論　文　　Technical Papers

対訳辞書からの概念項目の自動抽出

The Automatic Extraction of Conceptual Items from Bilingual Dictionries

徳永　健伸*　　田中　穂積*
Takenobu Tokunaga　Hozumi Tanaka

* 東京工業大学工学部情報工学科
Dep. of Computer Science, Faculty of Engineering, Tokyo Institute of Technology, Tokyo 152, Japan.

1990年1月9日　受理

Keywords: machine translation, machine readable dictionary, conceptual dictionary, word sense.

Summary

To improve the quality of machine translation systems, we should step toward the deeper analysis at the conceptual level. Developing the machine translation systems with deeper analysis requires the dictionaries including following information : the set of conceptual items, the mapping relation between the surface words and the conceptual items, and the mapping relation between the conceptual items of the source language and that of the target language.

There are several researches to compile such dictionaries. Japan Electronic Dictionary Research Institute (EDR) is now compiling such dictionaries on a large scale. Nirenburg, *et al.* at Carnegie Mellon University has proposed a systematic method to construct a conceptual dictionary. These attempts try to compile the dictionary by hand with the help of software tools. However this approach suffers from the problems such as huge amount of manual labor, the unstable result and so forth.

Unlike this approach, the paper proposes a method to extract the information about the conceptual items from a pair of machine readable bilingual dictionaries in an automatic way. It is very difficult to compile the complete dictionary in a fully automatic way. The results of the method may require some refinement and modifications by human. Our goal is rather to automate the compilation process as much as possible and to decrease manual labor.

In the paper, we make an approximation in that each word sense defined in the bilingual dictionary is considered as a conceptual item. Since each word sense has the proper translations in the bilingual dictionary, the above approximation is reasonable in terms of word choice in the translation, and we can easily get both the set of conceptual items and the mapping relation between the surface words and the conceptual items. The most difficult thing is to get the mapping relation between the conceptual items of the source language and that of the target language. The paper focuses on this issue.

We introduce three types of *translation circuits*. The translation circuit is a tuple which consists of four elements, that is, a headword of both the languages and one of the word sense of both the headwords, with the condition that the word sense of one language should have the headword of the other language as a translation. We assume that the word senses in a translation circuit represent the same concept, that is, there is a mapping relation between the conceptual items (word sense) in a translation circuit. The paper describes the outline of a preliminary experiment conducted to verify this assumption. The results of the experiment are promising and some remarks are also given.

We conclude the paper with pointing out the possibility by extending our method to construct the set of conceptual items which can be shared by more than two languages.

1. はじめに

機械翻訳システムにおいて翻訳の質を向上させるためには，表層語から概念の世界に踏み込んだ，より深い解析が必要となる．表層語だけを考慮したのでは，語の多義性のために，適切な訳語が選択できないからである[1][2]．また，概念にも言語に共通な概念と言語の

文化的背景に依存する言語固有の概念があるので，概念の世界を考える場合に原言語と相手言語の両方の概念を考慮することが重要である．

概念レベルの解析を行う機械翻訳では，

① 概念の集合

② 表層語と概念の対応関係

③ 原言語の概念と相手言語の概念の対応関係

を設定する必要がある．

概念レベルの解析を目指した機械翻訳用の辞書に関する大規模な研究として，日本電子化辞書研究所（EDR）が行っている概念辞書の開発があげられる．EDR では，上述の三つの要素に対応して，それぞれ，①概念辞書，②単語辞書，③言語間対訳辞書の開発を目指している[3]．ただし，EDR の概念辞書は概念の単なる集合ではなく，概念間の関係をも含むものである．また，言語間の対訳辞書としては，日本語，英語間の辞書を考えている．EDR では，これらの辞書の構築を計算機の支援によって，すべて人手で行うというアプローチをとっている[3][4]．また，Carnegie Mellon 大学の Nirenburg らも計算機の支援により概念を人手によって組織的に構築することを提案している[5]．このようなアプローチの問題点として，作業者の主観によって概念の設定に揺れが生じることがあげられる[6]．また，各言語の概念間の対応関係を設定する作業では，両言語の概念数の積に等しい対応関係の検査が必要となる．これらの作業を人手で行うことを考えると，その量は非常に膨大なものとなる．

これに対して本論文では，機械可読な対訳辞書の対から，できるだけ主観的な判断を排除し，機械的に，これらの三つの要素を抽出する手法を提案する．ここでいう対訳辞書とは，市販の英和辞典，和英辞典などの辞書のことである．もちろん，本論文で提案する手法によって完全な概念の集合や概念の対応関係が自動的に抽出できるわけではない．最終的には人間による精密化や修正が必要となるが，重要なことは機械的に処理できる部分は，できるだけ計算機で行う，という点である．例えば，概念間の対応関係を設定する場合に，作業者の負担という観点から考えると，人間が辞書を参照しながらゼロから設定するよりは，計算機で自動的に抽出した対応関係について，それが正しいかどうかを判断するほうが，負担ははるかに軽減される．計算機による自動的な処理によりどの程度の結果が得られるかについても本論文では検討する．

まず，概念の集合については，対訳辞書中で定義されている語義を概念の候補として考える．対訳辞書中の語義は，その見出し語が持つ意味の違いを表しており，各語義には適切な相手言語の訳語が割り当てられている．したがって，機械翻訳における訳し分けを考える場合に，語義を概念として設定し，それを経由した訳語選択を行うことには意味がある．また，対訳辞書の語義は見出し語に対する語義であるから，語義を概念として設定すれば，表層語と概念の対応関係は，自然に求めることができる．以下では，特に断らない限り，「語義」と「概念」を同義として使う．

最も困難なのは，原言語の語義と相手言語の語義の対応関係をどのようにして抽出するかという問題である．本論文では，2 言語に関する双方向の対訳辞書を使うことによって，この問題に対する一つの解を与える．このために，2 章では対訳辞書を翻訳グラフとしてモデル化し，翻訳回路という概念を導入する．翻訳回路は，直感的に，「両言語の見出し語の対について，両方向の辞書でこの見出し語をそれぞれ引いたとき，お互いに，何れかの語義が訳語として相手の見出し語を含む」ことに対応する．翻訳回路中に含まれる語義の対が対応関係にある，つまり，意味がほぼ等しい，というのが我々の仮定である．3 章，4 章では，既存の英和辞典，和英辞典を用いた語義対応の抽出実験について，その実験方法と結果について述べる．本論文で提案する手法を用いて，複数の言語について概念の集合を設定し，ある言語（これを中心言語と呼ぶ）とその他の言語の対について概念間の対応関係を設定したとしよう．中心言語からその他の言語への対訳辞書は一般に異なるので，語義の定義も異なり，複数の中心言語の概念の集合ができることになる．これらの中心言語の概念の集合の間で概念間の対応をとることは機械的にはできない．しかしながら，仮にこの対応がとれれば，つまり，中心言語の複数の概念の集合を一つにまとめることができれば，この概念の集合を通して他の言語の概念間の対応をとることができる．このようにまとめた中心言語の概念の集合，および中心言語の概念と他の言語の概念の対応関係は，言語に共通な中間言語の概念の集合を構築するための重要な手掛りとなる．5 章では，論文全体をまとめるとともに，この発展性についても述べる．

2. 対訳辞書の構造

2・1 対訳辞書と語義間の対応

ここでは，対訳辞書から 2 言語間の語義対応を抽出するための準備として，対訳辞書の構造について考察する．二つの言語 L^a と L^b について，L^a から L^b への対訳辞書 D_{a-b} と，L^b から L^a への対訳辞書 D_{b-a} を

考える．今，$D_{a\text{-}b}$の見出し語 $a_i \in L^a$ が m 個の語義 $a_i/1$, …, a_i/m を持ち，$D_{b\text{-}a}$の見出し語 $b_j \in L^b$ が n 個の語義 $b_j/1$, …, b_j/n を持つものとする．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（Fig. 1）．ここで，一つの語義が複数の訳語に翻訳されることもあるので，一般に，各語義からは，相手言語の複数の見出し語に向けた有向辺が張られる．Fig. 1 の例では，$D_{a\text{-}b}$の見出し語 a_i の語義 $a_i/2$ の訳語が b_j であることが示されている．このとき，訳語 b_j は n 個の語義を持つが，そのうちのどれが $a_i/2$ に対応する語義であるかはわからない．これは，対訳辞書が，語義と訳語（見出し語）との間の対応を示しているだけで，語義間の対応を示していないことに起因している．この対応を機械的に抽出するのが我々の目的である．

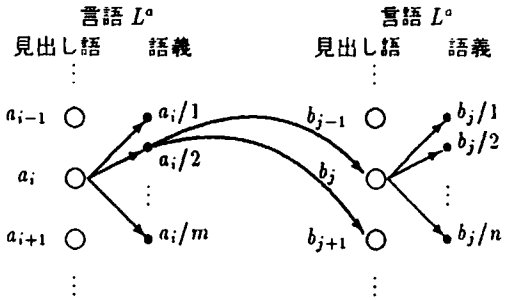

Fig. 1 The structure of a bilingual dictionary.

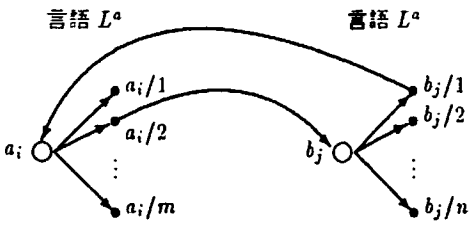

Fig. 2 An example of the translation circuit.

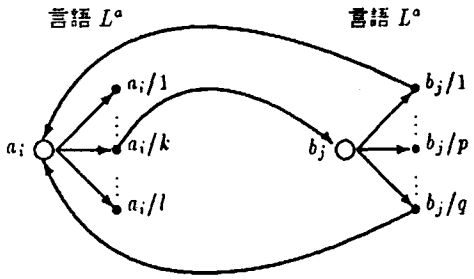

Fig. 3 An example of the type B word sense pair.

2・2 翻 訳 回 路

我々は，対訳辞書を一つの有向グラフとみなし（Fig. 1），このグラフを翻訳グラフと呼ぶ．言語の対を決め，言語 L^a から言語 L^b（ただし，$a \neq b$）への翻訳グラフを TG_{ab} と書く．添字 ab には方向性があることに注意．TG_{ab} は四つ組 $<H_a, S_a, T_b, E_{ab}>$ で構成される．ここで，H_a, S_a, T_b は節点の集合，E_{ab} は辺の集合である．$H_a \subset L^a$ は TG_{ab} の見出し語の集合，S_a は H_a の持つ語義の集合，$T_b \subset L^b$ は TG_{ab} の訳語の集合を表す．T_b の中には，TG_{ba} の見出し語に含まれないものも存在する．

TG_{ab} を用いた翻訳は，H_a の要素 a_i に対して一つの語義 a_i/k を選択し，そこに書かれた訳語 b_j を選択するというプロセスになる．この翻訳プロセスには，$a_i \rightarrow a_i/k \rightarrow b_j$ という経路が対応する．これを翻訳経路と呼ぶ．

ここで，二つの翻訳グラフ 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，閉路 $h_a(\in H_a) \rightarrow s_a(\in S_a) \rightarrow h_b(\in H_b) \rightarrow s_b(\in S_b) \rightarrow h_a$ を翻訳回路と呼ぶ．ここで，見出し語 $h_a \in H_a$ と $h_b \in H_b$ を含む回路が存在するとき，見出し語 h_a と h_b の間に回路が存在するという．

翻訳回路の例を Fig. 2 に示す．翻訳回路に含まれる語義の組によって，翻訳回路を表現する．例えば，Fig. 2 の翻訳回路は $<a_i/2, b_j/1>$ と表現する．

次に，見出し語対を固定したときに，その見出し語の間にできる翻訳回路を 3 種類に分類する．

定義：A 型の翻訳回路

「両言語の見出し語 h_a, h_b の間に唯一の翻訳回路が存在するとき，この回路を A 型の翻訳回路という．」

例えば，Fig. 2 において，見出し語 a_i と b_j の間に唯一の翻訳回路 $<a_i/2, b_j/1>$ が存在する．したがって，これは A 型の翻訳回路である．

定義：B 型の翻訳回路

「両言語の見出し語 h_a, h_b の間に複数の翻訳回路が存在し，これらの翻訳回路に含まれる語義の集合をそれぞれの言語について考えたとき，一方の言語の語義の集合が唯一の要素しか持たないならば，これらの回路を B 型の翻訳回路という．」

例えば，Fig. 3 の見出し語 a_i と b_j の間には，二つの翻訳回路 $<a_i/k, b_j/p>$ と $<a_i/k, b_j/q>$ が存在し，何れの回路も言語 L^a の語義については a_i/k しか含まない．したがって，$<a_i/k, b_j/p>$ と $<a_i/k, b_j/q>$ は，B 型の翻訳回路である．

定義：C 型の翻訳回路

「両言語の見出し語 h_a, h_b の間に複数の翻訳回路が存在し，これらの翻訳回路に含まれる語義の集合をそれぞれの言語について考えたとき，何れの言語の語義の集合も複数の要素を持つならば，これらの回路 C 型の翻訳回路という．」

Fig. 4 の例では，見出し語 a_i と b_j の間に四つの翻訳回路，$<a_i/k, b_j/p>$, $<a_i/k, b_j/q>$, $<a_i/l, b_j/p>$, $<a_i/l, b_j/q>$ が存在し，これらの回路は両言語

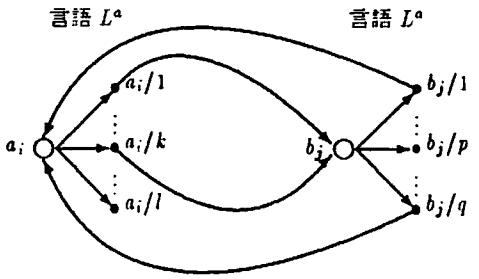

Fig.4 An example of the type C word sense pair.

の語義を二つずつ含んでいる．したがって，これはC型の翻訳回路である．

我々は翻訳回路中に含まれる語義は互いに意味が等しいと仮定している．このように意味の等しい語義の対を語義対応と呼ぶ．次章の実験によってこの仮定の妥当性を検証する．

3. 予 備 実 験

前章までの準備に基づき，実際のデータについて，予備実験を行った．本章では，この予備実験について説明し，次章では，この実験結果を踏まえ，本手法の問題点，限界について考察する．

実験用のデータとしては，研究社のライトハウス英和辞典[7]とライトハウス和英辞典[8]を用いた．これらの辞書を選んだのは，以下の理由による．

- 二つの辞書の編者が同じで，同じ編集方針に基づいて編集している．
- 語義の定義が明確で，語釈文が語義を機械的に抽出しやすい構造を持っている．

3・1 準 備

まず，2章で述べた語義対応の定義を実際の対訳辞書に適用するために，記法上の準備をする．

〔1〕 語義の定義

実際の対訳辞書を観察すると，英和，和英の何れについても，語義は以下のような階層構造を持っている．

- 英和辞典では，同じつづりで異なる意味を表す語は別の見出し語となっている．例えば，bank（土手）とbank（銀行）など．一方，和英辞典では，一つのひらがな表記の見出し語に複数の漢字表記が割り当てられている場合がある．
- 同一見出し語の中は意味の違いによって，いくつかの項目に分けられ，数字が割り当てられている．これを大項目と呼ぶ．
- 大項目の中は，意味の違いによって，いくつかの訳語が‘；’または‘，’で区切られている．このうち，‘；’のほうが意味の違いが大きいことを表しており，‘，’はほとんど交換可能な程度の意味の違いを表している．‘；’，‘，’で区切られたものをそれぞれ中項目，小項目と呼ぶ．

小項目のレベルに現れる訳語はほとんど同義と考えられるので，本実験では，中項目のレベルを一つの語義として定義する．したがって，一つの語義に複数の訳語が割り当てられることもある．また，何れの辞書においても，訳語中で，省略可能な要素（語）は‘（ ）’で，代替要素は‘［ ］’で囲まれている．何れの辞書についても省略可能要素は，省略したものと，省略しないものを訳語とした．また，代替要素については，和英辞典については，代替要素を置き換えたすべての場合を訳語としたが，英和辞典の場合には，代替要素を単に省略した．これは，英和辞典では，訳語（日本語）が分かち書きされていないので，どの部分が代替要素と置き換わるかが機械的に判別できないからである．

訳語の中には一つの語ではなく，句や，訳語に注釈文が付いた形で記述されているものがある．各語義について反対方向の辞書を引くことを考えると，このような句や注釈付きの語が見出し語として現れることはほとんどありえない．訳語に対する注釈は，語義の説明と考えられるので除去した．語義の説明は語義を区別することによって反映されるものと仮定する．

日本語のサ変名詞は，英和辞典では，訳語中に，（助詞＋）サ変名詞＋「する」の形で現れることが多いが，和英辞典の見出し語にはサ変名詞の形でしか現れない．英和辞典の訳語をプログラムによって抽出する際に，（助詞＋）サ変名詞＋「する」の形は，サ変名詞に変換した．その他の複数語からなる訳語については，このような訳語がどれくらいの頻度で存在するかを調べるのも実験の目的の一つなので，そのまま実験データとして含めた．

〔2〕 語義の表現

本実験では，語義を，大項目番号，品詞，中項目番号の三つ組で表現する．大項目番号は，辞書中で大項目に振られている番号をそのまま使用する．中項目番号は，大項目中での中項目の出現順に番号を0から割り当てる．実験の対象とした品詞は，名詞，動詞，形容詞，副詞の四つで，それぞれの辞書でTable 1のような記号を割り当てた．品詞については，英和と和英で若干扱いが異なっており，和英辞典では，すべての訳語の品詞が機械的に抽出できるわけではない．このような場合は，‘x’という記号を品詞として割り当てた．

Table 1 The correspondence of part of speech between dictionaries.

	英知辞典	和英辞典
名詞	n	n c（可算） u（不可算）
動詞	vi（自動詞） vt（他動詞）	v vi（自動詞） vt（他動詞）
形容詞	adj	adj
副詞	adv	adv
不明	—	x

3・2 実験方法および結果

残念ながら，ライトハウス英和辞典，和英辞典は何れも機械可読な形になっていない．したがって，まず，手作業で辞書を計算機に入力した．理想的には両方の辞書を全部入力することが好ましいが，経済的，時間的資源の制約上，これは不可能であった．そこで，対訳付きの英語のエッセイ集「最後のタヌキ[9]」から6編のエッセイを選び，その中に現れる自立語について辞書項目を入力した．日本語の場合は，単語の分かち書きが必要であるが，これも手作業で行った．このエッセイを選んだのは，一つのエッセイが300語程度の比較的容易な英語で書かれているので，一般的な語が多いであろうと予想したからである．最終的に，英和辞典については562見出し語，和英辞典については，779見出し語を入力した．入力に際しては，大項目を1行として入力する程度の加工は行ったが，基本的に辞書の字面をそのまま入力した．ただし，入力の手間を考え，例文は省略した．次にこれらの見出し語について，簡単なプログラムを作成し，見出し語，語義，訳語の組を機械的に抽出した．これは，2章で述べた翻訳経路に相当する．英和，和英についてそれぞれ，7 824，4 995の翻訳経路を抽出することができた．それぞれの辞書に関する品詞別頻度，見出し語当りの語義数，語義当りの訳語数をTable 2，Table 3に示す．

次に，抽出した翻訳経路をPrologの単位節に変換し，Prologインタプリタによって翻訳回路の数を数えた．結果をTable 4に示す．

Table 2 The statistics of the Englith-Japanese dictionary.
総見出し語数 ： 562
総訳語数（翻訳経路数）：7 824

品詞別頻度

名詞	動詞		形容詞	副詞
n	vi	vt	adj	adv
3 145	949	1 993	1 061	680

見出し語当りの語義数の頻度

語義数	1	2	3	4	5	6	7	8	9	10
頻度	60	72	54	60	39	31	32	32	24	18

11	12	13	14	15	16	17	18	19	20	21	22
12	12	14	5	13	11	9	6	5	3	5	6

23	24	25	26	27	28	29	30	31	32	33	35
3	2	1	3	3	5	1	5	2	2	3	1

36	37	40	41	43	54	66	69	平均
1	1	1	1	1	1	1	1	8.33

語義当りの訳語数の頻度

訳語数	1	2	3	4	5	6	7	平均
頻度	2 243	1 844	496	87	6	4	1	1.67

Table 3 The statistics of the Japanese-Englith dictionary.
総見出し語数 ： 779
総訳語数（翻訳経路数）：4 995

品詞別頻度

名詞			動詞		形容詞	副詞	その他	不明
n	c	u	vi	vt	adj	adv	—	x
487	594	159	378	952	471	232	186	1 082

見出し語当りの語義数の頻度

語義数	1	2	3	4	5	6	7	8	9	10
頻度	121	149	102	92	80	54	37	28	25	12

11	12	13	14	15	16	17	18	19	20	21	22
9	13	6	10	5	7	3	2	4	3	1	2

24	25	26	27	28	31	36	平均
1	1	2	1	1	3	5	5.16

語義当りの訳語数の頻度

訳語数	1	2	3	4	5	6	8	16	平均
頻度	3 266	623	80	43	3	4	2	1	1.24

訳語当りの単語数の頻度

単語数	1	2	3	4	5	6	7	平均
頻度	3 302	347	448	270	107	17	4	1.62

Table 4 The number of the extracted translation circuits.

型	正	誤	総数	正当率〔%〕
A型	237	7	244	97
B型	65	42	113	58
C型	18	64	82	22

4. 考　　察

4・1 辞書データ

Table 2, Table 3 から英和辞典のほうが，和英辞典よりも語義の数，訳語ともに多いことがわかる．この原因として，何れの辞書も日本語を母語とする人向けの辞書であるということが考えられる．和英辞典として英語を母語とする人向けの辞書を使うと結果が異なる可能性がある．この実験では，訳語が複数の語からなる場合に，語義対応をとることは考えていない．辞書の見出し語が複数の語からなることはほとんどないので，このような場合に，訳語から逆方向に辞書を引くと，ほとんど失敗する．しかしながら，Table 3 からわかるように，和英辞典については，73 %の訳語が一つの単語になっているので，1 単語の訳語だけを考慮しても第 1 近似としては，十分であると考えられる．

英和辞典の訳語は分かち書きされていないので，英和 (Table 2) については，「訳語当りの単語数の頻度」を調べていない．また，一般的に日本語について「単語」を定義することは容易ではない．この実験では，複数の単語からなる訳語を含む回路については考慮していないので，日本語訳を単語分割しても最終的に得られる語義対応の数には影響しない．

4・2 A 型の翻訳回路

A 型の翻訳回路は 244 個抽出できた．244 個中，7 個については，回路中に含まれる語義に対応関係はなかったが，残りの 237 個は，何れも正しい対応関係であった．したがって，97 %の正しさで語義対応が抽出できたことになる．A 型の翻訳回路については，機械的に抽出したものをそのまま語義対応として使用できると考えられる．誤りの主な原因は，対応する相手言語の訳語として適切な訳語が割り当てられていないためである．また，和英辞典では訳語の品詞が機械的に同定できない場合がある．実験では品詞不明の語義は何れの品詞とも照合できると仮定したが，これが誤りの原因となっている例もあった．

4・3 B 型の翻訳回路

B 型の翻訳回路は 113 個抽出できた．113 個中，正しいものが 65 個 (58 %)，誤ったものが 48 個 (42 %) であった．この実験に関しては，B 型の翻訳回路のうち，約半数が語義対応となることがわかる．一つの語義がいくつの語義に対応しているかの頻度を Table 5 に示す．正当率が約 50 %と低いので，B 型の翻訳回路から語義対応を選別するためには，人間の判断に頼らざるを得ないが，Table 5 からもわかるように，ほとんどの場合，語義が 1 対 2 あるいは 1 対 3 の関係であるから，適切なインタフェースを用意すれば人間の負担はさほど大きくないと予想できる．

Table 5 The number of word senses in type B word sense pairs.

組合せ	1 対 2	1 対 3	1 対 4	1 対 5	1 対 8
頻度	72	15	8	10	8

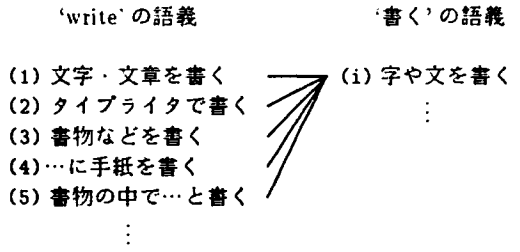

Fig. 5 The word sense pairs between 'write' and 'KAKU'.

語義の対応関係が 1 対多になる主な理由は，一方の辞書の語義が他方の辞書の複数の語義を含む場合があるためである．特に多くの語義を持つ見出し語は，より特殊な意味を持つ語義の前に，それらをすべて含むような一般的な語義を持つ場合がある．例として，見出し語'write' と'書く' の間で抽出できた五つの B 型の語義対応を Fig. 5 に示す．この例では，'write' の語義 (1) は'write' の一般的な意味を表し，'write' の他の語義も含んでいると考えられる．これに対して，'書く' の語義 (i) は一般的な'書く' を意味を表し，'write' の語義 (1) と対応する．'書く' のほうには'write' のように細かい語義の定義がないため，語義 (i) と語義 (2), …(4) の対応関係もそれぞれ抽出されてしまう．

4・4 C 型の翻訳回路

C 型の翻訳回路は 82 個抽出できた．82 個中，正しいものが 18 個 (22 %)，誤ったものが 64 個 (78 %) であった．C 型の翻訳回路では，見出し語の複数の語義どうしが対応しているが，対応する語義数の組合せを Table 6 に示す．正当率が低いため，C 型の翻訳回路から語義対応を選別する場合も，B 型と同様に人間の判断に頼らざるを得ない．さらに，語義の対応関係が多対多になると，判定すべき関係も組合せ的に多くなるため，人間の負担も B 型より重くなる．

Table 6 The number of word senses in type C word sense pairs.

組合せ	2 対 2	2 対 3	2 対 6	4 対 8
頻度	8	30	12	32

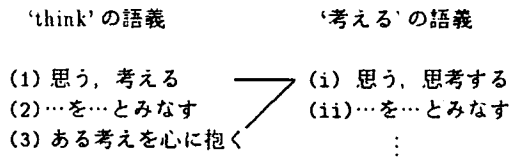

Fig.6 The word sense pairs between 'think' and 'KANGAERU'.

1対多の語義関係の中には，本質的には多対多関係であるが，語義に適切な訳語が割り当てられていないために回路ができず，1対多関係となっているものがある．注釈の解析などにより，より詳細な抽出を行えば，このような1対多の語義関係も多対多になる可能性がある．例えば，見出し語'think'と'考える'の語義の一部をFig. 6に示す．実線は実験により抽出できた語義対応を表す．この例では，語義(2)と語義(ii)が対応することがわかるが，語義(ii)の訳語として'think'が割り当てられていないため，(語義(2)，語義(ii))という対応が抽出できない．しかしながら，注釈などの情報を使って，この対応関係が抽出できるならば，これらの語義は多対多関係になる．

4・5 辞書の対称性

実験結果からわかるように，翻訳回路を構成しない翻訳経路の数が非常に多い．この実験では，辞書の見出し語の一部しか使っていないことも原因の一つであるが，その他にもいくつかの理由が考えられる．Byrdは，この理由として以下の四つをあげている[10]．ただし，Byrdの例は英語-イタリア語である．

① 語がほとんど派生形で使われる場合
② 特殊な語の訳語として一般的な語をあてている場合
③ 一般的な語の訳語としてより特殊な語をあてている場合
④ 辞書編集上の漏れによる場合

今回の実験に関して例をあげれば，次のようなものがある．矢印は翻訳の方向を表す．

dirt →わいせつな文章　②の例
fox →きつねの毛皮（全体で部分を指す）　③の例
machine →自動車（上位で下位を指す）　③の例

本論文では，双方向に語の翻訳ができる場合に限り，語義対応を考えてきた．1方向に語の翻訳ができれば，逆方向の翻訳も必ず可能である，という考え方もあるが，ここにあげた例については，特定の文脈が必要である．我々の目的は語義の対応を抽出することであるから，文脈に依存するような情報を使うことは好ましくない．しかし，1方向にしか語の翻訳ができない例は，言語に依存した概念の手掛りを与える場合もある．このような概念は本論文では対象としていないが，最終的に翻訳システムを考えるうえでは検討しなければならない問題である．

日本語-英語の間では，すでに指摘したように，訳語が必ずしも対訳辞書の見出し語として現れるような語でなく，複数の語からなる句となることも多い．例えば，'associate'の語義の一つに「仲間に加える」という訳語が割り当てられている．和英辞典には「仲間に加える」という見出し語はないので，この場合，翻訳回路はできない．このような訳語を扱うためには，まず，訳語の形態素解析を行い，見出し語として引ける単語に分割する必要がある．そのうえで対応がとれれば，概念レベルで構造変換を必要とするような概念の対応関係に関する情報が抽出できる可能性がある．複数からなる訳語をどのように扱うかは今後さらに検討を加える必要がある．

5. おわりに

本論文では，対訳辞書で定義されている各語義を概念の近似として用い，概念レベルの解析を行う機械翻訳システムに必要な情報の一つである2言語間の概念の対応関係を，機械可読な対訳辞書の対から，機械的に抽出する方法を示し，その実現可能性について検討した．また，予備実験から，本手法が有効であるという結果を得た．本手法では，まず，2言語間の対訳辞書を翻訳グラフでモデル化し，そこから，3種類(A型，B型，C型)の翻訳回路を抽出する．そして，抽出した翻訳回路に基づき，語義対応を求めるという手順をとる．

予備実験として，日本語と英語について，それぞれ約600見出し語を対象に抽出実験を行った結果，A型244個，B型113個，C型82個の翻訳回路を抽出できた．このうち，A型の翻訳回路は97％という非常に高い精度で正しい語義対応を与えることがわかった．B型，C型の翻訳回路の正当率はそれぞれ58％，22％であった．B型，C型の正当性の判断に人間の介入が必要であるが，適切なインタフェースを設ければ，人間の負担はさほど大きくはならないと考えられる．

今回の実験では，対訳辞書中の訳語が1語だけからなる場合を対象に実験を行ったが，実際には，訳語として句や節が与えられている場合がある．このようなものをどのように扱うか，今後さらに検討する必要がある．本論文では，対象を2言語に限定したが，この手法を複数の言語間で行えば，言語に共通な中間言語

を構築するための基礎データとなる可能性があることを以下で述べる．今，L^c，L^1，…，L^n の $n+1$ 個の言語について，以下のような手順を考える．

(1) 中心言語 L^c と L^1 の間の語義対応を抽出する．

(2) (1)で得られた語義対応に含まれる L^c の語義について，対応する L^2，…，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辞書 D_{2-c}，…，D_{n-c} を用いて $n-1$ 個の2言語間語義対応を抽出する．ここで，D_{a-b} は，言語 L^a から L^b への対訳辞書である．

このうち，(1)と(3)は計算機による自動化が可能であるが，(2)については，人間の判断に頼らざるを得ない．(2)については，中心言語の語義に人手で語を割り当てるのではなく，既存の辞書 $D_{c-i}(i=2, \cdots, n)$ を用い，機械的に語義対応を抽出することも考えられる．しかし，この場合，辞書 D_{c-i} の語義の定義はすべて異なるので，辞書間の語義の対応関係を人手で求める必要がある．どちらの方法がよいかは一概には言えない．何れにしても，このようにして作成した，中心言語の語義とその他の言語の語義の対応関係は，言語に共通な中間言語を構築するうえで重要な役割を果たすであろう．我々はこうして得られた語義とその間の対応関係をもとに，中間言語の概念項目の集合の核が構築できるのではないかと考えている．今回の予備実験では，辞書の一部しか使用しなかったため，実際に，対訳辞書の対からどれくらいの数の語義対応が抽出できるかについては，正確に予測することは困難である．しかし，予備実験の結果から単純に計算すれば，見出し語数の約40％の語義対応が完全に自動的に抽出できたことになる．これを単純に外挿すれば，例えば，見出し語6万語程度の辞書の対から，約24 000程度の語義対応が完全に自動的に抽出できることになる．B型，C型の翻訳回路から人間の介入によって抽出できる語義対応を含めればこの数はもっと増えるだろう．今後，辞書全体について実験を行うとともに，本手法で得られた情報を使った機械翻訳の実験システムを構築し，その妥当性を検証する予定である．

◇参 考 文 献◇

(1) 田中穂積，野村浩郷，編：機械翻訳，*bit* 別冊，pp. 39-46，共立出版 (1988)．

(2) 石崎 俊，井佐原均：文脈情報翻訳システムCONTRAST，情報処理，Vol. 30, No. 10, pp. 1240-1249 (1989)．

(3) 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，pp. 89-98，情報処理学会 (1988)．

(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) Nirenburg, S. and Raskin, V.：The subworld concept lexicon and the lexicon management system, *Computational Linguistics*, Vol. 13, No. 3-4, pp. 276-289 (1989)．

(6) 清野正樹：概念辞書における概念の安定化の方法，第3回人工知能学会全国大会，pp. 383-386 (1989)．

(7) 竹林 滋，小島義郎，編：ライトハウス英和辞典，研究社 (1984)．

(8) 小島義郎，竹林 滋，編：ライトハウス和英辞典，研究社 (1984)．

(9) Lummis, C. D.：*The Last Badger*. Syobunsya (1988)．

(10) Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, M. S. and Neff, J. L.：Tolls and methods for computational lexicology, *Computational Linguistics*, Vol. 13, No. 3-4, pp. 219-240 (1987)．

〔担当編集委員・査読者：河田 勉〕

著 者 紹 介

徳永 健伸（正会員）

1983年東京工業大学工学部情報工学科卒業．1985年同大学院理工学研究科修士課程修了．同年（株）三菱総合研究所入社．1986年東京工業大学大学院博士課程入学．1987年より同大学工学部情報工学科助手．自然言語処理，知識表現に関する研究に従事．情報処理学会，日本ソフトウェア科学会，認知科学会各会員．

田中 穂積（正会員）

1964年東京工業大学理工学部制御工学科卒業．1966年同大学院修士課程修了．同年電気試験所（現，電子技術総合研究所）入所．1983年東京工業大学工学部情報工学科助教授，1986年同大学教授．工学博士．人工知能，自然言語処理の研究に従事．情報処理学会，電子情報通信学会，認知科学会，計量国語学会各会員．