

発話を対象とした機械翻訳向け書き換え

下畑 光夫, 隅田 英一郎

ATR 音声言語コミュニケーション研究所
{mitsuo.shimohata,eiichiro.sumita}@atr.jp

松本 裕治

奈良先端科学技術大学院大学
matsu@is.aist-nara.ac.jp

1 はじめに

機械翻訳では、一般に入力文が長くなるにつれてその翻訳品質が低下する。これは、入力文が長くなることで形態素解析、構文解析、訳語選択の候補が増大し、その中から正しい訳文を選択する確率が少なくなるためである。

本論文では、長文を手で短文に書き換えて機械翻訳に与えることで翻訳性能がどのくらい向上するかについて報告する。書き換え方式は分割書き換えと要約書き換えの2種類の方式を用いている。分割書き換えは与えられた長文を複数の短文に分割するよう書き換える方式であり、要約書き換えは1つの短文に要約することで書き換える方式である。

実験は、旅行会話で用いられる発話¹を用いた日英翻訳で行った。そして、用例翻訳と統計翻訳という方式の異なる2種類の機械翻訳を用いて、書き換えによる性能向上の効果を検証した。

2 長文を対象とした書き換え

本論文で書き換え対象となる長文を明確にするために、文長を測る指標ならびにそれを用いた長文の定義について述べる。

2.1 文長を測る指標

本論文では、機械翻訳の性能低下を招くという観点において文長を測る指標として内容語の数を採用する。つまり、単語数を単位として機能語を除外している。機能語は内容語に付随して機能的役割を示すものであり、その数は文の意味の複雑さを直接反映しないと考えられるために除外した。特に発話では助詞の省略や文末表現が多様であることもあって、同等の意味を持つ文の間でもその数の変動は大きい。

¹本論文では、発話とは音声ではなく書き起こしたテキストを指す。

出現率 (%)

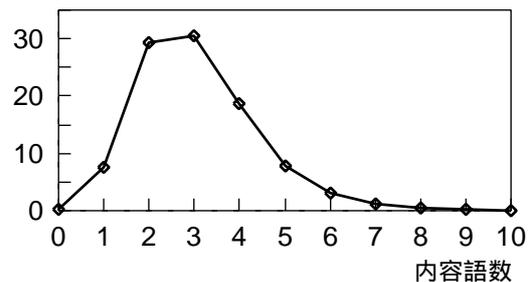


図 1: 基本表現コーパスの内容語数

この他にも文の長さを測る単位としては、文字数、文節数が挙げられるが、次の理由から不採用とした。まず、文字数であるが、機械翻訳は単語を単位として処理するために文字数を採用する根拠は小さい。そして、文節数の場合、他言語に展開する場合に「文節」という概念を適用することが難しいという問題がある。

なお、内容語と機能語は品詞で定義しており、名詞、動詞、形容詞、形容動詞、副詞などが内容語、助詞、助動詞、判定詞などが機能語に該当する。

2.2 長文の定義

実験に用いた機械翻訳はコーパスに基づく翻訳方式を採用しており、基本表現コーパス [1] をその基盤としている。そこで、基本表現コーパスを基に長文の基準を設定する。図 1 に基本表現コーパスの内容語数の分布を示す。5 語以上の内容語を持つ文は比率が低くなり、累計でも 12.9% に過ぎないことより、本論文では包含内容語数が 5 語以上の文を長文と定義する。

3 書き換え方式

長文を短文に書き換える方式として分割書き換えと要約書き換えの2種類を用いた。各書き換え方式については、以下の節で詳述し、図 2 に書き換え例を示す。図中、入力文中の内容語はボールド体で記述され

入力文 1	今日 予約をしている山田という者なんですけれどもチェックインできますか
分割書き換え	今日 予約をしている山田です. チェックインできますか
要約書き換え	予約をしている山田ですが, チェックインできますか
入力文 2	少し 大きいみたいなんですけどもう ーサイズ 小さいのはあるでしょうか
分割書き換え	少し 大きい みたいです. もう ーサイズ 小さい のはありますか
要約書き換え	もう ーサイズ 小さいのはありますか
入力文 3	この 小さい カバン は 持っ て入 り たい んです
分割書き換え	この 小さい カバン. 持っ て入 り たい の です.
要約書き換え	この カバン は 持っ て入 り たい んです.

図 2: 分割書き換えと要約書き換え

ている。なお、各書き換えでは、冗長な表現を削除したり(例: あちらの方に ⇒ あちらに)丁寧な表現を標準的な表現に変更する(例: でございます ⇒ です)ことで簡潔な表現に変更するという作業も加わっている。

3.1 分割書き換え

分割書き換え方式は、長文を複数の短文に分割する書き換え方式である。なお、文として分割することが困難な場合は、名詞句などの文でないものに分割してもよい。(図 2の入力文 3)

3.2 要約書き換え

要約書き換え方式は、長文を一つの短文に要約する書き換え方式である。含まれる内容語から 4 語以内で内容語を選択し、それらを含んだ文を作成する。内容語が並列関係にある文「A と B と C と D と E です」の場合でも重要な情報を選別して 4 語以下とする。また、述語は極力省略しない。

4 実験

実験の構成を図 3に示す。発話コーパスから、111 文の長文発話を抽出し、人手で分割書き換えならびに要約書き換えを行った。入力文の原文ならびにこれらの書き換え文を機械翻訳に与え、得られた 3 種類の訳文の訳質を評価することで書き換えによる機械翻訳の精度向上を検証した。また、cross-perplexity を用いて書き換え文の性質を分析した。

4.1 用いた機械翻訳

実験では、方式の異なる 2 種類の機械翻訳システムを用いた。一つは、用例翻訳 [2] であり、もう一つは統計翻訳 [3] である。これら 2 つの機械翻訳はいずれも

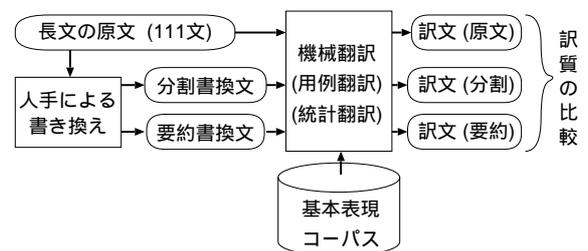


図 3: 実験の構成

コーパスベースの機械翻訳であり、どちらも基本表現コーパスを用いている。

用例翻訳は、用例コーパスと呼ばれるパラレルコーパスを備え、入力文と類似した文を用例コーパスから検索し、その訳文を修正することで、訳文を生成する。実験で用いた用例翻訳では、入力文との類似度を編集距離を用いて算出する。

統計翻訳は与えられた入力文に対する生起確率が最大となる訳文を算出することで翻訳を行う。生起確率は、ベイズの定理に基づき翻訳モデル(目的言語文から入力文が導かれる確率モデル)と言語モデル(目的言語文の尤度モデル)の積に分解される。学習コーパスとして与えられたパラレルコーパスから翻訳モデルと言語モデルのパラメータを学習する。

4.2 機械翻訳に対する効果

原文訳文と各書き換え方式による書換文訳文は、英語のネイティブスピーカーにより評価された。評価では、A(good), B(fair), C(acceptable), D(bad) の 4 種類のランクが付与される [4]。この内、A,B,C の評価の文を“適切な訳文”とした。入力文と原文訳文のペアならびに入力文と書換文訳文のペアを提示して評価さ

表 1: 書き換え適用による機械翻訳の性能

翻訳方式	書換方式	評価ランク (%)					正解率 (%)
		A	B	C	D	N	
用例翻訳	原文	7.2%	14.4%	10.8%	5.4%	62.2%	32.4%
	分割	20.7%	28.8%	20.7%	29.7%	0.0%	70.2%
	要約	17.1%	17.1%	14.4%	20.7%	30.6%	48.6%
統計翻訳	原文	21.6%	22.5%	16.2%	39.6%	-	60.3%
	分割	27.9%	20.7%	12.6%	38.7%	-	61.2%
	要約	20.7%	14.4%	22.5%	42.3%	-	57.6%

せており、書き換え文は提示していない。

実験結果を表 1 に示す。 図中、正解率は与えた入力文に対する適切な訳文が得られた入力文の比率を表す。また、用例翻訳では、入力文との類似文が用例コーパスに存在しない場合は訳文を出力しないので、その場合を N(no-translation) とした。

用例翻訳では、書き換えによる性能向上の効果が大きく現れている。原文訳文の正解率は 32.4% であるが、分割書き換えで 70.2%、要約書き換えでも 48.6% と大幅に向上している。各ランクの比率でも、原文訳文より書換訳文の方が適切な訳である A,B,C ランクの比率が大きくなっている。最も大きな違いは N ランクに現れている。原文訳文が 62.2% であるのに対し、分割書き換えは 0.0%、要約書き換えは 30.6% と大きく減少している。つまり、書き換えにより用例翻訳のカバレッジが拡大したことがこの性能向上の要因であると言える。

一方、統計翻訳では書き換えによる性能向上の効果はほとんどみられなかった。原文訳文の正解率は 60.3% であるが、分割書き換えでは 0.9% の改善、要約書き換えでは 2.7% の低下となった。分割書き換えでは A ランクの比率が増加しており、書き換えが訳質の向上に寄与したと考えられる。逆に、要約書き換えは C,D ランクの比率が増加しており、訳質の低下を招いたといえる。

4.3 書き換え文の cross-perplexity

cross-perplexity(CP) は、テストデータが学習データから作成された言語モデル²にどの程度適合しているかを示す指標である。CP が低いほど、テストデータと学習データの統計的性質の差が小さいことを表している。基本表現コーパスを学習データとし、原文、分

²本実験では bigram を用いた。

表 2: 書き換えによる cross-perplexity の差異

テストデータ	cross-perplexity
原文	61.7
分割	39.3
要約	45.8

割書き換え文、要約書き換え文をテストデータとした場合の CP の値を表 2 に示す。

分割書き換えおよび要約書き換えの文は、原文と比較するといずれも CP を大きく減少させている。つまり、原言語側だけの観点で考えると、書き換え文は原文より基本表現コーパスに統計的性質が近いといえる。しかし、4.2 節の結果に示されるように、統計翻訳ではその事が訳質の向上には反映されていない。

4.4 書き換えによる訳質の変化

4.2 節より、書き換え方式と翻訳方式により効果が変わることが判明した。この結果を、訳質が向上 / 低下した事例という観点で、より詳細に分析してみよう。図 3 に各書き換え方式による訳文と原文訳文の間での優劣の事例数を示す。評価の優劣は、A,B,C,D,N ランクにより判定している。まず、書換訳文は原文訳文と異なる評価ランクとなる事例が多いことが分かる。多い場合は 90%、少ない場合でも約半分の場合が異なる評価ランクとなっている。そして、書換訳文の訳質は、向上した場合と悪化した場合が混在することが分かる。用例翻訳の場合は向上する場合が悪化する場合より多く発生するが、統計翻訳の場合は拮抗している。

したがって、書き換えによる訳質向上を顕著にするためには、悪化した書き換え事例を抑制する必要がある。図 3 の「悪化事例抑制による正解率」に、書き

表 3: 原文訳文と書換訳文の訳質の優劣事例数

翻訳方式	書換方式	訳質比較			正解率 (%)	悪化事例抑制による 正解率 (%)
		書換 > 原文	書換 = 原文	書換 < 原文		
用例翻訳	分割	86	11	14	70.2%	79.2%
	要約	49	49	13	48.6%	57.6%
統計翻訳	分割	34	50	27	61.2%	76.5%
	要約	22	57	32	57.6%	72.0%

換えによる悪化事例を抑制した³場合の正解率を示す。これより、統計翻訳で原文訳文と比較して分割書き換えで 16.2%、要約書き換えで 11.7% の性能向上を達成できることが分かる。

このような書き換えによる訳質悪化事例を抑える方法には次の方法が考えられる。文全体を対象としている書き換え事例から局所的な書き換えパターンを抽出し、書き換えパターンの中で訳質向上に寄与するものだけを選択して用いるという方法である。書き換えパターンの抽出方法としては [5] を、書き換えパターンの選択方法としては [6] を利用することができる。この方法による効果的な書き換えについては、今後の課題である。

5 まとめ

発話に対する機械翻訳の性能を向上させる一つの方法として、発話を機械翻訳に適した文に書き換える方法が考えられる。本論文では、翻訳が困難な長文を短文に書き換えることによる性能向上を提案し、人手の書き換えによる実験について報告した。

実験より、書き換えによる効果は翻訳方式により大きく変わることが判明した。用例翻訳では書き換えによる性能向上が見られたが、統計翻訳ではほとんど見られなかった。原言語の観点では、書き換え文は原文より基本表現コーパスに統計的に近い性質を有しているが、そのことが機械翻訳に必ずしも改善をもたらすとは限らない。

ただし、統計翻訳でも訳質が向上した事例は多く見られており、書き換えによる効果を顕著に出すためには当該機械翻訳にとってプラスの書き換えとマイナスの書き換えとを選択する必要がある。この方策による書き換の効果の検証は、今後の課題である。

³書換訳文と原文訳文の内、評価の高い方を選択することを意味する

謝辞

本研究は通信・放送機構の研究委託「大規模コーパ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。

参考文献

- [1] 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 LREC*, pp. 147–152, 2002.
- [2] E. Sumita. Example-based machine translation using DP-matching between work sequences. In *Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 1–8, 2001.
- [3] T. Watanabe and E. Sumita. Example-based decoding for statistical machine translation. In *Proc. of the 9th MT Summit*, pp. 410–417, 2003.
- [4] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. Solutions to problems inherent in spoken language translation: The atr-matrix approach. In *Proc. of the 9th MT Summit*, pp. 229–235, 1999.
- [5] 下畑光夫, 渡辺太郎, 隅田英一郎, 松本裕治. パラレルコーパスからの機械翻訳向け同義表現抽出. 情報処理学会論文誌, Vol. 44, No. 11, pp. 2854–2863, 2003.
- [6] Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. Feedback cleaning of machine translation rules using automatic evaluation. In *Proc. of the 41st ACL*, pp. 447–454, 2003.