Vol.2014-NL-219 No.3
2014/12/16

# 同時音声翻訳における翻訳精度と遅延時間を同時に考慮した評価尺度

三重野 隆史[1,a)]　Graham Neubig[1,b)]　Sakriani Sakti[1,c)]　戸田 智基[1,d)]　中村 哲[1,e)]

**概要**：音声翻訳には，音声認識から翻訳された文を出力するまでにかかる遅延時間が存在する．近年注目されている同時音声翻訳では，文末を待たずに翻訳を開始することができ，遅延時間を短縮することができる．この同時音声翻訳において，翻訳精度と遅延時間を同時に考慮した評価方法は確立しておらず，システムの最適化を行うことが困難である．そこで本研究では，同時音声翻訳システムの自動評価や最適化を可能にするために， 人手評価に基づいて評価を行い，新しい評価尺度の提案を行う．

## 1. はじめに

音声翻訳は，ある言語の音声を異なる言語の音声に翻訳する技術であり，長年の研究によりその性能は改善しつつある．しかし，文単位で翻訳する従来の音声翻訳 [8] が講演のような発話が長い場面に使用される場合，発話開始から翻訳開始までの時間（以降，遅延時間）が長くなる．このため，翻訳内容と講演者の身振りやスライドなどの表示内容とのずれが生じ，講演全体が理解し辛くなる．

この遅延時間の問題を解決するために，同時音声翻訳の研究が行われている [1], [11] ．同時音声翻訳は文単位で翻訳する従来の音声翻訳とは異なり，文の途中で翻訳を開始するため，遅延時間を短縮することができる．同時音声翻訳で重要となるのは，翻訳精度をできるだけ維持しつつ，遅延時間を短縮することであり，従来の研究では遅延時間を短縮する様々な文分割法の提案がなされてきた．

しかしながら，遅延時間を減らせば減らすほど，翻訳に利用できる文脈情報も減るため，図 1 の例で示すように遅延時間を短縮すると翻訳精度が劣化することも知られている [4]．この中で同時音声翻訳における翻訳精度と遅延時間の相対的な重要度は比較的言及されてこなかった．つまり，人にとって遅延時間は翻訳精度，または翻訳精度は遅延時間と比べて，どの程度重要であるかということである．このことから，各遅延時間と翻訳精度を備えたどのシステムが，人にとって最適であるかは明らかではない．

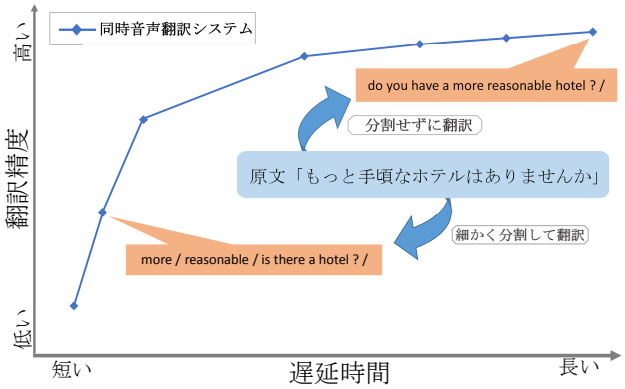


**図 1**　同時音声翻訳における遅延時間と翻訳精度の関係性の例

そこで本研究では，同時音声翻訳システムの自動評価や最適化を可能にするために，翻訳精度と遅延時間を同時に考慮した評価尺度の作成方法を提案する．具体的には，同一の入力動画に対して，精度の異なる複数の翻訳結果を作成し，動画に話者の実際の発話より遅れて提示する．こうすることにより，同一の動画に対して様々な翻訳精度と遅延時間を持った動画が得られ，これらを被験者に見せてランク形式で評価を行ってもらう．そして，遅延時間と翻訳精度を入力として人手評価結果を推定するランキング学習の問題として定式化し，評価関数を学習する．

実験では，評価の対象に TED 講演 [*1] を用い，英日方向で翻訳された結果を字幕出力として被験者に提示した．評価データの翻訳精度の素性として人手評価（5 段階評価）と自動評価を用いる．比較検証を行った結果，人手評価を用いた場合，遅延時間と翻訳精度を入力とした評価関数が

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] mieno.takashi.mh1@is.naist.jp
[b)] neubig@is.naist.jp
[c)] ssakti@is.naist.jp
[d)] tomoki@is.naist.jp
[e)] s-nakamura@is.naist.jp

[*1] http://www.ted.com

実際に動画を見た評価者の主観と最も高い一致率を示した．また，先行研究で仮定されてきた，翻訳精度と遅延時間のトレードオフを客観的に裏付ける結果にもなった．しかし，自動評価尺度を用いた結果は人手評価に比べて精度が低く課題は残る．

## 2. 評価関数

同時音声翻訳における翻訳精度と遅延時間を同時に考慮した評価を行うために，任意の同時音声翻訳結果が与えられたとき，主観評価と相関のある評価スコアを返す評価関数 $S$ を式 (1) のように定義する．

$$\boldsymbol{S} = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \qquad (1)$$

ここで，$\boldsymbol{\phi}$ は $\boldsymbol{x}$ から同時音声翻訳の評価に有用な素性を計算する関数である．本稿で $\boldsymbol{\phi}(\boldsymbol{x})$ を遅延時間と翻訳精度という 2 つの値を計算し，ベクトルとして返す関数とする．[*2]$\boldsymbol{w}$ はこの素性の相対的な重要度を表す重みベクトルである．本研究の目標は，この重みベクトルをデータに基づいて推定することで，同時音声翻訳において遅延時間と翻訳精度が聞き手の主観に与える影響を明らかにすることであり，次節以降にその具体的な手続きを説明する．

## 3. 評価データの収集法

本節では，前節で述べた評価関数の推定とメタ評価に利用するための，同時音声翻訳の翻訳精度と遅延時間を同時に考慮した人手評価データの収集法を記述する．

### 3.1 評価データの形式

2 節の自動評価関数は動画 $\boldsymbol{x}$ を受け取り，スコア $\boldsymbol{S}$ を返す．この関数を学習するデータを作成する方法として，まず，評価者に動画を視聴してもらい，スコア $\boldsymbol{S}$ を直接 5 段階評価などで評価付ける方法が考えられる．しかし，翻訳精度と遅延時間を総合的に評価する人手評価指標は確立しておらず，その設計が容易ではない．

そこで本研究では，$\boldsymbol{S}$ を直接付与する絶対評価ではなく，複数の候補を比較して評価する相対評価を採用することでこの問題を回避する．具体的には，同一の動画に対して，複数の異なった翻訳精度と遅延時間を持った翻訳結果を評価者に見せ，理解のしやすい順にランク付けを行ってもらう方法を用いる．表 1 に，評価の結果得られるデータの例を示す．

### 3.2 データの作成

1 つの動画を作成するために，まず平均文数 4〜5 文程度となるように動画の一部を選択し，切り出す．[*3] 文数は原

[*2] つまり，線形モデルに限定される．
[*3] 4〜5 文を利用する理由は，評価文数が多すぎる場合，被験者に負担がかかりすぎて評価が曖昧になることを回避するためである．

表 1 評価の結果得られるデータの例

| 事例 | ランク | | |
|---|---|---|---|
| | 1<br>遅延/精度 | 2<br>遅延/精度 | 3<br>遅延/精度 |
| 1 | 2/0.13 | 3/0.14 | 7/0.16 |
| 2 | 1/0.30 | 2/0.22 | 2/0.30 |
| 3 | 1/0.15 | 5/0.15 | 5/0.35 |

表 2 提示するデータの例

| | | |
|---|---|---|
| 1 | 時間 (sec)<br>提示文 | 00:00:00.100 - 00:00:03.000<br>去年 この 2 つのスライドをお見せして |
| 2 | 時間 (sec)<br>提示文 | 00:00:03.000 - 00:00:05.000<br>過去 3 百万年 アラスカとハワイを除く米国と |
| 3 | 時間 (sec)<br>提示文 | 00:00:05.000 - 00:00:07.000<br>同じ面積があった極域の氷河が |

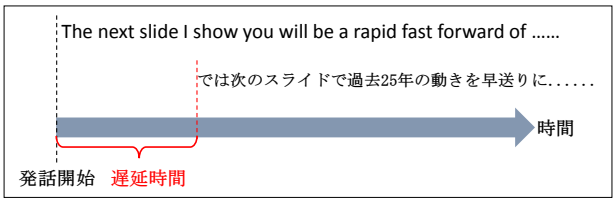


図 2 遅延時間の例

文である英文のピリオドを基準に算出する．選択する基準は，なるべくそれ以前の内容に依存せず，発話開始のタイミングが明確であることを重視する．

### 3.3 翻訳結果の提示

次に，実際に評価者に見せる動画を作成する．本研究では，翻訳結果の出力を字幕データとして，被験者に提示する．音声データではなく字幕データを用いた理由は，2 つある．1 つ目は，収録音声を用いた場合，不均一な声色とイントネーションにより，評価条件を均一に保つことが困難なことである．2 つ目は，合成音声を用いた場合，合成音の明瞭性や流暢性の低さから，対象となる翻訳精度と遅延時間の評価が困難となるためである．これらの要因の扱いは今後の重要な課題であるが本稿では評価データを作成する際に，音声の個人性や明瞭性などに評価が左右されない字幕データを用いる．

表 2 に，提示するデータの一例を示す．データにはそれぞれ字幕を表示する時間が，00 : 00 : 00.100−00 : 00 : 03.000，のように与えられている．更に，3.2 項で作成した翻訳結果の提示には，無作為に選択された遅延時間を付加し，切り出した動画の開始時点を遅延 0 秒として動画の上に表示する．本研究において遅延時間とは，講演者の発話開始から翻訳データの提示までに要した時間とすることに注意されたい．具体例を図 2 に示す．図から分かるように，仮に 20 秒の動画を選択した場合，翻訳結果の提示に遅延時間を 5 秒設けると，その動画は合計 25 秒の動画となる．た

だし，この場合，伸びた表示の時間だけ提示する動画の長さを伸ばすこととする．

### 3.4 動画の評価

動画の評価には，理解のしやすい順にランクを付ける方法を用いる．具体的には，ひとつの画面に異なる翻訳精度と遅延時間を持つ同一の動画を複数提示し，被験者に任意のタイミングで視聴して貰いランク付けを行う．このとき，正確な評価データを得るために同一の動画に関しては何度でも視聴し比較することは可能とする．

## 4. ランキング学習による重みの推定

前節で述べたデータを用いて重みベクトルを推定する．重みベクトルの推定にはランキング学習を使用する．ランキング学習の目的は，提示された動画から抽出された素性ベクトル（本稿では翻訳精度と遅延時間）に基づき，各動画に対するランキングを出力することである．ランキング学習の学習データは，動画から抽出された素性ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ と評価者により判定されたランク $y_i \in \{1, 2, ...\}$ のペア集合 $\{(\boldsymbol{\phi}(\boldsymbol{x}_i), y_i)\}_{i=1}^{m}$ により構成される．ランキング学習では，素性ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ のランクが高い（つまり，数字が低い）ほど大きな値を出力する関数 $f : \boldsymbol{S}$ を作成することが目標となる．

関数 $f$ を $f(\boldsymbol{\phi}(\boldsymbol{x})) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$ とすると，ランキング学習は各インスタンスのペア $(i, j)$, $\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}(\boldsymbol{x}_j)$ に関して，

$$\begin{aligned} y_i < y_j \quad &\Leftrightarrow \quad f(\boldsymbol{\phi}(\boldsymbol{x}_i)) > f(\boldsymbol{\phi}(\boldsymbol{x}_j)) \\ &\Leftrightarrow \quad \boldsymbol{w}^T(\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_i)) \end{aligned}$$

となる重みベクトル $\boldsymbol{w}$ を求めることになる．このようなベクトルを適切に学習するため，各素性ベクトルのペアを考え，新たに $\{(\boldsymbol{\phi}(\boldsymbol{x}_{i1}) - \boldsymbol{\phi}(\boldsymbol{x}_{i2})), z_i\}_{i=1}^{n}$ を作成する．ここで，

$$z_i = \begin{cases} +1 & y_{i1} < y_{i2} \\ -1 & y_{i1} > y_{i2} \end{cases}$$

であり，$n$ は全ての可能なペア数を表す．この新たなデータを学習データとして 2 クラス分類問題を解くことによって上記の大小関係を満たす関数を学習することができる．その際に，同じペアで順序を入れ替えただけのペアは境界からの距離が等しいという特徴を利用し，$z = +1$ のペアのみ学習に利用する．

## 5. 実験的評価

本節では，実験設定および実験結果について記述する．

### 5.1 実験設定

**【評価データ】**

評価データには TED 講演を使用し，被験者は同一の動画に対して英日方向に翻訳された結果を付与した 3 つの字幕付き動画を視聴しながら内容の理解しやすさを基準に 1 から 3 のランク付けを行ってもらった．ランク付けを被験者 10 人が行い，被験者は全て日本語を母国語とする．

表 3　評価データの翻訳精度

| | TED | S-rank | Travatar |
|---|---|---|---|
| BLEU+1 | 0.23 | 0.18 | 0.22 |
| RIBES | 0.71 | 0.59 | 0.68 |
| 人手評価 (5 段階) | 3.85 | 3.01 | 2.15 |

TED 講演を選んだ理由は，2 つ挙げられる．1 つ目は，リアルタイム性の高い動画であるからである．動画には講演者の身振りやスライドが含まれており，遅延時間が長くなると翻訳内容と講演者の身振りやスライドなどの表示内容とのずれが生じ易いため，同時音声翻訳の評価タスクに適している．2 つ目は，TED 講演が機械翻訳の性能を評価する際のテストセットとして頻繁に使用されているからである．

3.2 項のもと，20 秒から 30 秒程度の動画を 10 種類用意し，無作為に選ばれた翻訳結果と遅延時間を付加した字幕データを与えた．今回の評価データでは，1 動画の平均文数は約 4.3 文となった．また，動画にはスライドを含むもの（翻訳データと表示内容の遅延が分かりやすいもの）と，スライドを含まないものを同数用いた．

**【翻訳データ】**

翻訳データには，TED の字幕データ，S ランク（通訳経験年数 15 年）[13] の同時通訳者が同時通訳を行なった際の書起しデータ，機械翻訳システム Travatar[9] の翻訳データの 3 種類を用いた．字幕データの表示時間には，TED よりダウンロードできる字幕データを元に作成した．講演者の発話タイミングと字幕の表示タイミングに大きなずれが無いことをあらかじめ確認した．

翻訳精度には，自動評価尺度 BLEU+1[7] 及び RIBES[6]，人手評価の 3 つを用いた．人手評価には忠実性を 5 段階評価 [2] で被験者 5 人に評価を行ってもらい，その結果を加算平均して用いた．なお，学習の際にすべての翻訳精度が同じスケールになるように人手評価の 5 段階を 0-1 の間になるように正規化する．自動評価を計算する際に，日本語の単語分割には KyTea[10] を使用した．参照訳には TED の字幕データとは異なる翻訳者の翻訳結果を用いた．各評価データの翻訳精度を表 3 に示す．

**【遅延時間】**

遅延時間は秒単位で，$\boldsymbol{D} = \{0, 1, 2, 3, 5, 7, 10\}$ の 7 種類で与えた．3.3 項で示したように，今回の評価データにおいて遅延時間は発話開始からの時間とした．

**【学習・評価】**

学習器には LIBLINEAR[3] を用いた．正則化係数を調整したところデフォルトの 1 で最も高い精度となったため，

表 4　重み $\boldsymbol{w}$ と分類精度 Acc：遅延時間 (D)，翻訳精度 (A)

| 素性 | 評価尺度 | $\boldsymbol{w}$ | | $\boldsymbol{w}$ の比 | | Acc |
|---|---|---|---|---|---|---|
| | | 遅延 | 精度 | 平均 | 分散 | |
| D | - | -0.09 | - | - | - | 0.67 |
| A | BLEU+1 | - | 0.98 | - | - | 0.50 |
| | RIBES | - | -0.02 | - | - | 0.44 |
| | 人手評価 | - | 2.07 | - | - | 0.71 |
| D+A | BLEU+1 | -0.09 | 0.67 | 7.2 | 22.0 | 0.66 |
| | RIBES | -0.09 | 0.06 | 0.6 | 2.0 | 0.67 |
| | 人手評価 | -0.10 | 2.27 | 22.0 | 1.4 | 0.81 |

学習器の結果の評価方法を述べる諸設定はデフォルトのままとした．

### 5.2　実験結果

ランキング学習の結果，得られた翻訳精度と遅延時間の重み及びその比と分類精度 Acc を表 4 に示す．D は素性に遅延時間のみを用いた場合，A は素性に翻訳精度のみを用いた場合，D+A は素性に遅延時間と翻訳精度を用いた場合をそれぞれ表す．ここで，Acc はランク正解率を示しており，チャンスレートは 0.5 である．遅延時間及び翻訳精度の重みは，各動画の平均値を表しており，重み $\boldsymbol{w}$ の比は翻訳精度の重みを遅延時間の重みで割ったものである．

この結果からまず，遅延時間のみもしくは人手評価による翻訳精度のみを素性とした場合には，分類精度がチャンスレートを上回っていることが分かる．このことから，人手評価による翻訳精度と，遅延時間は素性として有効であるといえる．更に人手評価の場合，遅延時間と翻訳精度を同時に素性とすることにより分類精度が更に上昇することが確認された．これは，同時音声翻訳システムの評価の際に，翻訳精度と遅延時間を同時に考慮することの有用性を示している．

翻訳精度のみを素性としたとき，自動評価尺度の分類精度はチャンスレートを上回らず，動画に対する評価との相関が無いことが分かった．つまり，既存の自動評価尺度だけでは同時音声翻訳システムを主観に基づいた評価ができないことが明らかになった．これは，$n$-gram 一致率を用いて測るような既存の自動評価尺度では，意訳を考慮することができないため，訳出の意味が人にとっては妥当だとしても，参考文と異なった言い回しをしていれば不当に評価が下がるためだと考えられる．具体例を表 5 に示す．この表から分かるように，BLEU+1 及び RIBES では TED と Travatar の翻訳精度は同程度であるが，人手評価では TED の翻訳結果ほうが Travatar よりも良いと判断されている．

$\boldsymbol{w}$ の比で分かるように人手評価の場合，翻訳精度と遅延時間の間ではトレードオフの関係性が見られた．人手評価に 5 段階評価を用いたため，翻訳精度が 1 段階あがることは同じ翻訳システムに 5.5 秒の遅延時間を加えることと同じであると示している．今回の評価実験により得られた評価関数 $S$ をヒートマップとして図 3 に示す．ヒートマップの左上に行くに従い評価スコアが高くなり，右下に向かうほど評価スコアが低くなっていることが分かる．今後の同時音声翻訳に関する研究では，図 1 のようなシステムごとの遅延時間と翻訳精度を表したグラフを上記のヒートマップと照らし合わせることで，実際にどのシステムが最も主観的に良いかがある程度明らかになる．ただし，このようなヒートマップは言語対，分野，提示法に依存することも容易に考えられ，これらの影響を調べるのは重要な課題である．

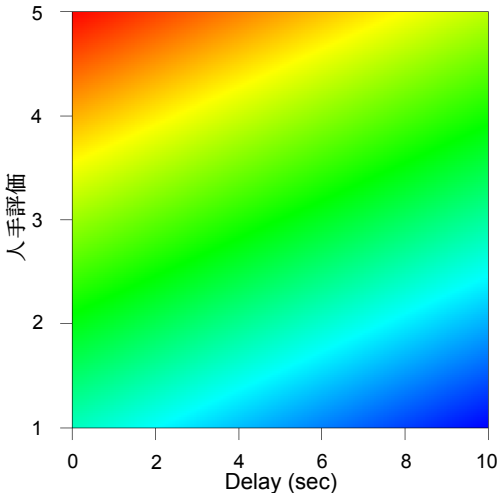


図 3　評価関数によって得られたヒートマップ (人手評価)
$S = -0.1 * \text{遅延時間} + 2.27 * (\text{翻訳精度} * 0.25 - 0.25)$

## 6.　関連研究

同時音声翻訳のための文分割位置に関する研究は，近年になっていくつか提案されている [1], [11], [12]．しかし，ここに挙げたいずれの手法も，同時音声翻訳における翻訳精度を維持したまま遅延時間を短縮する手法の提案に留まり，翻訳精度と遅延時間の関係性に関しては言及してこなかった．文献 [5] は，遅延時間と翻訳精度を同時に評価する関数を提案しているが，強化学習の報酬関数に用いる目的で考案されたものであり，主観評価に基づいてその妥当性が議論されているわけではない．そこで本研究では，主観評価に基づいた同時音声翻訳システムの翻訳精度と遅延時間を同時に考慮して評価のできる新しい評価尺度の提案を行った．

## 7.　おわりに

本研究では，同時音声翻訳システムの評価手法として，翻訳精度と遅延時間を同時に考慮した評価方法の提案を行った．その結果，人手で測る翻訳精度と遅延時間は両方，同時音声翻訳の結果を付与した動画を視聴した被験者の主

表 5　人手評価と自動評価の例

| | 例文 | BLEU+1 | RIBES | 人手評価 |
|---|---|---|---|---|
| 原言語文 | Now, it wasn't until this point that I realized that these photos were such a huge part of the personal loss these people had felt. | - | - | - |
| 参照文 | この写真は被害者が受けた個人的なダメージの非常に大きな一部であることに、その時になって気づきました。 | - | - | - |
| TED 字幕 | この時 私は初めて気付いたのですが、これらの写真は被災者が味わった個人的な、喪失感の大きな部分を占めていたのです。 | 0.12 | 0.64 | 3.60 |
| S-rank | でこの時点で、得た写真は 持ち主にとっては大変大切なんだろうと思いました。 | 0.09 | 0.72 | 2.20 |
| Travatar | 今、この写真を自己喪失感じました人のような大の部分があるのに気づきましたのは、ここまでありませんでした。 | 0.16 | 0.64 | 1.20 |

観評価との相関が見られ，精度と遅延を同時に考慮した方が，高い評価精度が得られた．また，5 段階の人手評価と遅延時間はトレードオフの関係にあり，評価が 1 段階上がるごとに 5.5 秒の遅延が許される結果となった．今後の課題としては，自動評価における精度改善，非線形なモデルへの適用，音声データを用いた評価などが挙げられる．

## 謝辞

本研究の一部は JSPS 研究費 24240032 の助成を受け実施したものである．

### 参考文献

[1] Bangalore, S., Sridhar, V. K. R., Golipour, P. K. L. and Jimenez, A.: Real-time Incremental Speech-to-Speech Translation of Dialogs, *Proc. NAACL* (2012).
[2] DARPA: Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations (2002).
[3] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, Vol. 9 (2008).
[4] Fujita, T., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation, *Proc. 14th InterSpeech* (2013).
[5] Grissom II, A., He, H., Boyd-Graber, J., Morgan, J. and Daum'e III, H.: Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation, *Proc. EMNLP*, pp. 1342–1352 (2014).
[6] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs, *Proc. EMNLP*, pp. 944–952 (2010).
[7] Lin, C.-Y. and Och, F. J.: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation, *Proc. COLING*, pp. 501–507 (2004).
[8] Matusov, E., Mauser, A. and Ney, H.: Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation, *Proc. IWSLT*, pp. 158–165 (2006).
[9] Neubig, G.: Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers, *Proc. ACL*, pp. 91–96 (2013).
[10] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. ACL*, pp. 529–533 (2011).
[11] Oda, Y., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Optimizing Segmentation Strategies for Simultaneous Speech Translation, *Proc. ACL* (2014).
[12] Ryu, K., Mizuno, A., Matsubara, S. and Inagaki, Y.: Incremental Japanese spoken language generation in simultaneous machine interpretation, *In Proc. Asian Symposium on Natural Language Processing to Overcome language Barriers* (2004).
[13] Shimizu, H., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Constructing a Speech Translation System using Simultaneous Interpretation Data, *Proc. IWSLT* (2013).