自 然 言 語 処 理 78－5
（1990．7．19）

多言語間で共用可能な概念の自動抽出

Hartono, 田中穂積
東京工業大学・工学部・情報工学科

概要
中間言語の概念をいかに設定すればよいかは大変重要な問題である．各言語の持つ語義が概念を表すという考え方がある．それによれば，対象となる言語の語義を抽出する必要があるが，複数の言語を対象として得られた語義の内，重複した語義をどのように発見するかという問題が生じる．また，一つの言語に数十万個の単語があり，各単語は複数個の語義を持ち，さらに多言語を考えれば，処理しなければならない語義の数はもはや人手で扱い切れなくなる．本稿では，対訳辞書の語義を利用して多言語間に重複する語義を機械的に抽出する方法を提案し，実験を試みる．まず対訳辞書の語義を説明し，抽出アルゴリズムを述べ，実験結果について考察する．

An Automatic Extraction of multilingual shareable concept

Hartono, Hozumi TANAKA
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-Ku, Tokyo-152 Japan

Abstract

Problem of the multilingual shareable concept establishment seems to be considerably complex and controversial in the interlingua construction. Those concepts may be representatively replaced by the word-senses belong to each language. In this case, extraction of the overlapping word-senses from among those intended languages should be taken into account. The problem, however, arises when the amounted millions of word-senses should be handled. In this paper, we propose a method of how those duplicated word-senses can be extracted automatically by using the bilingual-dictionaries. The experimental results which were obtained in the method described above are also shown to verify its effectiveness.

1 はじめに

翻訳処理を一つの中間言語に集中して行うことができる中間言語方式は，入力言語の文を解析して得られる中間言語から複数の目標言語を生成することができ，各言語対ごとの翻訳システムを構築する必要がなくなる．そのため，機械翻訳における中間言語方式が，多言語間機械翻訳システムの研究及びその実現と関連して注目されている．

しかし，中間言語そのものの設計について多くの問題を抱えているのが現状であろう．特に中間言語の概念どう設定すればよいか，またそれをどのように実現するかは重要な問題である．中間言語の概念には，次の二つの考え方がある [3]；

- 言語に依存しないプリミティブな概念の集合を設定し，それを組み合わせて語義を記述する
- 各言語が持つ語義の和集合

前者については，中間言語の普遍性を考慮すれば理想的には，語義をそれ以上細分できないプリミティブな概念の組み合わせで表現した方がよいとされている．しかし，対象となる言語で表現されているすべてのものを表現可能なプリミティブを抽出するのは極めて難しい問題であり，その実現の見通しもはっきりしないというのは現状である．

一方後者については，対象となる言語が共通に持つ語義と，各言語が個別に持つ語義とを，概念とする考え方である．それによれば，対象となる言語に含まれる語義を抽出しなければならないが，複数の言語を対象とした時，得られる語義の内，重複した語義をどのようにして発見するかという問題が生じる．重複した語義は，複数個の言語で共用される語義であるから，機械翻訳の立場からは，これを抽出することは，重要である．なぜならば，このような共用される語義は，中間言語方式による翻訳でいえば，中間言語の中核に位置するものと考えられるからである．

しかし，一つの言語に数十万個の単語があり，各単語は複数個の語義を持ち，さらに多言語を考えれば，処理しなければならない語義の数は膨大になる．この膨大な数の語義を手作業でいちいちマッチングを取って処理することはほとんど不可能であると考える．

われわれは，対訳辞書を通して単語の語義を見ることができる．この事実を利用して多言語が共通に持つ語義を機械的になんとか抽出できるのではないかと考えて研究を試みたのである．

本研究では，既存の対訳辞書の同義性に注目し，対訳辞書の語義を利用して多言語間に重複する語義を機械的に抽出する一つの方法を示す．また，この方法の有用性を確認するため実験を行った．

2 同義性と対訳辞書

われわれは，規則に従って単語を列挙して意義を表現し，意志疎通をはかる．列挙された単語には，それぞれ特定の語義が割り当てられ，表現されていると考える．われわれは，時には異なった複数の語義を同じ単語で，同じ語義を異なった語で表したりする (一般に前者は同音異義といい，後者は異音同義という)．その結果として，単語から時には複数の意味を想起でき，いくつかの単語が同じような意味を持つ可能性があるかどうかを判断することができる．上記の事実から Jones[10] は，「その同じような意味を持つと判断される両語の間には程度の差があるが，同義性が存在する」と結論付けている．

異なる言語を使って人間同志が翻訳を通して意志疎通が可能だという観点からもわかるように，各言語には共通語義もしくは互いに近似した語義が存在するということがわかる．つまり，同一言語内だけに同義性が存在するのではなく，異なった言語の間の中でも同義性が存在する．この多言語間の同義性を対訳辞書を通して簡単にみることができる．

本研究では，困難だとされている多言語間における共用語義を既存の対訳辞書を手立てに，機械的に設定可能であると考え，実験を試みた．

3 対訳辞書の見出し語と語義と訳語について

既存の対訳辞書を利用するに当たって，ここではまず対訳辞書の見出し語と語義と訳語の関係と，これから用いる用語／記号について説明する．

言語 L^Aから言語 L^Bへの対訳辞書を考えたときに，一般に Fig.1 で表現することができる．

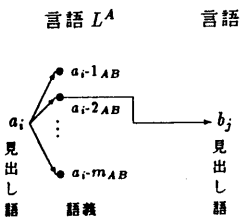

Fig.1 対訳辞書の見出し語と語義と訳語

ここに，a_iは言語 L^Aの見出し語で，a_i-1_{AB}...a_i-m_{AB}は a_iの語義で，添え字 ABは，この語義が言語 L^Aから言語 L^Bへの対訳辞書で定義されていることを示す．b_jは a_iの語義 a_i-2_{AB}の訳語 (言語 L^Bの語) であり (以下では，$a_i \rightarrow b_j$と書く)，単語で表したり，句または文で表現したりすることもある．

4 いままでの研究と問題点

概念項目の抽出に関しては，電子化辞書研究所 (EDR) [1] と田中ら [2] の研究がある．EDRでは，言語の見出し語に対してワークシートなどを使用してすべて手作

業で概念の抽出を行っている．また，日本語と英語のみが考慮されていて，それ以外の言語が併合される場合に，この膨大な数の概念をどういう手順で処理するかは，必ずしも明確でない．これに対して田中らは，二つの言語の対訳辞書を用いてアルゴリズムを与え，機械的に抽出しようとしている．ここで，田中らの手法を考えてみよう．二つの言語 L^A と L^B に対して L^A から L^B への対訳辞書（以下では T^{AB} と書く）と，逆の L^B から L^A への対訳辞書（T^{BA}）を考える．L^A の見出し語 a_i が m 個の語義 a_i-1,a_i-2,...,a_i-m を持ち，a_i のいずれかの語義の訳語が b_j であるとする．これに対して，L^B の見出し語 b_j が m 個の語義 b_j-1,b_j-2,...,b_j-m を持ち，b_j のいずれかの語義の訳語が a_i であったとする．このとき，両言語の語義間に対応が取れたとし，両語義の間に中間概念の候補として設定する，としている．しかし，他の言語の概念と併合，つまり三ケ国語以上で考える際に，この手法には以下のような大きな問題が残されている．

言語 L^A と L^B の対訳辞書 T^{AB} と T^{BA} で抽出した中間概念の候補と，言語 L^A と L^C の対訳辞書 T^{AC} と T^{CA} で抽出した中間概念の候補との対応を機械的に取ることができない(Fig.2)．これは，T^{AB} と T^{AC} に，それぞれ言語 L^A の見出し語 a_i を同じ個数の語義を持つとしても，辞書が異なっているから，それぞれの語義は，必ずしも同じではないからである．

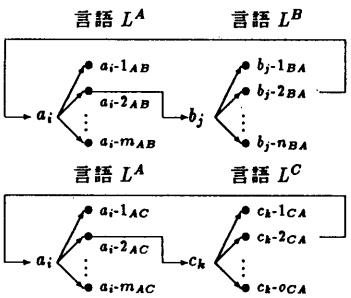

Fig.2 語義の対応

つまり，Fig.2 のように L^A, L^B と L^A, L^C の間にそれぞれ語義間に中立語彙項目候補が取れたとして，a_i-2$_{AB}$ と b_j-2$_{BA}$ の語義間で得られた中立語彙項目候補と，a_i-2$_{AC}$ と c_k-2$_{CA}$ の語義間で得られた中立語彙項目候補とは，なんの関係も結びつけられない．なぜならば，同じ a_i ではあるが，語義 a_i-2$_{AB}$ と a_i-2$_{AC}$ が同一であるという保証はまったくないからである．

本研究は，多言語間が共通に持つ語義を機械的に抽出する方法を示すものにあるが，三つの異なる言語間の対訳辞書を基軸にして抽出することによって上記の問題を解決できることを示す．

5 共用語義項目の抽出

ここでは，まず，対訳辞書の見出し語とその訳語それぞれが持つ語義の関係を理論的に説明する．次に，三つの対訳辞書を中心に用い，多言語が共通に持つ語義を機械的に抽出するアルゴリズムと，共用語義が抽出されるルートの種類について述べる．

この三か国語間の三つの対訳辞書とは，具体的には，言語 L^A, L^B, L^C を考えた場合に，T^{AB}, T^{BC}, T^{CA} のような組合せの対訳辞書のことを指す．

5.1 対訳辞書の見出し語とその訳語の語義関係

第二章ですでに述べたように，多言語間における語の同義性を対訳辞書を通して参照することができる．言語 L^A の語 a_i が言語 L^B の語 b_j に訳せるということは，語の多義性を考慮にいれても語 a_i と語 b_j との間には共通に表せる語義が存在すると考えられる．そこで，まず

> 「ある言語の語 a_i が，他言語の語 b_j に翻訳できると認められれば，a_i と b_j の両語間に共用語義が存在する」

という仮定を置く．

上記の仮定に基づいて，言語 L^B の b_j は，言語 L^A の a_i の訳語であるとした時，両語が持つ語義の集合は，以下の四つの相互関係で表すことができる．ただし，$S(a_i)$ と $S(b_j)$ はそれぞれ a_i と b_j が表す語義の集合で，$S(a_i) \neq \emptyset$ かつ $S(b_j) \neq \emptyset$ かつ $((S(a_i) \cap S(b_j)) \neq \emptyset)$ であるとする．

1. $(S(a_i) = S(b_j))$ ...**R1**
2. $(S(a_i) \supset S(b_j))$ ...**R2**
3. $((S(a_i) - (S(a_i) \cap S(b_j))) \neq \emptyset)$ かつ $((S(b_j) - (S(a_i) \cap S(b_j))) \neq \emptyset)$ ...**R3**
4. $(S(a_i) \subset S(b_j))$ ...**R4**

なお，a_i と b_j が共通に持つ語義は $(S(a_i) \cap S(b_j)$ で表されていると考える．

次に，以上の関係を考慮して共用語義項目を以下のように仮定する．

> 「a_i, b_j, c_k はそれぞれ言語 L^A, L^B, L^C の単語とし，T^{AB}, T^{BC}, T^{CA} の中から，$a_i \rightarrow b_j$ と $b_j \rightarrow c_k$ と $c_k \rightarrow a_i$ がそれぞれ認められれば，a_i, b_j, c_k が共通に持つ語義が存在するとし，その共用語義項目を［a_i, b_j, c_k］で表すものとする．」

共用語義項目は，Fig.3 のような回路によって抽出される．

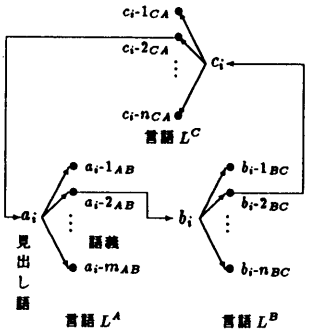

Fig.3 言語 L^A, L^B, L^Cのループ回路

• ループする条件について

a_i, b_j, c_kが，それぞれ言語 L^A, L^B, L^Cの単語で，T^{AB}から $a_i \rightarrow b_j$ と T^{BC}から $b_j \rightarrow c_k$が認められているとする．この時，a_iと c_kの語義関係は，以下のようになる．

1. [a_iと b_jの語義の集合は，**R1** または **R2** の関係の場合]
   $S(b_j) \subseteq S(a_i)$ であることと，$(S(b_j) \cap S(c_k)) \neq \emptyset$ であるので，$(S(a_i) \cap S(c_k)) \neq \emptyset$ が認められる．

2. [a_iと b_jの語義の集合は，**R3** または **R4** の関係の場合]
   $(S(b_j) - (S(a_i) \cap S(b_j))) \neq \emptyset$が存在するため，$(S(a_i) \cap S(b_j)) \cap (S(b_j) \cap S(c_k)) = \emptyset$ が有り得るから，つまり，$S(a_i) \cap S(c_k) = \emptyset$ が考えられる．

(1) については，$S(a_i) \cap S(c_k) \neq \emptyset$であるから，さらに新たな a_iと c_kの関係を結ばなくても共用語義の存在が認められるが，(2) の場合は，$(S(a_i) \cap S(c_k)) = \emptyset$ があるために，a_iと c_kがなんだかのかたちで関係を結ぶ必要がある．本研究では，$(S(a_i) \cap S(b_j) \cap S(c_k)) \neq \emptyset$ (同義性) が得られるように，$c_k \rightarrow a_i$を導入する．ただし，「$a_i \rightarrow b_j$」と「$b_j \rightarrow c_k$」と「$c_k \rightarrow a_i$」の関係はいずれも **R3** の場合は，理論上 $(a_i \cap c_k) \cap (b_j \cap c_k) = \emptyset$が考えられる．本研究では，この可能性は事実上小さいとみて共用語義項目の仮定を設定したが，今後実験の結果に応じてさらに検討を加える必要がある．

• 三か国語を基軸にすることについて

基本的に，a_iと b_jの共通語義は，$S(a_i) \cap S(b_j)$であることと，$(S(a_i) \cap S(b_j) \cap S(c_k)) \supseteq (S(a_i) \cap S(b_j) \cap S(c_k) \cap S(d_l))$ であるため，ループ中の言語が多ければ多いほど共通部分が小さくなる．つまり，語義は細分化されることがあっても，大きくなることはない．[1]この結果は対象となる言語間では共用可能な語義であることに変わりがないが，細かくなる共通部分と，既存対訳辞書の記述の限界とを考えれば，基軸にする言語が多ければ多いほど得られるループの数も少なくなる．実際に，共用語義として利用できるものも抽出されなくなる．

• 併合について

以上の方法を適用することにより，新しい言語の語義と，すでに得られた共用語義とを機械的に併合することができる．これは，すでに用いられた対訳辞書のうちのいずれかを利用することができるので，同じ辞書の同じ語義を利用することによって，多言語間の語の語義の対応を機械的に取れるからである．

具体的な例を見てみよう．言語 L^A, L^B, L^Cと対訳辞書 T^{AB}, T^{BC}, T^{CA}を用い，これらの言語間に Fig.3 のような共用語義項目 $[a_i, b_i, c_i]$が得られたとする．併合しようとする新しい言語を L^Dで，対訳辞書 T^{AB}, T^{BD}, T^{DA}を使用して，Fig.4 のような共用語義項目 $[a_i, b_i, d_i]$ が得られたとすれば，この時，d_iと，Fig.3 で抽出された $[a_i, b_i, c_i]$ とを併合することができる．これは，同じ対訳辞書 T^{AB}の同じ語義 a_i-2$_{AB}$を利用することによってができたから，結び付けられたものである．

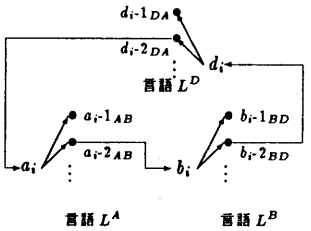

Fig.4 言語 L^A, L^B, L^Dのループ回路

5.2 抽出されるループの種類

前節で述べた共用語義項目の設定に基づいて，実際に抽出されるループを以下の二種類に分けることができる．

1. 語義が機械的に区別できるもの
2. 語義が機械的に区別できないもの

語義が機械的に区別可能なものについては，Fig.5 で示されているループがもっとも基本的なタイプである．

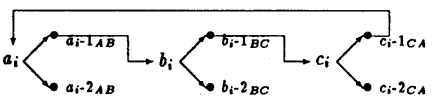

Fig.5

一つの語が複数の共用語義項目として使われる場合もある．例えば，Fig.6 では，a_iが共用可能な語義として，$[a_i, b_i, c_i]$ と $[a_i, b_j, c_j]$で使われている．

[1]例えば，「私たち」→「we」(英語)→「我 門」(中国語)から第一人称複数語義が作られる．インドネシア語の第一人称複数には，相手を含む語義と相手を含まない語義がある．この場合，第一人称複数語義を二つに分割する．さらに，例えば，男女を区別する言語があれば，さらに分割する必要がある．

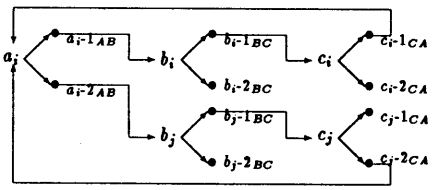

Fig.6

次に，語義が機械的に区別できないものには，以下のような種類がある．まず，Fig.7のような異なる複数の語義を一つの訳語で表すループが挙げられる．

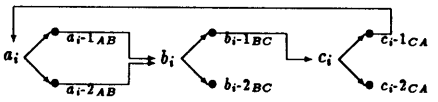

Fig.7

この場合，共用可能な語義 $[a_i, b_i, c_i]$ として認められるが，a_iのどの語義が使われているかは機械的に区別できないことから，併合を考える場合に語義 a_i-1$_{AB}$または a_i-2$_{AB}$を使用することができない．

最後に，同じく機械的に区別できないタイプに，Fig.8のようなループがある．

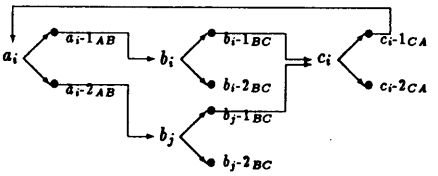

Fig.8

Fig.8のループでは共用可能な語義 $[a_i, b_i, c_i]$ と $[a_i, b_j, c_i]$ が，採用される．しかし，併合を考える際に，語義 c_i-1$_{CA}$ を利用する場合は，a_i-1$_{AB}$と a_i-2$_{AB}$との区別を考慮する必要がある．

6 実験と結果

- 対訳辞書の語義の定義
  既存の対訳辞書は，一般的に次のような記法を用いて語義を表す．
  - 番号付けで語義を分ける[2]．
  - 語義は意味の差異の度合いに応じて「,」とか「;」などの記号を用いて区分される．
    例えば，研究社ライトハウス英和辞典では，複数の訳語が存在する場合には，ほぼ同じ意味だと認められた時は「,」，ニュアンスの違いがあれば「;」を用いて列記する．

[2]配列は一般に頻度数の高い順に従う

しかし，われわれは，対訳辞書でいう「語義」をそのまま語義として扱うことはしない．対訳辞書には，例えば，「愛情」が同じ語義に「love ; affection ; attachment」の訳語を持っているが，訳語の意味(loveとattachment)の差は開き過ぎたり，また，複数の訳語を持つ語義が意味の差異を「,」とか「;」で分別される辞書は多いが，語の意味をめぐって辞書編集者によって認識が異なる．例えば，同じ語義内で「B1，B2」とする人もいれば，「B1；B2」として区別する人もいる[3]．

B1とB2は訳語であることにかわりがないことと，さらに利用される多種多様の辞書の対応をも考慮して本研究では語義を以下のように定義する．

「各訳語は，一つの語義を表す」

これらの語義はそれぞれ語義番号をつけて表現する．

- データ
  エントリのデータは「最後のタヌキ」C.ダグラス.ラミス著，中村直子訳の本[6]の中から「The Infected Pool (汚染されたプール)」，「The Hackneyed Future (陳腐な未来)」，「The Parable of the Word Processor(ワープロ寓話)」，「Atomic Rain (原子力の雨)」，「The Last Badger (最後のタヌキ)」と「Talking Machines (しゃべる機械)」の6章を使った．各章は，平均300単語(英語)からなる．重複する単語を排除した結果，最終的に712エントリが得られた．これらのエントリを見出し語にし，インドネシア語を引き，得られたインドネシア語のエントリを見出し語にし，日本語を引いた．さらに日本語から元の英語を引いた．対訳辞書はそれぞれCornell University Pressの「英語→インドネシア語」[7]，Japan Indonesia Associationの「インドネシア語→日本語」[4]，研究社ライトハウスの「日本語→英語」[5]を利用した．詳しい内容は以下の表を参照．(表の読み方:「→」の左側は，辞書中に現れた訳語の数(ただし，最初の712は，テキスト中の英語の単語数)．「→」の右側は，実際に対訳辞書にあった見出し語の数．「↓」の下は，訳語の数である．)

英語	712 → 565		
↓	↓		
インドネシア語	2233	→ 1232	
↓		↓	
日本語		3163	→ 2454
↓			↓
英語			5960

- 上記のデータを用いて第5章で示されたループを抽出する．最終的に，833個のループが得られた．

[3]例えば，同じ英語springを同語義内で(三省堂新コンサイス英和辞典)には「動機，原動力」，(研究社ニューコリダート英和辞書には「動機：原動力」と書いている

7 考察

●結果について

1. 英語の元のエントリー数712に対して得られたループの総数は833であるが，元の712エントリのうちループしなかったものが434エントリーあった．つまり，ループを持つエントリーの総数は278個である．実際に引けたのは565エントリーから考えれば約半数のエントリーがループしないという計算になる．これには，以下のような原因が考えられる．

●[英語→インドネシア語]:(英語の辞書引き)

英語の生テキストから712単語が得られたが，形態素等を考慮しなければ約250エントリーが辞書引きできない．辞書にないエントリーには，固有名詞，数字，動詞の変形，名詞の複数形などが含まれる．動名詞，過去形，分詞形，三人称動詞などは原形に，また名詞の複数形も単数形に戻して辞書引きをした結果，辞書引きできないエントリーを147に減らした．辞書引きできた英語の565エントリーから2233のインドネシア語の訳語(語義)を得た．このインドネシア語の訳語の種類は以下のようになっている．

略語	:	4
複合語または熟語	:	153
句または文	:	381
単語	:	1690

●[インドネシア語→日本語]:(インドネシア語の辞書引き)

上記のインドネシア語の訳の中から，複合語または熟語からは39エントリー，単語の方から1193エントリー，合計1232エントリーが辞書引きに成功した．訳が略語と文の場合については辞書引き不可能である．辞書引きできなかった複合語の中には，畳語[4]などが含まれる．1690個の単語の方からは501エントリーが引けなかった．その原因は主にインドネシア語の単語の形態素にある．インドネシア語の単語の多くは，kata dasar(基語)に接辞を付加して構成される．接辞には，接尾辞と接頭辞があり，これらが独立的に基語に付加する他に，接尾辞と接頭辞の組合せで基語に付加することもできる．これらをすべて辞書に登録していないことが大きな要因になっていると考えられる．辞書引きでたインドネシア語の1232エントリーから3155エントリーの日本語の訳語が得られた．

●[日本語→英語]:(日本語の辞書引き)

日本語の3155エントリーのうち辞書引きできたのは2454個．辞書引きできなかった709エントリーには，以下のようなものが含まれる．

- 「～な」,「～の」形容詞
- 「にする，～た，～ている」過去型または状態を表す動詞
- 「お／ご～」敬語／丁寧語
- 「～に，よく～」副詞
- 「によって，に対して」などの関係詞
- 句または文
  これには，否定文なども含まれる

その他の問題として，本研究で使った「インドネシア語→日本語」辞書の編集上の問題も見られた．例えば,「soal」を「試験の問題」と訳しているが，この単語自体は「試験問題」という特殊化された語義はなく，むしろ単に「問題」と訳した方が自然で正しいと思われる[5]．
辞書引きできた英語の2454エントリーから英語の訳5960エントリーが得られた．この訳が元の英語のエントリーにループできなかったのは以下のような問題が挙げられる．まず，5960エントリーの中に2050エントリーが複合語，句または文になっている．ループの出発は英語の複合語を含まない単語であったので，この三分の一のエントリーがループ対象の枠外となる．これらの複合語には，例えば，

- 限定詞が付くもの
  例：空→ the sky, the air，全体→ the whole，人民→ the people, the citizens など
- 合成語，熟語問題
  例：同型→ the same type，同色などは，一般に辞書のエントリーとして登録されない．
- 受身またその結果の状態の表現
  例：びっくりする→ be surprised, be astonished，依る→ be grounded，因る→ be based など

のような問題がある．一方2910の中からループを構成しないであろうと考えられる問題(辞書引きできないもの)は例えば，

- 複数形の問題
  例：樹木→ trees，周辺→ environs,outskirts，情勢→ conditions, circumstances など
- 形容詞問題
  例：進歩→ advanced, 活発→ animated, 賑やか→ crowded など

[4]インドネシア語は，語を繰り返して複数を表現する．例えば，kursi(chair)→ kursi-kursi(chairs)

[5]他の例では,「sampah(ゴミ)」を「乾いたごみ」と訳したりしている

などが挙げられる．

2. 得られた語義項目について
得られた833の語義項目を調べた結果，689項目(約83%)が共用語義として使えるという感触を得た[6]．

• 語義の曖昧性について
一般に中間言語で表現される語の意味は、曖昧性を含まないものでなければならない．意味構造の一部である中間言語の語彙項目の意味もユニークでなければならない．しかし，なにを基準に曖昧でない語を定義すればよいであろうか．ある言語(または，複数言語)の中で曖昧な意味を含まないとされている語でも，別の言語からみると複数の語義が混在しているということもある．例えば，日本語の「私達」と英語の「We」とを曖昧さを含まない三人称複数という語義を持っている．しかし，この語義は，インドネシア語からみれば曖昧なものである．インドネシア語では「相手を含む」と「相手を含まない」三人称複数の語義が明示的に識別され，両方を包含するような概念を持っていないからである．したがって，語の意味の曖昧性を論議するには，どういう言語の中で行われているのかを明示的に示す必要がある．つまり，中間言語の「概念」も，すべての言語において普遍であるというのではなく，限定されている言語のなかで論議されなければならない．本研究の手法で得られた「語義項目」の概念も中で考慮されている言語の間で共用可能なものであって万国共通なものではない．

• 拡張性について
5.1章で述べたように，本研究の手法は，多言語間の共用可能な語義の設定のために考慮したものである．他の新しい言語の語を共通の語義を利用して導入(併合)することができる．

基本的にすでに使われたどれかの同じ辞書を使用すればよいが，関連意味の劣化を防ぐために，どれかのペアの辞書を固定して拡張を計った方がよいであろう．例えば，「英語→日本語」を基軸にして「英語→日本語→インドネシア語」，「英語→日本語→タイ語」を取り入れればよい．

また，本研究では，同じ語義を通して併合を行うことができるとしているが，理論的に以下のような問題がある．5.1節のFig.3とFig.4で得られた$[a_i, b_i, c_i]$および$[a_i, b_i, d_i]$の語義については，$S(c_i) \cap S(d_i) = \emptyset$の場合を考えると，$S([a_i, b_i, c_i]) \cap S([a_i, b_i, d_i]) = \emptyset$となり，両語義の併合ができないのではないかという疑問が湧く．しかし，一方では，

- $(S(a_i) \cap S(b_i)) \supseteq S([a_i, b_i, c_i])$
- $(S(a_i) \cap S(b_i)) \supseteq S([a_i, b_i, d_i])$

が成り立つことも事実であることから，もし，$S(a_i) \cap S(b_i)$を基に語義を抽出するのであれば，この併合には，問題はないという考え方もできる．しかし，今のところ，どこまでが併合でき，どういうものが併合できないのか定かではない．今後，併合の実験を行い，その結果を検証する必要がある．

[6]その一部を付録に示す

8 おわりに

本稿では，「各訳語は一つの語義を表す」の定義を基に共用語義項目の抽出アルゴリズムを設定した．これによって得られた語義の間にお互いに近い関係を持ち，併合またはグループ化できると見られる語義が存在する．例えば，

[answer menjawab 答える]
[answer menjawab 返事する]
[answer menyahut 返事する]

を一つの語義としてまとめることも考えられる．これをまとめるためにはいくつかの方法が利用できる．その一つは，既存の対訳辞書の情報の利用が考えられる．上の例の辞書は，

answer → 1. menjawab, menyahut
→ 2. membuka
→ 3. ...
menjawab → 1. 答える,返事する

さらに，「答える」と「返事する」が「answer」に戻っている．「menjawab」と「menyahut」および「答える」と「返事する」がそれぞれ「,」(コンマ)で仕切り，両単語がほぼ同じ意味または同義であると表している．このような同義性を持つものを利用すれば，上記の三つの共用語義項目を一つにまとめることができるであろう．しかし，第6章にも述べたが，辞書編集者によってこれらの同義性の認定が多少異なるので注意する必要がある．

次に，得られた「語義項目」の概念を体系化する必要がある．膨大な数の概念を体系化する際に，いかに類似する概念をまとめるかが大きな問題である．すでに提案されているシソーラスなどを参考にして手作業でまとめることもできるが，本稿で考えている語義と上記で述べた同義性を考慮に入れれば，機械的に類似する概念を集めることもできると考える．

前節で述べたループ(共用語義項目)の抽出ができなかった原因の大部分を占めているのは，句または文で表現される訳である．これは，意味解析でも導入しなければ機械的に解決できないであろう．しかし，このようなものも，多少の人間の手を借りれば簡単に解決できる．例えば，Fig.9に示すように，

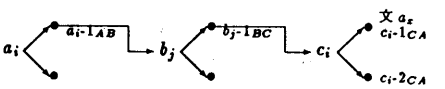

Fig.9

$a_i \rightarrow b_i \rightarrow c_i \rightarrow$ 文 a_z の場合は，$a_i \rightarrow b_i \rightarrow c_i$ が機械的に調べられるので，c_i-1_{CA} が a_i-1_{AB} と同義であることをわれわれが判断してやればこの語義が抽出可能となる．

本稿では，既存の対訳辞書に記述された語義に着目し，利用することによって多言語間における共用可能な語義を抽出する方法について検討し実験を行った．今回の実験では，三か国語しか使わなかったが，言語を増やして，またより多くの単語で実験し，検証する必要がある．しかし，現在のとこと，電子化された対訳辞書は少ないことがネックとなっている．実際，本実験で使用した辞書のデータはすべて手作業で入力したものである．

謝辞

本研究を行うにあたり，有益な示唆をいただいた徳永健伸助手に感謝致します．また，本研究に関して活発に討論をしていただいた田中研究室の諸氏に感謝致します．

参考文献

[1] 市山 俊治, 野村 直之. 多言語間機械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] 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.

[3] 特集「機械翻訳」. 座談会「機械翻訳における中間言語方式をめぐって」. 人工知能学会誌, Vol.4, No.6, 1989.

[4] 谷口 五郎. 標準インドネシア語・日本語辞典. 日本制作社, 1982.

[5] 小島 義郎, 竹林 滋. 日英辞書. 研究社, 1984.

[6] C. Douglas Lummis 著-中村直子訳. 最後のタヌキ晶文社セレクション, 1988.

[7] John M. Echols, Hassan Shadily. Kamus Inggeris Indonesia. P.T. Gramedia-Jakarta, 1979.

[8] Roy J. Byrd, N. Calzolari, M.S. Chodorow, J.L.Klavans, M.S.Neff, O.A.Rizk. Tools and Methods for Computational Lexicology. Computational Linguistics, Vol.13, No.3-4, 1987.

[9] Mary S. Neff, Branimir K. Boguraev. Dictionaries, Dictionary Grammars and Dictionary Entry Parsing. ACL, page 91-101, 1989.

[10] Karen Sparck Jones. Synonymy and Semantic Classification. Edinburgh University Press, 1986.

付録- 1

抽出されたループの一部

[ability,	kecakapan,	能力]
[about,	hampir,	ほとんど]
[about,	tentang,	大体]
[achieve,	mencapai,	達成する]
[adjust,	menyelesaikan,	まとめる]
[adjust,	menyesuaikan,	合わせる]
[all,	segala,	全て]
[all,	segala,	ことごとく]
[all,	semua,	全て]
[all,	semua,	全部]
[all,	semua,	ことごとく]
[all,	seluruh,	全て]
[all,	seluruh,	全体]
[always,	senantiasa,	絶えず]
[always,	senantiasa,	いつも]
[answer,	menjawab,	答える]
[answer,	menjawab,	返事する]
[answer,	menyahut,	返事する]
[answer,	membuka,	解く]
[appear,	kelihatan,	見える]
[appear,	menghadap,	出頭する]
[appear,	muncul,	現れる]
[appear,	terbit,	出る]
[appear,	nampak,	見える]
[appear,	tampak,	見える]
[appear,	tampak,	現れる]
[appear,	kelihatan,	見える]
[appear,	muncul,	現れる]
[argue,	memperdebatkan,	論争する]
[bad,	buruk,	悪い]
[bad,	jelek,	悪い]
[bar,	menghalangi,	妨げる]
[bar,	menghambat,	妨げる]
[baseball,	baseball,	野球]
[begin,	mulai,	始める]
[big,	besar,	大きい]
[bomb,	membom,	爆撃する]
[bring,	membawa,	もたらす]
[build,	membangun,	建設する]
[build,	mendirikan,	建てる]
[build,	mendirikan,	建築する]
[bury,	memendam,	埋める]
[but,	tetapi,	しかし]
[but,	tetapi,	それなのに]
[but,	kecuali,	以外]
[buy,	membeli,	買う]
[call,	memanggil,	呼ぶ]
[call,	menyebut,	名付ける]
[call,	menamakan,	いう]