日-韓機械翻訳における連語パターンを用いた変換手法

朴 哲済[†1] 李 鐘赫[†2]
李 根培[†3] 覧 捷彦[†4]

本稿では，直接翻訳方式を利用した日-韓機械翻訳システムにおいて，連語パターンによる変換手法を提案する．筆者らは，日本語での連語パターンを作成し，変換規則として用いることにより語彙の曖昧性の問題を解決した．また，意味素性をベースにした選択制限として，連語パターンの中で最もパターン類似度が高いものを選択する類似度のスコア計算の方法を提案する．我々は，このような変換方法を用いた日-韓機械翻訳システムを鉄鋼関係の特許文書翻訳のために構築し，その性能を評価法によって分析した．その結果，翻訳成功率は約94%であるし，意味的曖昧性を持つ用言と助詞に対して各々95.62%，94.96%の語彙の曖昧性を解決し，この手法が十分有効であることを確認した．

Collocation-based Transfer Method in Japanese-Korean Machine Translation

CHUL-JAE PARK,[†1] JONG-HYEOK LEE,[†2] GEUNBAE LEE[†3] and KATSUHIKO KAKEHI[†4]

In this paper, under the direct approach, we describe a lexical transfer method based on a collocational analysis in Japanese-Korean machine translation system. To resolve the lexical ambiguity and thus to make a correct choice of translation equivalents, we propose a collocation-based lexical transfer method in which surrounding contexts can be expressed by collocational patterns to specify selectional restrictions between words. Since we view selectional restrictions based on semantic features as preferences rather than all-or-nothing decision, a ranking scheme of collocational patterns is also proposed to pick the one that violates the fewest selectional restrictions. Finally, the evaluation results of translating patent materials on iron and steel subjects show how effectively the proposed system can work with the high translation rate of about 94%. For Japanese predicates and particles with multiple meaning, 95.62% and 94.96% of which can be transferred into Korean correctly.

1. はじめに

日本語と韓国語は言語系統上アルタイ語に属し，文法体系が似ている．特に文節単位では両言語の語順がほとんど一致していることから，文節を単位として両言語を1対1に対応して翻訳を行っても相当なレベルの翻訳結果が得られる．したがって日-韓翻訳システムの場合，その文法的類似性から実用化のためには，直接翻訳方式を採用している[1),8)~13)]．

しかし，語彙の曖昧性（lexical ambiguity）によって，両言語間で1対1に対応できないときも多い．したがって，日-韓機械翻訳システムが実用化されるためには，語彙の曖昧性の問題を解決する手法の提案が必要とされている[1),8)~11)]．

本稿では，直接翻訳方式を利用して構築した日-韓機械翻訳システムにおいて，連語パターン（collocation pattern）による変換手法について述べる．特に，意味的制約を持つ連語パターンと，対応される変換規則によって日本語形態素の曖昧性を解決し，妥当な訳語に変換する方法を提案する．

本論文の構成は次のようになっている．まず2章で，既存の研究における問題点について述べる．3章では，システム概要について述べる．4章では，本論文で提

†1 早稲田大学メディアネットワークセンター/現代情報技術株式会社情報技術研究所
Media Network Center, Waseda University/R&D Center for Information Technology, HIT
†2 浦項工科大学電子計算学科/浦項工科大学情報通信研究所
Department of Computer Science and Engineering POSTECH/Information Research Laboratories, POSTECH
†3 浦項工科大学電子計算学科
Department of Computer Science and Engineering, POSTECH
†4 早稲田大学情報学科
Department of Information and Computer Science, Waseda University

案する連語パターンによる変換処理，および，類似度計算によって最適訳語を選択する手法について説明する．そして，5章で，本手法による実験結果を提示し，まとめを行う．

2. 既存の研究での問題点

直接翻訳方式を利用して機械翻訳システムを構築するとき，直面する問題点として，語彙の曖昧性と，部分的な語順の調整問題があげられる[3]．この問題点の中で高品質の翻訳結果を得るためには，語彙の曖昧性の問題を解決することが重要である．

語彙の曖昧性の問題は大きくカテゴリの曖昧性，多義性，および，多訳性の3つに分けられる[3]．この中で，カテゴリの曖昧性は形態素解析の段階でヒューリスティックスを用いてほとんど解決できる．反面，多義性や多訳性の問題は構文と意味情報を基に解決しなければならない．語彙の曖昧性の問題は日–韓機械翻訳においても一番大きな問題点として残っている．

次は日–韓翻訳における曖昧性問題の一例である．

例） 彼は 金/父母 を失う．

例文の「失う」の場合，その意味が韓国語に翻訳されるとき，“**ilhta** (lose)” と “**yeuyta** (be bereaved)” の2つに分けられる．“**yeuyta**” の意味として使う場合は,「を」の前に人間の意味が含まれている体言がくる．このように同じ単語が別の意味として使えることを曖昧性と呼ぶ．

現在まで発表されている商用日–韓機械翻訳システム，たとえば，パルパル[12]やJ・ソウル[13]では語彙の曖昧性の問題を解決していないが，日–韓両言語間の類似性を生かして，ある程度の翻訳精度を得ている．しかし，高品質の翻訳結果を得るためには，両言語間の用言と助詞の使い方の違い等による語彙の曖昧性の解決が問題点として残っている．そして，直接翻訳方式で不合格と判定された文は，多訳性によるものを除き，ほとんどが用言と助詞の扱いに関するものである[9]．

用言における曖昧性の問題を解決する方法として文献8)では，格文法をベースにした単文の意味解析手法を提案している．これは，動詞と名詞の共起関係の選択制限を，意味素性の代わりに語と語の共起関係テーブルによって表現している．

助詞の曖昧性の問題を解決する方法としてATLAS-I[11]では，助詞の前に現れる単語の連接関係を用いている．また，文献9)では，各助詞ごとに前に接続する単語の意味や文法情報を考慮し訳語を登録した単語翻訳テーブルを用いている．しかし，その使い方が多様である助詞をより精密に処理するためには，名詞の意味素性と名詞間の意味関係を明確にする必要がある．

以上のような問題点を解決する方法として，我々は連語パターン（collocation pattern）に基づいた変換手法を提案する．連語パターンでは語と語の連接関係をすべての単語間（動詞–名詞，名詞–助詞など）にも選択制限として用いる．我々のシステムでは，連語パターンを人が簡単に理解できるように記号で定義し，それを直接辞書に記述して曖昧性の問題を解決している．また，用言の意味素性をもっと細かく分けることにより，連語パターンでは特殊な意味素性を持つ単語の記述も可能である．また，意味素性をベースにした選択制限として連語パターンの中で一番パターン類似度が高いものを選択する類似度計算の方法を提案する．言語構造が似ている日–韓翻訳においては，連語パターンで処理するだけで，本格的な構文解析や意味解析を行わず，実用的には高品質の翻訳が可能である．

3. COBALT-J/Kのシステム概要

我々は，日–韓機械翻訳システム COBALT-J/K（COllocation-BAsed Language Translator from Japanese to Korean）を構築した．

本システムでの形態素解析は，CYK法と日本語形態素間での接続可能性の検査によって行う．この方法で可能なすべての解析結果を得たのち，形態素解析段階での点数を求める．我々は品詞の接続テーブルを強い接続と弱い接続の2つに分けて，隣接する2つの形態素の品詞が強い接続なら高い点数を，弱い接続の場合は低い点数を与えた．隣接するすべての形態素に対して点数を求めた後，それを全部合わせて1つのパスに対する形態素解析段階での点数にした．これをヒューリスティックスとして利用して，解析結果を優先順位に従って整列する．この整列された形態素解析結果は変換部に渡され，いろいろな意味を持つ用言と助詞の訳語が決められる．

変換の前処理として助詞の格と用言との共起情報（co-occurrence information）を利用し，複文の場合は文分割を行い，分割された文を1つの処理単位として変換を行う．変換は日本語形態素の各訳語を選択する連語パターンと形態素解析結果をマッチングし，1つの訳語を決める．この連語パターンは，例外事例を持つことにより慣用句の処理を簡単にする．このときシソーラスの階層構造（thesaurus hierarchy）によるマッチングも行う．この訳語選択の結果の中で一番選好度が高いものが生成部に渡される．

生成部ではまず，1つに決められた変換結果から述部の意味素を様相類意味素テーブル（MFOLT: Modal-

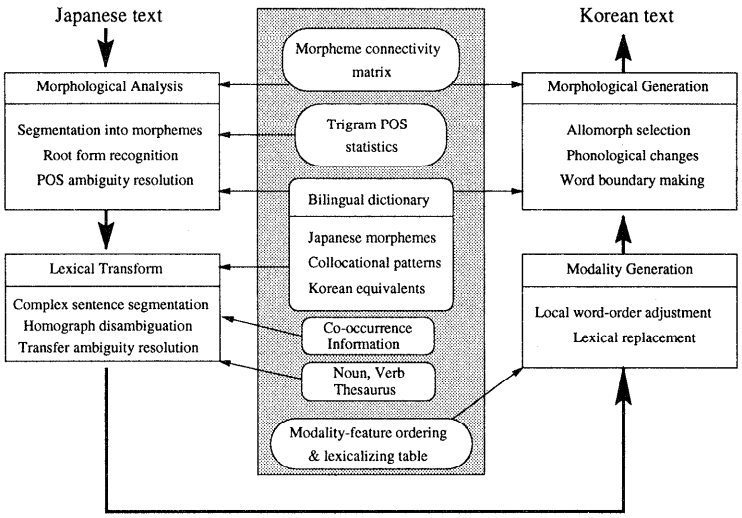


図1 システム概要
Fig. 1 System overview.

ity Feature Ordering and Lexicalizing Table）に活性化して，否定語，使役語等の処理を行う[10]．次に，MFOLTの順番関係に従って韓国語の述部を生成する．語尾等の異形態処理，音韻縮訳および不規則処理はこのとき行う．述部の生成が終わったら，韓国語の接続情報を用いて助詞の異形態処理を行い，全体の韓国語文を生成する．図1にシステムの概要を示す．

4. 連語パターンによる変換処理

変換部では，形態素解析のところから優先順位別にそなえた形態素解析結果（以下，パス（path）と呼ぶ）を入力として，各形態素間の意味を決めて該当する韓国語の訳語に変える．ただし，入力文が複文の場合は前処理として分割を行う．この分割された文を処理単位として，連語パターンを用いて各形態素を韓国語の訳語に変換する．

4.1 連語パターン

日本語の用言と助詞の意味決定において，一緒に使われた単語は重要な役割をなす．各単語の意味によって前後に使われる単語は構文的，意味的特徴を持つ．我々はこれを連語パターンとして規則化し，曖昧性の問題を解決する方法として用いた．連語パターンは文法的，意味的制約を与える連語関係項目（syntagmatic term）と，それらのあいだの関係を表す連語関係演算子（syntagmatic operator），そして括弧（bracket）で表現する．連語関係項目は日本語形態素そのものや，文の中でそれらの形態素の前後に現れる名詞，動詞のパターンを表す項目で構成した．項目間の関係は連語関係演算子によって表現する．括弧は複数のものの選択を表すオプション（option）や，用言において必要な格を表すセット（set）で構成した．用言に必要な格が複数のときはセットの中で順序別に並べて書く．我々はこれらの連語パターンを次のような記号を使って表現した．

◆連語関係項目（syntagmatic term）
- $ ： それ自身
- N ： 意味的制約を持つ名詞
- V ： 意味，構文的制約を持つ動詞
- 形態素：連語パターンで現れる日本語形態素

◆連語関係演算子（syntagmatic operator）
- * ： 連語関係項目の順序関係を表現，隣接に使用
- + ： 連語関係項目の順序関係を表現，隣接しなくてもよい
- - ： 否定的日本語形態素，現れてはいけない形態素を表現

◆括弧（bracket）
- [ ] ： オプション（“/”により区別）
- { } ： セット，用言において必要な格を表す（“,”により区別）

これらの記号を使ってシステムで用いた用言と助詞の連語パターンの一部を以下に示す．

(1) 用言の連語パターンの例：

- ころがす［転がす］

```
nemettulita {[Nperson]*[が/は],
             [Nperson]*を}+$
kwullita    {[Nperson]*[が/は],
             [Nmaterial]*を}+$
```

wuncenhata {車*を}+$

- うしなう［失う］

ilhta {[Nperson/Ngroup]*[が/は], [Nmaterial/Nconfidence/ Njob]*を}+$

nohchita {[Nperson/Ngroup]*[が/は], [Ntime]*を}+$

yeuyta {[Nperson]*[が/は], [Njob]*で, [Nperson]*を}+$

(2) 助詞の連語パターンの例：

- が

i/ka N*$+V

ul/lul N*$+[Vpossible/Vhope]

- から

pwuthe N*$+V

ulo [Nmaterial/Nelement]* $+[Vmake/Vachieve]

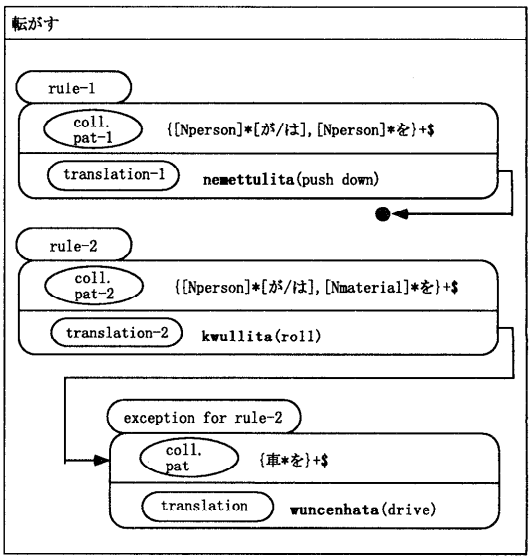


図2 「転がす」の変換規則
Fig. 2 Lexical transfer rule.

連語パターンと入力文がマッチングできるとき，意味的制約を持つ名詞と動詞はシソーラス階層構造の中で意味的類似度が計算され，日本語形態素の正確な訳語が決められる．"-" が付いている否定の日本語形態素は，入力文に存在してはならないものを表す．これは日本語形態素の前に使って文の中でその形態素と一緒に現れてはいけない形態素を表現する否定的共起関係を表現する．たとえば，日本語用言「組む」を意味する韓国語動詞のうち，"**phyensenghata** (organize)" は目的語を必要とするが，"**hanphayka toyta** (confederate)" は目的語が不要である．文の中で目的格が現れたら "**hanphayka toyta**" の意味になる可能性は少ない．

変換規則は連語パターン，対訳語および，例外事例で構成した．用言と助詞は変換規則を持ってそれを利用し，文の中での意味を決める．名詞の場合は変換規則を持たず，用言と助詞の意味解決段階で名詞の意味制約を利用して意味を決める．例外事例は連語パターンでマッチングしたとき，どちらか1つの規則に対応できるが，マッチングに成功した規則の意味ではないときのためである．例外事例の利用により変換規則を単純化し，規則に基づく機械翻訳（Rule-based Machine Translation）方法での欠点を補うことができた．変換規則は変換辞書に登録されている．**図2** に日本語用言「転がす」の変換規則を示す．用言「転がす」は韓国語 "**nemettulita** (push down)"，"**kwullita** (roll)"，"**wuncenhata** (drive)" の3つの訳語として翻訳可能である．ただし，"**wuncenhata**" は，"**kwullita**" と構文的意味的制約は同じであるが，「車」と一緒に熟語のように使われるため，"**kwullita**" の例外として処理する．

4.2 複文の分割

我々は連語パターンを利用し，解析部で構文や意味解析をしなくても語彙の曖昧性を解消することができた．しかし，連語パターンは，単文の中でのマッチングのみに対応することができる．複文に対しては，特に「+」演算での対象を誤認識する可能性がある．そこで，

(1) 複文は前処理して，単文に分解する．
(2) 個々の単文に連語パターンマッチングを適用する．
(3) 得られた結果を統合する．

という処置を施すことにした．

4.2.1 共起情報の適用

我々は複文の分割に助詞の格と用言との共起情報を利用した．共起情報はコーパスでよく現れる表現が最も自然に受け入れられることに基づいて，それを選択することで構文的曖昧性を解消する方法である．

まず，主格や目的格等の必須格と，用言との構文的曖昧性解消のため，ある必須格と特定の用言が係り受け関係になる確率 $P(\text{case}, \text{predicate})$ を定義した．全体コーパスの大きさ N の必須格の頻度 (Freq(case))，用言の頻度 (Freq(predicate))，必須格と用言が隣接に現れる頻度 (Freq(case, predicate)) 等を用いて式 (1) で計算した．式 (1) は共起する格と用言との相互関連性（word association）を客観的に算出するため，情

表1 必須格と用言との共起情報の例
Table 1 Co-occurrence information of object case with two predicates.

N	1,053
Freq(目的格)	7,210
Freq(従う)	356
Freq(目的格, 従う)	4
Freq(分析する)	173
Freq(目的格, 分析する)	73
P(目的格, 従う)	0.001641
P(目的格, 分析する)	0.061627

N：全体コーパスの大きさ（文書の総数）
Freq：全体文書での頻度

表2 文節の分類
Table 2 Classification of BUNSETSU (Japanese word).

No.	文節名称	属する種類
1	用言	動詞，形容詞，形容動詞
2	必須格結合体言	主格助詞，目的格助詞と結合した体言，助詞省略体言
3	連用格結合体言	連用格助詞と結合した体言
4	副詞	副詞
5	副助詞結合体言	副助詞と結合した体言
6	接続助詞結合体言	接続助詞と結合した体言
7	連体語	連体格助詞と結合した体言，連体詞
8	接続副詞	接続副詞
9	連体形用言	連体形語尾が結合した用言

報理論的概念である相互情報（mutual information）を利用したものである．

$$P(\text{case}, \text{predicate}) = \frac{N * \text{Freq}(\text{case}, \text{predicate})}{\text{Freq}(\text{case}) * \text{Freq}(\text{predicate})} \quad (1)$$

例文「**韓国語を依存文法に従って分析する**過程を示す.」に式(1)を適用すると，“目的格”と2つの用言「従う，分析する」について，それぞれ P(目的格, 従う)，P(目的格, 分析する) を求めて，一番高い確率を持つ用言を選択し,「韓国語を」との係り受け関係を設定する．**表1** にコーパスから求めた結果を示す．表1で P(目的格, 分析する) が最も高い確率である．したがって,「韓国語を」と「分析する」の係り受け関係が設定される．

4.2.2 分割アルゴリズム

日-韓翻訳では両言語の類似性を用いるため他の言語間での翻訳に比べて相当な部分の省略が可能である．しかし，複文の場合は，主語と述語の関係において一定の順序関係を用いずに現れることが多い．我々はこの問題を解決するため通常の構文解析を用いず，以下の簡単なアルゴリズムにより分割を行った．

本論文で提案する複文分割アルゴリズムは，2つの段階に分けられる．最初の段階では文を内包文単位に分割する．次の段階では，分割された内包文の中で連体化された内包文と被修飾内包文との係り受け関係を設定し，1つの文にする．連体化された内包文は，被修飾内包文との概念の関連性が高いため1つの内包文とした．また，本論文では文節の種類を**表2** のように定義した．

(1) 複文を内包文単位に分割する．アルゴリズムは以下のように行った．
  (a) 文の最後の文節から依存文法に基づいて解析を行う．各文節について(b)～(i)を反復する．
  (b) 現在の文節が用言であると，それを係られる語として設定する．
  (c) 現在の文節が必須格結合体言であると，それの右側にある文節のみ係り受け関係を探す．
  (d) (c)で係り受け関係がないと，現在の文節の右側にある用言とそれぞれの係り受け関係を探す．このとき，現在文節と係り受け関係がある用言を1つ選択する．
  (e) (d)で現在の文節が2つ以上の用言と係り受け関係になって構文的曖昧性があるときは，助詞の格と用言との共起情報を用いて，候補用言の中で1つを選択する．
  (f) 現在の文節が連用格結合体言であると，共起情報を用いて現在文節の右側にある用言の中で1つを選択し，係り受け関係を設定する．
  (g) 現在の文節が副詞または副助詞結合体言であると，右側で一番近くにある用言のみ係り受け関係を設定する．副助詞は未知格（unknown case）として主格，目的格，連用格等，すべての格としての機能が可能である．既存の構文解析では未知格のときは，可能なすべての係り受け関係を設定している．本アルゴリズムでは，可能な限りの構文的曖昧性を除去するために，係り受け関係が設定される可能性が一番高いもの，すなわち，右側で一番近い用言のみ係り受け関係を設定した．
  (h) 現在の文節が接続助詞結合体言，または，連体語の場合は右側で一番近い体言のみ係り受け関係を設定する．
  (i) 現在の文節が接続副詞であると，文頭にそれが現れると，最後の文節のみ係り受け関係を設定する．文の中で現れると，右側に隣接する文節のみ係り受け関係を設定する．
(2) 最後に2番目の段階では，前段階で分割された内包文の中で連体化された内包文と，被修飾内包文との係り受け関係を設定し，1つの文にした．アルゴリズムは以下のようである．

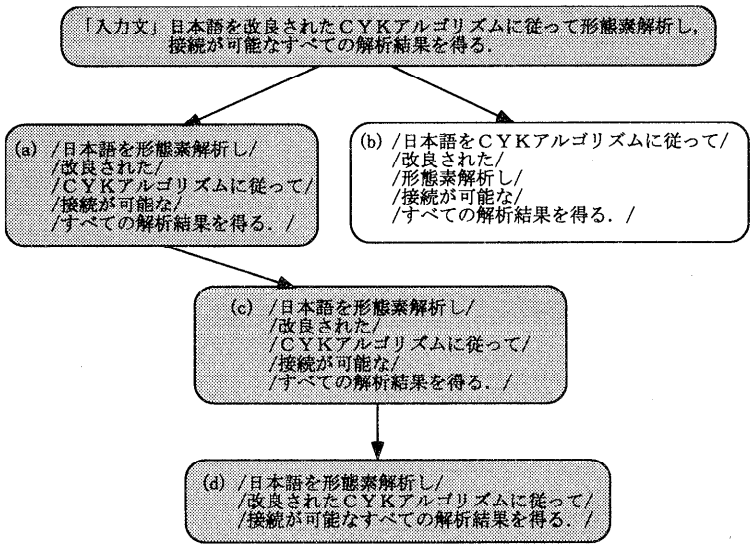


図3 複文分割の例
Fig. 3 Example of sentence segmentation.

(a) 連体化された内包文の係られる語である連体形用言の文節に対して，右側にある文節の中で係り受け関係がある一番近い文節と，係り受け関係を設定する．すなわち，概念的に関連性が高い連体化された内包文と被修飾内包文を1つの内包文にする．したがって，連体化された内包文の係られる語である連体形用言と，被修飾内包文との係り受け関係を設定する．

図3の例を見ると，入力文を内包文単位に分割する過程で図3(a)，図3(b)のように「日本語を」が,「従う」と「解析する」の2つの用言と係り受け関係の設定が可能である構文的曖昧性が発生する．これを解消するため“目的格”と「従う」，“目的格”と「解析する」の共起情報を利用する．実験結果〈P(目的格,従う)$=0.006133$〉,〈P(目的格,解析する)$=0.109535$〉になって，図3(c)で「日本語を」が「解析する」と係り受け関係になる．また，連体化された内包文と，被修飾内包文との係り受け関係の設定を通して，1つの内包文を作り，図3(d)の3つの文として分割される．

4.3 変換処理

変換部は，各形態素の意味を決定するところと，文のパスを選択するところの2段階として構成した．

形態素の意味を決める段階（曖昧性の問題を解決する段階）では，連語パターンを利用して日本語文の各形態素の意味を決定する．変換規則の連語パターンと形態素解析のところから優先順位別に出力された形態素解析結果をマッチング（2つが一致するか，しないかの比較処理）し，pSIM（pattern similarity）という類似度（similarity）点数を計算する．連語パターンで意味制約を持つ連語関係項目と，入力文の形態素間の意味的類似度を表すsSIM（semantic similarity）を計算する．そして，その点数を全部合わせて文全体の連語パターンに対するpSIMを計算した．

形態素の意味決定段階では，一番高い点数のpSIMを持つ連語パターンが選択され，そのpSIMの値が形態素変換の点数になる．そして選択された連語パターンに例外事例がある場合は，例外事例チェックのため，もう一度マッチングを行う．例外事例のpSIMが変換規則にある連語パターンのpSIM点数より高いときは，例外事例の意味に従って訳語が決まる．

パス決定段階では，1つのパスに対して変換段階での点数と形態素解析段階の点数を合わせて，最終結果として一番高い点数を持つ1つのパスを選択する．このパスは日本語の各形態素に対して意味が決められたもので，韓国語生成部の入力になる．

4.4 最適訳語の選択手法

4.4.1 シソーラスの階層構造

シソーラスは入力文の形態素と連語パターンの意味制約間の類似度を計算するときに使用する．動詞シソーラスは構文的，意味的属性によって40種に分類した[14]．名詞のシソーラスは意味によって1,000種に区分した[15]．シソーラスは階層構造になっている．名詞シソーラスは4レベルの階層構造で，動詞は2〜5レベルの階層構造である．**図4**に名詞と動詞のシソーラスの階層構造を示す．

noun						
nature	action	feeling	human	society	...	things
astronomy	movement	affection	person	region		material
weather	visit	thought	friendship	group		medicine
plant	observe	study	status	institution		food
animal	statement	intention	role	custom		building
.	.	.	.	.		.
.	.	.	.	.		.

verb				
action	feeling	representation	...	dynamic idea
give	emotion	agree		both relation
change	active emotion	relative rep.		.
move	expression	absolute rep.		.
.	.	.		.
.	.	.		.
.	.	.		interaction
.	.	.		あう

図4 シソーラスの階層構造
Fig. 4 Thesaurus hierarchy.

4.4.2 マッチングと変換点数計算

変換部での処理結果は入力文の各形態素に対して文の中での意味が決められた1つのパスである．日本語形態素の連語パターンの中で一番高いパターン類似度を持つものを選択することで各形態素の意味が決められる．変換の最初段階では，語彙の曖昧性を解決するため，各形態素について一番よくマッチングする変換規則を選択する．

連語パターンの各項目 P_i と入力文の各形態素 I_j 間の意味的類似度を表す $\mathrm{sSIM}(P_i, I_j)$ を定義した（**図5**参照）．類似度点数の計算は構文・意味的要素によって2つの方法でマッチングを行う．1つは完全マッチング，もう1つは近似マッチングである．

完全マッチングは連語パターンの構文・意味的要素が日本語の表層形態素のときに行う．マッチング点数は0または1になる．たとえば，$\mathrm{sSIM}(\text{が}, \text{が}) = 1$ になる．$\mathrm{sSIM}(\text{が}, \text{を}) = 0$ になる．否定的日本語形態素に対するマッチングは，マッチングに成功すると，1の代わりにペナルティとして -1 の値になる．

構文・意味的要素が名詞と用言の場合は近似マッチングを行う．意味的制約を持つ名詞と用言の意味的類似度は，シソーラス階層構造を利用して入力文の該当形態素と比較する．シソーラス階層構造では，親ノードをたくさん共有すればもっと類似である．意味的類似度を計算するために MSCA[5]（Most Specific Common Abstraction）を用いる．$(n+1)$ 階層のシソーラス構造で，階層の下から k 番目の階層にあるノードは (k/n) の値を持つ．

文献2)での意味的長さ計算方法は，事例に基づく機械翻訳（Example-based Machine Translation）方式での事例と入力を比較するために作成されたことから，入力と事例の意味的属性は必ず一番下の階層に存在する．本論文では，事例の一般化（generalized）された形態である連語パターンを用いるため，比較の対象 P_i がシソーラス階層構造で，より上位レベル（more abstract）にあるので，文献2)の方法はそのまま利用できない．本論文では，MSCA のパターンと入力の意味的属性レベルも一緒に考慮し，図5の式によって意味的類似度 $\mathrm{sSIM}(P_i, I_j)$ を求める．この式での $level(\mathrm{MSCA}(P_i, I_j))$ は，P_i と I_j の MSCA ノードの深さを表す．また，**図6** (a) のように入力文の形態素 I_j が連語関係項目 P_i の子孫ノードのときには，図6 (b) または (c) よりもっと類似する．これを区別するため "is-a penalty" を適用した．入力形態素が連語関係項目の子孫ノードの場合（図6 (a)）には，"is-a penalty" が1，それ以外（図6 (b), (c)）は0.5である．$\mathrm{sSIM}(P_i, I_j)$ は，連語関係項目 P_i と入力文の形態素 I_j がシソーラス階層構造の一番下レベルのときには，文献2)の方法と同じになる．

次に連語パターン P と入力文 I のマッチング類似度として $\mathrm{pSIM}(P, I)$ を定義した．$\mathrm{pSIM}(P, I)$ は完全にマッチングされたときの値で割り算することにより正規化した．この式での $S_{m,n}$ は，連語パターン P と入力文 I との最適マッチング点数を示す．$S_{m,n}$ は，連語パターンの否定的日本語形態素により負の値を持つ可能性がある．したがって，$\max(0, S_{m,n})$ を求めてマッチング点数が0以下になることを防ぐ．$S_{m,n}$ は，ダイナミックプログラミング技法により求める[6]．ただし，連語関係演算子 "*" と "+" の順序関係を反映するため，1つの連語パターンを "+" を基準に分

$$\mathrm{sSIM}(P_i, I_j) = \begin{cases} \dfrac{2\times\mathrm{level}(\mathrm{MSCA}(P_i,I_j))}{\mathrm{level}(P_i)+\mathrm{level}(I_j)} \times \text{is-a penalty} & (a) \\ 1 \text{ (if matched) or } 0 \text{ (if not matched)} & (b) \\ -1 \text{ (if matched) or } 0 \text{ (if not matched)} & (c) \end{cases}$$

(a) P_i is a semantic attribute
(b) P_i is a surface morpheme
(c) P_i is a negative surface morpheme

$$S_{i,j} = \begin{cases} 0 & \text{if } i=0 \text{ or } j=0 \\ \max \begin{cases} S_{i,j-1} \\ S_{i-1,j} \\ S_{i-1,j-1} + \mathrm{sSIM}(P_i, I_j) \end{cases} & \text{otherwise} \end{cases}$$

$$\mathrm{pSIM}(P, I) = \frac{\max(0, S_{m,n})}{\sum_i \text{Perfect-Matching-Score-of-}P_i}$$

$$\mathrm{mTS}(\mathrm{morph}) = \max_{P \text{ in morph}}(\mathrm{pSIM}(P, I))$$

$$\mathrm{sTS}(\mathrm{path}) = \frac{\sum_{\text{morph in path}} \mathrm{mTS}(\mathrm{morph})}{\#\text{ of morph with coll. patt.}}$$

図5 類似度の点数計算
Fig. 5 Formula of computing similarity.

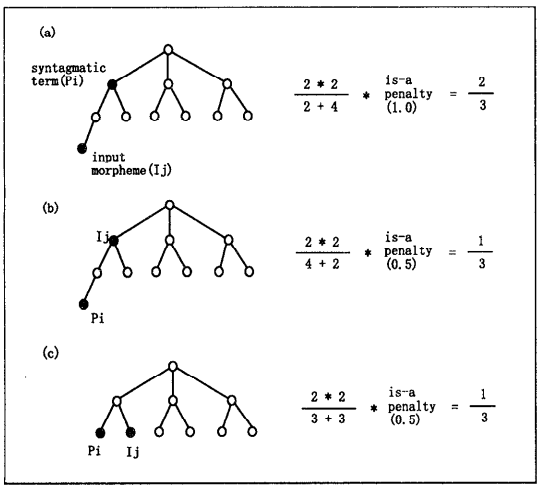


図6 Is-a ペナルティ
Fig. 6 Is-a penalty.

割して点数を計算する．また，連語パターンで隣接に表すことを表現する“*”演算子は1つの単位としてマッチングを行った．

たとえば，入力文$\langle I_1, I_2, I_3, I_4, I_5, I_6, I_7\rangle$に対して，連語パターン$\langle P_1 * P_2, P_3 * P_4 + \$\rangle$をマッチングする．このとき\$はI_6と完全マッチングであるとする．まず，“+”演算子を基準に連語パターンを分けて，入力文$\langle I_1, I_2, I_3, I_4, I_5\rangle$と連語パターン$\langle P_1 * P_2, P_3 * P_4\rangle$とのマッチングを行う．この場合$\langle P_1 * P_2\rangle$を1つの処理単位として，最初に$\langle I_1, I_2\rangle$とのマッチング点数を計算する．次は$\langle I_2, I_3\rangle$に進みながら$\langle I_4, I_5\rangle$までのマッチング点数を計算し，一番高い点数とマッチングされたとする．次は，そのマッチングされた2つの形態素を除いて同じ方法で$\langle P_3 * P_4\rangle$と入力文のマッチング点数を計算する．この方法により連語関係演算子“*”の隣接関係や，セットの中での非順番関係の表現が可能であった．

多様な単語に変換できる入力形態素は，各訳語に対して変換規則を持つ．入力形態素の意味を決め正しい訳語に変換するため，一番高いマッチング点数$\mathrm{pSIM}(P, I)$を持つ変換規則を選択した．この場合選択された変換規則のマッチング点数が図5で定義した形態素変換点数（図5のmTS(morph)）になる．

最後に，パス内部にあるすべての形態素について変換点数を合わせた後，その値を正規化してパスに関する変換点数（図5のsTS(path)）を求める．

4.5 変換処理の例

日本語文“彼が転がした車は，ブレーキが故障していた.”に対する変換処理の例を説明する．形態素解析の結果**図7**のように各形態素に分割される．図7に示すように，日本語文の各形態素に対する韓国語訳語が対応されている．また曖昧性を持つ多義語の場合は，その変換規則と訳語が対応されている．次に複文の分割を行う．連体修飾節が含まれている複文は，連体修飾節の後ろに出る名詞までを1つの変換単位として分割する．例文では「彼が転がした車は」と「ブレーキが故障していた.」の2つの文に分割する．これが変換部の入力になる．

変換部では各形態素の意味を決める．まず,「彼」の意味は1つであるから韓国語“**ku**”に変換する．次は，

	彼が転がした車は，ブレーキが故障していた.	
彼	[Nperson]	(ku)
が	rule-1 : N*$+V	(i/ka)
	rule-2 : N*$+[Vpossible/Vhope]	(ul/lul)
転が	rule-1 : [Nperson]*[が/は],[Nperson]*を+$	(nemettulita)
	rule-2 : [Nperson]*[が/は],[Nmaterial]*を+$	(kwullita)
	exception for rule-2 : 車*を+$	(wuncenhata)
し	$	()
た	$	sPAST
車	[Nmaterial]	(cha)
は	$	(un/nun)
,	$	(,)
ブレーキ	$	(puleyikhu)
が	rule-1 : N*$+V	(i/ka)
	rule-2 : N*$+[Vpossible/Vhope]	(ul/lul)
故障	$	(kocang)
し	$	( )
て	$	( )
い	$	(iss)
た	$	sPAST
.	$	(.)

図7 変換部の入力
Fig. 7 Input of the transfer phase.

彼が転がした車は，ブレーキが故障していた.

rule-1: $N * \$ + V$"$i/ka$"
sSIM(N, 彼) = 1/3×(is-a penalty:0.5) = 1/6
sSIM(V, 転が) = 1/3×(is-a penalty:0.5) = 1/6
sSIM(\$, が) = 1
pSIM = (1/6 + 1/6 + 1)/3 = 0.44

rule-2 : $N * \$ + [Vpossible, Vhope]$"$ul/lul$"
sSIM(N, 彼) = 1/3×(is-a penalty:0.5) = 1/6
sSIM(V[*possible*, *hope*], 転が) = 1/4×(is-a penalty:0.5) = 1/8
sSIM(\$, が) = 1
pSIM = (1/6 + 1/8 + 1)/3 = 0.43

彼	が	転が	し	た	車	は	,
(ku)	(i/ka)	(wuncenha)	()	sPAST & tDECL	(cha)	(un/nun)	(,)
ブレーキ	が	故障	し	て	い	た	.
(puleyikhu)	(i/ka)	(kocang)	()	()	(iss)	sPAST & tDECL	(.)

図8 変換処理の例
Fig. 8 Transfer example.

「が」を変換する．日本語形態素「が」は，曖昧性を持つ助詞で2つの変換規則を持っている．1つは韓国語形態素 "**i/ka**（主格助詞）" に対応し，もう1つは韓国語形態素 "**ul/lul**（目的格助詞）" に対応する（**図8**参照）．変換規則の連語パターンと入力文を，シソーラスの階層構造を用いて比較する．動詞「転が」が可能または感情を表す動詞でないことから図8のルール1が選択される．

次は,「転が」を変換する．用言「転が」は「車」と一緒に熟語のように使われているため，韓国語形態素 "**kwullita**" の例外として処理する．したがって,「転が」の意味は，例外規則により韓国語 "**wuncenhata**" として決められる．

次は,「車」を変換する.「車」は名詞で，名詞の場合は変換規則を持っておらず，動詞や助詞の意味を決める過程の中で，選択された変換規則の意味制約によって決める．したがって,「車」の意味は，助詞「が」の変換結果から韓国語 "**cha**" として決められる．以上のような過程を経て日本語形態素の韓国語訳語が決められ，その結果が韓国語生成部の入力になる．

5. 実験結果および考察

機械翻訳において評価すべき項目は，翻訳言語に依存しない汎用的な技術と，それら言語対に依存するものが多く存在する[16]．ここでは，全体の翻訳文書の質を言語対に依存するものを中心に評価する．本稿での実験環境は，SPARCStation 10（90 MIPS）であり，C言語で実装した．これは鉄鋼関係の特許文書を翻訳

表3 システムの翻訳結果
Table 3 Translation result.

分類	全体個数	荷重	成功個数	成功率（%）	スコア
形態素	8,319	0.5	7,831	94.13	47.07
単語	5,871	0.25	5,483	93.39	23.35
熟語	664	0.25	612	92.17	23.04
総合評価					93.46

するためのシステムとして開発した．

定量的評価は，NHKのニュース6カ月分のコーパスから49文，鉄鋼関係の特許文書約13万件のコーパスから223文を対象とした．評価に用いた文は平均30.58形態素であり，総文節数は2,343で，1文あたり8.61の文節であった．まず，日本語が理解できる3人で，辞書を参照して翻訳した結果を検討し「正解」を作った．我々は，これを「正解」としてシステムの翻訳結果を照合し成功率を求めた．システムによる翻訳結果を**表3**に示す．評価は，形態素，単語，熟語の3つに分類した．分類方法は入力文中の形態素がすべて形態素レベルで評価され，その中で単語であるものがさらに単語レベルで評価され，その中で熟語がさらに評価されるような方法である．これにより形態素が合わさって単語または熟語になって訳が別になるものについてのきめが細かいチェックになった．表3における「荷重」とは，総合評価のために分類別に与えた値[7]である．全体を1と考え，形態素の荷重を0.5にし，他のものを0.25にした．スコアとは，各分類内での翻訳成功率に荷重を掛け算した値である．このスコアを合わせてシステムの総合評価点数にした．我々は個々のレベル（形態素，単語，熟語）での成功率から翻訳システム全体を評価する方法として，この荷重により総合評価点数を求めた．システムは，8,319日本語形態素中7,831形態素について正しく韓国語に翻訳した（成功率94.13%），我々が付加した荷重による総合評価点数は93.46であった．これはシステムをトレーニングする前の翻訳結果としてはかなり高い．他のシステムとの翻訳性能の比較は，難しいのが現状である．しかし，翻訳性能は2つの観点からチェックすることができる．

1) 実際の文章でどの程度の翻訳率なのか？

2) 翻訳するのに難しい文章をテストコパースとして使って，どの程度の翻訳率なのか？

現在1)の方法では，我々のシステムと他のシステム[12),13)]はほぼ同じ精度であった．2)の方法では，我々のシステムの精度が最も高い．実際に2つのシステムを比較し，本システムの問題点を見つけるため比較評価を行った．その結果，翻訳誤りの数において文献13)のシステムは，我々のシステムより約2倍であった[4)]．

失敗原因の多くは，辞書の未整備にあった．我々はこれからシステムトレーニングを行い，総合評価点数を約97までアップする予定である．

表4 用言の曖昧性解消の評価
Table 4 Evaluation results of lexical ambiguation for predicates.

分類	出現個数	成功個数	成功率（%）
受ける	14	14	100
応じる	3	1	33.33
行う	10	10	100
合わせる	8	8	100
選ぶ	8	8	100
備える	20	20	100
示す	4	4	100
優れる	6	6	100
絞る	3	2	66.67
加える	4	4	100
込む	5	5	100
取り付く	15	13	86.67
対する	5	4	80.00
含む	24	24	100
施す	9	9	100
設ける	33	29	87.88
用いる	11	11	100
依る	69	68	98.55
合計	251	240	95.62

表5 助詞の曖昧性解消の評価
Table 5 Evaluation results of lexical ambiguation for particles.

分類	出現個数	成功個数	成功率（%）
の	391	385	98.47
に	244	227	93.03
が	126	118	93.65
を	364	357	98.08
は	61	58	95.08
で	61	51	83.61
と	189	168	88.89
から	41	39	95.12
も	14	12	85.71
とも	9	9	100
や	9	9	100
合計	1,509	1,433	94.96

表4，表5に，用言と助詞で多く発生する語彙の曖昧性解消の実験結果を示す．

表4，表5は，実験文から出現個数が3つ以上で，曖昧性を持っている用言と助詞の翻訳結果を調査した結果である．システムは，251個の用言形態素中240形態素について曖昧性が解消されて正しく韓国語に翻訳した（成功率95.62%）．助詞は1,509形態素中1,433形態素について曖昧性解消に成功した（成功率94.96%）．特に成功率が低い用言「応じる，絞る，対する」や，助詞「で，と，も」について失敗原因を分

析した．失敗原因は次の3つのいずれかが大半を占めた．

(1) 日本語の接続情報の辞書登録が間違っていて，接続不可能の判定になったもの

(2) 韓国語の接続情報の辞書登録が間違っていて，接続不可能の判定になったもの

(3) 連語パターンが間違っていて，誤訳になったもの

いずれの原因についても，接続情報や連語パターンを修正することにより解決できるものであった．これ以外の約3%を占める形態素解析による失敗や意味情報を持たなくては解決不可能な意味選択の問題もあった．我々は，13万件の特許文書を翻訳しながらシステムトレーニングをして，失敗原因を解決する予定である．

6. おわりに

本稿では，大規模で実用的な日-韓機械翻訳システム開発において，変換を中心に述べた．日本語文での連語パターンを変換規則として用いることにより曖昧性の問題を解決した．連語パターンと入力文とのマッチングは構文的，意味的に下位区分した名詞と，用言のシソーラス階層構造を利用した．変換規則に例外事例をおくことにより規則をより簡単にすることができた．

今後の課題として，日本語の慣用句が1つの韓国語単語に表現できるときの処理があげられる．たとえば，「手を加える」は韓国語 "**soncilhata** (handle with care)" に，「腹が下る」は "**selsahata** (have loose bowels)" になって3個の日本語単語が1個の韓国語単語に翻訳される．この場合のため例外事例の処理を改善し，単語を合わせて，1つの単語を作る処理が必要である．

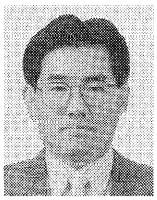
本研究で開発した日-韓翻訳システムは現在鉄鋼関係特許文書（約13万件）翻訳システムとして稼動し，検証を行っている．現在，形態素解析辞書には7万5千語が入り，連語パターンは全体用言の約30%程度が入っている．最初，特許文書を翻訳するために開発を始め，それに必要な専門用語辞書には，約1万語程度が入っている．形態素解析辞書と連語パターンは，一般的な辞典の文法情報を持って6名の辞書作業の専門研究員により構築を行っている．我々は現在このプロジェクトを2年の計画で大規模な辞書構築（鉄鋼関係単語：3万，一般単語：8万）とPCレベルでも使えるように，システムのポーティングを行っている．

謝辞 研究討論いただいた，浦項工科大学知識および言語工学研究室の諸氏に感謝いたします．

参考文献

1) Kim, E.J., Lee, J.-H., Lee, G.B.: A Lexical Transfer Model Using Extended Collocational Patterns in COBALT J/K, *Proc. ICCPOL'94*, pp.461–466 (1994).
2) Sumita, E.: Experiments and Prospects of Example-based Machine Translation, *Proc. 29th ACL*, pp.185–192 (1991).
3) Hutchins, W.J. and Somers, H.L.: *An Introduction to Machine Translation*, Academic Press, London (1992).
4) Jeong, J.-R., Kim, J.-I., Moon, K.-H., Lee, J.-H., Lee, G.B.: Evaluation of COBALT-J/K Japanese to Korean Machine Translation System, *Proc. 8th HKIP*, pp.338–345 (1996).
5) Kolodner, J. and Riesbeck, C.: Case-based Reasoning, *Tutorial Textbook of 11th IJCAI* (1989).
6) Baase, S.: *Computer Algorithms – Introduction to Design and Analysis*, Addison-Wesley (1988).
7) Shiwen, Y.: Automatic Evaluation of Output Quality for Machine Translation Systems, *Machine Translation*, Vol.8, No.1, pp.117–126 (1993).
8) 李 義東，中嶋正之，安居院猛：語と語の関係を用いた意味解析による日韓単文機械翻訳システム，電子情報通信学会論文誌，Vol.J72-D-I, No.10, pp.1689–1695 (1989).
9) 金 泰錫，浦 昭二：日韓機械翻訳における意味接続関係を用いた韓国語の生成方法，情報処理学会論文誌，Vol.33, No.12, pp.1578–1588 (1992).
10) 朴 哲済，文 敬姫，郭 鍾根，李 鐘赫，李 根培：連語パターンによる日韓機械翻訳システムの構築とその評価，情報処理学会自然言語処理研究会，NL-109-2 (1995).
11) 渡辺正己：ATLAS-Iにおける日韓機械翻訳システムの一考察，第38回情報処理学会全国大会論文集 (1989).
12) C.P.U.: 日韓自動翻訳システムパルパル：使用者ガイド（PC版），C.P.U., 東京 (1989).
13) 高電社：日韓翻訳システムJ・ソウル操作マニュアル，高電社，大阪 (1991).
14) 寺村秀夫：日本語の構文と意味I，法文社 (1988).
15) 大野 普，浜西正人：類語新辞典，角川書店，東京 (1982).
16) 中岩浩巳，森本康嗣，松平正樹，成田真澄，野村浩郷：JEIDA機械翻訳システム評価基準（開発者編），情報処理学会自然言語処理研究会，NL-96-10 (1993).

(平成8年5月10日受付)
(平成9年1月10日採録)

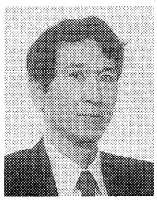


朴　哲済（正会員）

1986年韓国延世大学校数学科卒業．1991年早稲田大学大学院理工学研究科情報科学専攻修士課程修了．1995年同大学院博士課程研究指導認定退学．1986年POSCOシステム開発室．1987年Japan Knowledge Industry．1995年韓国浦項工科大学情報通信研究所委嘱研究員．1995年より早稲田大学Media Network Center特別研究員兼現代情報技術（株）情報技術研究所責任研究員．人工知能，自然言語処理，機械翻訳等の研究に従事．日本情報処理学会，言語処理学会，韓国情報科学会，情報処理学会等の会員．

Jong-Hyeok Lee received the B.S. in mathematics from the Seoul National University, Korea in 1980, and the M.S. and Ph.D. degrees in computer science from KAIST (Korea Advanced Institute of Science and Technology) in 1982 and 1988, respectively. After graduation, he spent a year as a postdoctoral researcher at KAIST, participating in the national project for English-to-Korean MT system. In the years 1989–1991, he joined the C&C Information Laboratory of NEC, Japan, working on the design and development of a Korean synthesizer in the PIVOT project for a multilingual MT system. Since 1991, he has been working for the Department of Computer Science and Engineering at POSTECH (Pohang Univ. of Science and Technology), Korea, and currently he is an associate professor. His research interests include the core technology for Korean language processing and its applications to machine translation, information retrieval, and text summarization. He is a member of the editorial board of Korea Information Science Society Review, and also a member of the AAAI, ACL, ALLC, CLCS, IPSJ, and JSAI.

Geunbae Lee was born in Chyeongna, Korea, in 1961. He received his B.S. and M.S. degrees in computer engineering from the Seoul National University, Seoul, Korea in 1984 and 1986, respectively and the Ph.D. degree in computer science from the University of California at Los Angeles in 1991. Currently, he is an assistant professor in the Computer Science and Engineering Department, Pohang University of Science and Technology, Pohang, Korea. His current research interests are morphological, syntactic and semantic analysis of Korean in the field of natural language processing, and its applications to speech recognition, speech synthesis and intelligent agent-based information retrieval. During the past 5 years, he has published more than 50 papers on Korean natural language processing in international journals and conference proceedings. He is both a member of the steering committee of the special interest group on artificial intelligence and on human computer interaction of Korean information science society (KISS). He is also a member of the ACM, AAAI, IEEE, and the Oriental Language Computer Society.



筧　捷彦（正会員）

1968年東京大学工学部計数工学科卒業．1970年同大学院修士課程修了．同大学助手，立教大学講師・助教授を経て，1986年より早稲田大学理工学部情報学科教授．プログラミング言語の設計・実現・環境構成等の研究に従事．国際委員会委員（IFIP TC2）企画調査会（SC22）．情報処理学会，日本数学会，応用数理学会，日本ソフトウェア科学会，ACM等の会員．