Ngramを用いた漸進的構文解析における曖昧性の解消

平石 智宣† 佐藤 健吾† 延澤 志保† 斎藤 博昭‡ 中西 正和‡
† 慶應義塾大学大学院 理工学研究科 計算機科学専攻
‡ 慶応義塾大学 理工学部 情報工学科
{tomonobu,shiho,satoken,hxs,czl}@nak.ics.keio.ac.jp

1 はじめに

音声での対話を念頭においた機械翻訳では，原言語の入力に従って順次解析を行い，目的言語に変換を行う漸進的解釈手法が必要となる．それらの代表的なシステムとして，Matsubaraらにより漸進的な英日話し言葉機械翻訳システムが提案されている [1]．

漸進的な翻訳システムでは，語の入力毎にそれまでの入力に対する解析木を随時作成する必要がある．それを実現するための構文解析の手法として，漸進的チャート解析法 [2] が提案されているが，この手法においては未入力部分の予測をもとに解析木を生成するため，構文解析結果に対して多くの曖昧性が生じ，作成される解析木の数の爆発が問題になる．これらの曖昧性を十分に解消できないと，文の解析結果として妥当でないものが多数得られてしまい，解析の効率を低下させてしまうことにもつながる．

これらの問題を解消する手法として，文献 [3][4][5] などが提案されているが，これらの手法はヒューリスティックによるルールベースな手法で対応しているため，それらの規則の構築・修正に多大な人的コストがかかり，保守・拡張性の観点からは問題がある．

そこで本稿では，Ngramを用いた統計情報を導入することにより，漸進的チャート解析法における構文的曖昧性を解消する手法を提案する．この手法により，曖昧さのある複数の候補の中から尤もらしい候補を選択できるのと同時に，可能性の薄い候補の枝刈りを行うことで，解析結果数の爆発を抑制し，また早期に正しい解析結果を選択することが可能になる．

2 統計情報の構築

2.1 統計情報を抽出するコーパス

インタラクティブなシステムにおいてはタスクが絞られやすく，統計情報を有用に反映させるため，本研究においてはタスクを限定して翻訳を行う．そのため，統計的情報を抽出する言語コーパスとして，翻訳タスクに依存したコーパスとタスクを有さないコーパス (以下，一般コーパス) を併用する．翻訳タスクは「旅行に関する会話」とし，コーパスとしてはATR対話データベース (ADD)[6]，ATR音声言語データベース (SLDB)[7] の2つを使用する．また，一般コーパスとしてはBrownコーパス，および電総研によって電子化された講談社和英辞典に含まれる例文 (以下，ETLコーパス) の2つを使用する．表1が各コーパスの文数，および単語数である．

表 1: 使用するコーパスの文数と単語数

タスク	コーパス名	文数 (文)	単語数 (語)
依存	ADD	24,802	263,739
	SLDB	15,187	184,652
一般	Brown	48,949	1,127,731
	ETL	30,287	262,802

2.2 統計的言語モデルの構築

漸進的チャート解析では，大域的な解析結果はあくまで予測の段階であるため，大域的な情報を必要とする統計的手法を導入することは適当ではない．そこで，本研究ではNgramモデルを基にして統計的言語モデルを構築する．

このNgramモデルは単語の連鎖をモデル化するために適した確率モデルであり，直前・直後の $(N-1)$ 単語にしか依存しない．つまり，現在解析されている単語の以後 $(N-1)$ 単語を予測するという局所的な状況下においては，最も強力で有効的な手法である [8]．局所的な入力単語が予測できれば，構文的曖昧性を解消し，かつ正しい構文解析結果を早期に選択することが可能になる．

本手法ではtrigramを用いて言語モデルを構築する．このNgramモデルの構築にはCMU-Cambridge Statistical Language Modeling Toolkit [9] を使用し，出力確率値の正規化を前処理として行った．

また，構文解析結果を選択する際の情報として，統計的言語モデルの英単語列に対する品詞列を利用するため，言語モデルを構築する英単語コーパスに対して予め品詞タグ付けを行い，その結果も統計情報に登録する．この品詞タグ付けには，Apple Pie Parser [10] を用いる．

2.3 統計情報の融合

以上の構築手法に従い，各々のコーパスから統計情報を構築し，それらを融合することで全体的な統計情報の構築を行う．それらの統計情報を融合する際に，翻

訳タスクに依存した統計情報の重みを増すことで，タスクに依存した統計情報が構築できる．

ここでは，タスク依存性の過多の問題を考慮して，タスク依存コーパスと一般コーパスを 2:1 の重みで融合する．これは，構築の際に使用するタスク依存コーパスと一般コーパスの文数をほぼ同数にするためであり，それにより統計情報全体に対する影響の均一化を図ったものである．

以上の処理により，統計情報のデータは (単語列，品詞列，生起確率) という 3 種類のデータを内包するようなデータ構造を持つ．

3 統計情報を用いた漸進的構文解析における曖昧性の解消

構築された統計情報を用いて，漸進的構文解析におけるすべての可能性のある候補に対して，統計情報の各要素とマッチングを行い，各解析結果に対する評価値を算出する．そして，その評価値に従って尤もらしい候補を選択するのと同時に，確からしい候補のみを残し可能性の薄い候補の枝刈りを行う．

3.1 マッチングによる評価値の算出

解析中の英単語 W が，言語モデルの英単語列の先頭単語 w と一致した場合，各構文解析結果に対して解析中の予測品詞列と言語モデル中の品詞列との類似性のマッチングを行い，該当する解析結果に対して類似性に基づいた評価値を加算する．ここでは，構築した trigram の情報を有効に利用するため，類似性のパターンで分類して，それぞれのパターンに対して評価値の算出を行う．

まず，マッチングアルゴリズムを定義する際に必要な表記・定義を以下で説明する．品詞 α が品詞 β に到達可能であることを $\alpha \rightarrow \beta$，品詞 α と品詞 γ の 2 つの品詞が共起して品詞 β に到達可能 (到達可能性の拡張) であることを $\alpha + \gamma \rightarrow \beta$，品詞 α と品詞 β が同型 (品詞が一致，もしくは左辺規則が α で右辺規則が β) であることを $\alpha \equiv \beta$ と表記し，解析中の品詞を T_1，解析中の予測品詞列を T_2, T_3，言語モデルの品詞列を t_1, t_2, t_3，言語モデルの確率値を P，i 番目の構文解析結果に対する評価値を V_i (初期値 0)，t_3 が到達可能な品詞を t と定義する．

以上の表記・定義に従いマッチングアルゴリズムを図 1 のように定義する．品詞間の類似性のパターンの分類は，直観的に以下の 2 つの場合であると理解できる．

- 各品詞列の 2 つ目の品詞がお互いに単独で対応する場合には，各品詞列の 3 つ目の品詞の到達可能性・同型による分類を行う．
- 各品詞列の 2 つ目の品詞がお互いに単独で対応していない場合には，言語モデル中の 2 つ目と 3 つ目の品詞が共起して，解析中の予測品詞列の 2 つ目の品詞への到達可能性による分類を行う．

このマッチングにおける評価値として，各該当言語モデル要素の生起確率の同時出現確率を使用している．

$$
\begin{aligned}
&\text{if}(W = w \text{ and } T_1 = t_1) \\
&\quad \text{if}(T_2 \equiv t_2 \text{ and } T_3 \leftarrow t_3) \\
&\qquad V_i = V_i + (-\log P) \\
&\quad \text{if}(T_2 \leftarrow t_2 + t_3 \text{ or } T_2 \leftarrow t_2 + (t \leftarrow t_3)) \\
&\qquad V_i = V_i + (-\log P)
\end{aligned}
$$

図 1: マッチングアルゴリズム

3.2 構文解析結果の枝刈り

算出された評価値を基にして，保存されている構文解析結果の整列を行い，設定した解析数の上限に応じて評価値が上位のものを採用し，それ以下の可能性が薄いと判断される解析結果の枝刈りを行う．

しかし，文は任意の場所で終了する可能性があり，それらの文が閉じることに対する有効的な統計情報を抽出することは難しい．本研究では統計情報として trigram を採用しているため，次の入力・もしくは 2 つ先の入力により文章が閉じる解析結果が採用されていない場合には，それらの解析結果を採用する候補に追加する．これにより，文章が閉じたことで解析が破綻してしまうことを防止する．

4 統計情報を用いた漸進的英日機械翻訳システム

4.1 システムの概要

提案した漸進的構文解析における曖昧性の解消手法を，漸進的な英日機械翻訳に導入したシステムを実現した．このシステムは入力文の任意の付加入力語に対して，漸進的構文解析・漸進的構文解析選択・漸進的翻訳・漸進的生成の 4 つのモジュールが同時進行的に働くことで，漸進的な翻訳を可能にする (図 2)．

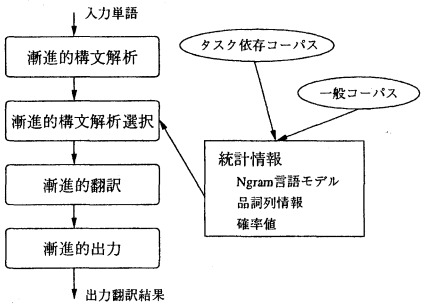

図 2: システムの翻訳処理過程

提案手法は漸進的構文解析選択モジュールで実現されており，他の 3 つのモジュールは Matsubara らのシステム [1] で提案されている 3 つのモジュールを再現したものである．以下では，本システムの各モジュールについての概要を説明する．

4.2 漸進的構文解析モジュール

このモジュールでは，漸進的チャート解析手法 [2] を用いて，語が入力される毎にそれまでの入力に対する解析木を生成する．

この構文解析において，構文的曖昧性のため可能性のある複数の解析結果が生成された場合には，このすべての候補を保存し，以後の翻訳解析処理を行う．

4.3 漸進的構文解析選択モジュール

このモジュールでは，提案した統計情報による構文的曖昧性の解消手法が実現されている．提案手法に従い各解析結果に対する評価値を算出し，可能性の薄い候補の枝刈りを行い，評価値が最大の候補を正しい解析であると仮定して以下の翻訳処理を行う．

しかし，入力された語に対する統計情報が必ず存在するとは限らない．また，統計情報が存在しても評価値が算出されない場合もある．それらの場合には，構文解析結果の正確性が確保できないと判断することができるため，解析結果の枝刈りは行わず，すべての解析候補を保持して処理を進める．またこの場合，尤もらしい候補の選択も不可能なため，統計情報により評価値が算出される単語が入力され構文解析の正確性が確保されるまで文字列の出力のタイミングを遅延することで，翻訳結果の適切な出力タイミングを図る．

4.4 漸進的翻訳モジュール

このモジュールでは，選択された解析結果に対して，文脈自由文法規則と 1 対 1 で定義されている翻訳規則を適応することで，構文解析結果により導かれる原言語構造を目的言語構造へと変換する．この翻訳規則は，日本語の言語的特徴である不適格表現 (繰り返し，語順の逆転，省略等) をルールとして定義したものであり，これらの特徴を有効に，かつ積極的に用いることで，自然な日本語単語列を生成することが可能になる．

4.5 漸進的出力モジュール

このモジュールでは，生成された目的言語構造を，システムの出力としての日本語文字列への変換を行う．ここで，日本語の言語的特徴である言い誤り・言い直しは，前入力単語によるシステムの出力を現入力単語のシステムの出力が訂正する場合に発生する．その際にまず言い淀みを生成し，ユーザに言い直しの発生を認識をさせることで，より自然な出力を得ることができる．

5 実験

作成したシステムを用いて，英語語彙 543 語，文法 138 規則の規模で実験を行った．システムの実装には C 言語を用いた．実験の対象として，ATR 音声言語データベース [7] の旅行申し込みをタスクとするバイリンガル電話対話の中の 8 対話について，そのすべての英語会話 432 文を用いた．平均の語の長さは 約 6.32 語であった．統計情報を構築するデータからは，これらの実験に用いたデータは除外してある．

そして，翻訳の即時性を満たすため，同単語で始まる統計情報に対して確率値が上位の 40 データを採用して実験に対する統計情報を構築した．構築された統計情報のデータ数は 12,796 データとなった．また，構文解析選択モジュールにおける枝刈りを行う上限値を 500 に設定した．

実験では，翻訳の正解率，及び入力単語毎に生成される解析結果数について，従来の解析手法 (本システムで構文解析選択モジュールを省略することで再現できる) と提案手法の比較を行った．

5.1 翻訳実験

従来手法による結果を表 2，提案手法による結果を表 3 に示す．

表 2: 従来手法の翻訳正解率

タイプ	文数	割合
(a) 正しい翻訳 (言い直しなし)	244	56.48%
(b) 正しい翻訳 (言い直しあり)	88	20.37%
(c) 不自然な翻訳・誤った翻訳	96	22.19%
(d) 解析の失敗	4	0.96%

表 3: 提案手法の翻訳正解率

タイプ	文数	割合
(a) 正しい翻訳 (言い直しなし)	280	64.81%
(b) 正しい翻訳 (言い直しあり)	91	21.06%
(c) 不自然な翻訳・誤った翻訳	11	2.55%
(d) 解析の失敗	50	11.58%

提案手法では表 3 に示すように，(a) または (b) に分類された 371 文が正解であり，85.87 % の翻訳正解率を得ることができた．これは，従来手法の結果の表 2 で (c) に分類されていた不自然な翻訳が，構文的曖昧性が減少したことで，正しい翻訳に分類されたためである．この結果より，提案手法の有効性を確認することができた．

また，提案手法のシステムの出力翻訳結果が著しく遅延することは確認できなかったため，漸進性を損なうことなくユーザへの適切な出力タイミングを実現できたと考えられる．

翻訳実験を Linux (メモリ: 384M CPU: Athlon 800MHz) 上で行ったところ，入力された英語 1 単語に対する翻訳処理平均時間は，従来手法 0.088 秒に対して，提案手法は 0.194 秒であった．付加処理により解析時間は増加しているが，この時間は英単語 1 単語の発話時間より短いと考えられるので，本システムは翻訳の実時間性という制約も満たしている．

5.2 生成される解析結果数の実験

従来手法と提案手法における，文の単語数毎に生成される解析結果数の平均値の比較を行った結果を図 3 に示す．

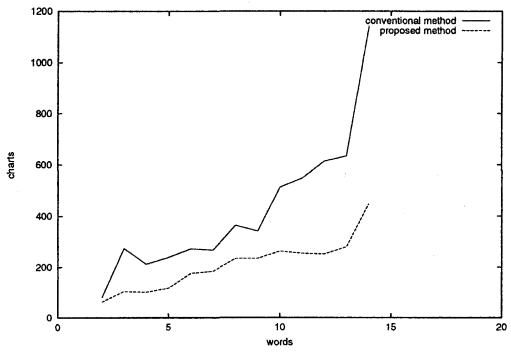

図 3: 解析結果数の比較

図 3 から，従来手法より提案手法が効率的に生成される解析結果数を削減していることがわかる．特に文が長くなった場合には，削減効果が大きいことが確認できる．これは，文が長くなると構文的曖昧性が増加し，それにともない作成される解析結果が多く作成されるためであり，提案手法が構文的曖昧性を解消し，解析結果数の抑制に有効であることが確認された．

6 翻訳解析失敗の考察

提案手法においては，Ngram を用いて構文的曖昧性を多く含むと判断された翻訳結果の枝刈りを行うため，枝刈りされた候補の中に正しい構文解析結果が存在することにより，解析が失敗する可能性がある．表 3 において，(d) に分類された 50 文のうち，この理由により解析が失敗したのは 46 文であった．それら翻訳の失敗した文に対する解析を行った結果，解析失敗の主な要因は以下に挙げた要因にあると考えられる．

- 複数の品詞を持つ単語の統計情報の偏りにより，その単語の特定の品詞に対する解析結果が構文解析選択モジュールで選択されない．
- 話し言葉における疑問文の割合が，統計情報を抽出したコーパス中の割合よりも大幅に大きいため，疑問文の構造に対する統計情報の妥当性が十分得られていないことにより，構文的曖昧性が解消できない．

これらを解消するためには，まず第 1 に統計情報を抽出するための学習データ数を増やすことで，妥当性のある候補を多く抽出することが重要である．その他には，統計情報を適応する状況 (肯定文・疑問文・タスク依存性) に応じて専門化・細分化することが考えられる．

7 まとめ

本稿では，英文を入力に従って順次解析を行う漸進的な英日話し言葉機械翻訳システムにおいて，漸進的に構文解析における曖昧性を解消する手法として，Ngram を用いた統計情報を導入することを提案した．

構文解析におけるすべての可能性のある候補に対して構築した統計情報の要素とマッチングを行い，算出された評価値に従って尤もらしい候補を選択するのと同時に，確からしい候補のみを残し可能性の薄い候補の枝刈りを行うことで，構文的曖昧性を漸進的に解消し，解析結果数の爆発の問題の抑制を可能にした．

提案手法を実装したシステムを用い，ATR 音声言語データベースに収録された英語会話 432 文に対して翻訳実験，および生成される解析結果数の実験を行い，従来手法と比較することで本提案手法の有効性を示した．

今後の課題としては，より大域的な情報を扱うことのできる統計的手法と融合する手法，および単語のクラスタリングによる単語クラスを用いた共起情報を用いる手法などの検討が必要がある．

謝辞

本研究に対していろいろな助言や，システム・データの提供をしていただいた名古屋大学言語文化部助手松原茂樹氏に深く感謝いたします．

参考文献

[1] S.Matsubara, Y.Inagaki: Incremental Transfer in English-Japanese Machine Translation, IEICE Transactions on Information and Systems vol.E80-D no.11, pp.1122–1129, 1997.

[2] S.Matsubara, S.Asai, K.Toyama, Y.Inagaki: Chart-based Parsing and Transfer in Incremental Spoken Language Translation System, Proceedings of The 4th Natural Language Processing Pacific Rim Symposium pp.521–524, 1997

[3] S.Matsubara, H.Ogawa, K.Toyama, Y.Inagaki: Incremental Spoken Language Translation bases on A Normal-Form Conversion of CFG, Proceedings of the 5th Natural Language Translation Pacific Rim Symposium pp.515–518, 1999

[4] 加藤 芳秀, 松原 茂樹, 外山 勝彦, 稲垣 康善: 漸進的構文解析における構文的曖昧性とその解消, 情報処理学会自然言語処理 vol.134 no.16 pp.117–122, 1999

[5] 村瀬 隆久, 松原 茂樹, 外山 勝彦, 稲垣 康善: 依存関係を用いた漸進的構文解析の効率化, 言語処理学会 第 6 回年次大会 pp.491–494, 2000

[6] 江原 暉将, 井ノ上 直己, 幸山 秀雄, 長谷川 敏郎, 庄山 富美, 森元 逞: ATR 対話データベースの内容,ATR Technical Report, TR-I-0186, 1990

[7] 浦谷 則好, 竹沢 寿幸, 松尾 秀彦, 森田 千帆: 音声言語データベースの構成, ATR Technical Report, TR-IT-0056, 1994

[8] 北 研二: 言語と計算–統計的言語モデル, 東京大学出版会, 1999

[9] http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[10] http://cs.nyu.edu/cs/projects/proteus/app/