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

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

1 はじめに

ルールベース型の機械翻訳システムでは一般に，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．

このような適切な訳語を選択する課題を解決するために，コーパスから獲得した統計的知識を利用する方法が，すでに提案されている．例えば，語義多義性の解消（訳語選択）に，翻訳対象単語の意味と原言語の共起単語との統計的知識を用いる手法が提案されている[1]．また，依存関係のある目的言語側の単語共起単語を用いて，訳語対に対する依存関係の強度と頻度情報[2]，あるいは訳語の組合わせの一番高い共起確率と二番目に高い共起確率の比の値[3]を選択基準値として，原言語単語の意味を特定し訳語を決定する方法が提案されており，これらの手法は語義多義性解消に効果がある．

しかし，単語・文対応済みの対訳コーパスや，単語間依存情報が人手付与済みの目的言語コーパスは，依存情報が未付与のコーパスよりも入手が困難である．そのため，網羅的な統計的知識の取得が難しい．またコーパス中の単語間の依存関係を，構文解析系によって得ることを考えた時，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正確性を低下させてしまう．それゆえ，誤りが少なく網羅的な統計的知識を取得するという点では，依存情報を前提としない目的言語コーパスを利用する手法が望ましい．

さらに，コーパスに出現する単語表記そのもののみを共起情報として収集すると，個々の件の共起頻度情報が低く,その頻度情報の信頼性は下がることが多くなる（データスパースネスの問題）このようにして構築した共起情報ＤＢを用いて訳語選択を行う場合，不適切な訳語候補が出力されやすいという問題がある．

本稿では，以上のような課題を解決するため，依存情報をもたないコーパスから得た目的言語の共起情報を用いて，共起現象不足を解決した機械翻訳システムの訳文品質を向上させる方法について述べる．ここでは特に，ルールベース型の機械翻訳（RBMT）で実際に失敗した翻訳文の品質を向上させることを目的とし，RBMTでの翻訳処理と相補的な関係を持つ共起情報を用いた手法を提案する．

2 RBMTの翻訳誤り傾向

NTTが研究開発したルールベース型日英機械翻訳システム ALT−J/E[5]を用いて，実際に翻訳させた訳文の品質を低下させている原因を分析しまとめた．辞書やルールをチューンしていない新聞記事文100文に対する翻訳文（ブラインドテスト）について人手で評価し品質低下原因を分析した．（図 1参照）

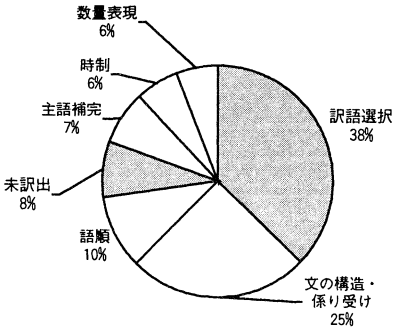

図 1：機械訳文の品質低下原因

この結果から，目的言語の共起情報の利用によって解決が期待できる品質低下原因を特定する．まず，訳語選択の不自然性による品質低下(38%)が，他の原因よりも顕著であったため，これを改善することの効果は大きい．一方，日本語のまま出力されてしまう未訳出の場合(8%)は，訳文品質が著しく低い．この未訳出には，翻訳辞書にもともとその訳語候補が登録されない未知語による原因(81%)と，入力日本語文の解析自体が失敗している単語分割誤りによる原因(25%)の２通りがある．

本稿では，これらの訳文品質を特に低下させる課題，２種類に対する解決法を提案する．

3 提案手法

英語コーパス中の単語共起情報を利用して，訳語選択と未知語の訳出を行う提案手法の処理フローを図 2に示す．

まず，英文コーパスから，コーパス中の単語共起情報を抽出し，英語共起情報ＤＢを構築する．次に共起利用訳語選択処理では，翻訳辞書，電子化辞書，およ

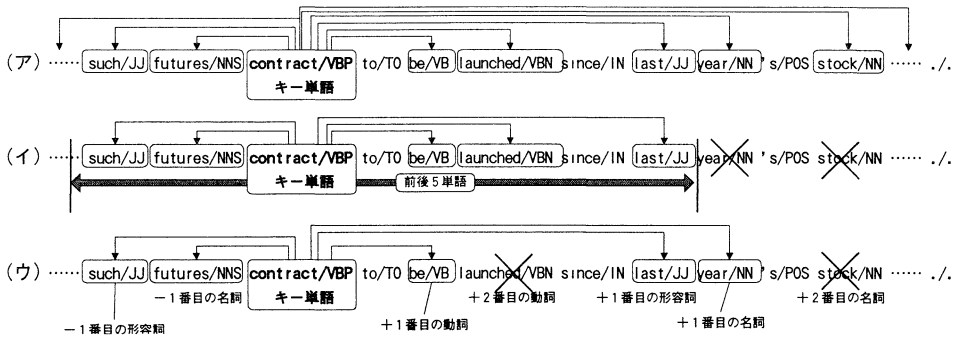

図 3：共起とみなす範囲

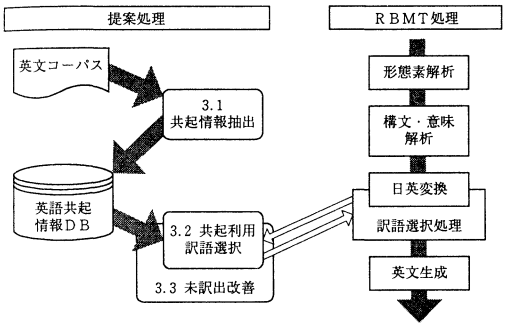

図 2：提案手法の処理フロー

び共起情報ＤＢを用いて入力日本語単語の訳語候補から適切なものを選択する．

また，未訳出改善処理では，翻訳文中に日本語のまま出現した未訳出文字列を検出し，共起利用訳語選択処理により，その文字列に対する最適な単語分割候補およびそれに伴う訳語候補列を決定する．

3.1 英単語共起情報の抽出

共起情報抽出処理では，英語共起情報を共起情報ＤＢとして保存する．共起情報ＤＢ内のエントリは，英語の共起単語対と，そのコーパス中の共起頻度からなる．ここでは，訳語選択で活用する単語として，英語コーパスからは，名詞，動詞，形容詞，副詞の単語を抽出する．

抽出時の共起範囲の設定

コーパスから共起単語対を抽出する際に，共起とみなす範囲を設定する．本稿では，一文中の共起範囲として以下に示すように三種類を選び，同じコーパスからそれぞれ別の共起情報ＤＢを構築し，比較を行う．

（ア） 全ての一文内共起
（イ） 前後５単語内共起
（ウ） 品詞別最近接共起

ここで，（ウ）品詞別最近接共起とは一文内の共起単語を探す際に，あるキーとなる単語（キー単語）に対して，この単語と共起する単語を，各品詞について，前後に最も近くに共起する単語を共起単語とみなす，と定義する．（図３参照）

例えば，同じ共起する名詞でも，遠い距離にある名詞は関連性が低いと仮定できるので，最も近くに出現する名詞だけを共起単語とすることができる．これによって，共起情報を収集する際に，依存関係の記述されていないコーパスを使用しながら，依存関係のない可能性の高い単語対を除外することができ，少ない共起収集計算量で，有効な共起情報の獲得が期待できる． 具体的には共起収集計算量は，一文中の単語の長さを n とすると，（ア）：$O(n^2)$，（イ，ウ）：$O(n)$ となる．

3.2 訳語選択の改善方法

訳語選択処理では，翻訳対象となる日本語単語列を入力し，上記英語共起情報ＤＢおよび翻訳辞書を用いて，各日本語単語に対する訳語候補から最適な訳語を選択する．

ここでは，２つの訳語候補どうしの共起強度として共起確率($p(e_1,e_2)=f(e_1,e_2)/N$)を用いる．

訳語選択処理の手順は，訳語候補列全体のスコアの計算を軽減するため，図４に示すように制約伝播法[3]を用いて行う.これは，訳語候補どうしの組み合わせの集合の中で,共起強度の最も高い訳語候補対から,訳語候補を決定していく処理方法である．

この時，訳語候補を検索する際に，機械翻訳システムが持つ翻訳辞書以外の電子化辞書を用いて，訳語候補を増やすことを行う．ただし，共起情報を利用して訳語が選択できなかった日本語単語は，RBMT の辞書・ルールによる制約条件を用いた従来手法で訳出する．また，実際の翻訳処理においては，訳語選択の結果を用いて各単語の語形変化，語順などを整えて最終的な翻訳文を生成する．

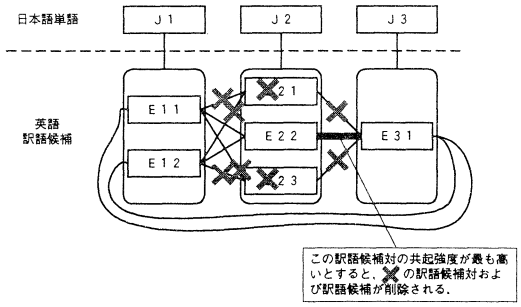

図 4：訳語選択手順

派生語・同義語・意味カテゴリの考慮

翻訳辞書から得られる英語訳語候補は，基本形で記述されており，一方でコーパス中の単語は，さまざまな活用変化を伴って出現する．これらを同一の単語とみなして，データスパースネスの問題を解消するため，訳語候補に対する変化形を，ALT−J/E の辞書（日英対照辞書：登録単語数約 40 万語）を用いて展開し，基本形として照合し頻度を加算することとする．

また同様にして，ある訳語候補の派生語，同義語，同一の意味カテゴリを持つ単語（以下，類義語とする）とも照合して，これら関連語の頻度もその訳語候補の共起頻度としてカウントする．

英語の派生語は ALT−J/E の英語辞書（登録単語数約 8 万語）を用いて，同義語は，WordNet[4]（WordNet 1.6：登録単語数約 12 万語）を用いて検索する．また，類義語は WordNet における上位語に近似する．

WordNet は，一つの英単語に対し複数の意味が対応しているので，本来ならば用いる際にその中から適切な語の意味を決定する必要がある（語義多義性解消の問題）．しかし，ここでは一つの英単語は全ての意味を持つと仮定し，対応する全ての同義語,類義語を収集することとする．

3.3 未訳出の改善方法

3.3.1 未知語による未訳出

未知語の訳出手段は以下の通りである．

[Step1] 日英機械翻訳システムが出力する翻訳文中に，日本語単語が混ざっている未訳出文字列を含む翻訳文を入力する．

[Step2] 未訳出文字列の訳語候補について，電子化辞書により訳語候補を検索し，その訳語候補と，翻訳文中で訳出されている英単語との組み合わせに対して，3.2節の訳語選択処理を行い，未訳出文字列に対する適切な訳語候補を出力する．

3.3.2 単語分割失敗による未訳出

解析誤りを原因とする未訳出の解決策として，以下の手順で単語分割候補とそれに伴う訳語候補を同時に決定し，未訳出の発生を解決する．

[Step1] 日英機械翻訳システムが出力する翻訳文中に，日本語単語が混ざっている未訳出文字列を含む翻訳文と元の日本語文を入力する．

[Step2] 次に，元の日本語文の複数の単語分割候補を RBMT の形態素解析処理から取得する．

[Step3] 各単語分割候補に含まれる日本語単語について，3.2節の訳語選択処理中の共起強度検出処理を行い，その共起強度に基づいて各単語分割候補に対する単語分割スコアを得る．ここで単語分割スコアは，単語分割中に含まれる訳語候補対の共起確率の積とする．

[Step4] 最も単語分割スコアの高い単語分割候補，およびそれに伴う訳語列を出力する．

4 評価

4.1 訳語選択実験

4.1.1 方法

本実験では，3 章で提案した共起情報ＤＢと訳語選択処理の有効性を調べる．収集する単語の品詞を限定するため，品詞情報を各単語に付与したタグ付き英文コーパスを利用する． コーパスは，Penn Treebank から Wall Street Journal のタグ付きコーパス（12.7 万文，263 万語）を使用した．ここから得られた共起単語対の数は，（ア）一文内共起：780 万，（イ）前後 5 単語内共起：310 万，（ウ）品詞別最近接共起：200 万，となった．

実験条件による優位性を測定するため，表 1 の実験条件を設定する．

表 1：実験条件

実験条件	利用する共起情報ＤＢ	訳語候補の追加	関連語の考慮
(1)	（ア）一文内共起	しない	しない
(2)	（イ）前後 5 単語共起	しない	しない
(3)	（ウ）品詞別最近接共起	しない	しない
(4)	（ア）一文内共起	する	しない
(5)	（ア）一文内共起	しない	する

実験対象は，2章で行ったブラインドテスト文翻訳結果で，訳語選択による品質低下原因と判断された文の集合から選ぶ．その中で，対訳辞書またはルール辞書の中には正解の訳語として存在したが，適切に訳語選択が行われなかった 29 単語を用いて検証した．訳語選択処理への入力は，ALT−J/E の日本語解析結果により，実験対象単語（列）と同じ文中にあり依存関係にある単語列である．

訳語選択結果の英語としての自然性を人手で判断し，評価基準は，ALT−J/E 出力の翻訳訳語列と相対比較して，次の基準値を用いる．

(品質向上率)
=((向上した数)−(低下した数))／(総数)

4.1.2 結果

訳語選択実験の結果を表 2 に示す

共起情報ＤＢ別　条件(1)においては，実験に用いた事例全体の 2 割弱が改善された．また(3)品詞別最近接ＤＢは，計算量・データ量とも減少したにもかかわらず，性能的に劣っていないことは特筆すべきである．

【適用例 1】 条件(1)～(3)
「製品の輸出拡大を管理する」
改善前：… export enlargement …
改善語：… export **expansion** …

訳語候補追加　(4)電子化辞書による訳語候補の追加によって，(1)と比較し，さらに訳文品質が改善した．適用例 2 は，(1)では失敗したが，(4)の条件では適切な訳語選択を行った例である．

【適用例 2】 条件(4)
「経営管理の企画策定」
改善前：… management management …
改善語：… management **administration** …

関連語考慮　(5)関連語を基本形のカウントに組み込む対処法においては，(1)に比べ全体的に品質向上率は低下したが，(1)～(4)全てにおいて失敗していた事例に対して，訳語選択が成功しているものがある（適用例 3）．

【適用例 3】 条件(5)
「最大処理性能」
改善前： … largest processing …
改善語： … **maximum** processing …

4.2 未訳出改善実験

3 章で提案した未訳出の改善方法の有効性を調べるため，英語共起情報を用いた2章で行ったブラインドテスト文翻訳結果で未訳出となった 5 文(37 語)を用いて訳出実験を行ったところ，単語単位で品質向上率約50%の改善が見られた．

一例として，実験文字列，「通信回線の大規模ユーザ」の適用例を以下に示す．

【適用例 4】
「通信回線の大規模ユーザ」
→「通信/回線/の/大規模/ユーザ」
改善前：… **通信 回線 大規模 ユーザ** …
改善語：… **communication line large scale user** …

4.3 検討

実験において，訳文品質が低下，または不適切な訳文となった事例の中には，正解ではないある特定の訳語候補対の共起強度が他と比べて強すぎるため，その制約が波及して訳語候補の選択失敗をしてしまう場合が多く見られた．この問題は[2]でも指摘されているが， RBMT の解析処理から得られるルールによる訳語候補の選択制約を新たに加味することで解決できると期待できる．

表 2：訳語選択実験結果

実験条件		品質向上率
(1) 基本条件		+17.2%
共起情報ＤＢ	(2) 前後５単語内	+6.9%.
	(3) 品詞別最近接	+17.2%
(4) 訳語候補の追加		+20.7%
(5) 関連語考慮		+13.8%

また，WordNet から同義語，同一意味カテゴリを持つ語を取得し，基本形に頻度情報を組み入れる処理で発生する単語と意味を対応づける際のノイズが，予想以上に悪影響を及ぼした．関連語に関しては，基本条件において共起強度を参照して信頼性が低いときだけ，考慮するという方針が望ましいと考える．

5 おわりに

本稿では，依存情報が付与されていない英語コーパスから抽出された共起情報を用いて，訳語選択処理と未知語の訳出処理を行い訳文品質を向上させる手法を提案し，その有効性を実証した．訳語選択実験では，RBMT で一度失敗した翻訳文において，訳文品質の向上が確認できた．共起現象のスパースネスを訳語候補の関連語を用いて解決する提案手法については，基本条件で失敗した幾つかの訳語選択事例を改善できた．

今後は，訳語候補対の共起強度，同義語，意味カテゴリに関する検討課題を解決していく予定である．また，未訳出の改善方法について，対象試験文を限定せず大規模な評価実験を行いたいと考えている．

参考文献

[1] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer. "Word-sense disambiguation using statistical methods." *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp.264-270, 1991.

[2] 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.

[3] I. Dagan, A. Itai. "Word sense disambiguation using a second language monolingual corpus." *Computational Linguistics*, Vol. 20, No. 4, pp.563-596, 1994.

[4] C. Fellbaum. "WordNet: an electronic lexical database." The MIT Press, 1998.

[5] 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, pp.1391-1398, 1997.