# 同時通訳コーパスの対訳アライメント手法とその評価

高木 亮† 松原 茂樹‡ 稲垣 康善†

†名古屋大学大学院工学研究科 ‡名古屋大学言語文化部

atakagi@inagaki.nuie.nagoya-u.ac.jp

## 1 はじめに

同時通訳システムの実現に向けて，同時通訳者の通訳テクニックやノウハウを調査し，それを活用することは有効な方法である．このような背景のもと，著者らは，同時通訳音声の収集を行っており，これまでに，数十万形態素規模のデータベースを構築した [4]．通訳者音声を詳細に分析するには，話者の発話とその通訳者の発話が比較的小さな単位で対応づけされていることが望ましい [1]．しかし，コーパスの規模が大きくなれば，それを人手で行うのは容易ではない．

本稿では，同時通訳コーパスの対訳アライメント手法を提案する．本手法では，対訳コーパスの自動アライメントに関する従来の手法 [2, 6] で用いられている語彙情報に加え，発話の開始時刻に関する時間情報 [5]，ならびに，発話の長さに関する情報を利用してアライメントを行う．また，対訳アライメント実験を行い，本手法の有効性を評価した．

## 2 同時通訳コーパス

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コーパスを構築した [4]．書き起こしでは，発話を 200ms 以上のポーズで単位分割し，各発話単位に対して，発話の開始時刻，及び終了時刻を付与している．図 1 に日本語独話音声とその同時通訳音声の書き起こしデータの一部を示す．左側が話者発話，右側がその通訳者発話である．

| 日本語話者発話 | 日英通訳者発話 |
|---|---|
| 0001 - 00:05:056-00:06:304 N: 英語講演のですね | 0001 - 00:07:744-00:09:157 I: (F uh) The simultaneous |
| 0002 - 00:06:768-00:07:984 N: 同時通訳システム | 0002 - 00:09:358-00:10:944 I: (W a) translation system of |
| 0003 - 00:08:432-00:09:488 N: いうことについて | 0003 - 00:11:448-00:12:447 I: English lecture |
| 0004 - 00:09:752-00:11:783 N: (F えー) ご説明いたします ⟨SB⟩ | 0004 - 00:12:896-00:14:895 I: is going to be elaborated ⟨SB⟩ |

図 1: 同時通訳コーパスの書き起こしデータ (一部)

| | |
|---|---|
| 英語講演のですね 同時通訳システム いうことについて | ご説明します。 |
| The simultaneous transration system of English lecture | is going to be elaborated. |

図 2: 対訳コーパスのアライメント

## 3 対訳アライメント手法

従来の対訳アライメント手法 [2, 6] では，多くの場合，対訳語情報を用いてアライメントを行っている．また，著者らはすでに，同時通訳コーパスに対して，話者発話と通訳者発話の開始時刻情報を用いたアライメント手法を提案している [5]．本手法では，さらに，発話の長さに関する情報を活用する．すなわち，本手法では次の 3 つの制約を仮定する．

**語彙的制約** 話者発話中の単語の対訳語が通訳者発話中に存在する．

**開始時刻制約** 話者より先に通訳者が発話を開始することはなく，かつ，通訳者は，話者が発話を開始した後，一定時間 $\theta$ 秒以内に発話を開始する．

**発話時間長制約** 通訳者発話の長さは，話者発話の長さに依存し，その長さの比は正規分布に従う．

話者発話 $S_i$ と通訳者発話 $T_i$ の対 $p_i = (S_i, T_i)$ に対して，次の式 (1) によって対応度 $h(p_i)$ を与える．ここで，$l(p_i)$, $t_1(p_i)$, $t_2(p_i)$ は，それぞれ語彙情報，開始時刻情報，発話時間長情報に基づいて式 (2),(3),(4) によって定める対応度を表す[1]．

$$h(p_i) = t_1(p_i) \times \{\alpha l(p_i) + (1-\alpha) t_2(p_i)\} \quad (1)$$

$\alpha$：重み係数 ただし，$0 \leq \alpha \leq 1$

$$l(p_i) = 2f_{ST}/(f_S + f_T) \quad (2)$$

$f_S$：$S_i$に含まれる自立語数, $f_T$：$T_i$に含まれる自立語数,

$f_{ST}$：$S_i$と$T_i$での対訳語数

$$t_1(p_i) = \begin{cases} 0 & (p_i\text{が開始時刻制約を満たす}) \\ 1 & (p_i\text{が開始時刻制約を満たさない}) \end{cases} \quad (3)$$

$$t_2(p_i) = e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (4)$$

$x$：$S_i$と$T_i$の発話時間長の比，$\mu$：$S_i$と$T_i$の発話時間長の比の分布の平均，$\sigma$：$S_i$と$T_i$の発話時間長の比の分布の標準偏差

対訳コーパス全体に対する対応度を部分的な対応に対する対応度の総和とし，それが最大となるように動的計画法によって最適化を行う [5]．

## 4 実験と評価

### 4.1 実験の概要

本手法の有効性を確認するために，名古屋大学 CIAIR 同時通訳データベース [4] に収録されている約 10 分間の日本語講演音声とその通訳音声からなる対訳テキスト 16 組を用いて，対訳アライメント実験を行った．正解データは人手で与えた．対応度を与える式 (4) での $\mu, \sigma$ の値

An Alignment Method of Simultaneous Interpreting Corpus and its Evaluation: Akira Takagi, Shigeki Matsubara and Yasuyoshi Inagaki (Nagoya University)

[1] 式 (4) は正規分布の密度関数の値 $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ を，それが 0 から 1 の範囲になるように正規化したことによる．

表 1: 単語レベルの判定による評価 (16 講演平均)

| 手法 | 適合率 | 再現率 | $F$ 値 |
|---|---|---|---|
| 語彙のみ | 69.2(%) | 27.5(%) | 39.0(%) |
| 語彙 + 開始時刻 | 75.0(%) | 34.9(%) | 47.3(%) |
| 本手法 | 73.9(%) | 39.4(%) | 51.2(%) |

表 2: 区切りレベルの判定による評価 (16 講演平均)

| 手法 | 適合率 | 再現率 | $F$ 値 |
|---|---|---|---|
| 語彙のみ | 20.8(%) | 46.8(%) | 28.7(%) |
| 語彙 + 開始時刻 | 23.5(%) | 49.5(%) | 31.8(%) |
| 本手法 | 24.9(%) | 37.5(%) | 29.5(%) |

は，上記のデータから計算し，$\mu = 1.19, \sigma = 0.55$ とした．また，開始時刻制約の $\theta$，式 (1) の重み係数 $\alpha$ を実験において評価値が最大となるように，$\theta = 8.0, \alpha = 0.8$ とした．比較のため，対訳語の一致による語彙情報のみを用いたアライメント実験 (すなわち，$t_1(p_i) = 1, \alpha = 1$)(以下，「**語彙のみ**」)，語彙情報，及び発話開始時刻情報を用いたアライメント (すなわち，$\alpha = 1$) (以下，「**語彙 + 開始時刻**」) 実験を併せて行った．

## 4.2 評価手法

書き言葉の対訳コーパスでは，対応先がないケースは稀であるが，同時通訳の場合，訳文が省略されたり，フィラーなど対訳語が出現しない現象が多数存在する．そこで本研究では，単語レベルの判定と区切りレベルの判定によって評価した．

**評価方法 1: 単語レベルの判定** 話者発話に含まれる単語と対応する通訳者発話に含まれる単語を対として，正解データと比べたときの単語対の数の一致度によって判定する [3].

**評価方法 2: 区切りレベルの判定** 対訳テキストの区切りの位置 (図 2の実線) を正解データと比べたときの一致度によって判定する．

区切りレベルでの判定を考慮することにより，対訳対応先がない発話に対するアライメント精度を評価結果に反映することができる．評価指標として，適合率，再現率，ならびにそれらの調和平均である $F$ 値を用いた．

## 4.3 結果

評価方法 1 による結果を表 1に示す．本手法では，$F$ 値が，「語彙のみ」に対して 12.2%，「語彙 + 開始時刻」に対して 3.9% 上昇しており，語彙情報，開始時刻情報，発話時間長情報を併用する本手法の優位性が示された．

一方，評価方法 2，すなわち，区切りレベルでの判定では，「語彙 + 開始時刻」に対して，本手法における再現率が 12% 低く，結果としてF値が 2.3% 低下するに至っている．本手法では，発話時間長の比が正規分布に従うと仮定したが，対応先がない発話については，当然ながらその仮定に反する．発話時間長の制約を対応度の計算に用いることによって，対応する通訳発話に対訳語が存在しない場合でも対応付けが可能となるがその反面，そもそも対応先がない発話についても不要な対応付けを行う場合がある．話者発話に現れる単語の対訳語が通訳者発話中に存在しないとき，その原因が，対訳辞書データが不十分であることによるものか，あるいは，対応する発話が存在しないことによるものかを判定することが重要となる．

| | |
|---|---|
| 0146 - 04:18:760-04:20:367 N: 実験的なシステムですが<br>0147 - 04:20:688-04:21:231 N: ライナス (L と)<br>0148 - 04:21:800-04:24:432 N: (F えー) 呼ぶシステムの実現を行いました ⟨SB⟩ | 0105 - 04:20:752-04:24:096 I: (F ah) (W The) there is experimental system called LINAS⟨SB⟩<br>0106 - 04:24:528-04:28:391 I: And we (W are;have) (D ak) already implemented those experimental system called LINAS⟨SB⟩ |

図 3: 人手の結果において対応先の判定が異なる例

実験では，人手で行ったアライメント結果を正解データとして評価したが，人間が行っても対応先の判断が難しい箇所が多数あり，例えば，通訳者が言い直しを行ったときに，言い直された部分 (図 3, 英語発話 ID:0105) をどのように対応づけるかは作業者間で一致せず，適切な正解データの作成は容易ではない．そこで，対訳コーパス中で人間が見て明らかに区切りであると判断できる箇所に対する本手法の再現性を調べるために，同一講演に対し，正解データを作成した人と異なる 2 人により対訳アライメントデータを作成した．なお，人手によるアライメントは，本学大学院国際言語文化研究科の大学院生に依頼して実施した．人手で行った合計 3 つの対応づけ結果において，いづれも同じ対応づけである部分を正解データとしたとき，本手法の評価方法 2 に基づく再現率は，49.1% であり，表 2に示した結果に対して 11.6% 上昇した．「語彙 + 開始時刻」と比べてもその差は 1.7% に縮小しており，必ずしも劣っているわけではないことが示された．

## 5 おわりに

本稿では，同時通訳コーパスの対訳アライメント手法とその評価について述べた．本手法では，語彙情報，開始時刻情報に加えて，発話時間長情報を用いてアライメントを行う．また，人手で行ったアライメント結果との比較を通して本手法の評価を試みた．

## 参考文献


[1] Y.Aizawa, S.Matsubara, N.Kawaguchi, K.Toyama, Y.Inagaki, "Spoken Language Corpus for Machine Interpretation Research", *ICSLP-2000*, Vol.III, pp.398-401, (2000).

[2] N.Collier, K.Ono, H.Hirakawa, "An Experiment in Hybrid Dictionary and Statistical Sentence Alignment", *COLING-ACL'98*, Vol.1, pp.268-274 (1998).

[3] P.Langlais, M.Simard, J.Veronis, "Methods and Practical Issues in Evaluating Alignment Techniques", *COLING-ACL '98*, Vol.1, pp.711-717 (1998).

[4] 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).

[5] 高木亮, 松原茂樹, 稲垣康善, "同時通訳コーパスにおける発話対応関係の推定", 情報処理学会第 63 回全国大会講演論文集 (2), pp.249-250 (2001).

[6] 宇津呂武仁, 松本裕治, "対訳辞書および統計情報を用いた二言語対訳テキスト照合", コンピュータソフトウェア, Vol.12, No.5, pp.12-21 (1995).