Vol.2010-NL-199 No.9
2010/11/18

言い換えを用いた技術マニュアルの類似文検索

矢 舗 将 隆[†1]　難 波 英 嗣[†1]　竹 澤 寿 幸[†1]

技術マニュアルなど専門的な文書の翻訳作業を効率化するために，過去の翻訳事例を可能な限り再利用する翻訳メモリが広く用いられている．しかし，翻訳メモリ上から翻訳事例を探す際，ユーザが異表記の用語を用いて検索した場合には，その事例を発見することが困難であるという問題が生じる．そこで，本稿では，表層的な文字列の一致だけでなく，言い換えを用いて，原文と類似度の高い文を検索する手法を提案する．提案手法の有効性を確認するため，マツダエース株式会社が業務に用いている技術マニュアル文書データを用いて実験を行った．実験の結果，再現率において 9.2 ポイントの改善が得られ，提案手法の有効性を確認した．

Retrieving Example Sentences in Technical Manuals Using Paraphrases

Masataka Yashiki ,[†1] Hidetsugu Nanba [†1]
and Toshiyuki Takezawa[†1]

Recently, the Translation Memory (TM) has been used widely. The TM reuses a past translation examples as much as possible to promote efficiency of the translation work of professional documents such as technical manuals. However, human translators can not detect past translation examples using different expressions. In this paper, we explore the use of paraphrasing methods for retrieving sentences that resemble a search query sentence. To confirm the effectiveness of our method, we conducted some examinations using the data of technical manuals of Mazda Ace Co., Ltd. As a result, the improvement of 9.2 points was obtained in recall and we confirmed the effectiveness of our method.

†1 広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

1. はじめに

自動車などの製品を複数国で販売するには，技術マニュアルを販売国で使用される言語に翻訳する必要がある．多くの場合，この作業は翻訳の専門家に依頼される．人手による翻訳作業を効率的に行うために，翻訳メモリという概念がある．翻訳メモリは，過去に翻訳者本人が作成した翻訳を事例としてデータベースに蓄積しておき，新たな技術マニュアルを翻訳する時には，可能な限り過去の翻訳データを再利用するものであり，近年翻訳メモリを利用して翻訳をする機会が増えてきている．しかし，翻訳メモリを用いて翻訳事例を探す際，翻訳メモリ上に同一内容の事例が存在していても，翻訳者が異表記の用語（同義語）を用いて検索した場合には，その事例を見つけることができない．その結果，翻訳にかかる費用がかさむ，という問題が生じる．この問題点を改善するため，本研究では，表層的な文字列の一致だけではなく，言い換えや表記揺れも考慮することにより，翻訳メモリ上から原文と類似度の高い文とその対訳文をを抽出する手法を提案する．

本論文の構成は以下の通りである．次節では本研究の関連研究を示し，3 節では，言い換えを用いた翻訳支援について詳細に述べ，本研究における提案手法を述べる．4 節では，実験内容及び実験結果を，5 節ではその考察について述べ，6 節で本稿をまとめる．

2. 関連研究

本節では，テキスト評価と類似文検索，同義語および言い換え抽出について述べる．

2.1 テキストの自動評価

これまでに，テキスト要約の自動評価手法として，人間が作成した要約（以下，参照要約）とコンピュータの作成した要約（以下，システム要約）との類似性による自動評価手法がいくつか提案されている．この手法は，参照要約とシステム要約との間の一種の類似度を計算するものであり，参照要約との類似度が高いほどより良い要約であるという考えに基づく．以下に，代表的な評価手法の一つである ROUGE について説明する．

ROUGE[4] は Lin が提案した要約評価用の尺度であり，様々な種類のものが存在するが，そのうちの一つである ROUGE-N は現在，要約システムの自動評価法として最も広く用いられている手法である．ROUGE-N は，参照要約とシステム要約の間で一致する n-gram（ある文字列から切り出した一定個数の文字集合）の割合を以下の式を用いて計算する．

$$ROUGE(C,R) = \frac{\sum_{e \in n\text{-}gram(R)} Count(e)}{\sum_{e \in n\text{-}gram(C)} Count_{clip}(e)} \quad (1)$$

ここで，$n\text{-}gram(C)$ はシステム要約に含まれる n-gram，$n\text{-}gram(R)$ は参照要約に含まれる n-gram 集合を表す．$Count_{clip}(e)$ は，システム要約に含まれる n-gram のシステム要約における出現頻度 $Count(e \in n\text{-}gram(C))$ と参照要約における出現頻度 $Count(e \in n\text{-}gram(R))$ の小さい方の値を採用する．Lin らは n を 1〜4 まで変化させ，マニュアル評価との相関を調べた結果，$n = 1, 2$ が最も高い相関であったと報告している．本研究では，$n = 1$ を比較対象として採用している．

2.2 同義語および言い換えの自動抽出

テキストから同義語や言い換えを自動的に抽出する研究は，近年多岐に渡って行われている[2] が，本研究と関連のある代表的なものとして，統計的機械翻訳の技術を用いた研究がある．

海野らは，対訳コーパスから統計翻訳を用いて言い換え表現を自動獲得し，これを従来の情報検索の枠組みに取り入れることにより新しいクエリ拡張手法を提案している[5]．彼らは日英対訳コーパスを用意し，同一の訳語とアライメントのとれた 2 つの語句を言い換えと見なしている．例えば，日本語の「二酸化炭素」と「炭酸ガス」は “carbon dioxide” と対応付けが行われる確率が高いことから，“carbon dioxide” をピボット（軸）として,「二酸化炭素」と「炭酸ガス」が言い換え表現になっていると見なすことができる．海野らの提案する言い換えの自動抽出手法を，本研究における言い換え知識抽出の方法のベースとして使用する．

2.3 言い換えを用いたテキストの自動評価

上述の統計的機械翻訳技術による同義語抽出手法を，テキストの自動評価に利用した研究として，Kauchak らのものがある[3]．この研究は機械翻訳の評価において，文脈を考慮した言い換え手法を提案している．Kauchak らは参照テキスト（人間の被験者が作成した正解翻訳）の言い換えのうち，システムテキスト（システムが自動的に生成した翻訳）に表れている語のみを言い換え候補としている．そして，言い換え候補を参照要約に適用する際に，文脈的に適切かどうかを判断している．適切と判断された言い換えを用いて，複数の参照テキストを生成し，自動評価における新たな参照テキストとして用いる．この手法により，最初の参照テキストのみを用いた（言い換えを用いて生成された参照テキストを用いない）評価と比べ，人手により近い評価が行えることを示している．

Zhou らは，英中対訳文データから統計的機械翻訳技術を用いて言い換え表現を自動的に抽出し，テキスト要約の評価に用いる ParaEval という手法を提案している[6]．この手法では，参照要約とシステム要約の比較を，大域的には最適マッチング，局所的には最長マッチングを行うことにより段階的に言い換えマッチングを行ない，テキストを評価する．すなわち，第一段階では動的計画法に基づき，フレーズ対フレーズによる言い換えマッチングを行う．そして第二段階で，残りの単語に対して Greedy 法に基づいて単一語対フレーズ，または単一語対単一語による同義語マッチングを行う．第三段階では第一段階，第二段階で言い換えに一致しなかった残りの単語に対して，ROUGE と同様の語彙マッチングを行う．Zhou らは，ParaEval の評価と人間の評価との相関が ROUGE のそれと似ているということを示し，提案手法の有効性を確認している．

平原らは，日本語テキスト要約に対して言い換えを用いた自動評価を行っている[1]．上述の Zhou らの ParaEval では言い換えの適用手順に関して検討していないことを指摘し，ParaEval と同様に言い換えの適用を行った後で語彙マッチングを行う「ParaEval 手順」と，語彙マッチングを行った後で言い換えの適用を行う「逆 ParaEval 手順」について検討した．また，言い換え知識として Zhou ら，海野らの手法の他に，分布類似度，WordNet，NTT 日本語語彙体系など，複数の言い換え知識を利用した．実験の結果，従来のテキスト評価手法に比べ，自由作成による要約に対して提案手法がより有効であると報告している．

3. 言い換えを用いた類似用例文検索

本節では，言い換えを用いた類似用例文検索について述べる．3.1 節では，技術マニュアルの翻訳における課題について言及し，3.2 節で言い換えを用いた類似用例文検索における本研究での提案手法について述べる．3.3 節では，提案手法での言い換え知識について述べる．

3.1 技術マニュアルの翻訳における課題

コンピュータを利用した翻訳技術のうち，実際の訳例を多数集めてデータベース化し，翻訳の現場でそれを再利用する「翻訳メモリ」と呼ばれる技術がある．翻訳メモリの主な機能は，

- 翻訳者によって書き起こされた翻訳を，その原文とともに，専用のデータベースに登録する．
- 過去にデータベースに登録された翻訳を，同一または類似の原文が出現したときに自動的に「類似文」として表示する

である．これらの機能によって，

- 同じ文章を繰り返し翻訳する
- 文章を手作業で複製し貼り付ける

などの，これまで翻訳者に任されていた単純作業を自動化し，さらに同じ文章や類似した文章の翻訳における表現の統一も自動化することが可能になるため，翻訳品質の向上も期待できる．

技術マニュアルの翻訳においても翻訳メモリを用いる現場が多いが，技術マニュアルなどの局所的で専門的な文書は，次の理由で翻訳支援が困難であると考えられている．

- 技術マニュアルは厳密性が要求されるため，一文が非常に長いものもある．また，部品名などの一文が非常に短いものもある．文が極端に長い，あるいは短いと，文の構造解析が失敗する可能性が高くなり，翻訳も解析に依存するため失敗する．
- 技術マニュアルは新規性・専門性の高い用語が多く使用されており，新しい概念や用語が次々に出てくるため，辞書登録が追い付かず訳語が欠落する．
- 従来の技術マニュアルの作成は，翻訳を人手で行っているため翻訳者によって表記が異なり，多数の類似表現が作成される．言い回しの数に伴い他言語翻訳費用が増大する傾向にある．

本研究ではこのような問題を踏まえ，実際の現場では主流である翻訳メモリを用いた翻訳支援に対して，3.2 節，3.3 節に述べる手法を用いて言い換えを考慮し，より多くの類似文検索を行う．

3.2 言い換えを用いた類似用例文検索

本研究では，3.1 節で述べた翻訳メモリを用いて，原テキスト（以下，クエリテキスト）と翻訳メモリ上のテキスト（以下，メモリテキスト）との類似度を測る際に言い換えを考慮することで，より類似度を上げることが可能になると考えられる．従来の自動評価手法では難しい，表記揺れや言い換えを含むテキストを評価するために，本稿では平原らの “ParaEval 手順” と同様の手順を用いて言い換えを考慮する．これは，メモリテキストとクエリテキストを比較する際，従来手法と同様の語彙マッチングを行う前に，互いのフレーズの間に言い換えが含まれていれば，それを同じ単語と見なすことで言い換えを考慮する．以下にアルゴリズムを示す．

- ParaEval のアルゴリズム
  （1） テキストを走査し，句と句から成る言い換えの一致を Greedy 法に基づいて検出する．
  （2） (1) で一致しなかった語に対して，単一語対句，または単一語対単一語を走査し，単語レベルの言い換えや，表記揺れによる言い換えの一致を Greedy 法に基づいて検出する．
  （3） (1) と (2) で一致しなかった語に対して，語彙マッチングを行う．

3.3 言い換え知識の抽出

本研究では，Zhou ら[6]，海野ら[5] らと同様に，統計的機械翻訳により生成されるフレーズテーブル（翻訳モデル）を用いて言い換え辞書を作成する．この手法は,「もし，X と Y の翻訳が同一であれば，X と Y は言い換えと見なすことができる」という考えに基づいている．本研究では日英対訳文として，マツダエース株式会社で業務に用いられている技術マニュアル文書 39,731 文対を，また，統計的機械翻訳用のツールとして GIZA++[⋆1] を，それぞれ用いている．なお，得られたフレーズテーブルから言い換え辞書を作成する際，品詞情報を考慮し，先頭の品詞と後尾の品詞が一致したフレーズ[⋆2] のみを同義語と定義した．また,「オイル プレッシャ」など品詞の並びが「名詞-名詞」などの複合名詞は「名詞」に縮約して扱うこととした．

この手法により，最終的に 38,815 対の言い換え知識を獲得した．これらの言い換え知識は，自立語，付属語問わず全ての品詞を含み，フレーズ長も任意である．

4. 実　　験

3 節で述べた手法の有効性を確認するために実験を行った．

4.1 実 験 方 法

本節では，実験方法として，実験に用いたデータ，言い換え知識の作成，評価方法について述べる．

■実験に用いたデータ

本研究では，マツダエース株式会社で通常業務に用いられている日本語技術マニュアル文書 39,731 文を，クラスタリングツール bayon[⋆3] を用いて分割した．その結果，734 クラスタに分割された．

bayon により分割された 734 クラスタのうち，ランダムに選択した 200 クラスタ内の任意の 80,478 文対を実験データとして用いた．この 80,478 文対に対して，クエリテキスト

⋆1 http://www.fjoch.com/GIZA++.html
⋆2 例えば,「名詞-助詞-動詞」から構成されるフレーズと,「名詞-名詞-動詞」から構成されるフレーズは同義語と見なす．
⋆3 http://code.google.com/p/bayon/wiki/Tutorial_ja

とメモリテキストの比較を行い，全ての文対に基準 0，基準 1，基準 2 の計 3 種類のタグを人手により付与し，これを正解データとした．このとき，メモリテキストの情報（日英文対）が存在すると仮定した場合に，クエリテキスト（日本語文）が英文に翻訳できるかどうかの度合いにより判断した．この 3 種類のタグの付与基準について，以下にその詳細を述べる．また，基準 2 の例を示す．

- 正解データのタグ詳細
  - 基準 0
    メモリテキストの情報だけではクエリテキストを翻訳できないと判断した際に付与（メモリテキストとクエリテキストは類似していない）
  - 基準 1
    メモリテキストの情報からクエリテキストを翻訳することが可能であると判断した際に付与（メモリテキストとクエリテキストがほぼ一致している）
  - 基準 2
    メモリテキストの情報を元にクエリテキストをある程度翻訳することが可能であると判断した際に付与（クエリテキストとメモリテキストが類似している）
- 基準 2 の正解データのタグ付け例
  - **クエリテキスト：** O リングに損傷を与えないように注意する。
    メモリテキスト： スプリングに傷が付かないよう気をつける。

■実験に用いた言い換え知識

3.3 節で述べたように，本研究では統計的機械翻訳の過程で生成されるフレーズテーブルから言い換え知識を獲得した．対訳コーパスとして，マツダエース株式会社で業務に用いられている技術マニュアル 79,673 文対を用い，最終的に言い換え知識 61,442 対を集積した．

■評価手法 評価実験において，ROUGE，“ParaEval 手順” により算出されるスコアは 0〜1 の値をとり，この値が 1 に近いほどクエリテキストとメモリテキストが類似していると判断できる．これらのスコアと上述した正解データのスコアとの一致度を測ることにより，提案手法の有効性を評価する．一致度は，以下の式を用いて算出する．

$$\text{精度} = \frac{\text{システムが評価した正解数}}{\text{人手で作成した正解数}} \quad (2)$$

$$\text{再現率} = \frac{\text{システムが評価した正解数}}{\text{システムが評価した数}} \quad (3)$$

表 1 ParaEval 手順：「基準 1」と「基準 0 と 2」の一致度
Table 1 ParaEval: Corresponding rate of “criterion 1” and “criteria 0 and 2”

		人手	
		基準 1	基準 0 と 2
システム	基準 1	0.1%	0.1%
	基準 0 と 2	0.0%	99.8%

表 2 ROUGE：「基準 1」と「基準 0 と 2」の一致度
Table 2 ROUGE: Corresponding rate of “criterion 1” and “criteria 0 and 2”

		人手	
		基準 1	基準 0 と 2
システム	基準 1	0.1%	0.0%
	基準 0 と 2	0.0%	99.9%

なお，システム評価をマニュアル評価と同じ 3 種類の基準とするための閾値 μ は，200 クラスタ中 100 クラスタについて，システム評価とマニュアル評価の一致度を上記 (2), (3) 式を用いて算出し，その調和平均である F 値が最も高くなる値を選択した．この予備実験により，$\mu = 0.7$ のとき最も F 値が高くなり，この値を用いて残りのクラスタに対して同様の実験を行い，評価する．すなわち，$\mu < 0.7$ の場合基準 0 に，$0.7 \leq \mu < 1$ の場合基準 2 に，$1 = \mu$ の場合基準 1 とし，残りの 100 クラスタに対して同様に評価を行う．また，本実験で使用する文対（100 クラスタ）は評価量として大きいものではないので，評価結果の妥当性，再現性を保持するため 4-fold cross validation を行った．

4.2 実験結果

4.1 節の実験結果を以下に示す．評価実験として，システム評価とマニュアル評価の一致度を「基準 1」と「基準 0 と 2」,「基準 1 と 2」と「基準 0」のそれぞれについて測る実験を行った．なお,「基準 1」と「基準 0 と 2」での一致度は，完全に一致した文対の抽出の優劣をつける尺度であり,「基準 1」と「基準 0 と 2」での一致度は，類似文対の抽出の優劣をつける尺度である．

100 クラスタで ParaEval 手順，ROUGE に閾値 $\mu = 0.7$ を適用した場合の,「基準 1」と「基準 0 と 2」の実験結果を**表 1** に,「基準 1 と 2」と「基準 0」の実験結果を**表 2** に示す．また,「基準 1 と 2」と「基準 0」の実験結果を**表 3** に,「基準 1 と 2」と「基準 0」の実験結果を**表 4** に示す．

表 1〜表 4 より (2), (3) 式を用いて精度と再現率を算出した．このとき，データ量が少

表 3 ParaEval 手順：「基準 1 と 2」と「基準 0」の一致度
Table 3 ParaEval: Corresponding rate of “criteria 1 and 2” and “criterion 0”

		人手	
		基準 1 と 2	基準 0
システム	基準 1 と 2	4.1%	3.5%
	基準 0	4.5%	87.9%

表 4 ROUGE：「基準 1 と 2」と「基準 0」の一致度
Table 4 ROUGE: Corresponding rate of “criteria 1 and 2” and “criterion 0”

		人手	
		基準 1 と 2	基準 0
システム	基準 1 と 2	3.5%	1.7%
	基準 0	5.1%	89.7%

表 5 「基準 1」と「基準 0 と 2」での精度，再現率
Table 5 Precision and Recall of “criterion 1” and “criteria 0 and 2”

	ParaEval	ROUGE
精度	40.8%	**77.6%**
再現率	82.9%	**100.0%**

表 6 「基準 1 と 2」と「基準 0」での精度，再現率
Table 6 Precision and Recall of “criterion 1 and 2” and “criteria 0”

	ParaEval	ROUGE
精度	54.1%	**64.6%**
再現率	**52.3%**	42.8%

なく結果に偏りが予想されるため，同 100 クラスタに対して 4-fold cross validation による評価を行った．その結果を**表 5**，**表 6** に示す．なお，表中の太字は比較結果における最大値である．

5. 考　　察

表 5 において，ParaEval の精度，再現率が ROUGE を下回った．これは，3.3 節で述べた同義語辞書に原因があると考えられる．この同義語辞書は自動的に収集されたものであるため，例えば「ガラスの」に対して「ガラスの端末」のような，一方が他方を包括する言い換え知識が含まれていることが多い．しかし，本実験で用いた “ParaEval 手順” では，クエリテキストとメモリテキスト内にそれぞれ「ガラスの」および「ガラスの端末」という語が含まれていれば，包括関係にある言い換え知識を用いて，この 2 語が言い換え関係にあると判断されてしまう．この問題は，主にクエリテキストとメモリテキストが完全一致すると人手で評価された場合に起こりやすい傾向が確認された．また，この問題は類似文であると人手で評価された場合にも確認された．

表 6 において，ParaEval の再現率が ROUGE を 9.2 ポイント上回っており，単純な語彙マッチングである ROUGE よりも，言い換えを用いる提案手法の方がより多くの類似文を検索できることが分かる．すなわち，ROUGE では抽出できなかった類似文が ParaEval では抽出できたことを示しており，提案手法の有効性が確認できた．

次に，表 6 において，ParaEval の精度が ROUGE を 10.5 ポイント下回ったことに関する考察を行う．これは，ParaEval が誤って類似文として評価したと考えられる．以下に誤って評価された例を示す．

クエリテキスト： 作業開始前のキーは使えなくなるので、作業開始前に新品のキーを 2 本以上準備する。

メモリテキスト： キー・シリンダを交換する場合は、作業開始前のキーは使えなくなるので，作業開始前に登録用キーを 2 本以上準備すること。

このように，上述した包括関係にある言い換え知識が多く用いられている場合，すなわち，言い換え知識として「キー」と「キー・シリンダ」の言い換え,「作業開始」と「作業開始前の」,「作業開始前に」の言い換えが包括関係にある場合，ParaEval がうまく機能せず，正しい評価が行えなかったことが原因である．これは，言い換えの適用と語彙マッチングの適用順序を変更することで対処が可能になると考えられる．

6. おわりに

本研究では，技術マニュアルなどの機械翻訳が困難な文書に対して，言い換えを考慮することによりクエリ文書と類似度の高い文書を検索する手法を提案した．評価対象が一般的な文書ではなく，技術マニュアルという専門的な文書であり対訳文が存在すること，また，言い換え知識を獲得する際に品詞情報を考慮しているという点が異なる．提案手法の有効性を確認するため，マツダエース株式会社の技術マニュアル文書データを用いて実験を行った．

データの偏りを除去するため 4-fold cross validation を採用し実験を行った結果，再現率において 9.2 ポイントの改善が得られ，提案手法を用いることによりより多くの類似文を検索できることが確認された．

また，本研究の結果，現在の自然言語処理技術を用いて獲得できる言い換え知識を適用しても，重要な点で反意語を用いているために十分な評価を行うことが不可能であり，翻訳メモリから抽出できない類似文が存在することが明らかになった．この問題については，あくまでも翻訳支援，すなわち翻訳作業コストの低減という観点から，現在の技術の可能性と限界を吟味する態度が重要であろう．

謝辞 本研究で用いた技術マニュアルのデータを提供して下さったマツダ株式会社およびマツダエース株式会社に深謝致します．

参 考 文 献

1) 平原一帆，難波英嗣，竹澤寿幸，奥村学：言い換えを用いたテキスト要約の自動評価，情報処理学会論文誌データベース，Vol.3，No.2，pp.91–101 (2010).
2) 乾健太郎，藤田篤：言い換え技術に関する研究動向 (2004).
3) Kauchak, D. and Barzilay, R.: Paraphrasing for Automatic Evalutaion, *Proc. the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL)*, pp.455–462 (2006).
4) Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Proc. ACL workshop on "Text Summarization Branches Out"*, pp.74–81 (2004).
5) 海野裕也，宮尾祐介，辻井潤一：自動獲得された言い換え表現を使った情報検索，言語処理学会第 14 回年次大会，pp.123–126 (2008).
6) Zhou, L., Lin, C.-Y., Munteanu, D.S. and Hovy, E.: ParaEval: Using Paraphrases to Evaluate Summaries Automatically, *Proc. the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL)*, pp.447–454 (2006).