対訳コーパスを用いた同時翻訳単位の検出

笠 浩一朗† 松原 茂樹† 稲垣 康善‡
†名古屋大学情報連携基盤センター ‡豊橋技術科学大学
ryu@el.itc.nagoya-u.ac.jp

1 はじめに

近年の音声処理技術の進展にともない，対話を対象とした音声翻訳システムの開発が盛んに行われている [1, 2, 6, 10]．これらのシステムの多くは，ターンや文を処理の基本単位として用いているため，文の終了を待ってからでないと訳出を開始することができず，話者の待ち時間が長くなる．それに対して，同時通訳者のように話者の発話に追従した翻訳（以下，**同時翻訳**）を導入することが考えられる．実際，通訳者を介した異言語間対話において，逐次通訳で訳出した場合と同時通訳で訳出した場合を比較した研究では，同時通訳の方が対話の効率や円滑さが大幅に向上することが指摘されている [8]．

本論文では，同時翻訳のための翻訳単位を提案する．本研究では，対訳データを用いて翻訳単位を定める方法を導入する．すなわち，文対応のついた対訳データを，単語対応に基づいて単位分割することにより，翻訳単位を定める．

同時翻訳システムが入力に追従した訳出を実現するために，翻訳単位は，文に比べて十分に細かく，かつ，入力と同時進行的に検出可能である必要がある．本論文では，提案する翻訳単位の粒度の細かさと同時的な検出可能性について検証する．翻訳単位の検出には，節境界単位の種類，形態素の種類，ポーズの有無に着目した手法を用いた．日本語対話文を用いて分割実験を実施した結果，精度にして 80.8%，再現率にして 74.3% を達成しており，提案する単位の検出可能性を確認した．

2 同時翻訳単位

本節では，同時翻訳のための翻訳単位（以下，**同時翻訳単位**）について検討する．

2.1 同時翻訳単位の性質

同時翻訳単位は，発話の入力と同時進行で検出でき（**検出可能性**），かつ，同時的な訳出を実現するのに十分に細かいこと（**細粒度性**）が求められる．

一方，同時翻訳の処理単位が，他の単位とは独立に翻訳でき（**翻訳の独立性**），かつ，その単位を検出したら即座に訳出可能 (**訳出の即時性**) であれば，システムの出力タイミングを制御する必要がないという利点がある．

2.2 関連研究

同時翻訳に関する従来研究として，Ryu ら [9] は，句を翻訳単位として採用している．しかし，句の訳出は他の句に依存せざるを得なく，翻訳の独立性を満たさない．また，日本語と英語のように構造が大きく異なる言語間での同時翻訳では，語順の問題があり，訳出の即時性を満たさない．

一方，Kashioka ら [3] は，日英翻訳の処理単位として節を用いることを提案している．節は，節境界解析（例えば，[7]）を使用することにより入力と同時的に，かつ，高精度に検出できる．また，意味的・統語的にまとまった単位であるため翻訳の独立性が高い．しかし，節の生起順序は両言語間で必ずしも同一ではなく，即時性を満たさない場合がある．

主な言語単位の同時翻訳単位としての適切さを表 1 に示す．このように一般的な言語単位において翻訳の独立性と訳出の即時性をともに満たすものはなく，このことは同時翻訳のための新たな単位を定める必要があることを示唆している．

3 訳出に基づく翻訳単位への分割

本論文では，日英対話における同時翻訳単位として，翻訳の独立性と訳出の即時性を考慮して翻訳単位を新

表 1: 同時翻訳単位に必要な性質と言語単位との関係

処理単位	検出可能性	細粒度性	翻訳の独立性	訳出の即時性
単語	◎	◎	×	×
句 [9]	◎	◎	×	×
節 [3]	◎	○	○	×
文	◎	×	◎	◎

たに定め，それが検出可能性，及び，細粒度性を満たすことを示す．本研究の以下では，この単位を，訳出されたデータに基づいて決まる単位であることから，**訳出単位**と呼ぶ．訳出単位は，日英対訳データにおける，以下の性質：

- 原文の内容を聞き手が自然に理解可能 (**訳質**)
- 原文の語順に準じている (**漸進性**)
- 意訳や省略を含まない (**直訳性**)

を満たす訳文（以下，**逐語訳**）とその対訳対応関係を用いて，次の手順で定める．

Step1: 独立に翻訳できる単位に日本語文を分割する．

Step2: 分割された単位を，英語文との対訳対応関係を考慮して，語順の入れ替わりが生じない最小の単位でまとめる．

例として，日本語対話文

(J1) 今のところ予定通りですが出発が遅れる可能性がありますのでご了承くださいませ．

とその英語逐語訳

(E1) For now, it is on time, but the departure might be delayed. Please understand it.

に対して，上記で示した手順で分割することを考える（図 1 参照）．まず，Step1 で,「今のところ」,「予定通りですが」,「出発が」,「遅れる」,「可能性がありますので」,「ご了承くださいませ」の 6 つに分割する．次に，Step2 で,「遅れる」と「可能性がありますので」の語順が入れ替わるため，それらを一つにまとめ，最終的に「今のところ」,「予定通りですが」,「出発が」,「遅れる可能性がありますので」,「ご了承くださいませ」の 5 つが訳出単位となる．

表 2: データの規模

対話	216
文数	8736
訳出単位数	13510
形態素数	57016

4 同時翻訳単位としての利用可能性

4.1 対訳コーパスを用いた分割データの作成

訳出単位が，細粒度性と同時的な検出可能性を満たすことを検証するために，対訳コーパスを用いてデータを作成した．コーパスとして，名古屋大学同時通訳データベース [11] に収録された旅行対話データの日本語話者発話データを利用した．逐語訳データとして，日本語話者発話データに対して，翻訳者が作成したものを利用した．

日本語対話文を訳出単位に分割したデータの例を図 2(a) に示す．また，図 2(a) に対応するように逐語訳を分割したものを図 2(b) に示す．図 2(a) と図 2(b) の左側の数字の組は，左側の数字が文 ID を示しており，右側の数字が各文における提案単位の ID を示している．

訳出単位の基礎統計量を表 2 示す．

4.2 細粒度性の分析

訳出単位の平均形態素数は 4.22 形態素であり，文の平均形態素数 6.53 の 64.6%である．15 形態素以上の文と訳出単位の長さの分布を図 3 に示す．図 3 より，文に比べ訳出単位では 15 形態素以上の長さのものが少なく，粒度が細かくなっていることがわかる．

4.3 検出可能性の分析

訳出単位の検出では，形態素境界が訳出単位の境界になる確率を算出し，その確率が閾値以上の場合に訳出単位の境界であると判定する．確率計算のモデルとして，最大エントロピー法に基づくモデルを用いた．

4.3.1 最大エントロピー法に基づくモデル

出力値 y は形態素境界が提案単位の境界になる $(y = 1)$ か否か $(y = 0)$ であるとし，また，j 個の素性 $g_i (1 \leq$

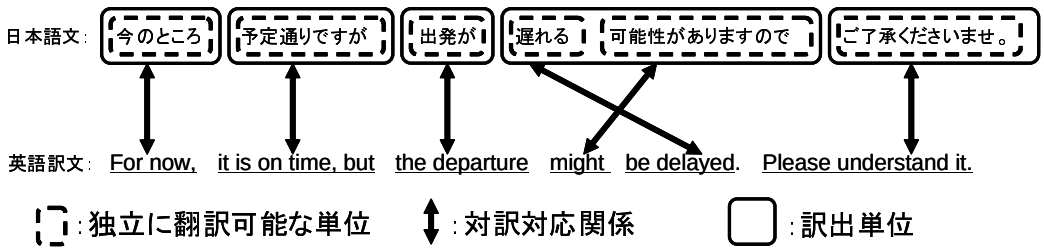

図 1: (J2.1) の訳出単位の分割例

1-1 お店は 1-2 道路沿いではないんですけれども 1-3 林ビルの二階にあります。
2-1 林ビルはすぐ見つけていただけると思います。
3-1 テレビ塔という大きなタワーのすぐ横ですので。
4-1 多分 4-2 日本語と思いますので 4-3 今からお書きします。
5-1 そうですね。
6-1 せっかくお越しいただいてるので 6-2 名古屋城を見られたらと思いますね。

(a) 訳出単位に分割された日本語対話文

1-1 The restaurant 1-2 is not on the street but 1-3 It's on the second floor of Hayashi building.
2-1 You can find Hayashi building easily.
3-1 It's just next to a tall tower called TV tower.
4-1 Perhaps 4-2 it is written in Japanese. 4-3 So I write it for you.
5-1 I see.
6-1 As you took a trouble to come here. 6-2 You should see Nagoya castle.

(b) 訳出単位に対応する逐語訳

図 2: 訳出単位に分割された日本語対話文とその逐語訳

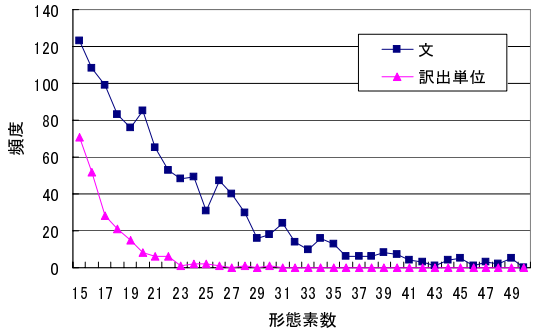

図 3: 文と訳出単位の形態素数の分布

$i \leq j$) を考えるとき，文脈 x において形態素境界が訳出単位境界になる確率 $p^*(1|x)$ を，以下のように計算する．

$$p^*(1|x) = \frac{exp[\sum_{i=1}^{j} \lambda_{1,i} g_i(x,1)]}{exp[\sum_{i=1}^{j} \lambda_{1,i} g_i(x,1)] + exp[\sum_{i=1}^{j} \lambda_{0,i} g_i(x,0)]} \quad (1)$$

ここで，$\lambda_{y,i}$ は素性関数 $g_i(x,y)$ のパラメータを表している．また，$g_i(x,y)$ は文脈 x において素性 f_i が観測され，かつ，出力値が y となるときに 1 を返す素性関数である．

4.3.2 検出実験

前節で述べた検出手法を用いて，日本語対話文を訳出単位に分割する実験を実施した．

実験に用いたデータは，4.1 節で述べた 216 対話である．モデルの学習用として 180 対話，素性選択のテスト用として 18 対話，実験のテスト用に残りの 18 対話を用いた．

条件付き最大エントロピー法のツールとしては文献 [4] を用いた．このツールでは，パラメータの学習の繰り返し数を設定する必要があり，実験では経験的に十分な大きさと思われる 50 に固定した．さらに，パラメータ λ_i の推定には L-BFGS[5] を用いた．訳出単位境界確率の閾値を，0.5 に設定した．素性選択実験により，素性には前後の 3 形態素の出現形と品詞（品詞，活用形，活用型）と発話単位境界の有無を選択した．

実験結果を表 3 に示す．精度にして 80.8%，再現率にして 74.3%を達成しており，同時翻訳単位の検出可能性を確認した．

表 3: 実験結果

手法	精度	再現率	F 値
本手法	80.8%(329/407)	74.3%(329/443)	77.4

5 おわりに

本論文では，同時翻訳処理単位の性質として，翻訳の独立性，訳出の即時性，細粒度性，検出可能性を挙げた．また，翻訳の独立性と訳出の即時性に基づき対訳データから分割して得られた提案単位が，細粒度性と検出可能性を満たすことを確認した．

謝辞

本研究の一部は，科研費基盤研究 (B)（課題番号 20300058）によります．

参考文献

[1] R. Frederking, A. Blackk, R. Brow, J. Moody, and E. Stein-brecher, “Field Testing the Tongues Speech-to-Speech Machin Translation System,” Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp. 160-164, 2002.

[2] R. Isotani, K. Yamada, S. Ando, K. Hanazawa, S. Ishikawa and K. Iso, “Speech-to-Speech Translation Software PDAs for Travel Conversation,” NEC Research and Development, 44, No.2, pp. 197-202, 2003.

[3] H. Kashioka, T. Maruyama, “ Segmentation of Semantic Unit in Japanese Monologue,” Proceedings of Oriental COCOSDA2004, pp. 87-92, 2004.

[4] Z. Le, Maximum Entropy Modeling Toolkit for Python and C++, 2004.

[5] D. C. Liu, J. Nocedal, “On the Limited Memory BFGS Method for Large Scale Optimization,” Mathe. Programming, pp.503-528, 1989.

[6] F. Liu, Y. Gao, L. Gu and M. Picheny, “Noise Robustness in Speech to Speech Translation,” IBM Tech Report RC22874, 2003.

[7] 丸山岳彦， 柏岡秀紀，熊野正， 田中英輝, “ 日本語節境界検出プログラム CBAP の開発と評価” 自然言語処理, 11, 3，pp. 39-68, 2004.

[8] 大原誠，松原茂樹，笠浩一朗，河口信夫，稲垣康善，“同時通訳を介した異言語対話の時間的特徴逐次通訳との比較に基づく対訳コーパスの分析”，通訳研究，No.3，pp.34-52, 2003.

[9] K. Ryu, S. Matsubara, Y. Inagaki, “Simultaneous English-Japanese Spoken Language Translation Based on Incremental Dependency Parsing and Transfer,” Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp.683-690, 2006.

[10] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo and S. Yamamoto, “A Japanese-to-English Speech Translation System:ATR-MATRIX,” Proceedings of 5th International Conference on Spoken Language Processing, pp. 957-960, 1998.

[11] http://slp.el.itc.nagoya-u.ac.jp/sidb/