同時通訳コーパス表示システム

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

1.　はじめに

講演などの通訳では講演者の発話とほぼ同時に翻訳していく同時通訳が求められる。講演における一文は比較的長いため、通訳者は文の終了を待たず、途中の適当なタイミングで「翻訳単位」を見つけ出し訳出を開始している。そのような通訳者の翻訳単位を分析することは講演を対象とした機械翻訳システムを構築する上で参考となるだろう。

我々は、同時通訳コーパスのデータから発話のどの部分をどのタイミングで訳出しているかを表示するシステムを作成した。システムは、講演者と通訳者の発話をポーズごとに分割したフレーズを入力とし、自動的にフレーズ単位の訳出部分を求めて表示する。本稿では、ツールとしての同時通訳コーパス表示システムを紹介し、フレーズ単位に訳出部分を求める方法については文献[1]を参照されたい。

2.　同時通訳コーパス

我々の同時通訳コーパスは、NHKの解説番組「あすを読む」の日本語とその同時通訳(英語)、それぞれの音声データから書き起こした日本語テキスト・同時通訳テキストと、日本語テキストを翻訳した翻訳テキストの３種類のテキストからなる。番組は一人の話者による10分間の解説である。コーパスは250番組からなり、番組あたりの文数の平均は、日本語が60文、同時通訳が76文、翻訳が60文である。番組が扱う話題は、政治・経済・社会・文化と多岐にわたる。(同時通訳の収録条件などについては文献[2]を参照)

日本語と同時通訳のテキストは、ポーズにより発話をフレーズに分割し、それらに時間情報(10ミリ秒単位の開始・終了時刻)を付与したものとなっている。図１は番組の冒頭部分のコーパスの例である。各行がフレーズで、左欄の数値が開始・終了時刻である。

翻訳テキストの作成では、日本語の一文を「節」に分割し[3]、翻訳者は「節」を意識して翻訳した。そのため翻訳テキストには翻訳のどの部分が日本語のどの「節」に対応してるかという情報が示されている。図２は翻訳テキストの例で、図１と同じ番組部分である。日本語の各行が「節」を表し、英語の行頭の番号はどの番号の日本語「節」の翻訳かを示している。

同時通訳テキストと日本語テキストのフレーズの間には、翻訳テキストの「節」番号ような翻訳の対応関係はつけられていない。

```
2232  -  2583 今晩は。
3913  -  7333 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
7676  -  10295 私<watakushi>達の経済ですとか社会の仕組みを今
10554 -  12058 大きくかえ始めています。
12708 -  13442 そこで今晩は
13888 -  15892 経済の動きを中心にしまして
16117 -  17372 アメリカの例もみながら
17558 -  19099 今後の動きを探ってみようと思います。


3028  -  3647  good evening./
6633  -  9848  Internet information network based on computer
10762 -  11284 is
11367 -  12667 dramatically changing
12890 -  16129 the way of life and the way, the society works today./
16647 -  17062 so
17220 -  17726 today
17806 -  18845 take an look at {||the}
18947 -  22268 the experience in the United States, especially in the field of
economy./
```

図１同時通訳コーパス

```
1, 今晩は。/文末/
2, コンピューターを利用しました/連体節/
3, インターネットという情報ネットワークが私達の経済ですとか/並列節トカ/
4, 社会の仕組みを今大きくかえ始めています。/文末/
5, そこで/談話標識/
6, 今晩は/主題ハ/
7, 経済の動きを中心にしまして/テ節/
8, アメリカの例もみながら/ナガラ節/
9, 今後の動きを探ってみようと/引用節/
10, 思います。/文末/

1, good evening./
3, an information network
2, based on computer
3, called Internet is
4, dramatically changing
3, our economic and
4, social systems./
5, so
6, this evening
9, I would like to take a look at the trends toward the future,
7, focusing on the economic trends and
8, referring to the cases in the United States./
```

図２節分割と翻訳テキスト

3.　同時通訳コーパス表示システム

我々の表示システムは、コーパスの日本語テキストと同時通訳テキストを入力とし、両者の対応関係をフレーズ単位で求め、訳の対応を表示する。表示システムの設計においては、松原らによるコーパス構築支援システム[3]を参考にした。

3.1　日本語と同時通訳の対応づけ

目標とする表示システムを実現するためには、日本語と同時通訳のフレーズの対応関係を求めなければならい。その方法として、単語レベルの日英アライメントを取り、アライメントの良いフレーズどうしを対応づけるという方法が考えられるが、意訳や省略などで対応づけが難しい場合がある。一方、翻訳テキストは、通訳に比べると日本語を忠実に訳しているので対応関係を求め易い。そこで、まず日本語のフレーズを翻訳テキストと対応づけ、翻訳テキストから通訳のフレーズへ対応づけることで、日本語と通訳のフレーズの対応関係を求めることにした。表示システムでは次のような処理を行う：

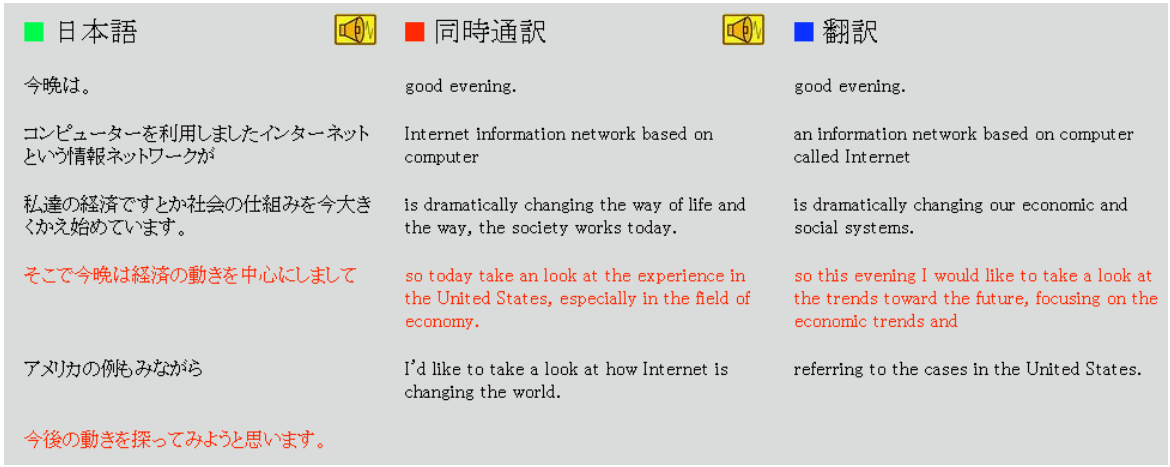

図3テキスト翻訳を介した日本語・同時通訳の訳出関係の表示

(1) 翻訳テキストから日本語フレーズの対訳となる部分を、日英単語アライメントを求めて推定する。

(2) 同時通訳テキストと翻訳テキストの同じ翻訳部分を推定する

(3) (1)(2)から翻訳テキストを介して日本語フレーズと同時通訳フレーズの対応関係を求める

(1)では、5万語規模の和英辞典などを使って節ごとに日英単語アライメントを行っている。翻訳テキストでは「節」単位に翻訳部分が分かっているので単語アライメントも節単位に行うことができる。(2)では、表層形のDPマッチングなどで同じ翻訳部分を求めている。(3)では「節」分割とフレーズ分割は必ずしも一致しないことに注意しながら目的の対応関係を求めている。これらの処理の詳細については文献[1]を参照されたい。

図3は、システムに日本語と同時通訳の対応づけを表示させたものである。同じ行に配置されているテキストが対応している部分である(連続したフレーズが同じ翻訳部分に対応づけられている場合は、一行にまとめらている)。ただし、複数対複数の対応もあるので、マウスカーソルをテキストに置くと、その翻訳部分がハイライト表示するようになっている。図3では「そこで今晩は―」の部分の日本語の2つの行が通訳の一行に対応していること表示している。また、日本語と英語で順序が逆になる場合は、日本語の順序に合わせて英語をずらして表示する。

3.2 時間情報の表示

表示システムでは、フレーズの発話時刻が見えるようにテキストを配置表示するモードも用意した。図4はその表示例で、実際の発話時間を縦方向の帯で表現しいる(図3の表示と同じく、対応する翻訳がハイライト表示される)。

3.3 コーパスの編集

日本語と通訳の自動対応づけは必ずしも正確ではないため、人手により誤りを修正するGUIを用意した。修正結果を保存することで同時通訳コーパスを更新する。

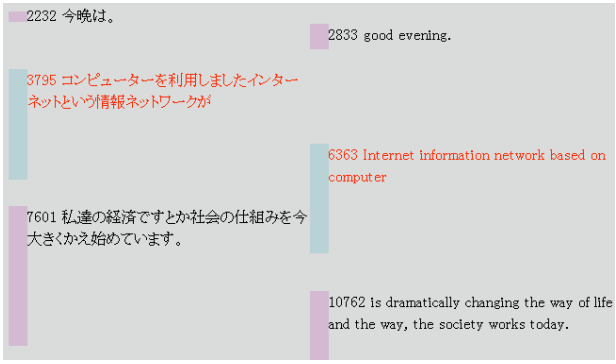

図4同時通訳訳の時間情報の表示

4. おわりに

本稿では、我々が実現した同時通訳コーパスのための表示システムについて述べた。日本語と通訳との自動対応づけは、通訳データベース作成の支援になると考えている。また、そうして作成されたものは同時通訳者の育成支援に使用することも期待できる。課題としては、日英単語アライメントが上げられる。「あすを読む」の語彙は広範囲に渡るため、現在の規模の辞書では単語アライメントがうまく取れていない場合が多く、辞書の拡充が必要である。

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。「あすを読む」はNHKとの共同研究として使用させていただいている。

参考文献

[1] 加藤, 渦原“日英同時通訳者の翻訳単位”言語処理学会第11回年次大会 (2005)

[2] 柏岡“講演同時通訳アライメント”言語処理学会第8回年次大会 pp188-191 (2002)

[3] 丸山, 熊野, 柏岡“日本語における独和の特徴と文分割”言語処理学会第7回年次大会 pp429-432 (2001)

[4] 松原ほか“同時通訳コーパスの設計と構築”通訳研究 No.1 pp85-102 (2001)