

Internal Use Only (非公開)

TR-SLT-0090

会話文書き換えコーパスの構築
Building a Paraphrase Corpus

下畑 光夫
Mitsuo Shimohata

隅田 英一郎
Eiichiro Sumita

2005年3月24日

概要

本稿では、会話文とそれを人手により書き換えた文から構成される会話文書き換えコーパスについて報告する。会話文は SLDB,LDB,MAD1 を利用し、日本語、英語の両言語を書き換え対象としている。会話文は包含する内容語の数により短文と長文に分けられ、短文には簡潔書き換えを、長文には分割書き換えと要約書き換えを行った。また、会話文書き換えコーパスは、テストデータと学習データから構成されている。テストデータは数百文程度の分量であるが、1 発話文について 10 種類以上の書き換え文を備えている。学習データは数万文の分量からなる。

(株) 国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619-0288 「けいはんな学研都市」 光台二丁目 2 番地 2 TEL : 0774-95-1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone: +81-774-95-1301
Fax : +81-774-95-1308

©2005 (株) 国際電気通信基礎技術研究所
©2005 Advanced Telecommunications Research Institute International

目次

1	会話文書き換えコーパスの目的	2
2	会話文書き換えコーパスの概要	3
3	会話文書き換えコーパスの元文について	5
4	書き換え作業の概要	5
5	書き換え方式	7
5.1	簡潔書き換え	7
5.2	要約書き換え	7
5.3	分割書き換え	8
6	会話文書き換えコーパスの所在, フォーマット	8
6.1	ディレクトリ	8
6.2	ファイル	8
6.3	フォーマット	9
7	まとめ	10
A	簡潔書き換えの作業文書	12
B	要約書き換えの作業文書	14
C	分割書き換えの作業手順文書	16

1 会話文書き換えコーパスの目的

音声翻訳システムでは、音声発話を音声認識器で認識したテキストが翻訳対象となる。この音声認識から出力されたテキストは話し言葉の性質を反映しており、その性質が翻訳にとって不利となる事も多い。例えば、話し言葉は書き言葉と比較して非文法的な文や話し言葉特有の表現が多く発生することや、冗長な表現も頻発することが知られている。したがって、話し言葉のテキストをそのまま機械翻訳に適用すると訳質が低下してしまう場合が多くなる。

これを解消する一つの方策として、「入力文を、その主要な意味を損なわずに翻訳に有利に書き換える」ということが考えられる。入力文が持つ発話特有の表現を中立的なものに変換したり冗長な情報を削除することにより発話の性質を軽減し、機械翻訳しやすい文に変形する。また、この方策は入力文に対する前処理として実装するものであるため、目的言語に依存せずに効果を出すことも期待できる。

この方策の有効性を確かめるために、我々は以前にMAD1(機械翻訳支援による旅行対話データ)のテストセットを用いて予備実験を行った[1, 2]。その予備実験では、会話文を3方式の書き換え方式(本資料で紹介している方式と同じ)で人手で書き換え、用例翻訳(D3)と統計翻訳(SAT)を用いてその効果を検証した。結果としては、用例翻訳には顕著な効果がみられたが、統計翻訳にはあまり効果がみられなかった。しかし、統計翻訳においても訳質が全く変らなかったわけではなく、訳質が向上したものと低下したものが同程度発生したため、これらを合算すると全体効果が少ないという結果であった。したがって、適用する機械翻訳に合わせて書き換えを適切に制御し、訳質を低下させる書き換えを抑制することができれば、この方策は有効であると考えられる。

この予備実験のデータは全部で683文と少ないため詳しい分析や最適な書き換えの自動制御の研究まで進めるのは困難であり、さらなる研究開発のために会話文書き換えコーパスの分量を大幅に拡大することが望まれていた。このような目的の下、音声言語研究所が持つ会話データであるLDB, SLDB, MADに対して人手により3方式の書き換えを行った本コーパスが作成された。

なお、音声言語コミュニケーション研究所では名称が類似している日本語パラフレーズデータベース(TR-SLT-0036)[3][4]を構築している。この日本語パラフレーズデータベースは、ある英語文が与えられた時に日本語の訳文はどれくらいの種類が発生しうるかを網羅的に調査したデータである。パラフレーズデータベースは表現の多様性の分析が目的であり、入力文の表現の多様性に対してロバストな機械翻訳の開発に有用である。一方、本コーパスは、入力文である会話文を機械翻訳に有利なように自動的に書き換える方法のためのデータであり、日本語パラフレーズデータベースとは目的が異なる。

2 会話文書き換えコーパスの概要

会話文書き換えコーパスは、音声言語コミュニケーション研究所が保有する発話データである LDB, SLDB, MAD に対して 3 方式の書き換えを行ったコーパスである。対象言語は、日本語と英語である。本コーパスの概要図を図 1 に示す。ここで、書き換えの元となる会話文を元文と呼ぶ。

本コーパスは、テストデータと学習データから構成されている。テストデータは、日英とも 1,300 文程度と少量であるが、日本語では 12 種類¹、英語では 10 種類の書き換え文を備えていることが特徴である。書き換えは、個人差によるばらつきがある程度発生すると考えられるが、これら多数種の書き換え文があることにより、個人差によるばらつきを比較分析することが可能である。これら多数種の書き換え文により、書き換え文の自動評価も実現できると考えられる。また、テストデータは学習データ作成を担当する作業者の選別にも利用している。テストデータ作成を担当した作業者の中から、平均的な書き換え (他の作業者とも共通する書き換え文を多く作成) を行った作業者を選定し、学習データ作成を行っている。学習データは、日英とも約 4 万 5 千文と多くの文を元文としているが、種類は少なく、日本語で 2 種類、英語で 1 種類である。

元文は、包含する内容語の数が 4 語以下を短文、5 語以上を長文に分け、短文には簡潔書き換えを、長文には分割書き換えおよび要約書き換えが適用している。短文と長文で書き換え方式を換えているのは次の理由からである：機械翻訳では一般に入力文が短いと訳質がよく、長いと訳質が低下することが知られている。したがって、短文は元々の訳質が高いために書き換える場合は大幅に書き換える必要はないが、長文を書き換える場合は大幅に書き換える必要がある。

また、日本語の書き換え文については、書き換え文を英訳した文を作成している。この英訳文は、書き換え文そのものの機械翻訳に対する効果を検証に利用できる。

¹ここでいう“種類”とは異なる作業者が作成した文の数という意味であり、異なる文の数ではない。異なる作業者が偶然に同一文を作成することも多々生じている。

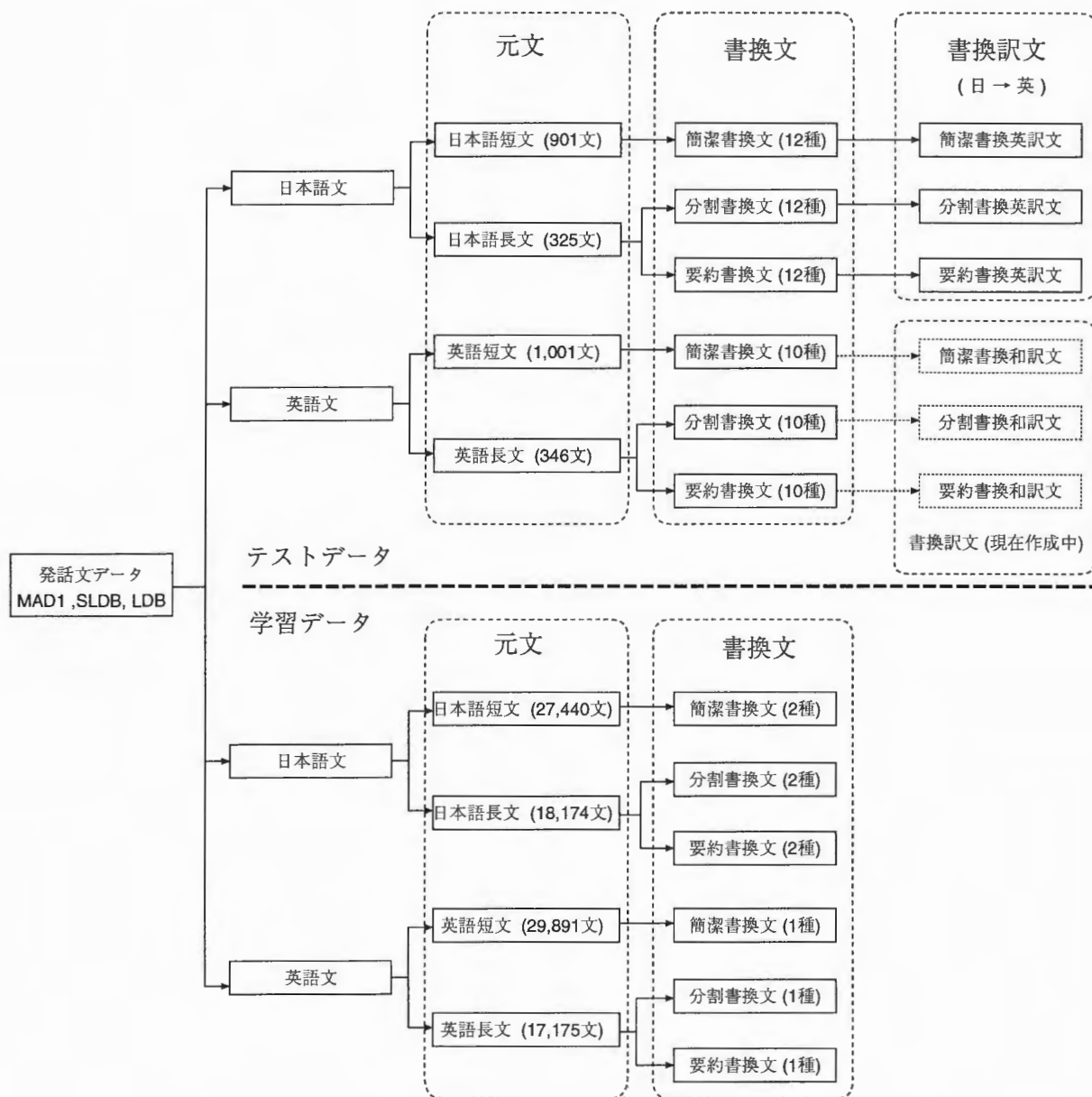


図 1: 会話文書き換えコーパスの元文, 書き換え文, 書き換え訳文の構成図

3 会話文書き換えコーパスの元文について

書き換えの元となる文(元文)は, 会話テキストから取得した. 具体的には, MAD1 [5], SLDB, LDB-JE [6] である. 言語研での音声認識や機械翻訳の処理は発話単位である. 発話は, 概していえばある発話者の意味的なまとまりのある単位を意味しており, 複数の発話文が連結して1発話を構成することが多い. しかし, 会話文書き換えコーパスの作成においては, 文単位(“|”で分割される)を採用した. これは, 構成する複数文の意味的連続性が弱い発話が散見され, それらの複数文をまとめて書き換えるのが難しいことや, まとめて書き換える意義は大きくないと判断したためである. 下に, 意味的連続性が弱い複数文からなる発話の例を示す.

百五十ドルになります.| お支払は現金ですか, それともカードですか.
誰か修理に行かせます.| お客様のお部屋番号は何番ですか.

また, MAD1, SLDB にはテストセットが設定されており, これらのテストセットは本コーパスでもテストセットとして用いている. SLDB のテストセットは SLTA1 と呼ばれているデータである. 表 1 に, MAD1, SLDB, LDB-JE から取得したのべの元文数を示す.

次に, 元文となる会話を長文と短文に二分した. 長文とは内容語を 4 語以上含む文と定義し, 短文とは内容語を 3 語以下含む文と定義した. 図 2 に, BTEC および MAD の包含内容語のグラフを示す. BTEC とは旅行会話用のフレーズブックでよく用いられる表現を集積したコーパスである [7]. この BTEC は, 自然言語処理研究室が開発したすべてのコーパスベース機械翻訳の学習データとなっている. BTEC では, 内容語 3 語を含む文が最も多く, 続いて 2 語, 4 語を含む文となっている. 内容語 5 語以上を含む文は BTEC では急激に少なくなっており, BTEC を学習データとした場合にこれらの文はあまりカバーされないことが予想される. そこで, 内容語 4 語をしきい値として採用した. ちなみに, BTEC と比較すると MAD データは包含する内容語の数は多くなっている.

なお, 同一の文が重複して現れる場合も多く存在するため, 元文全体について uniq を取り, 同じ文については書き換え作業は一度しかしないようにした. 表 2 に, ソース側で uniq を取った上で, 元文を短文, 長文に二分した時の元文の数を示す.

4 書き換え作業の概要

会話データを長文と短文に分け, 長文には要約書き換えと分割書き換えの 2 方式を, 短文には簡潔書き換えの 1 方式により, 人手で書き換えを行った. 各々の書き換え方式については 5 節を参照されたい.

MAD1/learning	4,239
MAD1/test	683
SLDB/learning	20,892
SLDB/test	877
LDB-JE	42,678
合計	69,369

表 1: 元文となった会話文の数 (のべ)

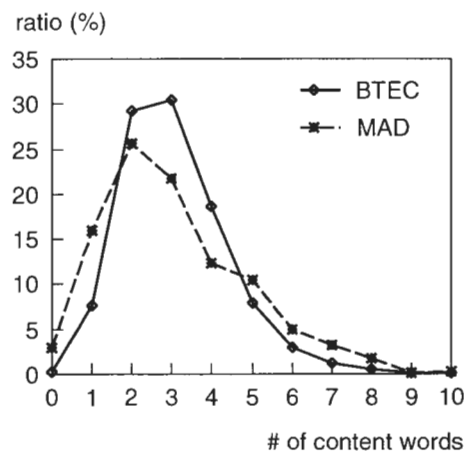


図 2: BTEC および MAD の包含内容語数

データ種別	文長	日	英
テストデータ (MAD1/test + SLTA1)	短文	901	1,001
	長文	325	346
学習データ	短文	27,440	29,891
	長文	18,174	17,175

表 2: 元文となった会話文の数 (異なり)

テストデータは、書き換え文の個人差を見ることや書き換えの自動評価に利用することを想定し、多人数により書き換えを行った。各書き換え方式において、日本語では 12 種類、英語では 10 種類の書き換え文を作成した。日本語の書き換え作業では、3 方式の書き換え $\times$ 12 種類の文 = 36 という組み合わせにおいて、すべて異なる作業者を付した。したがって、日本語の書き換え作業では、36 名の異なる作業者を付している。英語の書き換え作業では、3 方式の書き換え $\times$ 10 種類の文 = 30 という組み合わせにおいて、1 人の作業者は 3 方式の書き換えをすべて担当している。したがって英語の書き換え作業では、10 名の異なる作業者を付している。なお、日本語の作業者は、過去に編集や文章作成などの仕事を経験していることという条件を付して選定している。英語の作業者は、アメリカ英語を母語とするという条件を付しているが、それ以外の条件は付けていない。

さらに、日本語の書き換え文については、これを原文とした英訳を行った。英訳は、日本人が英訳を行い、後で英語ネイティブがチェックするという形で行った。英語の書き換え文に対する和訳作業は、現在作業中である。(作業完了時期未定)

学習データは、日本語部分は 2 種類、英語部分は 1 種類の書き換え文を作成した。学習データの書き換え作業者は、テストデータ作成に関わった作業員から日本語で 4 名、英語で 2 名を選別した。作業員の選別は、他の作業員と比較して最も平均的な書き換えを行ったかどうかという観点で選んだ。具体的には、自身の書き換え文が他の作業員の書き換え作業員と重なる率が高い作業員を選別した。学習データを前半部と後半部に二分し、各作業員はいずれかの部分を担当した。したがって学習データ全体としては、日本語は 2 種類、英語は 1 種類の書き換え文が得られる。

5 書き換え方式

本節では、会話文書き換えコーパスで採用している 3 方式の書き換え方式、簡潔書き換え、要約書き換え、分割書き換え、について述べる。

5.1 簡潔書き換え

簡潔書き換えは、短文に適用するものである。簡潔書き換えでは、冗長な情報を削除することと、表現を平易なものに修正することが主眼である。冗長な情報とは、それがなくても会話の進行に支障をきたさないと考えられる情報である。この“冗長な情報”の判断は書き換え作業員に委ねられる。簡潔書き換えでは、他の要約、分割書き換えにあるような内容語に関する条件はない。

下に、簡潔書き換のサンプルを示す。

元文	I was hoping you'd have a triple room.
書き換え文	I'd like a triple room.

書き換え作業員に与えた書き換え作業マニュアルを付録 A に記す。

5.2 要約書き換え

要約書き換えは、長文に適用するものである。長文に対する書き換えでは、翻訳に有利な短文に置き換えることを目的としている。要約書き換えでは、重要度の低い内容語を削除することで一つの短文に縮退させる。重要度が低いとはいえ元文に含まれる情報が一部欠落してしまうが、短くすることで文全体の訳質が上がるのであれば、その方が音声翻訳全体としての性能が向上する。

下に、簡潔書き換えのサンプルを示す。下線部のある単語は内容語を表しており、書き換えにより内容語が5個から3個に減少している。

元文	Let me <u>just check</u> my <u>computer</u> and <u>get back</u> to you on that.
書き換え文	I will <u>check</u> and <u>get back</u> to you on that.

書き換え作業者に与えた書き換え作業マニュアルを付録Bに記す。

5.3 分割書き換え

本書き換え方式は、長文に適用するものである。分割書き換えでは、与えられた長文を複数の短文(句でもよい)に分割し、翻訳に有利な短文に誘導する。下に、簡潔書き換えのサンプルを示す。下線部のある単語は内容語を表しており、書き換えにより内容語を5つ含む元文が、内容語を3つ含む短文と2つ含む短文に分割されている。

元文	the <u>twin room</u> <u>facing</u> the <u>ocean</u> is <u>three hundred dollars</u> per night
書き換え文	there is a <u>twin room</u> <u>facing</u> the <u>ocean</u> . it is <u>three hundred dollars</u> per <u>night</u>

書き換え作業者に与えた書き換え作業マニュアルを付録Cに記す。

6 会話文書き換えコーパスの所在、フォーマット

6.1 ディレクトリ

会話文書き換えコーパスのデータは以下のディレクトリに納められている。
 /DB/LNG/LNG.04004/[en,ja]/paraphrase/[learning, test]/[UTX, SEG, OrgID] 上記のパス名と格納されているファイルの対応を下に記す。なお、本データのメンテナンスは音声言語処理研究室が担当しており、上記のパス中の“LNG.04004”はそのバージョンを表す。したがって、最新のデータを利用する場合はバージョン番号を確認の上、使用すること。

対象言語	en	英語
	ja	日本語
データ種別	learning	学習データ
	test	テストデータ
ファイル種別	UTX	テキストファイル
	SEG	形態素ファイル
	OrgID	会話文書き換えコーパスと元テキストの対応を収録したファイル

6.2 ファイル

前節で述べたディレクトリに格納されているファイルについて、そのファイル名とフォーマットについて記す。

UTX と SEG ファイルのファイル名は、「書き換え方法+作業員 ID. 言語. ファイル種別」という構成をしている。例えば、“concise_14.ja.UTX” というファイルは、簡潔書換、作業員 ID=14、日本語、UTX、という事を表す。OrgID ファイルのファイル名は、「書き換え方法. 言語.OrgID」という構成をしている。例えば、“concise.ja.OrgID” というファイルは、日本語の簡潔書換における書き換え文と実際のデータ ID(MAD や SLDB など)における ID の対応関係を示すファイルである。

6.3 フォーマット

書き換え文は,UTX, SEG, OrgID の3種類のファイルに情報が格納されている. UTX と SEG は BTEC,MAD といった言語研の他のデータとフォーマットが同じである.

UTX ファイル

```
paraphrase\learning_concise\000110\ja\14\\\ それはどうもすみません
paraphrase\learning_concise\000120\ja\14\\\ よかったです
paraphrase\learning_concise\000130\ja\14\\\ 日本の履物のようなものですね
paraphrase\learning_concise\000140\ja\14\\\ 電話に出ないようです
```

第1フィールド	会話文書き換えコーパスはすべて ‘‘paraphrase’’ が記載
第2フィールド	learning/test と 書換方式を記載
第3フィールド	文 ID
第4フィールド	言語
第5フィールド	書き換え作業者 ID
第6,7フィールド	空欄
第8フィールド	書き換え文

SEG ファイル

```
paraphrase\learning_concise\137210\ja\31\\00010\\|||UTT-START|||
paraphrase\learning_concise\137210\ja\31\\00020\ また|また|マタ|また|接続副詞|||
paraphrase\learning_concise\137210\ja\31\\00030\ 連絡|連絡|レンラク|連絡|サ変名詞|||
paraphrase\learning_concise\137210\ja\31\\00040\ し|し|シ|する|補助動詞|サ変|連用||
paraphrase\learning_concise\137210\ja\31\\00050\ ます|ます|マス|ます|助動詞|特殊サ|基本||
paraphrase\learning_concise\137210\ja\31\\00060\ .|.|.|記号|||
paraphrase\learning_concise\137210\ja\31\\00061\\|||UTT-END|||
```

第1フィールド	会話文書き換えコーパスはすべて ‘‘paraphrase’’ が記載
第2フィールド	learning/test と 書換方式を記載
第3フィールド	文 ID
第4フィールド	言語
第5フィールド	書き換え作業者 ID
第6フィールド	空欄
第7フィールド	形態素 ID
第8フィールド	形態素情報 (フォーマットは一般の SEG ファイルと同様)

OrgID ファイル

```
paraphrase\learning_concise\000070\ja\\|||DEN94B_23010841VF.JMOR-000000250-1| ←
↪ DEN94B_23011146VF.JMOR-000000260-1
paraphrase\learning_concise\000080\ja\\|||DEN94B_23011041VF.JMOR-000000360-1
paraphrase\learning_concise\000090\ja\\|||mad432.org.JMOR-043200110-3
```

第1フィールド	会話文書き換えコーパスはすべて‘‘paraphrase’’が記載
第2フィールド	learning/test と 書換方式を記載
第3フィールド	文 ID
第4フィールド	言語
第5,6,7フィールド	空欄
第8フィールド	対応するデータの文 ID. 複数の箇所で出現する文の場合は, 該当する複数 ID が‘‘ ’’で連結して記される. また, ID は‘‘-’’で結合された3フィールドから構成されている. 第1フィールドはファイル名, 第2フィールドはutterance-ID, 第3フィールドはutterance内文番号を表す. utterance内文番号とは, utteranceが複数文から構成される場合の文の出現順を表す.

7 まとめ

本稿では, 会話文書き換えコーパスについて述べた. 会話文書き換えコーパスは, MAD1, SLDB, LDB に収録されている会話文を元文とし, それらを機械翻訳に有利になるように簡潔, 分割, 要約という3種の書き換えを適用した文からなるコーパスである. 本コーパスは元文と書き換え文が対になったデータとしては世界的に見ても類のないものであり, しかも各書き換え方式と各言語について2万文近い分量が学習データとして収録されており, テストデータには10種類以上の書き換え文が収録されている.

今後の課題として, 機械翻訳の性能を向上させる方式の開発が挙げられる. フィンチらは, 会話文書き換えコーパスを利用して入力文から必要でない語を削除する書き換え方式について発表を行っている [8]. 英語の要約書き換えの学習データを基に, 会話文において削除可能な語を前後2語の情報を参照して決定するものであり, 入力文を冗長な情報を除いて短くすることで翻訳精度向上を図っている. この書き換え方式により, ATR の翻訳システムにおいて若干の精度向上が認められた. 今後も, このような翻訳性能向上の書き換え方式開発に活用されることを期待する.

また, テストデータは10種類以上の書き換え文を備えており, 人による書き換えの傾向の分析や書き換え文の自動評価への利用が期待できる. こちらのデータの活用も期待している.

参考文献

- [1] M. Shimohata, E. Sumita, and Y. Matsumoto. Building a paraphrase corpus for speech translation. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 453–457, 2004.
- [2] 下畑光夫, 隅田英一郎, 松本裕治. 発話を対象とした機械翻訳向け書き換え. 第10回言語処理学会年次大会, pp. 177–180, 2004.
- [3] F. Sugaya, T. Takezawa, G. Kikui, and S. Yamamoto. Proposal of a very-large-corpus acquisition method by cell-formed registration. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 326–328, 2002.
- [4] 菅谷史昭, 金城由美子, 竹澤寿幸, 菊井玄一郎, 山本誠一. 大規模言語データベース収集法の一提案. 情報処理学会第64回全国大会, 第2巻, pp. 81–82, 2002.

- [5] T. Takezawa and G. Kikui. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. In *Eurospeech-2003*, pp. 2757–2760, 2003.
- [6] T. Takezawa. Building a bilingual travel conversation database for speech translation research. In *Proc. of the 2nd international workshop on East-Asian resources and evaluation conference on language resources and evaluation*, pp. 17–20, 1999.
- [7] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 147–152, 2002.
- [8] フィンチアンドリュース, 下畑光夫, 隅田英一郎. 機械翻訳のための文簡易化. 情報処理学会研究報告, 2005-NL-165, pp. 73–78, 2005.

付録

A 簡潔書き換えの作業文書

● 作業概要

・与えられた発話を、下の2つの要件を満たす文に書き換える。

- (1) 簡潔な分かりやすい表現、文体であること
- (2) 冗長または不要な情報が除かれていること。

● 書き換えのガイドライン

・簡潔で分かりやすい表現、文体とは？

- ・構文が複雑でない
- ・使われる表現が標準的、簡潔な表現である。

ex) でございます => です

ex) にしていただけないでしょうか => にしてください

・文体は「ですます」調が望ましい。くだけすぎたり、粗雑な表現は使わない

ex) ○ パスポートを見せてください

× パスポートを見せろ

- ・文として自然であること。
- ・目安としては、日本語初級者に理解しやすい文であること

・不要な情報とは？

- ・その情報がなくても聞き手の次の行動、発話に変化しないと思われる情報
- ・発話状況から容易に類推できる情報。

ex) (今日は) 満室です

ex) (近くの) レストランを教えてください。

ex) (あなたの) 名前を教えてください

ex) × (ご主人の) 名前を教えてください (ご主人は類推できない)

・価値が低い情報。(従属節ごと削除してもよい)

ex) (あいにく) 満室です

ex) (暑いので) 窓を開けてください。

ex) (ちょっと) 教えてください

ex) (隣の部屋がうるさいので) 部屋を換えてください

・挨拶、呼びかけ、応答など

ex) お待たせしました、すみません、はい、いいえ etc.

・発話状況から容易に類推できる情報。

・価値が低い情報。

・挨拶、よびかけなど

ex) (近くの) レストランを教えてください。=> 不要

(レストランを) 教えてください。=> 重要

(あなたの) 名前を教えてください => 不要

(名前を) 教えてください => 重要

そっち(の方)にあります。=> 不要

- ・判断基準は各人に委ねる。記載している例に則する必要はない
- ・原文がそもそも簡潔であると判断した場合は、書き換えしなくてもよい。

- ・体言止めなど、述語を省略した文にはしないこと。(原文で省かれている場合は対象外)
なお、書き換え文の述語と原文の述語は代ってもよい。

ex) [十三日] の [ジャル] [一二三便] の [エコノミー] で [お願い] します。

→ × [十三日] の [ジャル] [一二三便] の [エコノミー]。

ex) 座席はできたら通路側にしてもらえるとありがたいのですが。

→ 座席は通路側にしてもらえますか

※ 述語を「ありがたい」から「もらえる」に変えてもよい

● 作業ファイルのフォーマット

Org 原文

: 簡潔文

※ 作業効率化のために、要約文の部分にあらかじめ原文をコピーしてある

● 例

===== 書き換え作業前 =====

Org はいお部屋はツインにお替えしてよろしいんでしょうか

: はいお部屋はツインにお替えしてよろしいんでしょうか

===== 書き換え作業後 =====

Org はいお部屋はツインにお替えしてよろしいんでしょうか

: 部屋をツインに替えていいですか

- 以上 -

B 要約書き換えの作業文書

● 作業目的

- ・与えられた発話文（内容語を5語以上含む文）を、内容語が4語以下の簡潔な文に書き換える。

● 書き換えのガイドライン

- ・原文を見て、簡潔で分かりやすい文に書き換える。（簡潔書き換えの手引き参照）
- ・上記に加え、内容語を必ず4語以下に限定する。
もし、簡潔書き換えの段階で内容語が4語以下になった場合は、その段階で終えてよい。

※ 内容語とは？

- ・独立した意味を持つ語、名詞、動詞、形容詞、形容動詞、副詞のこと
- ・作業ファイルでは、文中の内容語がカッコ書きで示されている。
- ・表示されている単語は機械処理で付与されたものであるために、誤りを含む場合がある。
特に1単語を分割してしまうことが多い。
その場合は各自で適宜判断して作業を進める。
語を合成した段階で4語以下になった場合は、簡潔書き換えのみを行う。

ex) ボーナス払いをお願いします：[ボーナス] [払い] [お願い]

=> [ボーナス払い] と解釈すると、内容語数は2語

- ・本作業の目的は、実質的な内容語を4語以下にすることであり、表示された内容語を減らすことではない。

したがって、内容語を言い換えても相当する語がある場合は、内容語数は変化しない。

ex) [早く][チェックイン]したいんです

→ [至急][チェックイン]したいんです ※ 「早く」を「至急」と変えても内容語数は同じ

- ・体言止めなど、述語を省略した文にはしないこと。（原文で省かれている場合は対象外）
なお、書き換え文の述語と原文の述語は代ってもよい。

ex) [十三日]の[ジャル][一二三便]の[エコノミー]で[お願い]します。

→ × [十三日]の[ジャル][一二三便]の[エコノミー]。

ex) 座席はできたら通路側にしてもらえるとありがたいのですが。

→ 座席は通路側にしてもらえますか

※ 述語を「ありがたい」から「もらえる」に変えてもよい

- ・合成語を用いて内容語を減らしても構わない。
ただし、合成語は一般的な語であること。

ex) [ジャル]の[一二三便]が[出る][時間]は[いつ]ですか。 ※ 内容語数5

→ [ジャル]の[一二三便]の[出発時間]は[いつ]ですか。 ※ 内容語数4

また、省略形への書き換えはしないこと（英語）

ex) I will take this one. => × I'll take this one.

● 内容語選別のガイドライン

・複文からなっている場合、副文すべてを削除するというのも一つの方法である。

ex) この辺のことはよく知らないのですが、近くに中華料理店はありませんか。

→ 近くに中華料理店はありませんか。

・複数の内容語が並列に述べられている場合は、優先順位を決めて選別する。

ex) 宿泊は、12日から15日までの3泊でシングルを1部屋ですね。

→ 12日から15日までシングルを1部屋ですね。

● 作業ファイルのフォーマット

=====

Word [内容語 1] [内容語 2] [内容語 3] [内容語 4] [内容語 5]

Org 原文

: 要約文

=====

※ 作業効率化のために、要約文の部分にあらかじめ原文をコピーしてある

● 作業例

Word [エアコン] [スイッチ] [入れ] [動か] [直し] ※ (内容語が5語ある)

Org エアコンのスイッチを入れても動かないんですけど直していただけますか

: エアコンが動かないんですが直してくれますか ※ (内容語を3語に絞る)

- 以上 -

C 分割書き換えの作業手順文書

● 作業目的

- ・与えられた発話文（内容語を5語以上含む文）を、内容語が4語以下の短文（複数か単数）に書き換え
る。

● 書き換えのガイドライン

- ・原文を見て、簡潔で分かりやすい文に書き換える。（簡潔文書き換えの手引き参照）
- ・上記に加え、内容語4語以下の短文に分割する。文は句点"."で区切ること。
短文は複数になってもよいし、1つだけになってもよい。

※ 内容語とは？

- ・独立した意味を持つ語、名詞、動詞、形容詞、形容動詞、副詞のこと
- ・作業ファイルでは、文中の内容語がカッコ書きで示されている。
- ・表示されている単語は機械処理で付与されたものであるために、誤りを含む場合がある。
特に1単語を分割してしまうことが多い。
その場合は各自で適宜判断して作業を進める。
語を合成した段階で4語以下になった場合は、簡潔書き換えのみを行う。

ex) ボーナス払いをお願いします: [ボーナス] [払い] [お願い]

=> [ボーナス払い] 1語と解釈すると、内容語数は2語

● 分割のガイドライン

- ・分割は、文だけでなく、句や単語という形で分離してもよい。

ex) 19日から21日の2泊でシングルをお願いします。

=> 19日から21日の2泊。シングルをお願いします。

=> 19日から21日の2泊です。シングルをお願いします。（動詞を補ってもよい）

ex) 黒い皮でできたシャネルのバッグを飛行機の中に忘れてしまいました。

=> 黒い皮のシャネルのバッグ。（名詞句）飛行機の中に忘れました。（文）

- ・単文を連続して読んだ時に、原文の意味が再現できること
- ・接続詞、動詞などを適宜補ってもよい。
- ・単語の順番は適宜入れ替えてもよいが、語順はなるべく保つように工夫する。

● 作業ファイルのフォーマット

=====

Word [内容語 1] [内容語 2] [内容語 3] [内容語 4] [内容語 5]

Org 原文

: 要約文

=====

※ 作業効率化のために、要約文の部分にあらかじめ原文をコピーしてある

● 作業例

Word 【エアコン】 【スイッチ】 【入れ】 【動か】 【直し】

Org エアコンのスイッチを入れても動かないんですけど直していただけますか
: エアコンがスイッチを入れても動きません. 直してください

※ 内容語が4語と1語の2文に分割

- 以上 -