同時通訳コーパスを用いた通訳単位の統語的分析

笠 浩一朗† 松原 茂樹‡§ 河口 信夫‡§ 稲垣 康善♮

†名古屋大学大学院工学研究科 ‡名古屋大学情報連携基盤センター

§名古屋大学統合音響情報研究拠点 ♮愛知県立大学情報科学部

k_ryu@matubara.net

1 はじめに

これまでに、いくつかの音声通訳システムが開発されているが [3, 4]、それらは、文単位での通訳処理に基づくものがほとんどである。そのようなシステムでは、1文全体が入力された後でないと通訳結果の出力を開始できないため、対話の円滑さが大きく損なわれることになる。話者の発話途中で出力が可能になれば、ユーザもまた、早い段階で話者の発話内容を理解し、応答することができ、スムーズな多言語間対話の実現が期待できる。

同時通訳における重要な課題として、同時通訳に適した通訳単位の獲得がある。通訳単位とは、通訳システムに入力された原文においてシステムが処理可能な単位であり、通訳システムは、原単語列の入力に対して通訳単位を漸進的に認識する必要がある。著者らがすでに提案している通訳単位の獲得手法 [2] は、対訳データを利用するものであるため、原単語列のみから通訳単位を決定することはできない。

そこで本稿では、通訳システムが原単語列の統語情報を利用した通訳単位の認識可能性を判断するために、著者らの手法によって獲得した通訳単位と、人手により付与した統語単位との関係を調査したので、その結果について述べる。

2 同時通訳コーパスと通訳単位

本節では、分析対象となる対訳データと、そのデータから獲得した通訳単位について述べる。

2.1 対訳データ

著者らは、日英双方向の同時通訳音声を収録した音声対訳データベースを構築している [1]。その話者音声の書き起こしデータに対し、逐語的な対訳文の作成を行った。対応が細かくなるように以下の制約を満たす逐語訳を作成した。

- 可能な限り原文の語順に順じた訳を付与する。
- 文脈への 依存を避けるため、意訳や省略は原則として避ける。
- 原文の内容が聞き手にスムーズに理解できるような訳文を作成する。

Simultaneous Interpretation Corpus-based Syntactic Analysis of Translation Unit: Koichiro Ryu, Shigeki Matsubara, Nobuo Kawaguchi (Nagoya University), and Yasuyoshi Inagaki (Aichi Prefectural University)

表 1: 逐語訳データの基礎統計

項目	英語	英日	日本語	日英
単語数	69424	68675	70225	62605
発話単位数	10858	10387	13176	12516
文数	7889	8108	8725	8231

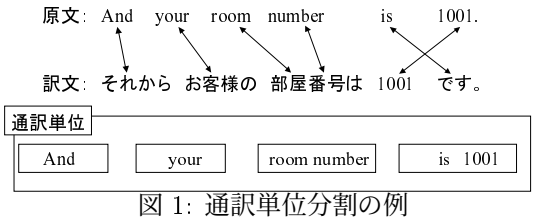

図 1: 通訳単位分割の例

分析対象は、話者音声の書き起こしデータであり、通訳単位に分割するために上記の手法で作成した逐語訳データを利用する。

作成した対訳データの英日、日英それぞれの単語数(日本語は形態素数)、発話単位数、文数を表 1 に示す。

2.2 同時通訳の通訳単位の獲得手法

同時通訳システムに適した通訳単位への要求として、その通訳単位ごとに訳を出力できることが挙げられる。そこで、原文と訳文との間で単語対応を取り、単語対応が交差する、すなわち、原文と対訳文との間で対訳語の生起順序が異なる部分を一つの通訳単位とする。また、語順が同じ部分は最小単位(英語は単語、日本語は文節)に分割したものを通訳単位とする。これにより、漸進的に訳を出力することができる単位を獲得することができる。以下に、英日間における通訳単位の獲得手順を説明する。また、その例を図 1 に示す。

Step 1 原文と通訳文を、英文は単語、日本文は文節単位で分割する。

Step 2 Step1 で分割した単位ごとに、原文と通訳文間で人手で対応を取る。(図 1 の矢印が各単語対応を表している)

Step3 英語と日本語の文法上における構造の違いにより、矢印が交差する現象が起きた場合、交差している対応の原単語列を 1 つの通訳単位としてまとめる。交差しない場合は、その対応の原単語を一つの通訳単位とする。

上記の手順によって、図 1 の下方にある通訳単位を獲得することができる。

表 2: 通訳単位の基礎統計

項目	英日	日英
通訳単位数	16338	16516
1 文あたりの通訳単位数	2.07	1.89
1 通訳単位あたりの単語数	4.25	4.25

この手法により、前節で述べた対訳コーパスから通訳単位を獲得した。獲得した通訳単位数、1 文あたりの通訳単位数、1 通訳単位あたりの単語数を表 2 に示す。

3 通訳単位と統語的分析

3.1 統語単位の付与

原文データに対して、以下の 3 つの統語レベルにおける全 9 種類の統語単位を階層的に人手で付与した。語レベルは、原則として 1 単語からなる単位であり、句レベルは、2 単語以上からなり、かつ、述部を含んでいない単位であり、節レベルは、2 単語以上からなり、かつ、述部を含んでいる単位である。

語レベル: 名詞、形容詞、副詞

句レベル: 名詞句、形容詞句、副詞句

節レベル: 名詞節、形容詞節、副詞節

3.2 通訳単位と統語単位の関係の調査

通訳単位の自動認識に統語的な情報が利用できるかを調査するために、通訳単位と統語単位の比較を行った。統語単位の総数、統語単位が通訳単位と一致する回数、統語単位の途中で通訳単位の境界が出現する回数を調査した。その結果を表 3,4 に示す。表の括弧内は、それぞれの総数に対する割合を示す。統語単位と通訳単位の一致する割合から、

- 英日では、名詞類、副詞類が通訳単位と一致する可能性が比較的高い
- 日英では、副詞類が通訳単位と一致する可能性が比較的高い
- 英日、日英とも、副詞節が通訳単位になる割合は 5 割以上である

ということ、また、統語単位の途中で通訳単位の境界が出現する割合から、

- 統語単位の途中で通訳単位の境界が出現する可能性は非常に低い

ことが観察できる。以上より、統語単位の種類によって通訳単位のなりやすさが異なり、さらに、言語によりその傾向に違いがあることがわかった。これらは、通訳単位を認識する手がかりとして利用できる。

さらに、 通訳単位を認識するときに、原単語列の統語単位を漸進的に認識できれば、統語単位内で通訳単位の境界が存在するかを判断する必要はないことが明らかになった。

4 おわりに

本稿では、原単語列の統語情報から同時通訳に適した通訳単位の認識可能性を調査するために、対訳データから獲得した通訳単位と人手により付与した統語単位との関係を調査した。調査結果より、通訳システムが通訳単位を認識するために原単語列の統語情報が重要な手がかりになることがわかった。今後は、通訳単位を認識する手法と、それを用いた同時通訳手法について検討する予定である。

表 3: 通訳単位と統語的な単位の関係（英日）

統語レベル	統語単位	総数	通訳単位と一致した統語単位数	通訳単位境界を含む統語単位数
語	名詞	6484	881(13.6%)	0(0.0%)
	形容詞	1168	34(2.9%)	0(0.0%)
	副詞	579	88(15.2%)	0(0.0%)
句	名詞句	5400	1293(23.9%)	32(0.6%)
	形容詞句	712	53(7.4%)	2(0.3%)
	副詞句	1235	246(19.9%)	5(0.4%)
節	名詞節	43	3(7.0%)	1(2.3%)
	形容詞節	59	3(5.1%)	0(0.0%)
	副詞節	87	45(51.7%)	1(1.1%)

表 4: 通訳単位と統語的な単位の関係（日英）

統語レベル	統語単位	総数	通訳単位と一致した統語単位数	通訳単位境界を含む統語単位数
語	名詞	6751	264(3.9%)	3(0.0%)
	形容詞	672	16(2.3%)	1(0.1%)
	副詞	321	56(17.4%)	0(0.0%)
句	名詞句	4197	259(6.2%)	23(0.5%)
	形容詞句	886	42(4.7%)	1(0.1%)
	副詞句	1219	195(16.0%)	7(0.6%)
節	名詞節	39	1(2.6%)	0(0.0%)
	形容詞節	43	2(4.7%)	0(0.0%)
	副詞節	63	42(66.7%)	2(3.2%)

参考文献

[1] 松原茂樹, 相澤靖之, 河口信夫, 外山勝彦, 稲垣康善: 同時通訳コーパスの設計と構築, 通訳研究, No.1, pp. 85-102, (2001).

[2] K.Ryu, S.Matsubara, N.Kawaguchi and Y.Inagaki: “Bilingual Speech Dialogue Corpus for Simultaneous Machine Interpretation Research, Proceedings of Oriental COCOSDA 2003, pp. 217-224 (2003).

[3] T.Takezawa, T.Morimoto, Y.Sagisaka, N.Cambell, H.Iidaand, F.Sugaya, A.Yokoo, and S.Yamamoto. “Japanese-to-English Speech Translation System: ATR-MATRIX”, Proceedings of 5th ICSLP, pp. 957-960 (1998).

[4] T.Watanabe, A.Okumura, S.Sakai, K.Yamabana, S.Doi and K.Takahashi, “An Automatic Interpretation Software for Travel Conversation”, Proceedings of 6th ICSLP, Vol.IV, pp.444-447 (2000).