化辞書の検索は, キーと見出しの単語の照合によって行われ, 紙に印刷された辞書と同様に制限されている.

本稿では翻訳支援システムにおいて, 辞書や蓄積された翻訳結果のデータベースなどからの《テキスト》の《柔軟な検索》が有効であることを論じ, その実現のために《規則に基づいた構文の照合》という新しい枠組みを提案する.

2. 翻訳支援システムとテキストの柔軟な検索

翻訳支援システムで適当な目的言語の表現の選択に迷ったとき, 原言語のテキスト(熟語, 慣用句, 特殊な言い回し, 文など複数の単語からなるもの)に類似したテキストとその対訳が検索できれば, ユーザはそれらを比較して, 最適なテキストを選択し, その対訳を修正することによって容易に翻訳できるようになる(図1).

従来の電子化辞書は単語単位で検索するので、上に示したような複数の単語からなるテキストの検索は困難であり、類似したテキストの検索は不可能であった.

本論文で提案する《規則に基づいた構文の照合》による検索の概要は以下の通り.

1) 単語だけではなく, 句, 文等任意の《テキスト》を検索のキーとして入力し,
2) 検索のキーと辞書中の《全見出し》と照合し,
3) A) 照合が成功の場合には終了する.
   B) 照合が不成功の場合には《規則》に従って, キーを一般化して照合を繰り返す.

このようにしてキーに《構文的によく似た》見出しの辞書項目を選択する.

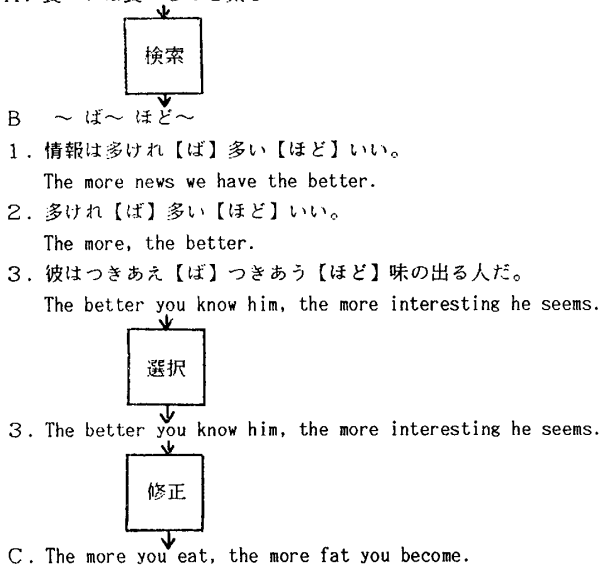

図1 翻訳支援システムと柔軟なテキスト検索

3. 実験システムETOC

著者らは, 上で述べたテキストの柔軟な検索の有用性を確かめるために日英の対訳を集めた辞書[keene他]をデータとして, 翻訳支援のための検索システムETOC (Easy TO Consult)を作成した. 図2に本システムの構成を示す. 本システムは次の各要素－3つのデータ1) 検索キー, 2) 辞書, 3) 一般化規則と3つのモジュール4) 解析部, 5) 検索部, 6) 一般化部からなる. 次に, 解析部, 一般化規則, 高速化の工夫について述べる.

解析部の目的は検索キーと検索される辞書の見出しを同じ水準で構造化することである. 一般化規則及び一般化部も同じ水準で作成される. 今回の実験システムでは, 形態素解析[maruyama他]を採用し, 十分に有用であることがわかった.

4. 一般化

検索キーと辞書の全ての見出しがマッチしないときに, 以下に述べる規則に従って, 検索キーの重要な情報とそうでない情報を区別し, 後者を削除したり, 緩和したりして, 辞書の見出しとマッチするように一般化される. 現在次の規則が実現されている.

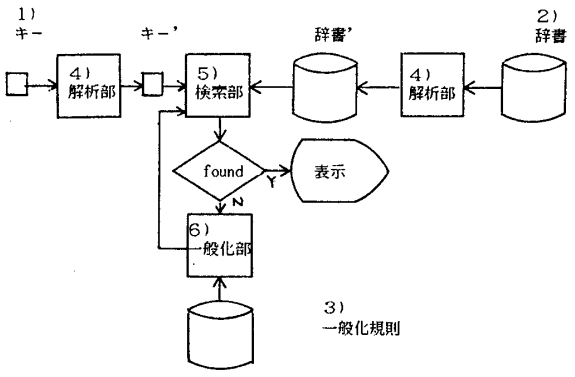

図2 ETOCのシステム構成

A translation aid system using flexible text retrieval based on syntax-matching (overview)
Eiichiro SUMITA and Yutaka TSUTSUMI
IBM Research, Tokyo Research Laboratory

1）語順の入れ換え
日本語では単文の範囲において，格要素（名詞＋格助詞）の順番は重要でないので，格要素の順番を入換えて検索する．
2）代名詞の置き換え
代名詞は代替性が大きいので任意の名詞に置き換えて検索する
3）修飾語の削除
修飾語の付いている句は中心語で置き換える
4）名詞の置き換え
5）動詞，形容詞の置き換え
6）格助詞の置き換え
意味の近い格助詞（例『に』≒『へ』）を置き換えて検索する
7）格要素の削除
自由格のみならず必須格も省略して検索する

上の規則の基本方針は『自立語によって表現される各テキストの個別性を順次無視し，最後に付属語で表現される文の骨格や法情報のパタンにまで一般化する』である．
この基本方針は，ユーザのテキスト検索時の要求に合致する．ユーザの要求は以下のように考えられる．複数の語からなるテキストを検索するときは，必ずしも全ての単語が平等に重要ではない．テキスト全体と一致するものが辞書に無い場合は，ユーザは普通，機能語より内容語を検索条件から外す．

5．高速化

辞書検索システムは，検索速度が十分速くないと，誰も使いたがらない．前節までで述べたように類似したテキストが見つかるまで辞書全体の検索を繰り返すので、単純に実現すると現実的な反応時間をえることは不可能である．
高速化のために本システムでは，早見表というデータ構造を採用した．
早見表を利用し，探索空間を十分制限し、詳細な照合を行うので，検索時間を小さくできる．
早見表には、全ての単語についてその出現した見出しの番号を登録する(図３)．
この早見表は，システムを作るときに，あらかじめ作成し，新たにレコードを追加する度に更新される．キーが複数の単語からできている場合は，それぞれの単語の集合の積集合の大きさに探索空間は制限される．例えば，『僕は学校に行く』を検索する．ここで早見表から『僕』『は』『学校』『に』『行く』が含まれていた見出しの集合をもとめ，それらの積集合を求める．この大幅に狭まった検索領域の中だけで，詳細な照合を行えば良いので，検索時間を小さくできる．
本実験システムでの数値を示す．Tを早見表，T(W)を単語Wを含むレコードの集合とする．辞書中の全レコード数は16870，#(T(が))は4153，#(T(を))は5984，#(T(が)&T(を))は909である．従って『～が～を～』というテキストを検索する場合には，探索空間は909/16870即ち５．４％に制限される．
日本語では，これらの格助詞に代表される機能語は表現の仕方を決定する重要な要素であり，我々の一般化規則でも出来るだけ多くの機能語を一般化せずに残そうとしている．従って，機能語による探索空間の制限が有効に働くことは重要である．
本実験では早見表をハッシュ化し，lispの主記憶上のワークスペースで実現した．この早見表は，Ｂ木などの従来のデータベース技術を利用できるので，システムの規模を問題なく拡張できる．

(a)

#	見出し	内容（対訳）
1	彼女は医者に行った．	She went to see the doctor.
2	彼は東京に行った．	He went to Tokyo.
3	彼は医者になった．	He became a doctor.

単語	#
医者	1，2
行く	1，2
彼女	1
彼	2，3
た	1，2，3
東京	2
なる	3
に	1，2，3
は	1，2，3

図３ 早見表の例

6．あとがき

本論文では，翻訳のための計算機を使った支援ツールとして辞書などからの《テキスト》の《柔軟な検索》が有効であることを論じ，そのための枠組みを提案した．キーと見出しのテキストを同じ方法で解析し，キーと全ての見出しを照合する．照合が成功するまで，構文要素の重要度を決める規則に従って，重要でない要素から削除してキーを一般化する．こうしてキーと《構文的に近い》見出しとその対訳が検索できる．ユーザはこれを使って，必要な訳文を容易に作成できる．
本方法には以下の長所がある．
(1)ユーザは形式的な言語を使わずに，単にテキストを入力すればよい．
(2)特定のドメインに典型的な原文とその対訳のペアを集めることによって専用のシステムが容易に作成できる．
全件検索を高速化するために早見表というデータ構造を採用し現実的な反応時間をえた．
この枠組みの有効性は日英翻訳のための実験システムＥＴＯＣによって確認された．

今後の課題としては以下のものが考えられる
・複数の言語，複数の分野のデータを収集し，正確で有用な一般化規則の集合を開発し評価する．
・構文解析など高度な処理を取り入れて，検索の精度をあげる．
・言語教育に於ける教材収集に応用する．
・ユーザとのセッションのログから，各ユーザの検索の傾向を学習し，検索を最適化するメカニズムを検討する．
・この枠組を拡張して、例文を使って自動的に翻訳する方法[nagao]を検討する．

参考文献

[melby] Melby A.: "On human-machine interaction in translation", Machine Translation, pp.145-154, 1987
[wachowicz] Wachowicz K.: "On intelligent dictionaries", CaT, vol.1, no.4, pp.225-233, 1986
[keene他] ドナルド・キーン,羽鳥博愛: "会話作文英語表現辞典",朝日出版社, 1982
[sumita他] Sumita E. and Tsutsumi Y.: "A translation aid system using flexible text retrieval based on syntax-matching", Proceedings of the second international conference on theoretical and methodlogical issues in machine translation of natural languages, CMU, Pittsburgh,1988, also available as TRL Research Report, TR87-1019
[maruyama他] Maruyama N., Morohashi M., Umeda S., and Sumita E.: "A Japanese sentence analyzer", IBM Journal of Research and Development, vol.32, no.2, pp.238-250, 1988
[nagao] Nagao M.: "A framework of a mechanical translation between Japanese and English by analogy principle", Artificial and Human Intelligence (A. Elithorn and R. Baneriji. Ed.), pp.173-180, 1984