原言語例文とその対訳に関する質問に基づく意味構造変換ルールの獲得

秋葉 泰弘
NTT コミュニケーション科学基礎研究所
京都府相楽郡精華町光台 2-4
akiba@cslab.kecl.ntt.co.jp

白井 諭
ATR 音声翻訳通信研究所
京都府相楽郡精華町光台 2-2
shirai@itl.atr.co.jp

1 はじめに

トランスファー方式の機械翻訳システムが必要とする意味構造変換ルールは，辞書作成の専門家がその条件付けを行なう事により，作成されている．本稿では，専門家が行なうこの作成作業を如何に支援するかと言う問題を取り上げる．実用的なシステムを構築するためには，変換ルールを大量に作成する必要があり[7]，作成を支援する手段が必要であった．

変換ルールの作成を支援する従来法には，帰納学習によるアプローチ[6; 4; 2; 3]がある．このアプローチは翻訳用例を沢山必要とするため，このアプローチにより作成可能な変換ルールは限られていた．特に，テキスト中の出現頻度が低い表現に対応する変換ルールは，大量に作成される必要であるにも関わらず，このアプローチでは作成が難しかった．

そこで本稿では，支援システムがそれ自身と専門家のやり取りを通して変換ルールの条件を獲得する支援手法を提案する．システムは，シソーラスを利用して例文を生成し，その例文が翻訳例として正しいか否かを専門家に質問し，専門家の応答により適切な条件を探索する．

提案手法を評価するために，専門家が作成したルールの条件を提案手法で獲得出来るかを試したところ，十分な質問が可能な時には提案手法はその正しい条件を獲得出来た．

以下，2節では本稿で取り上げるタスクを明確にする．3節で提案手法について述べる．4節で実験結果を示し，その結果を議論する．5節でまとめる．

2 獲得タスク

トランスファー方式の機械翻訳システムでは，動詞と名詞の共起関係に着目し，原言語の共起関係パタンとそれに対応する目的言語の共起関係パタンの対からなる意味構造変換ルール (パタン対と呼ぶ) が中心的な役割を果たす．例えば，NTT で研究開発している日英機械翻訳システム，ALT-J/E[1]は，このようなパタン対を備えた，トランスファー方式のシステムである．

図1に ALT-J/Eのパタン対の例を示す．IF 部は日本語パタンで，THEN 部はそれに対応する英語パタンである．このパタン対は，「入力文の日本語動詞が'焼く' の場合，'焼く' の主語 (N1) が⟨人⟩の語義を持ち，'焼く' の目的語 (N2) が⟨パン⟩または⟨菓子⟩の語義を持てば，日本語動詞 '焼く' に対する適訳は英語動詞 'bake' で，'bake' の主語は N1 の英語訳，'bake' の目的語は N2 の英語訳である．」と示唆している．⟨人⟩，⟨パン⟩，⟨菓子⟩が入っている各スロットには，意味カテゴリと呼ばれる名詞の語義が入る．ALT-J/Eは，意味カテゴリを約2700個持っており，これらの意味カテゴリは図2に示す様な最大の深さ12段の階層構造 (シソーラスと呼ぶ) を成す．ALT-J/Eの日本語辞書中の各名詞 (全部で約40万語) は，いずれかの意味カテゴリ (1つとは限らない) を語義に持つ．

パタン対を作成する際，専門家は各スロットに入れるべき適切な (抽象的過ぎず，特殊過ぎない) 意味カテゴリを探索する．以下本節では，専門家が入力すべき意味カテゴリを特徴付け，本稿で取り上げるタスクを明確にする．そのために，まず幾つかの記号と概念を導入する．

今，専門家があるパタン対のあるスロット (対象スロットと呼ぶ) に入れるべき意味カテゴリを探索しているとする．そして，(1) そのパタン対の各スロットにマッチする原言語名詞 (具体語と呼ぶ) の組，(2) 各具体語の意味カテゴリ，を思い浮かべているとする．対象スロットに対する具体語の意味カテゴリがシソーラスのリーフであれば，専門家が対象スロットに入れるべき意味カテゴリは対象スロットに対する具体語の意味カテゴリ自身又はその上位概念であるから，入れるべき意味カテゴリの候補はシソーラスの頂点と対象スロットの具体語の意味カテゴリを結ぶパス上にある．

そこで，次のリスト (候補リストと呼ぶ) を考える．

$$(C_1, C_2, \cdots C_i, \cdots C_M)$$

ここで，i番目の要素は，上述のパス上のルートから数えてi番目の意味カテゴリーである．従って M は，対象スロットに対する具体語の意味カテゴリの深さである．また，シソーラス上にある C_i の兄弟を $S(i,j), (1 \leq j \leq N_i)$ と記述する．ただし，N_i はシソーラス上にある C_i の兄弟数である．

例えば，対象スロットに対する具体語が '太郎' で，その意味カテゴリが⟨男⟩であれば，図2に従って，(⟨名詞⟩, ⟨具体⟩, ⟨主体⟩, ⟨人⟩, ⟨人間⟩, ⟨人間 < 生物学的特徴 >⟩, ⟨人間 < 男女 >⟩, ⟨男⟩) が候補リストとなり，$M = 8$ となる．また，$C_3 =$ ⟨主体⟩

IF	J-Verb	= '焼く'
	N_1 (J-Subj)	≡ ⟨人⟩
	N_2 (J-Obj)	≡ ⟨パン⟩ or ⟨菓子⟩
THEN	E-Subj	= N_1
	E-Verb	= 'bake'
	E-Obj	= N_2

図 1: パタン対の例 (日本語動詞 '焼く' に対する)

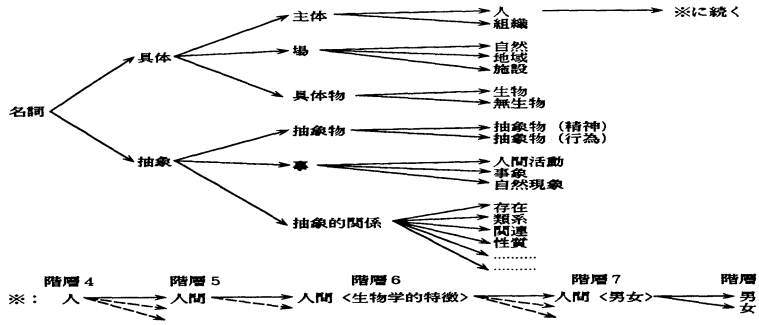

図 2: ALT-J/Eにおける意味シソーラスの一部

で，$N_3 = 2$，$S(3,1) = \langle$場$\rangle$，$S(3,2) = \langle$具体物$\rangle$ となる．なお，リーフには子ノードはないが，説明の簡単のために，$N_{M+1} = 1$，$S(M+1,1) = C_M$ と定義する．従って，先の例の場合，$N_9 = 1$ で，$S(9,1) = C_8$ となる．

C_i に対する有効事例率：(1) 前述の具体語の組を利用して生成される文，(2) 対象スロットに対する具体語を $S(i+1,j)$，$(1 \leq j \leq N_{i+1})$ 配下のリーフを語義に持つ各原言語名詞で置き換えて生成される文を考える．例えば，$S(3,1)$，$S(3,2)$ 配下の部分木は，図3において $S(3,1)$，$S(3,2)$ 配下の左右の三角形に対応するので，底辺が丁度リーフに対応する．置き換えに使われる名詞は，いずれかの底辺上のリーフをその語義に持つ名詞である．この時，ある文は原言語文として受理されるが，ある文は受理されない．置き換えによって生成される文の内，原言語文として受理される文の割合を，意味カテゴリ C_i に対する**有効事例率**と呼ぶ．

C_i に対する成立事例率：また，上述の原言語文として受理される文の内，ある文は獲得対象パタン対による翻訳が適切であり，ある文は適切でない．受理される文の内，獲得対象パタン対による翻訳が適切な文の割合を，意味カテゴリ C_i に対する**成立事例率**と呼ぶ．

更に，対象スロットに指定する意味カテゴリの最低限持っていて欲しい有効事例率，成立事例率を，順に**最低有効事例率**，**最低成立事例率**と呼ぶ．また，有効事例率，成立事例率が共に最低有効事例率，最低成立事例率以上である意味カテゴリを**成立カテゴリ**と呼び，成立カテゴリでない意味カテゴリを**不成立カテゴリ**と呼ぶ．

上述の用語を用いると，専門家が入力するべき意味カテゴリはシソーラス上最上位の成立カテゴリと特徴付けられる．なぜならば，適切な意味カテゴリは，対象スロットにマッチすべき名詞をなるべく多く覆い，マッチすべきでない名詞の覆いを最小限度に押える必要があるからである．

そこで本稿で取り上げる獲得タスクを，次の様に定める．あるパタン対のあるスロットに入れるべき意味カテゴリを獲得する時，

入力：(1) 獲得対象パタン対のスケルトン
(2) IF 部の各スロットの条件にマッチする具体語の組
(3) (2) の各具体語の意味カテゴリ

出力：シソーラスの頂点と対象スロットに対する具体語の意味カテゴリを結ぶパス上にある，最上位の成立カテゴリ

例えば，図1に示すパタン対の J-Subj スロットに指定すべき意味カテゴリを獲得したい場合，スケルトンとして，図1の J-Subj スロット及び J-Obj スロットを空欄にした雛型が入力されてる．IF 部のスロットの具体語として，J-Subj スロットに対し‘太郎’，その意味カテゴリとして ⟨男⟩，また，J-Obj スロットに対し‘アップルパイ’，その意味カテゴリとして，⟨ケーキ⟩が例えば入力される．出力される意味カテゴリは，図2のルート⟨名詞⟩とリーフ ⟨男⟩を結ぶパス上の 8 個の意味カテゴリの内，最上位の成立カテゴリである．

3 提案手法

本稿では，前節で述べた獲得タスクを解く支援システムを通して，専門家を支援する手法を提案する．支援システムは，専門家とのやり取りを通して，探索リストの中から上述の最上位の成立カテゴリを探索する．システムは後述の 3 つの探索アプローチのいずれかで探索を行なうが．その際各探索ポイントで専門家に質問をする．まずその質問戦略を説明する．

3.1 質問

探索ポイントが C_i であるとき，システムは以下の質問を生成し，専門家から回答を入力してもらう．

質問生成

システムは，各 $S(i+1,j)$，$(1 \leq j \leq N_{i+1})$ について，シソーラス上の $S(i+1,j)$ 配下のリーフに位置する意味カテゴリを語義に持つ名詞を利用して，前節の有効事例率の定義で説明した通りに文を生成する．例えば，初期入力時が前節の例の場合，C_2，⟨具体⟩が探索ポイントであれば，$N_3 = 2$，$S(3,1) = \langle$場$\rangle$，$S(3,2) = \langle$具体物$\rangle$ なので，⟨場⟩と ⟨具体物⟩各々について，文を生成する．⟨土地⟩が ⟨場⟩配下のリーフであって，‘高地’がこれを語義とする名詞であれば，‘高地’を使った文‘高地 がアップルパイを焼く’を生成する．生成される他の文は，下線部の名詞が置き変わった文である．

ここで問題となるのは，システムが生成する文の量である．リーフの意味カテゴリを語義に持つ名詞は非常に多いため，システムが生成可能な全ての文を調べ尽くす事は現実的でない．そこで各リーフについて，そのリーフを語義に持つ特徴的な名詞を，そのリーフの代表語としてシステム側で予め準備する．その上でシステムは，置き換えに利用する名詞を決められた個数だけ，システ

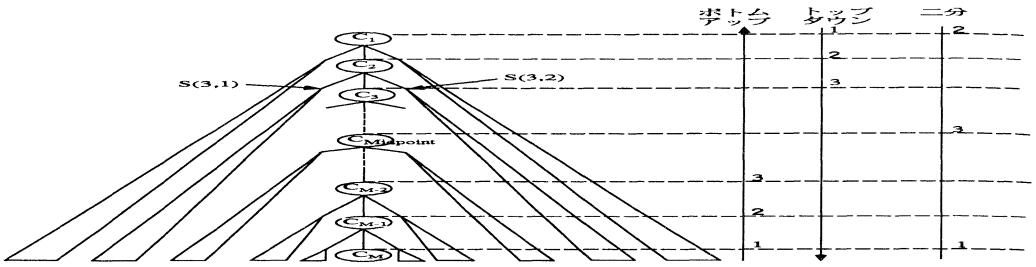

図 3: 提案手法の探索イメージ

ムが注目中のリーフの代表語中からランダムに抽出する事とする．

質問提示とその回答

システムは生成した質問を専門家に提示する．専門家は，提示された生成文が原言語文として受理されるかどうか，もしそうなら獲得対象のパタン対による翻訳が適切かどうかを判断し，その回答を全てシステムに入力する．成立事例率や有効事例率が下限に近く微妙な場合には，専門家は必要に応じてより多くの事例を提示してくれるようシステムに要求する．

3.2 提案手法における探索アプローチ

以下，システムが取る3つの探索アプローチを説明する．これらの違いは候補リスト中のどの意味カテゴリから上述の質問戦略を適用するかである．最初の2つボトムアップアプローチとトップダウンアプローチが，learner search あり，最後の二分アプローチが binary search である．図3に各探索アプローチのイメージを示す．

ボトムアップアプローチでは，システムはパス上の最下位の意味カテゴリ (候補リストの末尾) から順に，即ち C_M，C_{M-1}，$\cdots$ の順に上述の質問戦略を適用する．不成立カテゴリに到達した時点で探索を終了し，不成立カテゴリに到達する前に到達した最後の成立カテゴリを最終出力とする．

トップダウンアプローチでは，システムはパス上の最上位の意味カテゴリ (候補リストの頭) から順に，即ち C_1，C_2，$\cdots$ の順に上述の質問戦略を適用する．成立カテゴリに到達した時点で探索を終了し，その成立カテゴリを最終出力とする．

二分アプローチでは，システムはまず，リーフ，頂点の順に質問戦略を適用する．継いで，探索すべき候補を丁度二分する意味カテゴリに，もし二分する意味カテゴリが無ければ大体二分する下よりの意味カテゴリに，上述の質問戦略を適用する．質問戦略を適用された意味カテゴリが成立カテゴリであるか否かにより，候補リストを binary search 流に更新し，次回はこの更新された候補リストを基に探索を繰り返す．従って，候補リスト (初期候補リストを除いて) の第一要素は常にが不成立カテゴリ，最終要素が成立カテゴリとなる．探索を繰り返すと，最後候補リストの大きさが2となる．この時システムは最終要素を最終出力とする．

4 実験

以下の2点について提案手法を実験的に評価した．

- 専門家が正しいとする，適切な意味カテゴリを提案手法により獲得するには，最低有効事例率，最低成立事例率を幾つに設定すれば，各探索アプローチや獲得対象の意味カテゴリに依存せず，獲得性能が高くなるか？
- 獲得性能が高くなる値に最低有効事例率，最低成立事例率を固定した場合，システムは各探索アプローチによりどの位の精度で専門家が正しいとする適切な意味カテゴリを獲得出来るか？また，もし獲得出来ないならば，正解から上下にどれだけずれた意味カテゴリーがどの位の割合で獲得されるのか？

実験で正解とした意味カテゴリは，人手で作成されたALT-J/Eのパタン対[5]に指定された以下のもので，日本語動詞'読む'のパタン対中の⟨主体⟩，⟨精神⟩，⟨抽象物(精神)⟩，日本語動詞'選ぶ'のパタン対中の ⟨長⟩，日本語動詞'入賞する'のパタン対中の⟨式・行事等⟩[1]である．これらが適切な指定であることは事前に確認した．また，実験に利用したシソーラスは先に説明した ALT-J/Eのシソーラスで，代表語はALT-J/Eのシソーラス作成時に参考にした名詞である．提案手法のシステムの操作は，ALT-J/Eのパタン対を作成している熟練作業者2名が当たった．

なお，最低有効事例率と最低成立事例率の値としては，前者には1%，10%，20%，30%を，後者には100%，90%，80%を選んだ．有効事例率は生成文の生起確率に対応するため，その値を余り大きく指定するとシステムに無視される生成文が多くなり問題である．また，最低成立事例率にあまり小さな値を指定すると，獲得される意味カテゴリがカバーすべきでない名詞を多くカバーするようになり，問題が多い．

4.1 最低有効事例率の比較

実験方法

上述した一番目の点を評価するために，最低有効事例率が幾つである時に，提案手法により適切な意味カテゴリが獲得出来る条件付き確率が最大になるかを調べた．具体的には，最低有効事例率を固定し，最低成立事例率，獲得する意味カテゴリを色々変えて各探索アプローチで提案手法により1度つづ獲得試行を行ない，各最低有効事例率毎に，獲得される意味カテゴリが正解に一致する割合を集計した．質問のための文は各 $S(i+1,j)$ につき10個生成させた．

[1]この意味カテゴリの方が日本語動詞'入賞する'のパタン対に[5]で指定されている意味カテゴリより適切なため，この意味カテゴリを正解とした．

表 1: 各探索アプローチの平均質問数

アプローチ	ボトムアップ	トップダウン	二分
平均質問数	116.0	62.7	72.0

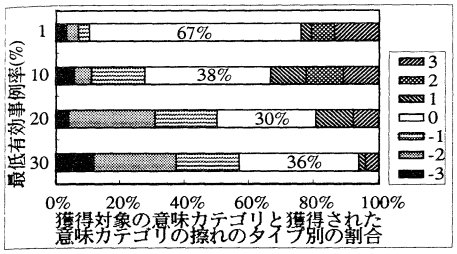

図 4: 最低有効事例率の比較

実験結果

実験結果を図4に示す．凡例の数字は，獲得された意味カテゴリの，正解からのずれ及びその方向を示す．例えば，1，-1は，各々正解から上に1，下に1ずれた事を示す．百分率は各事象(一致した，1つ上にずれた等の)の全試行に対する割合である．最低有効事例率が1%の時，正解に一致する条件つき確率が67%と一番高い．

4.2 最低成立事例率の比較

実験方法

上述の結果を基に最低有効事例率を1%に固定した場合，最低成立事例率が幾つである時に，提案手法により適切な意味カテゴリが獲得出来る条件付き確率が最大となるかを調べた．質問のための文は先と同様に生成させた．

実験結果

その結果を図5に示す．凡例の数字と百分率の意味は先と同じである．最低成立事例率が90%または80%の時，正解に一致する条件付き確率が69%と一番高い．さらに誤差一段まで見ると，最低成立事例率が90%の時，誤差が一段以内である率は85%と一番高い．

以上より，一番目の点に関しては，最低有効事例率と最低成立事例率の値が各々1%と90%の時に，提案手法により適切な意味カテゴリが得られる可能性が高い事が分かった．

4.3 探索アプローチの違いによる性能比較

実験方法

最低有効事例率を1%，最低成立事例率を90%に固定した場合，提案手法が獲得する意味カテゴリが正解の意味カテゴリと一致するか，また一致しないならどれだけずれたかを，上述の5つ意味カテゴリについて各探索アプローチ毎に実験した．実験に際し，各獲得対象カテゴリの獲得を各探索アプローチで3度行なった．質問のための文はこれまでと同様に生成させた．

実験結果

図6と表1に，各探索アプローチによる提案手法の正解率と平均質問数を示す．二分アプローチは，正解率66%，平均質問数62回と他の探索アプローチより性能，質問数共に勝る．

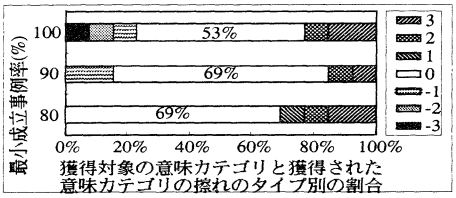

図 5: 最低成立事例率の比較

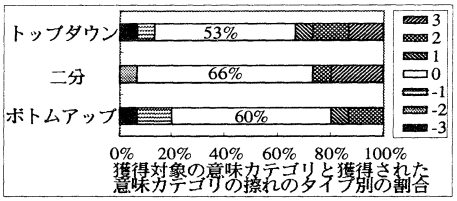

図 6: 各探索アプローチ毎の性能比較

5 おわりに

本稿では，支援システムが辞書作成の専門家とのやり取りを通して変換ルールの条件を獲得すると言う，変換ルールの生成支援手法を提案した．提案手法を利用すれば，翻訳システムの語彙体系を熟知していない作業者でも支援システムに聞かれる質問に答えるだけで適切な意味カテゴリの指定が可能となった．今後は提案手法の実験評価を更に進める．また，生成文の作り方（代表語の選び方など）を工夫し，より少ない質問数で有効事例率と成立事例率を精度よく推定する方法を開発する．

謝辞

本論文をまとめるに当たり種々御協力頂いた，ＮＴＴアドバンステクノロジ(株)の関係各位並びにＮＴＴコミュニケーション科学基礎研究所の中岩浩巳氏に感謝致します．

参考文献

[1] Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. Toward an MT system without pre-editing–effects of new methods in ALT-J/E. In *Proc. of MT Summit-3*, pp. 101–106, 1991.

[2] Hang Lie and Naoki Abe. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, Vol. 24, pp. 71–88, 1998.

[3] Hideki Tanaka. Decision tree learning algorithm with structured application to verbal case frame acquisition. In *Proc. of Coling-96*, pp. 943–948, 1996.

[4] アルモアリムフセイン, 秋葉泰弘, 金田重郎. 木構造属性を許容する決定木学習. 人工知能学会誌, Vol. 12, No. 3, pp. 421–429, 1997.

[5] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 東京, 1997.

[6] 秋葉泰弘, 石井恵, アルモアリムフセイン, 金田 重郎. 人手作成ルールと事例に基づく英語動詞選択ルールの学習. 自然言語処理, Vol. 3, No. 3, pp. 53–68, 1996.

[7] 白井諭, 池原悟, 横尾昭男, 井上浩子. 日英機械翻訳に必要な結合価パタン対の数とその収集方法. 情処研報 NL-110-7, pp. 43–50, 1995.