7G-5 対訳辞書からの中間概念の抽出

徳永健伸，HARTONO，田中穂積
東京工業大学 工学部

1 はじめに

多言語間機械翻訳の方式として中間言語方式が有望視されているが，残念ながら，これまで中間言語に関する研究が十分なされてきたとはいえない．中間言語を考える時には，(1) 中間言語の語彙(言語に依存しない中立的な概念項目の集合)の設定，(2) それらの語彙項目間の関係の設定，という2つの問題がある．

本稿では，この第1の問題に対する解決の糸口を探ろうとするものである．同様な試みは，電子化辞書研究所 (EDR)でもおこなわれているが，EDRではこの語彙項目をすべて人手で設定することを考えている[2,4]．一方，我々は，中間言語の語彙項目を既存の対訳辞書から機械的に抽出することを目指している．本稿では，このための基礎的考察をおこなう．

2 対訳辞書の構造

2つの言語 L^a と L^bについて，L^aから L^bへの対訳辞書と，L^bから L^aへの対訳辞書を考える．今，L^aの見出し語 a_iが m 個の語義 a_i-1, ... a_i-m を持ち，L^bの見出し語 b_jが n 個の語義 b_j-1, ... b_j-n を持つとする．そして，a_iの語義 a_i-2 の訳語が b_jであったとしよう(図 1)．語義 a_i-2 の訳語が複数個ある場合もあるので，語義 a_i-2 からは，言語 L^bの複数個の見出し語に向けた有向辺が張られることもある．

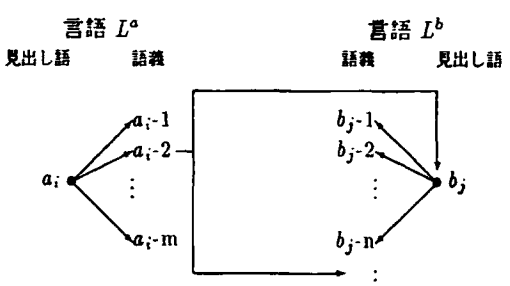

図 1 対訳辞書の基本構造

我々は，対訳辞書を1つの有向グラフと見なし(図 1)，このグラフを翻訳グラフ(Translation Graph : TG)と呼ぶ．言語 L^a から言語 L^bへの TG を TG_{ab}と書く．添字 ab には方向性があることに注意．この時，明らかに，TG_{ab}の見出し語は言語 L^a に属し，訳語は言語 L^bに属する．TG_{ab}は$< A, AS, B >$から構成される．ここに A は TG_{ab}の見出し語の集合，ASは A の持つ語義の集合，Bは L^bにおける訳語の集合である．ここで，ASの訳語中には，TG_{ba}の見出し語に含まれないものが存在する．1つの理由として，訳語が辞書の見出し語として現れない句や文で表現される場合があるということが挙げられる．しかし，訳語が L^bの語で表現されていても，それが TG_{ba}の見出し語として含まれな場合もある[5]．

TG_{ab}を用いた翻訳では，A の要素 a_iに対して1つの語義 a_i-k を選択し，その訳語 b_jを選択する．この翻訳プロセスには，$a_i \rightarrow a_i$-k $\rightarrow b_j$という経路が存在するが，これを翻訳経路 (Translation Path)と呼び，以下では$[a_i, k, b_j]$と記す．

ここで，2つの翻訳グラフ TG_{ab}と TG_{ba}の和 TG_{ab+ba} $(= TG_{ab} \cup TG_{ba})$を考え，これを双方向翻訳グラフと呼ぶ．TG_{ab+ba}において，A の要素が始点と終点以外では2度と同じ頂点を通らず，L^aと L^bの見出し語をそれぞれ n 個ずつ含むものを双方向翻訳 n 次回路(n 次回路と略す)と呼ぶ．

$n = 1$ の場合には見出し語の対が1つに固定されるが，$n \geq 2$ の場合には，複数個の見出し語の対を考えなければならない．ここで，見出し語 $a_i \in A$ と $b_j \in B$を含む回路が存在する時，見出し語の対 a_iと b_jの間に回路が存在すると言う．1次回路の例を，図 2 に示す．図 2 の1次回路を，$< a_i$-2, b_j-1>と略記する．

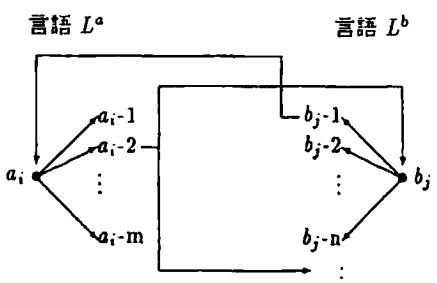

図 2 1次回路の例

3 語義対応の自動抽出

3.1 語義対応手続き

語義対応について以下の仮定をおく．

> 「ある見出し語の対の間に，唯一の1次回路が存在するとき，1次回路中に含まれる両言語の語義間に双方向対応を取ることができる．2次以上の回路からは，語義間の対応が取れない．」

たとえば，図 2 において，見出し語 a_iと b_jの間に唯一の1次回路$< a_i$-2, b_j-1>が存在し，この回路が語義 a_i-2 と b_j-1 を1度だけ通る．これは，語義 a_i-2 と b_j-1 とが双方向対応していることを示唆している．一方，2次回路以上では，誤りの混入する確率が高くなるので，本稿では語義対応の定義から除外しておく．

次に，見出し語間に複数の1次回路が存在する場合について考察する．まず，見出し語 $a_i(\in L^a)$と $b_j(\in L^b)$との間に，次の翻訳経路が存在する場合を考えよう(図 3)．

The Extraction of Interlingual Concepts from Bilingual Dictionaries
TOKUNAGA Takenobu, HARTONO, TANAKA Hozumi
Tokyo Institute of Technology

$[a_i, k, b_j]$, $[a_i, l, b_j]$, $[b_j, p, a_i]$, $[b_j, q, a_i]$,
ただし, $k \neq l, p \neq q$.

この場合には, < a_i-k, b_j-p>, < a_i-k, b_j-q>, < a_i-l, b_j-p>, < a_i-l, b_j-q>という 4 つの 1 次回路が存在し, 語義 a_i-k と a_i-l に対して, b_j-p と b_j-q のいずれを対応させるべきかが決まらない. 逆に語義 b_j-p と b_j-q に対して, a_i-k と a_i-l のいずれを対応させるべきかが決まらない. このように両側からみて, 語義間の対応に曖昧性がある場合には, 語義 a_i-k , a_i-l と語義 b_j-p , b_j-q 間の対応は考えない.

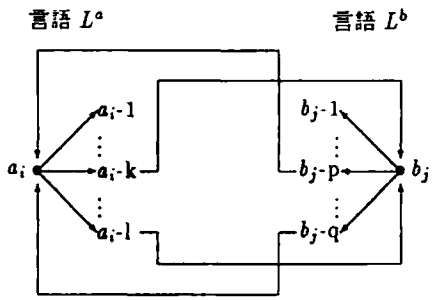

図 3　語義間の対応がとれない翻訳経路例

一方, a_iから b_jに向けて唯一の翻訳経路しか存在しない場合, この例でいえば, $[a_i, l, b_j]$が存在しない場合には, < a_i-k, b_j-p>, < a_i-k, b_j-q>という 2 つの 1 次回路が存在する. a_i-k の側からみれば, a_i-k は b_j-p と b_j-q のいずれに対応するか決まらず曖昧である. ところが b_j-p と b_j-q の側から見ると, 一意に a_i-k が対応していることがわかる. したがって, この場合には, 語義 a_i-k と語義 b_j-p, 語義 a_i-k と語義 b_j-q との間にそれぞれ双方向の対応が取れると仮定する.

以上の考察から, 語義対応手続きを次のように定義する.

[語義間の双方向対応手続き]
見出し語 $a_i \in L^a$と $b_j \in L^b$との間に存在するすべての翻訳経路について,

(1) a_iから b_jへの翻訳経路がただ 1 つ存在する.

(2) b_jから a_iへの翻訳経路が存在する.

が成り立つ場合, (1)の翻訳経路中に現れる a_i側の唯一の語義と, (2)の翻訳経路中に現れる b_j側のすべての語義との間にそれぞれ双方向の対応を取る.

3.2　語義対応グラフ

前節で述べた手続きにより両言語の語義間に双方向の対応が抽出できたとしよう. ここで, 語義間の対応にのみ着目し, この対応を有効辺として表現した, 語義対応グラフ(Sense PairGraph : SPG)を考える. TG では, 語義からは訳語に向けた有向辺が張られているだけであったが, SPG では, 各語義にはその語義を持つ見出し語が付加されているものとする.

4　中間言語の語彙項目の抽出

以上の手続きで, 異なる 2 言語間の語義の間の対応が得られる. 我々は, 語義は中間言語の語彙項目の候補であると考え, 特定の言語の語義を中心に置き, この言語とその他の言語との 2 言語間の語彙対応を手がかりに, 最終的な中間言語の語彙項目を設定する, というアプローチをとる.

ここで, 問題となるのは, 2 言語間の語義対応を, 各 2 言語間の双方向の対訳辞書を用いて抽出したのでは, 中心言語において別の辞書から抽出された同じ語義の統合ができないことである. この点は人間が介入せざるをえない. しかし, 個々の言語間との辞書から得られた中心言語の語義を総当たり的にまとめるのでは, 作業が膨大になってしまう. そこで, 我々は, 次のような手順をとる. 今, $L^c, L^1, \ldots L^n$の $n+1$ 個の言語を考える.

(1) 中心言語 L^cと L^1の間の語義対応を抽出する.

(2) (1)で得られた L^cの語義に相当する $L^2, \ldots L^n$の語を L^cの語義に割り当てる.

(3) (2)で割り当てた情報と $L^2 \rightarrow L^c, \ldots L^n \rightarrow L^c$ の辞書を用いて $n-1$ 個の 2 言語間語義対応を抽出する.

(4) 中心言語 L^cの語義を分割, あるいは統合する.

これらのうち, (1)と(3)は計算機による自動化が可能である. (2)については, 人手に頼らざるを得ないが, (4)については, 一部, 自動化の可能性がある.

このように, 中心言語を中心に 2 言語間の語義対応をとると, 語義の対応が 1 対多, あるいは, 多対 1 になる場合が考えられる. 手順(4)はこのような場合にどのように対処するかという問題であるが, これについては, 現在, 検討中である.

5　おわりに

本稿では, 対訳辞書から自動的に語義対応を抽出する方法を示し, これを発展させて機械翻訳のための中間言語の語彙項目を設定する手法について検討した. 現在, 英語と日本語について手順(1)についての実験を計画している.

類似の研究として, [1]があるが, 市山らの提案するダイヤモンド構造は, 我々が既に[3]で提案し, 本稿でも述べた 1 次回路と同じであり, 我々の手法と基本的に同じである. また, [1]では, 計算機処理と人手による処理をはっきりと切り分けておらず, 3 言語以上が関与する場合についても明確な指針を与えていない.

参 考 文 献

[1] 市山俊治, 野村直之. 多言語間機械翻訳用辞書の開発手法. 情報処理学会 自然言語処理研究会, NL73-14, 1989.

[2] 内田裕士. 電子化辞書の開発. ,「自然言語処理技術」シンポジウム論文集, 89-98 ページ, 情報処理学会, 1988.

[3] 田中穂積, 徳永健伸, Hartono, 岩山真. 翻訳用辞書からの中間概念の自動抽出に関する基礎的考察. 情報処理学会 自然言語処理研究会, NL72-3, 1989.

[4] 電子化辞書研究所. 単語辞書 (第 2 版). TR-006, 電子化辞書研究所, 1988.

[5] R. J. Byrd, N. Calzolari, M. S. Chodorow, and M. S. Klavans, J. L. Neff. Tools and methods for computatinal lexicology. *Computational Linguistics*, 13(3-4):219-240, 1987.