4Q-01

話し言葉から書き言葉への変換における対訳単位分割

下田 裕晃† 住田 一男‡

† 東京工業大学 大学院総合理工学研究科 ‡ 東芝 研究開発センター

1 はじめに

話し言葉では書き言葉には無い表現があり，可読性が良くないという特徴がある．既存の翻訳ソフトは書き言葉を想定して開発されているので，音声認識結果を機械翻訳する場合には話し言葉を書き言葉に整形することで翻訳精度が向上すると期待される．話し言葉やその書き起こしを整形する研究としては文献[1]や文献[2]などがある．

本論文では，統計的機械翻訳(SMT)を用いて日本語の話し言葉を書き言葉に変換する．この場合，話し言葉と書き言葉間の対応データが SMT のための対訳コーパスとなる．文献[2]では文字単位で除去・置換などを行っているが，話し言葉から書き言葉への変換においては，不要な話し言葉を文節単位で除去することが必要になる場合がある．また，文末の言い回しや助詞の省略や変形など文節間の変換で対応可能な現象が多い．したがって，話し言葉と書き言葉間での対訳コーパスの単位は文節単位程度が望ましいと考えられる．一方，対訳コーパスを短い単位とした場合，話し言葉と書き言葉の文節間で対応が付かない状況が発生する．提案手法では，これへの対処として話し言葉と書き言葉の文節間で対応しない文節に対して特定の記号を割り当てる．文献[1]ではフィラーの除去のみが行われていたが，書き言葉に不要である話し言葉表現はフィラー以外にも存在することを考慮し，提案手法ではフィラー以外で書き言葉に不要な話し言葉表現の除去も行う．以下，本提案手法ならびに日本語話し言葉コーパス(CSJ)[1] から作成した対訳コーパスに基づいて行った実験結果について述べる．

2 提案手法

話し言葉と書き言葉との対訳コーパス作成においては，以下の要領で発言単位対訳を作成する(発言単位対訳コーパス)．

- 口語的表記や誤った表記，くだけた文末表現をフォーマルな話し言葉に修正する．
- 助詞が省略されている場合は補完する．
- 意味上は無視できる接続助詞を削除する．
- 語順や数字の表記は特に変更しない．

この対訳コーパスを元にして，発言よりも細かい単位で対訳を分割した対訳コーパス(非発言単位対訳コーパス)を新たに作成する．

2.1 非発言単位対訳コーパス

以下に示す手順で非発言単位対訳コーパスを作成する．また，図 2.1 に非発言単位対訳コーパスの作成手順の例を示す．

(1) 発言単位対訳コーパスから文単位対訳コーパスを作成する．
(2) CaboCha[2] を利用して文単位の対訳コーパスを文節ごとに区切り，そこから名詞ごとに区切った文節単位対訳コーパスを作成する．
(3) 文節単位の対訳コーパスから n 文節ずつ連結して n 文節単位対訳コーパスを作成する．この時，文をまたいでの文節連結はしない(n は正の整数)．

話し言葉	書き言葉
格好とかは何か背広とか着てますか	格好は．背広などを着ていますか.

↓文単位に分割

格好とかは	格好は.
何か背広とか着てますか	背広などを着ていますか.

↓文節単位に分割・名詞ごとに区切る

格好とかは	格好は.
何か	B(ブランク記号)
背広とか着てますか	背広などを着ていますか.

↓n 文節ずつ連結(ここでは $n = 2$)

格好とかは	格好は.
何か 背広とか着てますか	B 背広などを着ていますか.

図 2.1：非発言単位対訳コーパス作成手順の例

上記(2)において，書き言葉には不要である話し言葉表現が存在する場合がある．そのような

Split of Parallel Translation Unit for Conversion of Spoken Sentences to Written Sentences
Hiroaki SHIMODA †, Kazuo SUMITA‡
† Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
‡ Corporate R&D Center, Toshiba

[1] http://www.ninjal.ac.jp/corpus_center/csj/

[2] https://code.google.com/p/cabocha/

話し言葉の存在を明示するためにブランク記号を導入する(図 2.1 の網掛け部)．ここではブランク記号にアルファベットの”B”を用いている．ブランク記号を 1 つの文節として扱うことで，全ての話し言葉が書き言葉と対応付けされるようになる．

2.2 言語モデル

SRILM[3] を使用して，対訳コーパスにおける書き言葉側のコーパス(書き言葉コーパス)から言語モデルを作成する．1 つの書き言葉コーパスから 1 つの言語モデルを作成するので，発言単位書き言葉コーパスから言語モデルを 1 個作成し，非発言単位書き言葉コーパスから言語モデルを 4 個作成する．

2.3 変換における前処理・後処理

変換における前処理として，入力文を発言単位から文単位・文節単位・2 文節単位のいずれかに分割してから変換を行う．変換後は，出力文を発言単位に戻しブランク記号を削除する．

3 評価実験

3.1 実験内容

CSJ から抽出した 8995 発言(529KB)を用いて，評価実験を行った．400 発言を評価用データとし，残りの 8595 発言を学習用データとして利用した．使用した対訳コーパスは 6 種類であり，各コーパスについて，言語モデル 5 種類と前処理方法 4 通り(前処理しない場合も含む)との組み合わせ 20 通りのスコアを算出し，その最大値を比較した．変換精度の評価尺度は，NIST と BLEU を用いた．表 3.1 に，学習に用いた対訳コーパスの対訳単位とその対訳数を示す． n 文節単位対訳コーパスは，n = 1，2，3 の場合について調べた．また，ブランク記号の有無による変換精度への影響を調べるために，ブランク記号を除去した 3 文節単位対訳コーパスを用意した(B 無し 3 文節単位対訳コーパス)．翻訳モデルの生成には GIZA++[4] を，デコーダは Moses[5] を使用した．

3.2 実験結果と考察

図 3.1 に，実験結果のグラフを示す．文節単位対訳コーパスと 3 文節単位対訳コーパスにおいて，発言単位対訳コーパスを上回る変換精度を実現することが出来た．また 3 文節単位対訳コーパスと B 無し 3 文節単位対訳コーパスとを比較すると，前者の方が変換精度が高い．このことから，ブランク記号が変換精度の向上に貢献していると言える．

表 3.1：学習用対訳コーパスの対訳数

対訳単位	対訳数
発言(従来法)	8595
文	9777
文節	30307
2 文節	17626
3 文節	13996
B 無し 3 文節	13357

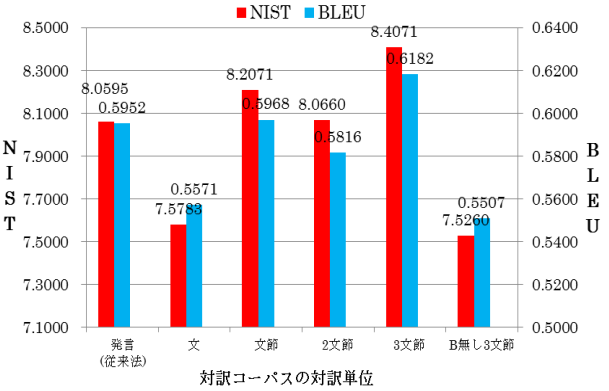

図 3.1：対訳コーパス別の変換精度

4 おわりに

対訳コーパスにおける対訳単位の分割を行い，話し言葉から書き言葉への変換を行った．その結果，対訳単位分割において導入したブランク記号によって不要な話し言葉表現を除去し，変換精度を向上させることができた．今後の課題としては，ブランク記号の割り当てをより正確なものにするために n 文節単位対訳コーパスの生成アルゴリズムを改良することである．また，n 文節単位対訳コーパスの n を更に大きくした場合の変換精度を調査し，対訳単位の長さと変換精度との関係を検証することも課題である．

参考文献

[1] 下岡 和也，南条 浩輝，河原 達也，”講演の書き起こしに対する統計的手法を用いた文体の整形” 自然言語処理，vol. 11，No. 2，pp. 67-83，2004.

[2] 佐々木 彬，水野 淳太，岡崎 直観，乾 健太郎，”機械学習に基づくマイクロブログ上のテキストの正規化”，人工知能学会第 27 回全国大会，4B1-4，June 2013.

[3] http://www.speech.sri.com/projects/srilm/
[4] https://code.google.com/p/giza-pp/
[5] http://www.statmt.org/moses/?n=Main.HomePage