日産科学振興財団研究報告書, 15 (1992)

対訳辞書からの中間概念の自動抽出法とその機械翻訳への応用に関する研究

The automatic extraction of conceptual items from bilingual dictionaries and its application to machine translation

代表研究者　東京工業大学教授　田中穂積
Prof., Tokyo Inst. of Tech.
Hozumi TANAKA

協同研究者　東京工業大学講師　徳永健伸
Lecturer, Tokyo Inst. of Tech.
Takenobu TOKUNAGA

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 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, this 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 the three types of *translation circuit*. 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. And the word sense of the 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 concepts, 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 are shared by more than two languages.

研究目的

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[3),6)]。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おこなう機械翻訳では，

(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挙げられる。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[5)]。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，すべて人手で行なうというアプローチをとっている[5),7)]。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[10)]。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ゆれが生じることが挙げられる[2)]。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。

これに対して本研究では，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，本研究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おこなう，という点である。たとえば，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方が，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研究では検討する。

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おこな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，とくにことわらない限り，「語義」と「概念」を同義として使う。

もっと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。本研究では，2 言語に関する双方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時，お互いに，いず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。

本研究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もできる。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がかりとなる。

研究経過

2.1 対訳辞書の構造の分析

2.1.1 対訳辞書と語義間の対応

まず，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 $D_{a\to b}$ と，L^b から L^a への対訳辞書 $D_{b\to a}$ を考える。今，$D_{a\to b}$ の見出し語 $a_i \in L^a$ が m 個の語義 $a_i/1, \ldots a_i/m$ を持ち，$D_{b\to a}$ の見出し語 $b_i \in L^b$ が n 個の語義 $b_j/1, \ldots b_j/n$ 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（図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。

図1の例では，$D_{a\to b}$ 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。この時，訳語 b_j は n 個の語義を持つが，その内のどれが $a_i/2$ 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。

2.1.2 翻訳回路

我々は，対訳辞書を一つの有向グラフとみなし（図1），このグラフを翻訳グラフと呼ぶ。言語の

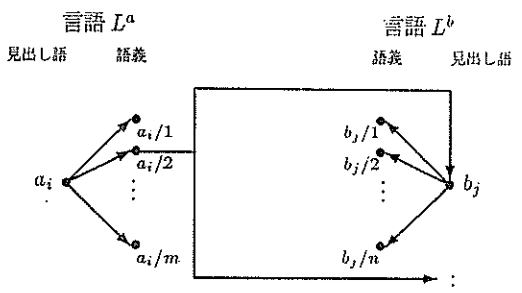

図1. 対訳辞書の基本構造.

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 $\langle H_a, S_a, T_b, E_{ab}\rangle$ 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，S_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，$a_i \to a_i/k \to b_j$ 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 $h_a(\in H_a) \to s_a(\in S_a) \to h_b(\in H_b) \to s_b(\in S_b) \to h_a$ 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。

次に，見出し語対を固定した時に，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。」

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，各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時，一方の言語の語義の集合が唯一つの要

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。」

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，各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時，いずれの言語の語義の集合も複数の要素を持つならば，これらの回路をC型の翻訳回路という。」

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。

2.2 実験

以上の議論に基づき，実際のデータについて，予備実験をおこなった。本節では，この予備実験について説明し，次節では，この実験結果をふまえ，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[4]とライトハウス和英辞典[1]を用いた。残念ながら，ライトハウス英和辞典，和英辞典はいず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，これも手作業でおこなった。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，翻訳経路に相当する。英和，和英についてそれぞれ，7824，4995の翻訳経路を抽出することができた。これらの翻訳経路をもとに翻訳回路の数を数えた。結果を表1に示す。

表1. 翻訳回路の抽出結果

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%

表2. B型の語義対応に含まれる語義数

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

2.3 考察

2.3.1 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，いず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，品詞不明の語義はいず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。

2.3.2 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表2に示す。正当率が約50パーセント低いので，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，表2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，適切なインターフェイ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。

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。

2.3.3 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同士が対応しているが，対応する語義数の組合せを表3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。

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，より詳細な抽出をおこなえば，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。

2.3.4 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1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挙げている[8]。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) 辞書編集上の洩れによる場合

今回の実験に関して例を挙げれば，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

本研究では，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，ここに挙げた例については，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がかりを与える場合もある。このような概念は本研究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上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。たとえば，'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おこない，見出し語として引ける単語に分割する必要がある。その上で対応がとれれば，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。

研究成果

本研究では，対訳辞書で定義されている各種義を概念の近似として用い，概念レベルの解析をお

表3. C型の語義対応に含まれる語義数
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

こなう機械翻訳システムに必要な情報のひとつである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おこなった結果，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，適切なインターフェイスを設ければ，人間の負担はさほど大きくならないと考えられる。

今後の課題と発展

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おこなったが，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。

本研究では，対象を2言語に限定したが，この手法を複数の言語間でおこなえば，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，$L^c, L^1, \ldots L^n$ 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 $L^2, \ldots L^n$ 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辞書 $D_{2\to c}, \ldots D_{n\to c}$ を用いて $n-1$ 個の2言語間語義対応を抽出する。ここで，$D_{a\to b}$ 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えない。(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 $D_{c\to i}$ $(i=2, \ldots n)$ 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 $D_{c\to i}$ 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。いずれにしても，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対応関係は言語に共通な中間言語を構築する上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，たとえば，見出し語6万語程度の辞書の対から，約2万4千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おこなうとともに，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) 小島義郎，竹林　滋，編集者：ライトハウス和英辞典，研究社，1984.
2) 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.
3) 石崎　俊，井佐原均：文脈情報翻訳システムCONTRAST，情報処理，30, (10), pp. 1240–1249, 1989.
4) 竹林　滋，小島義郎，編集者：ライトハウス英和辞典，研究社，1984.
5) 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.

6) 田中穂積, 野村浩郷, 編集者: 機械翻訳 bit 別冊, pp. 39–46, 共立出版, 1988.
7) 電子化辞書研究書: 単語辞書 (第 2 版), TR-006, 電子化辞書研究所, 1988.
8) R. J. Byrd, N. Calzolari, M. S. Chodorow, and M. S. Klavans: J. L. Neff, Tools and methods for computational lexicology, *Computational Linguistics,* **13**, 3–4, pp. 219–240, 1987.
9) C. D. Lummis: *The Last Badger, Syobunsya,* 1988.
10) S. Nirenburg and V. Raskin: The subworld concept lexicon and the lexicon management system, *Computational Linguistics,* **13**, 3–4, pp. 276–289, 1989.

発表論文リスト

・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.
・Tokunaga, T. and Tanaka, H.: The Automatic Extraction of Conceptual Items from Bilingual Dictionaries; Proceedings of the Pacific Rim International Conference on Artificial Intelligence '90, Nagoya, pp. 304–309, 1990.
・徳永健伸, 田中穂積: 対訳辞書からの概念項目の自動抽出; 人工知能学会誌, 6(2) pp. 228–235, 1991.