日英同時通訳者の翻訳単位

加藤　直人　　渦原　茂
ＡＴＲ音声言語コミュニケーション研究所
{naoto.kato, shigeru.uzuhara}@atr.jp

1　はじめに

講演は１つの発話が長く、その翻訳には発話が終わってから行う逐次翻訳よりも、発話途中でも同時進行で行う同時通訳が適している。同時通訳では、この“同時進行”を実現するために、あるまとまった単位（翻訳単位）で翻訳が行われる。翻訳単位は、通訳者の記憶容量の制約にもよるが、翻訳対象となる言語ペアに強く依存すると考えられる。日本語から韓国語への通訳では、両言語は構文構造が近いので頭から順に翻訳しやすい。一方、日本語から英語への通訳では、動詞出現箇所のように構文構造が大きく異なるので、ある大きな翻訳単位が必要となる。

講演を機械翻訳する場合でも、同時通訳的な翻訳が望ましい。したがって、原文を翻訳単位に分割する必要がある。通訳者の場合には、翻訳単位を通訳者養成学校等で訓練し、そのノウハウを身に付けている。また、言語学的な立場からは、船山[1]が日英同時通訳を対象に翻訳単位に関する分析を行っている。しかし、これらを計算機上にインプリメントすることは困難である[1]。工学的な立場からは、丸山ら[2]が、日本語の節解析を利用した手法を提案している。この手法では原言語（日本語）のみの情報で行っている。また、遠山ら[3]らは英日同時通訳のために、訳出パターンを人手により分析し作成している。しかしながら、通訳に十分な訳出パターンを人手で作成することは困難である。むしろ、現在の自然言語処理技術で行われているように、大量の翻訳単位データを集め、機械学習の手法を使って、翻訳単位を学習するのがよいと考える。しかしながら、日英同時通訳における翻訳単位のデータは、例えば文献[4]のようなものがあるものの、非常に少ない。

本稿では、日英同時通訳における翻訳単位データの作成に向け、同時通訳コーパスから翻訳単位を自動推定する手法について述べる。本手法では、通訳者が連続して発声している発話（英語）を翻訳単位と考え、それに対応する原発話（日本語）を推定する。本手法で使用している同時通訳コーパスは、講演として日本語による解説番組を対象とし、通訳者により日英同時通訳を行ってもらい，さらに元の日本語の講演を人手で翻訳したものである。

本手法で得られた翻訳単位データは、通訳者の具体的なデータであるので、同時通訳の分析研究や通訳者の訓練に使うことも期待できる。

2　同時通訳コーパス

我々の同時通訳コーパスは、原文（日本語）、通訳文（英語）、翻訳文（英語）で構成されている。同時通訳コーパスの例を図１に示す。

原文はＮＨＫの解説番組「あすを読む」である[2]。時事問題に関してＮＨＫ解説委員が解説する10分の番組である。話す内容はある程度原稿として用意されているので、一般的な講演ほど「話し言葉」的ではない。しかし、ニュースのように原稿をそのまま読むのではない。

通訳文は、録画した放送を見てもらいながら、同時通訳者に通訳してもらったものである。これら原文や通訳文の詳細な分析については文献[5][6]に譲る．

翻訳文は、原文を人手で翻訳したものである。ただし、翻訳の際には、なるべく原文の情報が欠落しないようにしてもらい、さらに通訳結果も参考にしてもらっている。したがって、翻訳文と原文とでは内容的にほとんど同等であり、通訳文と翻訳文では共通する単語も多い。このような性質により、翻訳文と原文、通訳文と翻訳文で、それぞれにおいて単語アライメントがなるべく取れるようにしている。なお、通訳者と翻訳者は異なる。現在、同時通訳コーパスは250ほどの番組から構成されている。

[1] もちろん、機械翻訳と通訳者では翻訳単位が異なるという考え方もある。

[2] ＡＴＲではＮＨＫとの共同研究により使用している。

J1:今晩は。
J2:コンピューターを利用しましたインターネットという情報ネットワークが [PAU] 私達の経済ですとか社会の仕組みを今 [PAU] 大きくかえ始めています。
J3:そこで今晩は [PAU] 経済の動きを中心にしまして [PAU] アメリカの例もみながら [PAU] 今後の動きを探ってみようと思います。
J4:なじみのない方のために手短にインターネットについて説明しておきましょう。
J5:こちらにあります。
J6:インターネットで言いますのは情報のネットワークなんです。
J7:言いかえると情報のプールと言ってもいいんじゃないでしょうか。
J8:私達電話線を通して接続業者 [PAU] プロバイダーと言われているんですが、ここに結びますと、こ
J9:こからインターネットに入っていきます。
J10:電話をかけるのと同じように [PAU] 例えば友達と [PAU] 手紙のやりとりをしたりあるいは写真を送ったり [PAU] そういうことができます。
J11:それからここにホームページというのがあります。
J12:企業ですとかあるいは政府 [PAU] こちらはホームページという [PAU] 掲示板を作って [PAU] ここに様々な情報を載せています。

原文（日本語）

S1:good evening.
S2:Internet information network based on computer [PAU] is dramatically changing the way of life and the way, the society works today.
S3:so today take an look at the experience in the United States, especially in the field of economy.
S4:I'd like to take a look at how Internet is changing the world.
S5:let me firstly explain Internet.
S6:the Internet is a information network. in a sense, Internet is a pool of information.
S7:through the telephone circuits we are connected to providers.
S8:and from them we can enter into the Internet.
S9:just like using telephone [PAU] with your friend [PAU] you can exchange letters, or you can exchange [PAU] information.
S10:and here is the web site.
S11:government or companies [PAU] prepare a web site as a bulletin [PAU] where they can post a lot of information.

通訳文（英語）

T1:good evening.
T2:an information network based on computer called Internet is dramatically changing our economic and social systems.
T3:so this evening I would like to take a look at the trends toward the future,
T4:focusing on the economic trends and referring to the cases in the United States.
T5:first let me explain the Internet briefly for those who are unfamiliar to it.
T6:here it is.
T7:the Internet is an information network.
T8:in other words, we may say it is a pool of information.
T9:we can enter into the Internet by connecting ourselves with a connecting service company, called provider, through the telephone lines.
T10:just like using the phone, for instance, you can exchange letters with your friends and send photos.
T11:and here is a web site.
T12:companies and governments prepare a web site as a bulletin where they can post a variety of information.

翻訳文（英語）

図１　同時通訳コーパスの例
（[PAU]はポーズを表す。ただし、文末にポーズがある場合には改行のみで表した。）

図１で、“[PAU]”はポーズ（無音区間）を表している。我々はこのポーズを通訳者の翻訳単位の切れ目であると仮定した。すなわち、ポーズとポーズの間の（通訳者が連続して発声している）発話を翻訳単位であるとした。例えば，図1通訳文中のS2を見ると、“Internet information network based on computer”でポーズが置かれているので、この英語節を通訳者の翻訳単位と考える。推定したいのは、英語節に対応する日本語節である。今の例では、この日本語節は、「コンピューターを利用しましたインターネットという情報ネットワークが」となる。以下では、「英語節」や「日本語節」のように、ある意味的なのまとまりを便宜的に「節」と呼ぶ。

３　日英翻訳単位の自動推定

本手法では、通訳文中からポーズ単位で切り出された英語節に対して、単語アライメントと構文的な制約に基づいて、対応する節番号を原文の単語上に徐々に決定していく。

単語アライメントを行う際には、原文(日本語)と通訳文（英語）とで直接行うのではなく、翻訳文(英語)を介在して、原文と翻訳文の単語アライメント、翻訳文と通訳文の単語アライメントというように行った。これは、柏岡ら[7]が報告しているように、通訳文と原文とでは直接アライメントできる単語が少なかったからである。もちろん利用する対訳辞書の語彙数にもよるが、通訳では意訳する（逐語訳しない）場合が少なくないからでもある。翻訳文は原文をなるべく忠実に翻訳しているので、原文と翻訳文とで日英の単語アライメントを行ったほうがより多くの対応付けを行えることが期待できる。さらに、翻訳文を作成する際には通訳文を利用しているので、翻訳文と通訳文では共通の単語が含まれている。単語アライメントは、単語間の類似度に基づいている。原文と翻訳文の（日英の）単語アライメントでは対訳辞書（ニューアンカー英和・和英辞典データベース[8]）を使っている。一方、通訳文と翻訳文の（英英の）単語アライメントでは、単語の表層的類似度を使っている。最後に通訳文中のポーズ単位に日本語文を切り出し、日本語の翻訳単位とする。

３．１　通訳文の節単位への分割

通訳文における英語節は、基本的にはポーズ単位で切り出しているが、英語節内の単語が一単語の場合には前後の節に吸収させる場合もある。例えば、単語“the”の前後でポーズがあった場合が存在したが、節とせずに後の英語節に吸収した。これは、一単語の前後にあるポーズは多くの場合、通訳者の言い淀み等により生じたものであると考えられたからである。一単語の場合が英語節となるか否かは、構文構造から判断した。

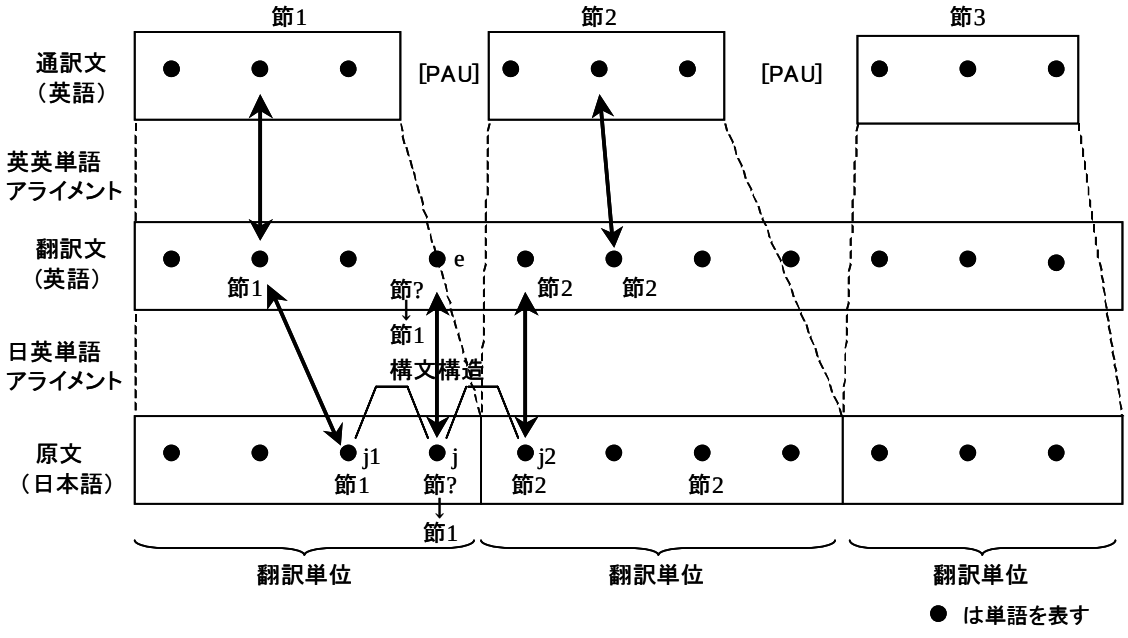

図２　同時通訳コーパスによる翻訳単位の自動推定

３．２　翻訳単位自動推定アルゴリズム

翻訳単位を自動推定するアルゴリズムを下記に、本アルゴリズムによって節番号が決まる様子を図２に示す。

STEP1では単語アライメントを行い、通訳文とアライメントが取れた場合には、翻訳文中も原文中もそれぞれの単語にその節番号を与える。これが初期値となる。STEP2、STEP3、STEP4ではこの初期値に基づいて、構文構造を手がかりとして、周囲の節番号徐々にを決めていく。STEP6はデフォルトの処理である。これまでで原文中に節番号が決まらない単語があれば、構文構造上で最も近い単語の節番号に決める。ただし、一文中のすべての単語で節番号が未決定であるならば、その文の節番号はSTEP6でも決まらない。そのような日本語節は、対応する英語節がない（すなわち、通訳者が通訳しなかった）と考え、節番号は決まらないままとした。

【翻訳単位自動推定アルゴリズム】

[STEP1] ２つの単語アライメント結果より翻訳文、原文のそれぞれの単語が属する節を決める。ただし、原文の場合には、通訳文→翻訳文→原文と辿ることにより、節番号を決める。

[STEP2] 原文中で節番号が未決定な単語（図２中のj）の中で、自分の右と左に節番号が決まっている単語（同j1, j2）があれば、構文構造でより近い方の単語の節番号を、その単語（同j）の節番号とする。さらに、この単語が翻訳文と単語アライメントされているならば、翻訳文中の単語（同e）の節番号も同じものにする。

[STEP3] 同様に翻訳文に対してSTEP2と同じことをする。すなわち、翻訳文中で節番号が未決定な単語の中で、自分の右と左に節番号が決まっている単語があれば、構文構造でより近い方の単語の節番号、その単語の節番号とする。さらにこの単語が原文と単語アライメントされているならば、原文中の単語の節番号も同じものにする。

[STEP4] 節番号が新たに決まらなくなるまで、STEP2とSTEP3を繰り返す。

[STEP5] 原文中で、これまでに節番号が決まらなかった単語に対して、構文構造上一番近い単語の節番号にする。文中に１つも単語アライメントが取れなかった場合には、対応する節がないものとする。

４　おわりに

同時通訳コーパスより、単語アライメントと構文的な制約を用いて、日英翻訳同時通訳における翻訳単位を自動推定する手法について述べた。

予備実験として、アンカーの対訳辞書[8]を使った自動推定を行ってみたがを、日英アライメントを取ることができる単語が十分でなかったために、周囲に拡大して節番号を決めるステップ（STEP4 、STEP4、STEP4）がすぐに収束してしまった。その結果、デフォルト（STEP5）で節番号が決まることが多かった。今後は辞書をさらに拡大して評価実験する予定である。また、翻訳単位データを使った表示システムも開発している[9]。

翻訳単位は一種のチャンクである。今後は、チャンクを同定する学習アルゴリズムの中で、得られた翻訳単位を学習データとして使いたい。そして、日本語を翻訳単位に処理する、同時通訳的な日英機械翻訳を実現したい。

謝辞

本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1]船山仲他．同時通訳における処理単位について．通訳理論研究 10，pp. 4-13，1996.
[2]丸山岳彦，柏岡秀紀．「文節訳の出現順序」を考慮した翻訳単位の決定手法．言語処理学会第８回年次大会，pp. 188-1191，2002.
[3]遠山仁美，松原茂樹．同時通訳コーパスを用いた通訳者の訳出パターンの分析．信学技法，思考と言語，TL2003-26，pp. 61-66，2003.
[4]インタースクール教材開発チーム．インターの「自信がつく」シリーズ、国際政治・紛争表現集、国際経済・外交表現集．(株)インターグループ，2003.
[5]柏岡秀紀．講演の同時通訳データ作成と分析．信学技法，思考と言語，TL2000-33，pp. 61-66，2000.
[6]柏岡秀紀，田中英樹．講演の同時通訳データの分析． 言語処理学会第７回年次大会，pp. 433-436，2001.
[7]柏岡秀紀．講演同時通訳データのアライメント．言語処理学会第８回年次大会，pp. 188-1191，2002.
[8]ニューアンカー英和・和英辞典データベース．学習研究社，1999.
[9]渦原茂，加藤直人．同時通訳コーパス表示システム．言語処理学会本年次大会， 2005.