E_021

節構造の推定に基づく統語森解析の高精度化

釜谷聡史, 知野哲朗, 降幡建太郎
東芝 研究開発センター

1 はじめに

音声翻訳など, 自然発話を受け付けるシステムでは, 任意の文を扱えることがユーザにとって非常に重要である. 我々の音声翻訳システム [1] は, 発話の断片が入力されることを前提として, 全ての統語的可能性を一括解析, 最尤の各断片をまとめる節と, 節間の関係を推定することで, 高い性能を実現している. これは, 話し言葉が断片的な発話の連続になり易い一方で, ひとたび節としてまとまれば, その節内構造は, 整った書き言葉に近いという知見に基づく. しかし, 個々の断片, 節を独立して翻訳するべきか, 関連付けて翻訳するべきかを判断することは非常に困難な問題であった. これは, 全ての音声アプリケーションに共通する課題であり, 断片的, 連続的に入力される情報を, いかに各処理系に適した単位に区切るかが鍵となる.

2 統語森駆動翻訳方式

我々の音声翻訳システム (図 1) は, 文法的尤度を伴って解析するよう拡張した一般化 LR 解析法により, 全ての解釈を圧縮共有統語森 [3] (以下, 統語森と呼ぶ) として導出, 選好される構造を抽出して変換することで, 入力に対する最尤構文候補からの翻訳を実現し, 高精度化を図っている. また, 訳文生成には, 特に市場での成功例に富むトランスファ方式で使われる翻訳知識を活用し, 高い品質とカヴァレッジを実現している.

本方式では, 統語森の中から, 共起的, 意味的に選好される構造を抽出する手法として, 統語森係り受け解析 [2] を用いている. これは, 構文木ベースの意味解析手法に基づきながら, 統語森上で一括して解析し, 付与した係り受け尤度に基づき, 最尤構造を選出する手法である. 同方式では, 二単語間の共起確率を評価の中心に据え, 統語森上に表現された, 二単語間の統語的構造距離, 表層距離, 及び, 文法的に推定される節構造に基づいて, 尤度を計算する. 節構造は, 独自開発した口語文法規則により導くが, 種々の発話に対する頑健さを保つため, 大まかに区別しているのみであった. ゆえに, 各節の特徴や構成形態素に応じた評価に課題がある.

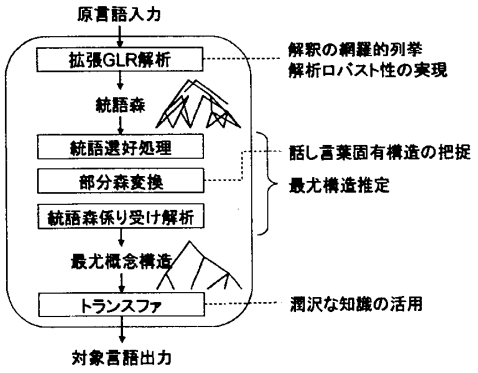

図 1: 音声翻訳処理の流れ

3 節情報を用いた係り受け解析

3.1 節とその種類の判定

日本語の節は, 局所的な形態素系列を調べることによって, 高精度に抽出できることが知られている [4]. 本研究においても, 入力形態素系列のみを手掛かりとするパタンマッチにより節とその種類を推定, 同節が文法的に選択された場合の振る舞いを規定する. 次いで, 推定した節とその種類の情報を係り受け解析における尤度計算時に反映する方式を採る.

表 1 に開発した節推定規則の一部を示す. 規則の開発には, 南 [5] の分析を基礎とした. 南は, 節境界の違いから従属節を複数種に分類し, その自立性の度合いと関連付けている. これを参考として, 例えば, 図 1 において (1) の規則は, 接続助詞 “が” に素性 “cp_ga” が「任意の系列＋接続助詞 “が”」なる表層パタンに基づいて付与され, 同素性と素性 “cm_ga” が競合するとき, 尤度-0.9 が与えられることを示している.

3.2 統語森係り受け解析への適用

図 2 を用いて, 解説する. まず, 節推定規則を用いて, 入力形態素に素性を振る. 図 2 の例では, 格助詞 “を” に素性 “kj_ga” が, 接続助詞 “が” に素性 “sj_ga” が付与されている. 次いで、節点 c に注目する. 同節点は、“彼” と “会わ (ない)” との共起関係を表しており、その共起に基づく尤度は文献 [2] と同様, 次式で求める.

$$DepScore(x,y) = 1 - \frac{1}{1 - exp(\mu - p(x,y))} \quad (1)$$

ここで,$p(x,y)$ は EM アルゴリズムにより学習した, 単語 x, y の共起確率,μ は全ての $p(x,y)$ の平均値である.

表 1: 節推定規則

	素性名	パタン	尤度
(1)	cp_ga	-1.sf=*&0.pos=“接続助詞”&0.sf=“が”	cm_ga;-0.9
(2)	cm_ga	-1.pos=名詞&0.pos=格助詞&0.sf=“が”	cm_ga;-1.0

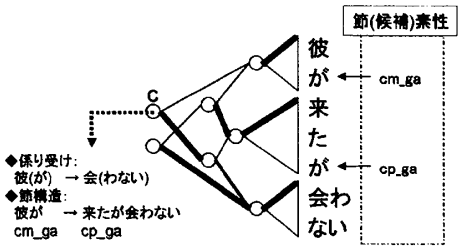

図 2: 節点における尤度計算

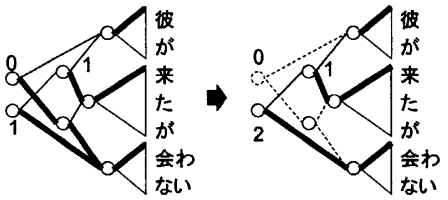

図 3: 統語森係り受け解析による最尤構造推定

また, 節点 c が支配する係り側構造の素性 "cm_ga" と, 同受け側構造の素性 "cp_ga" が競合していることも分かる. そこで、節点 c は, $DepScore$(彼, 会) と節推定規則 (1) に基づく尤度-0.9 の合計値を与える. 以降, 各節点における部分森の尤度を計算し, それらの最大値を統語森の葉から根方向に足し合わせ, 尤度の高い順に根から葉方向に部分構造を選択する. 同過程は, 文献 [2] と同様であるから, 図 3 を示すことで省略する.

以上により, 断片的な入力系列から推定した節構造の尤度, 口語解析文法に基づく統語的尤度, 共起関係に基づく意味的尤度の三者を, 過剰な仮定を置くことなく結合させ, 解析精度の向上を図ることが可能となる.

4 評価

4.1 実験方法

話し言葉翻訳向けに自作したコーパスから, 未知の発話 200 文 (平均文長 17.3 文字, 平均単語数 9.4 単語) をランダムに取り出して評価対象とし, 統語森係り受け解析による, 統語森内の構文木候補の削減効果と, 翻訳品質の変化を評価した. ここで, 正解係り受け関係は, 事前に人手で与えた. また, 今回の実験用に開発した節推定規則の数は 51 である. 構文木候補の削減効果を評価する指標としては, 次式に示す, 内包率を定義した。ここで,ρ は、統語森が正解係り受けのみを有する構文木候補を持つ場合は 1 を, それ以外は 0 をとる.

$$\text{内包率} = \rho \cdot \frac{\sum_{t \in \mathcal{F}} \frac{\text{構文木 } t \text{ 中の正解係り受け数}}{\text{付与した正解係り受け数}}}{\text{統語森 } \mathcal{F} \text{ 中の構文木 } t \text{ の数}}$$

翻訳品質の評価に際して, 構文木候補が複数残った場合は, 統語森中でより早期に解析された構造を 1 つ選んで翻訳した. 訳文の評価は, 発話意図の伝達可能性を重視しながら, 部分訳, 主語/ロール誤りを厳しく評価し, 評価者 3 人による目視多数決方式で実施した.

表 2: 評価結果 (200 文)

節推定	なし	あり
構文木候補	3.27	2.80
内包率	0.851	0.875
翻訳精度	77.8%	78.8%

4.2 実験結果とその分析

実験結果を表 2 に示す. 表には, 本稿で導入した節推定方式を利用した場合と利用していない場合の二つを合わせて掲載している.

まず, 解析精度の面では, 拡張一般化 LR 解析による受理時の平均構文木候補数が 5.58 であったのに対し, 節推定を導入したことで, 候補削減数した上に, 内包率においても向上が確認できる. より詳しくは,27 文で内包率の改善が見られ,1 文で悪化が見られた. 唯一悪化があった例文も文法規則の不備に起因しており, 改善は容易であり, 総じて節構造推定と統語森係り受け解析の融合効果の高さを確認できる.

一方, 翻訳品質の面では, 節推定の導入によって,21 の訳文が変化した. 内,9 文が不正解から正解に, 10 文が不正解ながら品質が改善された. 悪化した 2 文は, 文法規則及びトランスファ規則の不足に起因しする. また, 訳文に改善が見られなかった 4 文においても, 正解の解析が得られており, トランスファ規則の追加で, 訳文の向上が期待できる. すなわち, 総じて悪化無く翻訳品質の改善が見られたといえる.

5 おわりに

本稿では, 入力表層による節推定情報を利用して, 統語森係り受け解析を高精度化する手法を提案し, 実験によって, 解析精度の向上と, 翻訳品質の改善を確認した. 今後の課題は, 現在人手で与えている, 節推定規則と尤度を機械学習等で与え、網羅性を高めることである.

参考文献

[1] Tetsuro Chino, Satoshi Kamatani, "Partial Forest Transfer for Spoken Language Translation", In *Proc. of RANLP2005* pp.157–161,(2005).

[2] Satoshi Kamatani, et al. "Syntax Forest Based Syntactic and Dependency Analysis towards Robust Spoken Language Processing", In *Proc. of PACLING 2005* pp.192–199,(2005).

[3] M. Tomita (Eds.), "Generalized LR Parsing", Cluwer Academic Publishers, (1991).

[4] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝, "節境界自動検出ルールの作成と評価", 言語処理学会第 9 回年次大会発表論文集, pp.517–520,(2003).

[5] 南不二男,"現代日本語の構造", 大修館書店,(1972).