自然言語処理 135－22
（2000．1．28）

翻訳ルールの意味制約と目的言語共起情報を併用した訳語選択

麻野間 直樹　中岩 浩巳
NTT コミュニケーション科学基礎研究所
{asanoma,nakaiwa}@cslab.kecl.ntt.co.jp

ルールベース型機械翻訳システム（RBMT）の訳語選択処理において，用言と格要素により訳語選択条件が記述された翻訳ルール（結合価パタン対）の意味制約と目的言語側の単語共起情報によって訳語候補の絞り込みを行う訳語選択手法を提案する．パタン対の格要素名詞の訳語を選択する評価実験では，パタン対と共起情報を併用することにより，+6%の品質向上率が得られた．また依存関係がある単語対を品詞タグ付きコーパスから効率よく取得方法を提案しその有効性を示す．

Selecting alternative translations using semantic restrictions in valency dictionaries and word co-occurrence extracted from target language corpora

Naoki Asanoma　Hiromi Nakaiwa
NTT Communication Science Laboratories

This paper proposes a method to improve the translation selection in rule-based machine translation systems (RBMT) to select relevant translation candidates using semantic restrictions in valency dictionaries and word co-occurrences extracted from the target language corpora. According to a preliminary evaluation of the method, the translation quality improves by +6% as a whole. Moreover, we propose an efficient method to acquire word co-occurrences with their dependent relations without using syntactic parser and show its effectiveness.

1 はじめに

ルールベース型機械翻訳システム（RBMT）では一般に，ルールや辞書に記述された目的言語の訳語候補から，訳語選択条件の優先順位と変換ルールの制約等によって，訳文の中で用いる訳語を選択する．その後，入力文中の各複数単語に対して，選ばれた訳語を並べて合成（要素合成）し訳文を生成する．このようにして得られる翻訳結果は，訳語を生成する際の単語並びとしての適切性を十分に考慮していない場合が多い．

このような適切な訳語を選択する問題に対しては，語義曖昧性解消の問題（WSD）として，数多くの解決法が提案されている．その代表的なものは，テキストコーパスから獲得した統計的知識を利用する方法である．この手法は利用するコーパスの種類によって次の二つに分けられる．

- 二言語コーパスを利用［Brown91，Doi93，Rapp95］
- 目的言語コーパスを利用［Nomiyama91，Dagan94，Tanaka96，Kikui98］

後者で用いる目的言語コーパスは，多量に流通し入手が容易であるため，後者の二言語コーパスより統計的知識を獲得しやすい．

一方，テキストコーパスを用いて語義曖昧性を解決する際に，関係の強い単語組を得るために，依存関係のある単語対を活用することがよく行われる．この単語間依存情報を取得するには，人手で付与したコーパスを用いるか，あるいは依存情報未付与のコーパスを構文解析するかどちらかの手段がとられていた．

しかし，一般に前者のコーパスは入手が困難で十分な量得ることは難しい．また後者のコーパスは，解析失敗による誤った依存情報が，それより得られる統計的知識の信頼性を低下させてしまう問題が内在する．

さらに，実際の翻訳処理に統計的な指標を適用することを考える場合，訳語選択の精度だけでなく，統計的知識の構築に必要な時間的・金銭的コストも考慮に入れる必要がある．

上記のような課題をふまえ，我々は単語間依存情報を前提としない目的言語コーパスを利用して，訳語選択処理を改善する方法を提案した［Asanoma99］．ここでは英語側の共起情報を取得する際の共起とみなす単語を工夫し，英単語の共起情報を得た．本稿では，この共起取得法の詳細な検討を行う．

さらに現在ある RBMT の翻訳辞書の資産を有効活用するため，翻訳ルールの基本的性能を活かしながら，共起情報も用いた訳語選択方法を提案する．その後、実際の新聞記事文を用いて訳語選択実験を行う．

以下本稿では，原言語を日本語，目的言語を英語とした機械翻訳に限定して検討を進める．

2 RBMT の訳語選択

2.1 訳語選択の方法

ここでは NTT が研究開発している日英機械翻訳システム ALT−J/E［Yamaki97］を例に取り，従来のルールベース型機械翻訳システム（RBMT）における訳語選択処理の概要を述べる．

ALT−J/E で用いられる辞書（日本語語彙大系［Ikehara97］）のうち，構文体系辞書の結合価パタン対（以下，パタン対）には，用言に対する格要素への意味的制約が，約 3000 に分類整理された意味カテゴリによって記述されている．また日英対照辞書[Shirai97]には，日英の対訳関係に上記と同じ意味カテゴリが付与されている．ALT−J/E では，このように体系化された翻訳辞書によってきめこまかな翻訳を可能としている．

以下に具体的な翻訳の流れについて述べる．

翻訳対象文は，形態素解析，構文解析を行い，文節間の係り受け関係が記述された日本語構文構造を得る（図 1）．

次の意味解析では，この日本語構文構造と，構文体系辞書のパタン対との照合を行い，最もよく一致するパタン対を選択する．これにより，パタン対の多義および単語の意味の多義を解消し，1 用言に対して 1 パタン対を決定すると同時に格要素の意味も決定する．意味解析後，翻訳対象文 1 文に対して，意味的な構造を表す構文意味構造に変換される．

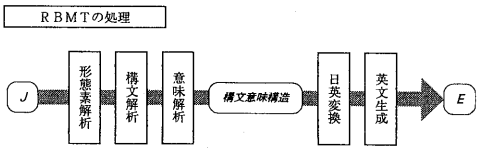

図 1：RBMT 翻訳処理の流れ

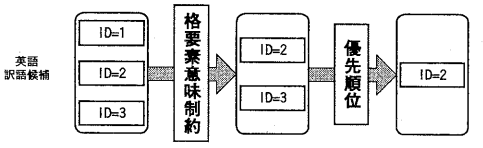

図2：訳語候補絞り込みの流れ

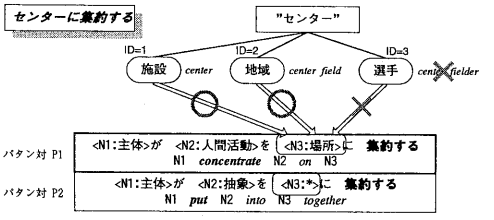

図3：格要素意味制約と名詞の意味とのマッチング

さらにパタン対が格支配する名詞の訳語は次のようにして決定する（図2）．

① 格要素に対する意味制約とマッチしない名詞の訳語を候補から削除する

② 訳語候補が複数残っている時，辞書に記述されている優先順位の最も高いものを訳語として決定する

例えば，「センターに集約する」という例（図3）では，構文意味辞書にある「集約する」に対する2つのパタン対のうち，最もよく一致するパタン対 P1 が選ばれる．そしてここでは，P1 の格要素 N3 の意味制約＜場所＞と“センター”の3種類の意味が照合され，マッチした2つの意味がこの文の“センター”の意味として残り，“センター”の＜選手＞の意味を持つ訳語候補は削除される．

このように意味制約によって訳語候補を一つに絞り込めるかは，パタン対の網羅性と意味制約の厳密性に依存しており，パタン対自身の記述内容が訳語候補の絞り込みにおいて重要となる．

2.2 訳語選択能力

このパタン対の訳語候補の絞り込み性能を検証するため，新聞記事文（150文）から，用言と格関係にあり，複数の訳語候補を持つ名詞（88件）を抽出し，パタン対の意味制約による候補の絞り込み性能を調査した．

表 1は，意味制約によって絞られる訳語候補数と，名詞に対する理想訳（正解）が訳語候補に含まれていた件数（正解含有数）との関係をまとめたものである．この結果から，抽出した名詞の26%（23 件）は1つの訳語候補に決定でき，そのうちの74%（17 件）は正解の訳語を選択しており，パタン対による訳語選択精度が高いことがわかる．また6%は，1つにはならないが絞り込みはできた．この5件は全て正解を含んだまま絞り込みが行えていることから，パタン対が絞り込みに有効な方法であることがわかる．

ただし，残りの 68%（60 件）はパタン対によって絞り込むことはできなかった．

表 1：意味制約による絞り込み

絞り込み後訳語候補数	件数	正解含有数	平均訳語候補数	
			絞り込み前	絞り込み後
1つ	23 (26%)	17	2.3	1
減少する	5 (6%)	5	3.6	2.6
変わらない	60 (68%)	60	2.2	2.2
全体	88 (100%)	82	2.3	1.9

このように，パタン対により絞り込めなかった訳語候補に対しては，別の尺度を用いることにより絞り込むことが必要となる．

3 提案方式

絞り込めなかった訳語候補を絞り込むために，目的言語側の単語並びとしての適切性を考慮した手法を提案する．訳語候補絞り込みの尺度としては英語側の単語共起情報を用いる．以下に英語共起情報を用いる訳語選択の手順を述べる．

最初に，英語コーパスから単語共起現象を抽出して，英単語対とその共起頻度の集合からなる共起情報ＤＢを作成しておく．共起頻度は，事前に設定した範囲に同時に生起する単語対をカウントして得る．

次に作成された共起情報ＤＢを用いて訳語選択処理を行う．この訳語選択処理についての詳細は 3.2 節で説明する．

以下，目的言語における単語共起の取得方法と，その単語共起情報を用いた訳語選択の方法について説明する．

3.1 単語共起取得法

コーパスから共起単語対を取得する際には，「どのような場合に共起する単語対と認定するか」を決める必要がある．この共起取得法として

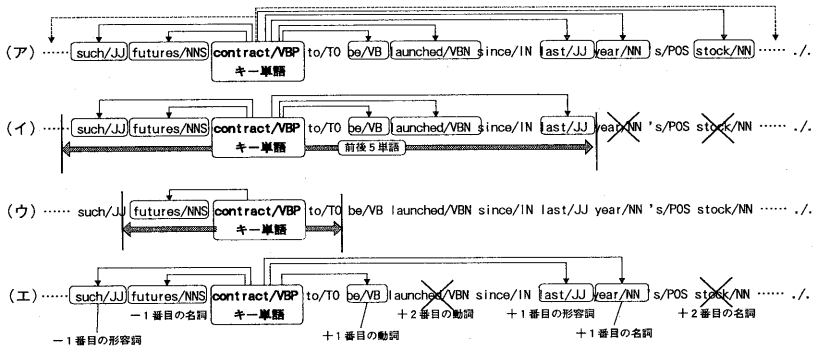

図4：共起とみなす単語

は，従来から以下の3種類がよく用いられる（図4）．

（ア） 一文内共起：一文内に同時に出現する単語すべてを共起単語とする．

（イ） 前後5単語：前後5単語内に同時に出現する単語を共起単語とする．

（ウ） 単語 bigram：隣接する単語を共起単語とする．

従来の共起取得法は，ノイズが少なく，十分な量の依存関係のある単語対（依存単語対）を効率良く収集するという点を重視していなかった．例えば，（ア）（イ）は依存関係のない単語対を多くとる傾向にあり（適合率が低い），（ウ）は逆に依存単語対をとりこぼす傾向が強い（再現率が低い）．

そこでこの問題を解決するため，我々は[Asanoma99]において係り受け関係のある単語対を近似的に収集する共起収集法として，品詞別の最近接共起単語を用いた取得方法を提案した．これは「同時に出現する単語のうち，同じ品詞の単語に関しては，最も近い単語が最も関連性が高く，依存関係のある確率も高い」という仮定を基にしている．

（エ）品詞別最近接：**品詞毎**にそれぞれ最も近くに共起する単語を共起単語とする．サーチは一文内で前後両方向に行う．

この方法は，依存関係が付与されていないコーパスを使用しても，単語に付与された品詞情報に基づいて，依存単語対を多く収集することが期待できる．例えば図4では，"contract"と共起する名詞の中でも，遠い距離にある"stock"は排除し，最も近くに出現する"year"が共起単語となる．

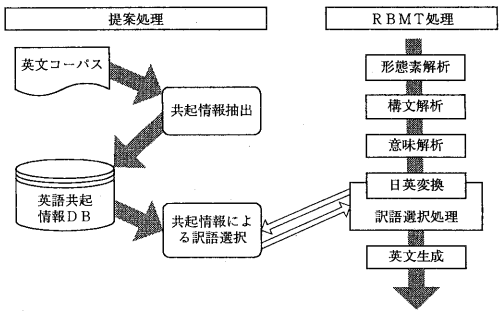

図5：従来の RBMT の処理と提案方式

3.2 訳語選択処理

以下に，単語共起情報を用いた訳語選択処理について詳細を述べる．本手法では，「原言語側で共起する単語の適切な訳語の組は，目的言語側でも共起する可能性が高い」ことを前提として訳語の選好を制御している．

提案する訳語選択処理（図5）では，2節で示した訳語選択手順のパタン対による意味制約の絞り込み①の後に，次の処理を行う．

①' 目的言語側の単語共起情報を用いて訳語候補を絞り込む

この①'の処理で，単語共起情報を用いて選択される訳語列 $\hat{E}$ は，以下の(1)式で決定する（図6）．

$e_{xy}(y=1\cdots n_x)$：x 番目の日本語に対する y 番目の訳語候補

$p(e_a,e_b)$：単語共起確率

$$\hat{E}=\arg\max_{a_1\cdots a_n}\prod_{i,j}p(e_{ia_i},e_{ja_j}) \qquad (1)$$

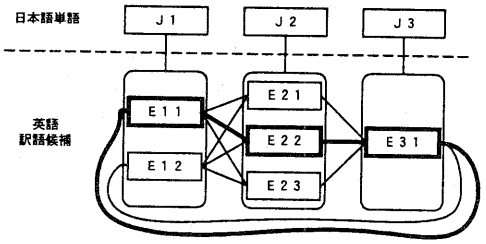

図6：共起強度による訳語選択

共起情報によって絞り込んでも第1位の候補が複数になってしまう場合[1]は，その後の処理②によって，日英対照辞書に記述してある優先順位によって選択する．

4 共起取得法の評価

4.1 方法

従来と比較して，依存単語対をどのくらい効率良く収集できるかという観点から，提案した共起取得法を評価する．

理想的な共起情報DBの性質として，以下の3つの条件を仮定する[2]

（条件1）コーパスで出現した依存単語対を多く含む
（条件2）高頻度の共起単語対を多く含む
（条件3）作成されたDBのサイズが小さい

この各条件に対する共起取得法の比較を行うため，以下の2つの評価を行った．なお本評価では，対象とする品詞を，普通名詞，動詞，形容詞，副詞に限定した．

検証1

依存単語対の特定実験は，1988年のWall Street Journalコーパスからランダムに10の英文（平均単語数：22.2 words）を選び，その中の前述の対象とする品詞以外の単語を削除する．これに人手で依存関係情報（平均依存関係数：12.2）を付与して，この依存単語対を正解とする．このテスト文から各共起取得法によって共起単語を収集し，下記の再現率，適合率により評価する．

（再現率）＝（抽出された正解の依存単語対の数）／（正解の依存単語対の数）
（適合率）＝（抽出された正解の依存単語対の数）／（抽出された単語対の数）

本評価により（条件1）を満たす共起情報DBが構築できるかを評価する．なお共起情報DBのノイズを少なくすることを優先させると，適合率が高いことが望ましく，依存単語対を網羅的に取得することを考えると，再現率が高いことが望ましい．

検証2

1988年のWSJ（98万文）から，各共起取得法で共起単語を抽出し，共起単語総計と共起単語対の数を集計する．ここで共起単語総計とは，コーパスから取得した全ての共起単語対の頻度の総計を示し，（条件2）を満たすかを判断するためのパラメータとなる．また共起単語対の数は，コーパスから取得した共起単語対の数の総和であり，（条件3）を満たすかを判断するためのパラメータとなる．

4.2 結果

検証1の結果を表2に示す．ここから，再現率，適合率でバランスよく高い値が得られているのは，品詞別最近接（エ）であるといえる．また，検証2の結果を表3に示す．この結果においても，条件2，条件3をバランスよく満たしているのは（エ）であるといえる．単語bigram（ウ）の適合率は高いが共起頻度総計が比較的少なく，十分な量のデータを得るには相当量のコーパスが必要となるため問題である．

以上から我々の提案した品詞別最近接の共起取得法が有望であることがわかった．

表2：検証1の結果

共起取得法	再現率	適合率
(ア)一文内共起	100%	14.3%
(イ)前後5単語	95.3%	21.1%
(ウ)単語bigram	54.0%	48.9%
(エ)品詞別最近接	82.7%	27.9%

[1] 複数訳語候補が残るのは，目的言語側での共起現象が全くなかったか，あるいは全く同じ値の共起強度を持つ訳語候補ペアが複数ありどちらとも決定できない場合である．

[2] コストを考えると，DB構築にかかる時間も考慮したい．本実験により経験的に，構築にかかる時間はDBのサイズにおおよそ比例することがわかった．

表 3：検証 2 の結果

共起取得法	共起頻度総計	共起単語対の数の比率
(ア)一文内共起	6310 万	1
(イ)前後 5 単語	1990 万	0.39
(ウ)単語 bigram	440 万	0.03
(エ)品詞別最近接	2280 万	0.24

4.3 考察

ここで得られた実験結果をもとに我々が提案した手法の性能改善について検討してみる．品詞別最近接の共起取得法は，キーとなる単語から一番近い単語のみを品詞別に取得したが，共起とみなす単語の抽出方法を工夫することで性能が向上する可能性がある．例えば，（名詞，動詞）の組と比較して（名詞，副詞）の組は関連性が薄いというふうに，品詞の組によって関連性の強さが異なることが予想されるので，品詞組を限定することが考えられる．また，一番目に出現するものだけでなく，二番目以降の単語も活用するように範囲を変化させながら，共起単語として取得することにより適合率の向上が見込まれる．

5 訳語選択の評価

本章では英語側の共起情報を用いた訳語選択の評価について述べる．

5.1 方法

提案手法ではパタン対による意味制約を考慮しているので，本評価のタスクは，日本語側で，用言と格要素名詞の関係を持つ名詞の訳語選択とする．また，辞書中に登録されている訳語候補の中から最適なものを選択することに問題を限定する．

共起情報 DB

共起情報は，WSJ 1987 年～1988 年分（185 万文）から取得し DB 化した．収集する単語の品詞を限定するため，コーパスはあらかじめ Tagger [Brill92] によって品詞タグを付与しておく．

本評価では，訳語選択を行う際にパタン対の主用言および格要素となりうる，名詞，動詞，形容詞，副詞に限って単語を抽出し保存する．

実験対象文・単語

実験対象文としては，人手で英訳を付与した新聞記事文からランダムに選択した 150 文を用いた．これを日英機械翻訳システム ALT－J/E に入力して，以下の条件を満たす 88 の日本語単語を実験対象単語として抽出した．

- 複数の訳語候補を持つ普通名詞
- 他の文節との構文解析結果は正しい
- パタン対の中の格要素となっている
- 対訳中の理想訳が訳語候補に含まれる（正確な評価のため）

実験対象単語とパタン対，およびそれらの訳語候補の例を表 4 に示す．

正解訳語

訳語選択結果の正誤判定については，選択された訳語と，対訳コーパスから人手で抽出された日本語単語に対する理想訳とが，一致した場合を正解とする．

訳語選択条件による性能比較

前述の実験単語を，以下の 3 種類の訳語選択条件によって訳語を選択させ，結果の評価を行う．

(A) **パタン対辞書のみを用いる訳語選択**：従来の ALT－J/E

(B) **英語共起情報を用いる訳語選択**：日本語構文構造が 1 つに定まった後，格関係にある名詞と用言に対する訳語候補の全ての組み合わせのうちから，共起情報によって訳語を決定する

(C) **パタン対辞書と英語共起情報を併用する訳語選択**： パタン対による意味制約で絞り込めなかった訳語候補を，さらに共起情報によって絞り込む

なお本性能比較で，活用する共起情報 DB は品詞別最近接共起で収集した共起情報 DB を用いた．

共起取得法による性能比較

前述の実験単語を，以下 2 種類の共起情報 DB をそれぞれ用い，パタン対辞書と単語共起情報を併用した手法（C）で訳語選択させて結果の評価を行う．

（ア）一文内共起

（エ）品詞別最近接共起

評価尺度

次式で求められる品質向上率によって性能を比較評価する.

(品質向上率) =((向上数)−(低下数))／(総数)
(向上数) = (誤り訳から正解訳に変化した数)
(低下数) = (正解訳から誤り訳に変化した数)

5.2 結果

5.2.1 訳語選択条件による比較

訳語選択条件ごとの実験結果を表5に示す. また表4に格要素名詞の訳語選択の例を示す. 結果の通り (B) (C) の英語共起情報を用いることによって, (A) のパタン対のみによる条件に対して, 品質を向上することができた. さらに (C) については (A) との間に, 有意水準1%で母平均に有意な差のある向上が見られた.

本評価により, 目的言語の共起情報を用いることで, RBMT 単独による訳語選択と比べてより適切な格要素名詞の訳語を選択できることがわかった.

5.2.2 共起取得法による比較

共起情報DBを共起取得法別に変化させて行った実験結果を表6に示す. 結果の通り (エ) 品詞別最近接共起の DB を用いた場合は, (ア) 一文内共起の条件と同程度品質を向上することができた. (ア) (エ) の条件両方で, 上記 (A) のパタン対のみ用いた場合との間に, 有意水準1%で母平均に有意な差のある向上が見られた.

この結果から, (エ) 品詞別最近接は, 少ない共起単語対でも (ア) 一文内共起を用いた時の訳文品質向上率をほぼ維持していることがわかった.

5.3 考察

表7は, 意味制約によって絞られる訳語候補数と, 訳語選択後の正解訳数との関係を訳語選択の手段別にまとめたものである.

パタン対によっても絞り込めず訳語候補数が「変わらない」場合でも, 正解訳数が増加しており, 共起情報を用いることの有効性が示された. 「1つ」の訳語候補に絞り込まれた23例について, 共起情報のみを用いた場合(B)の正解訳数をみると, 17から15に減少しており, 逆に成績が悪くなっている. この結果から, パタン対によって訳語候補を一つに絞り込める場合は, その結果を優先させるべきであることが言える.

結論として, パタン対で訳語候補を絞り込んだあとに, 共起情報を用いてさらに絞り込むという, 本稿で提案する適用順序の妥当性を示すことができた.

6 おわりに

従来の RBMT 訳語選択処理に, 目的言語側の単語共起情報による訳語候補の絞り込み手法を提案し, 本手法により訳語選択の品質は向上することが示された.

品詞別最近接共起の共起取得法は, 一般的によく行われる一文内共起などの共起取得法に比べて, より効率よく依存単語対を収集することができ, またそれによって得られた目的言語共起情報を用いて, 訳語選択処理の改善を行っても実現コストを軽減しつつ同等レベルの性能が得られることが判明した.

共起強度に関しては, 本稿で用いた単純な共起確率の代わりに, より適切な共起強度を用いることで訳語選択の性能改善が期待できる. 今後は, 改良した Dice 係数[Kitamura97], あるいは共起ベクトル [Kikui98] などを参考にし, 共起強度を表す統計量について検討を進める予定である.

また今回行った実験では, 訳出結果の機械的な照合を可能とするため, 正解訳語はただ一つに設定したが, 実際はいくつかの訳語が正解ということもありうる. 英語の単語並びとしての適切性を追求する意味では, 今後は結果としての翻訳文全体に対する人手の評価を考えている. また, 人手によらない客観的な評価基準・手段についても検討したい.

参考文献

[Brown91] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer. "Word-sense disambiguation using statistical methods." *Proceedings of ACL-91*, pages 264-270 (1991).

[Doi93] S. Doi, K. Muraki. "Evaluation of DMAX criteria for selecting equivalent translation based on dual corpora statistics." *Proceedings of TMI-93*, pages 302-311 (1993).

[Rapp95] R. Rapp. "Identifying word translations in non-parallel texts." *Proceedings of ACL-95*, pages

320-322 (1995).
[Nomiyama91] 野見山. "目的言語の知識を用いた訳語選択とその学習性." 情報処理学会研究会資料, NL86-8 (1991).
[Dagan94] I. Dagan, A. Itai. "Word sense disambiguation using a second language monolingual corpus." *Computational Linguistics*, Vol. 20, No. 4, pages 563-596 (1994).
[Tanaka96] K. Tanaka, H. Iwasaki. "Extraction of lexical translations from non-aligned corpora." *Proceedings of COLING-96*, pages 302-311 (1996).
[Kikui98] G. Kikui. "Term-list translation using mono-lingual word co-occurrence vectors." *Proceedings of COLING-ACL-98*, pages 670-674 (1998).
[Asanoma99] 麻野間, 中岩. "目的言語の単語共起情報を利用した訳語選択と未知語の訳出." 言語処理学会第 5 回年次大会発表論文集, pages 442-445 (1999).
[Kitamura97] 北村, 松本. "対訳コーパスを利用した対訳表現の自動抽出." 情処論文誌, Vol.38, No.4, pages 727-736 (1997).
[Ikehara97] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編集). 日本語語彙大系. 岩波書店 (1997).
[Yamaki97] 八巻, 大山, 白井, 横尾. "日英機械翻訳システム ALT－J/E の研究開発." NTT R&D, Vol. 46, pages 1391-1398 (1997).
[Shirai97] 白井, 横尾, 内野, 松尾. "日英変換技術と意味辞書." NTT R&D, Vol. 46, pages 1405-1410 (1997).
[Brill92] E. Brill, "A simple rule-based part of speech tagger." *Proceedings of 3rd ANLP*, pages 152-155 (1992).

表 4：実験単語，パタン対，正解訳語，および実験による結果一部

日本語単語（訳語候補）	パタン対用言（訳語候補）	正解訳語	(A)パタン対	(C)パタン対＋共起
回復 (improvement, recovery, restoration)	目指す (aim, head)	recovery	improvement	recovery
分野 (field, sector)	進出する (enter, advance, participate, start)	field	field	field
支出 (expenditure, expense)	達する (rise, reach, attain)	expenditure	expenditure	expense

表 5：訳語選択手段と格要素名詞の品質向上率の関係

訳語選択手段	正解率（件数）	向上数	低下数	品質向上率
(A)パタン対	73%(64)	－	－	－
(B)共起情報	77%(68)	16	12	+5%
(C)パタン対＋共起	78%(69)	12	7	+6%

表 6：共起取得法と格要素名詞の品質向上率の関係

共起取得法	正解率（件数）	向上数	低下数	品質向上率
（ア）一文内共起	80%(70)	11	5	+7%
（エ）品詞別最近接	78%(69)	12	7	+6%

表 7：意味制約後の訳語候補数と訳語選択手段毎の正解訳数

意味制約後訳語候補数	件数	正解訳数		
		(A) パタン対	(B) 共起情報	(C) パタン対＋共起
1つ	23	17 (74%)	15 (65%)	17 (74%)
減少する	5	5 (100%)	5 (100%)	5 (100%)
変わらない	60	42 (70%)	48 (80%)	47 (78%)
全体	88	64 (73%)	68 (77%)	69 (78%)