1F−5

用例に基づいた翻訳

隅田英一郎、飯田仁、幸山秀雄

ATR自動翻訳電話研究所

1 はじめに

従来の規則に基づいた機械翻訳(MT)には、文法規則、訳語選択規則など複雑で大規模な規則のデータベースが必要である。これらの知識の構築にかかるコストが実際的なシステムを作成する際の大きな問題である。又、規則の追加の影響が単純に予測できず翻訳の質の向上も困難である。

これらの問題点を解決するために著者らは用例(原文と対訳)のデータベースを直接利用する翻訳方式を提案する。システムは入力に類似した用例をデータベースから検索し、その用例を用いて入力に対する訳を決定する。この枠組みをEBMT (Example-Based Machine Translation) と呼ぶ。

用例を利用する提案は長尾[1]が最初であり、佐藤と長尾の動詞の訳し分けの実験[2]がある。本稿では訳し分けの手法が確立していない「N_1のN_2」という形の日本語の名詞句を英語の名詞句に翻訳する問題を例にEBMTの実現性を示す。

2節でEBMTとMTの比較、3節でEBMTのメカニズムの説明、4節で「N_1のN_2」の実験の詳述、5節で種々の考察を行う。

2 EBMTとMT

2.1 計算時間 MTは、構文、意味、構造変換、訳語選択、生成などのための大規模な規則のデータベースを使うシステムであるので、MTは計算時間がかかる。一方EBMTは、長い規則の連鎖を通して推論するのではなく、検索された用例から訳を直接生成するので少ない計算時間で済む。

2.2 システムの改良 MTの場合、翻訳の質の向上は、相互に依存した規則の修正によって行われるので、規則の一貫性を保って質の向上をはかるのは難しい。EBMTでは、適切な用例をデータベースに追加することだけで済むので、質の向上が容易である。

2.3 データベース作成のコスト MTのための規則を作成することは困難な作業であり、言語学的な訓練を受けた専門家が必要である。対照的にEBMTに必要な用例は簡単に入手できる。さらに、電子化出版が盛んになるにつれて、機械可読な用例も増加している。

2.4 ロバストな翻訳 MTはexact match、EBMTはbest matchの推論を用いる。MTは入力に正確に照合する規則が存在しないと翻訳に失敗する。EBMTは類似した用例が存在すればよく、本質的にフェイルセイフに翻訳する。

2.5 信頼係数 EBMTでは、入力と検索された類似用例との距離に基づいて、翻訳結果に信頼係数を付与できる。MTにはそのための機構がない。

2.6 知識の独立性 EBMTの知識はMTの場合のように特定のシステムのための規則ではなく、言語事実である。従って、他のシステムで使ったり、種々の言語分析の材料とすることが可能である。

3 EBMTのメカニズム

3.1 構成 EBMTは用例とシソーラスの二つのデータベースと解析、用例検索、用例適用の三つ翻訳モジュールからなる。例えば入力「京都での会議」に対して表1のような用例が検索される。これに基づいて訳「the conference in Kyoto」が出力される。以下では用例検索について述べる。

d	日本語	英語
0.4	東京での滞在	the stay in Tokyo
0.4	香港での滞在	the stay in Tokyo
1.0	大阪の会議	the conference in Osaka

表1 検索例

3.2 用例の検索 入力に類似した用例をデータベースから検索するために、入力と用例の距離を定義する。入力と用例は同じデータ構造すなわち属性値のリストとして表現されているとする。リストをI, E、i番目の属性値をI_i,E_iと書く。

3.2.1 用例の距離 全体の距離d(I,E)は各属性値の距離d(I_i,E_i)と属性値の重みw_iを用いて次の式で計算する。

(1) $d(I,E) = \Sigma d(I_i,E_i) \times w_i$

3.2.2 属性値の距離 意味属性以外は値が一致するか否かに従って0か1とする。意味属性では部分一致を認め、0から1の実数を割当てる。シソーラスの上での最小の共通の上位概念(MSCA)[3]の位置に比例した値を割当てている。意味の距離の計算に共起情報の利用がよく行われるが[2]、必要な用例の数が膨大になるので採用していない。

3.2.3 属性値の重み 属性値の重みは、その値が訳の選択に与える影響の大きさをあらわす。ここではMemory-Based Reasoningで用いられている式[4]を採用した。

(2) $w_i = \sqrt{\Sigma(E_i = I_i$である翻訳パターンkの頻度$)^2}$

この計算は重いが、データベースでの静的な頻度にのみ依存しているので、システム作成時に前もって計算でき、実行時にコストはかからない。

4 「N_1のN_2」の実験

Translating with Examples

Eiichiro SUMITA, Hitoshi IIDA, and Hideo KOHYAMA
ATR Interpreting Telephony Research Laboratories

4.1 「N_1のN_2」の翻訳の重要性 「N_1のN_2」の形の名詞句は頻度の高い表現である。その英語への翻訳は多様でその選択は難しい。実際デフォルトと考えられている「N_2 of N_1」の頻度は20-40%に過ぎず、他の前置詞が使われたり、前置詞なしで翻訳される。EBMTでは「N_1のN_2」に関する先行研究と異なり深い理解の機構を用いない[5]。

4.2 「N_1のN_2」のシステム 用例は「国際会議に関する対話」のデータベース[6]から抽出し、シソーラスは大野、浜西の体系[7]に準拠している。今回の実験で使った属性は、名詞に対しては、品詞の下位分類(サ変、普通…)、接頭語·接尾語の存在、シソーラスの意味コード、連体の格助詞に対しては、その種類(の、での、からの…)である。

ここで、重みの計算を例を用いて説明する。表2は特定の属性値を持つ用例の翻訳パターンのデータベース中での分布の一例である。「N_1」の意味コードが「地名」の場合は「in」、「from」など多様な前置詞が使われたり、「N_1」の形容詞化が起こったり、訳の選択との相関は弱い。格助詞が「での」の場合は全て前置詞「in」で表現されていて、強い相関がある。このような相関が、3.2.3節の式(2)で数値化される。

k	頻度
B in A	12/27
†^A B	4/27
B from A	2/27
…	…
B to A	1/27

E_1 = 地名

k	頻度
B in A	3/3

E_2 = での

それぞれの重みは
式(2)により
$W_1 = 0.49$
$W_2 = 1$

表2 属性値の重み(†^AはAの形容詞)

4.3 実験結果 現在の用例数は約700であり、約5,000まで増やす予定である。700の用例を二つのグループに分けた。(1)100の用例の日本語部分を実験の入力とし、(2)残りの600を用例データベースに登録した。翻訳の失敗率は42%とかなり高かったが、この失敗のうち約90%(42件のうち38)は類似した用例がないことが原因であり、適当な用例を追加することで容易に改善できる。

EBMTは用例データベースの特徴を素直に反映する。例えば対話に頻出する丁寧な表現である「Nの方」の英語への翻訳では常に「方」が略され、「N」だけになる。このような翻訳が規則を用いずに、単に用例データベースを収集すれば可能になる点がEBMTの長所の一つである。

5 考察

5.1 「N_1のN_2」以外の現象 EBMTに適した日英翻訳に現れる言語現象を以下に例示する。(1)「だ」文。「N_1はN_2だ。」という形で、単純な場合は「N_1 be N_2.」と訳されるが、これも多様な翻訳がありえて、「N_2」が動詞化されたり、省略されている動詞を補う必要があったり、翻訳が困難なことが知られている。しかしながら、その構造は「N_1のN_2」と同様で、二つの名詞と機能語からなり、EBMTで処理できる。(2)日本語のアスペクト表現は「動詞+アスペクトマーカ」という形をしており、これも動詞のシソーラスを用いてEBMTで処理できる。(3)慣用表現はその要素の翻訳からは合成できない(例「孫の手→a back scratcher」)。これは規則に基づいたパラダイムは適さず、EBMTに適していることを意味する。(4)佐藤と長尾[3]は一つの動詞とその要素としての幾つかの名詞からなる単文が同様のメカニズムで翻訳出来ることを示している。

5.2 規則に基づいたパラダイムとの統合 全てのプロセスがEBMTで扱えるか否かはまだ明らかではない。著者らはEBMTが翻訳のコントロールを握り、類似した用例が検索できなかった場合に規則に基づいた一般的なシステムを呼び出す方法が全体のスループットを上げるという点で優れていると考える。

5.3 高速化 EBMTは、best matchな用例を得るために、用例データベース全体を調べることになる。単純な実現では、遅くなることが避けられない。索引と並列計算が解決策として考えられる。索引は構文的な類似検索のシステムにおいて実績が有り[8]、また検索は本質的に他の検索と独立なので並列計算も有望である。またexact matchを最初に探し、見つかった場合に他のプロセスを抑制するバイパスも自然に導入できる。

6 おわりに

従来の機械翻訳の問題点を解決するために、著者らは用例(原文とその対訳)のデータベースを直接利用する翻訳方式EBMTを提案し、「N_1のN_2」という形の日本語の名詞句を英語の名詞句に翻訳するシステムを実現した。

謝辞 本研究を行う機会を与えて頂いたATR自動翻訳電話研究所 樽松明社長に感謝致します。

参考文献

[1]M. Nagao: "A framework of a mechanical translation between Japanese and English by analogy principle", in Artificial and Human Intelligence, ed. A. Elithorn and R. Banerji, pp. 173-180, North-Holland, 1984.
[2]佐藤理史, 長尾真: "実例に基づいた翻訳 ", 自然言語処理研究会NL70-9, 情報処理学会, 1989.
[3]J. Kolodner and C. Riesbeck: "Case-Based Reasoning", tutorial textbook of 11th IJCAI, 1989.
[4]C. Stanfill and D. Waltz: "Toward Memory-Based Reasoning", CACM, 29-12, pp. 1213-1228, 1986.
[5]A. Shimazu, S. Naito, and H. Nomura: "Semantic structure analysis of Japanese noun phrases with adnominal particles", pp.123-130, Proc. of 25th ACL, 1987.
[6]橋本一男, 小倉健太郎, 江原暉将, 森元逞: "対話データベースを用いた各種言語現象の検索", 情報処理学会大40回全国大会予稿集, 1990.
[7]大野晋, 浜西正人: "類語新辞典", p.932, 角川, 1984.
[8]E. Sumita and Y. Tsutsumi: "A translation aid system using flexible text retrieval based on syntax-matching", Proc. of the second international conference on theoretical and methodological issues in machine translation of natural languages, CMU, Pittsburgh, 1988.