日本語対話文の節分割に基づく同時的な翻訳単位に関する考察

丁 喆*，笠 浩一朗，松原 茂樹，吉川 正俊 (名古屋大学)

Simultaneous Interpreting Unit of Conversational Japanese based on Clause Boundaries

Zhe Ding*, Koichiro Ryu, Shigeki Matsubara, Masatoshi Yoshikawa (Nagoya University)

1 まえがき

音声・言語処理技術の進展により、音声翻訳システムの開発が現実的になりつつある。これまでに作成されたシステムの多くは対話音声を対象としており、翻訳精度の向上を目指して研究が進められてきた。しかしながら、システムを介して遂行される異言語間対話の効率及び円滑さを考慮すると精度だけでなく、翻訳の同時性も重要となる [1]。

そこで本稿では、同時的な日英対話翻訳における翻訳単位について述べる。本研究では、日本語の言語単位の１つである「節」に着目する。日本語の節入力に対して即座に出力可能でかつ対訳対応関係にある英語句を取り出し、それらの言語的特長について分析を与える。分析には、名古屋大学 CIAIR 同時通訳コーパス [3] を使用した。

2 節を単位とした同時的な日英翻訳

同時翻訳では、訳文の生成タイミングが重要であり、そのパフォーマンスは設定された翻訳単位に依存する。文よりも短い単位として、単語や句などを翻訳単位とする手法が英日翻訳において検討されてきた [2]。一方、日英翻訳では、両言語間における述語の出現位置の異なり具合を考慮すると、「節」（述語を中心としたまとまり）を単位とすることは１つの方法である。たとえば、日本語対話文

- ホテルの予約をしてこなかったのでこちらでホテルを紹介して頂きたいんですけれども

の英語対訳文は、

- I haven't made any hotel reservation so could you introduce me any nice hotel?

であり、いずれも２つの節から構成され、それらは互いに対訳関係にある。したがって、各節を翻訳単位として定めることができ、節「ホテルの予約をしてこなかったので」が入力された段階で、それに対応する"I haven't made any hotel reservation"を出力することができる。

しかしながら、対話文を構成するあらゆる節が、上述のような対応関係にあるとは限らないため、翻訳単位になりうる節を区別できる必要がある。

3 音声対訳コーパスを用いた分析

節と翻訳単位との関係を調べるため、名古屋大学対話同時通訳コーパス [3] に収録された旅行対話データを用いて分析した。分析には、全 11 対話に含まれる日本語対話 519 文のうち、複数の節に分割された 207 文とその英語対訳文を使用した。節分割には、節境界解析プログラム CBAP[4] を用いた。

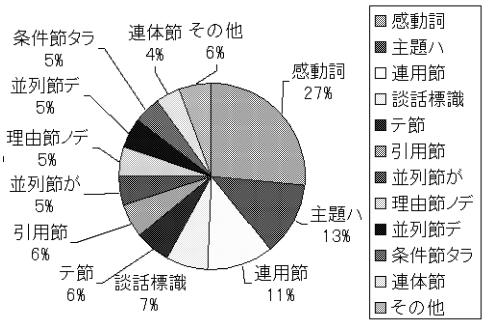

図 1: 節境界の種類と出現頻度

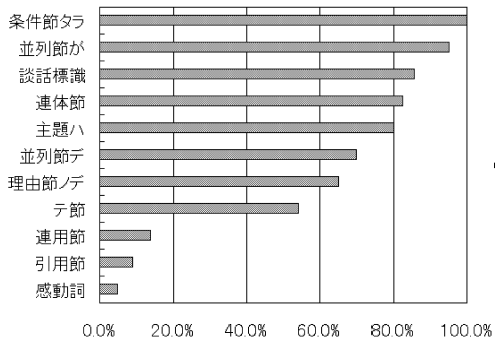

図 2: 節境界の種類と分割可能な割合

207 文の 288ヶ所に節境界が存在した（文末を除く）。節境界の種類による内訳を図 1 に示す。上位 11 種類の節境界で全体の 94.4%に占めている。288ヶ所の節境界のうち、前節の例文のように、対応する英語文で同様に分割可能な節境界は全体の 46.2%(181/392) であった。

上位 11 種類のそれぞれにおける分割可能な節境界の割合を図 2 に示す。この図から、節境界の種類によって分割可能性が異なることがわかる。割合の高い上位８種類の節境界で日本語対話文を分割し、それを翻訳単位とみなすとき、その認識性能は、精度にして 78.9%(157/199)、再現率にして 86.7%(157/181) であり、節境界の種類が同時翻訳単位の認識に重要な手がかりになることがわかった。

文献

(1) 大原．他： 通訳研究，No.3, pp. 34-52, 2003.

(2) K. Ryu, et al.: Proc. of Asian Symposium on Natural Language Processing, pp. 91-95, 2004.

(3) H. Tohyama, et al.: Proc. of Eurospeech, 2005.

(4) 丸山．他：自然言語処理, Vol. 11, No. 3, pp. 39-68, 2004.