

多言語リアルタイム会話における日本語動詞語尾崩れの検出

鄭育昌*, 長瀬友樹*, 上柿稔†, 岡田伊策†

株式会社富士通研究所 スピーチ&ランゲージテクノロジー研究部*

富士通株式会社 S I 技術サポート本部 システム技術統括部†

{cheng.yuchang, nagase.tomoki, uegaki.minoru, isaac-okada} @jp.fujitsu.com

1 はじめに

インターネットが地球規模で使えるようになり、機械翻訳のサービスとインスタントメッセージング (IM) を利用して、言語の壁を越えて外国語話者とのリアルタイムコミュニケーションを行う人が増えている ([1],[2])。企業においてもインスタントメッセージングを使用してリアルタイムの業務連絡を行うことも増加している。しかし、機械翻訳の品質が不十分の場合、意思疎通の齟齬で不要な会話が増加し、迅速さが損なって業務連絡に支障が出る恐れがある。

機械翻訳を用いた業務連絡における意思疎通の齟齬を防ぐため、我々は企業内でインスタントメッセージングを用いた業務連絡 (日本語) を分析、業務連絡会話の特徴を観測し、更に日英・日中翻訳ソフトを用いて業務連絡の会話ログを英語に翻訳して、業務連絡会話の特徴が機械翻訳の品質に対する影響を観測した。その結果、図 1 の入力文「もっと買わなきゃだ」のように、「日本語動詞の語尾崩れ」が機械翻訳の品質を低下させる重要な要素であることがわかった。

我々は上記の調査からの知見に基づき、IM 会話において、機械翻訳品質の低下要因である日本語動詞の語尾崩れの自動検出システムを提案、構築した。日本語の動詞語尾崩れの修正には以下のタスクがあると考えられる: (1) 機械翻訳の失敗を引き起こす動詞語尾崩れの検出; (2) 動詞語尾崩れを同じ意味を持つ一般表現での置換; しかし、(2)の処理は極めて困難のため、本稿では、(1)の処理に焦点を絞った。

日本語動詞語尾表現崩れの検出は、文中の動詞 (語幹と語尾表現) の検出と動詞の語尾表現が崩れたかどうかの検証の順序が含まれる。インスタントメッセージの大半は短文であるため、SVO 語順を取る日本語では動詞が文の終わりに置かれているとみなせる。検出された動詞に対し、動詞語尾を一般表現で置換と置換前との翻訳品質の変化を測ることで、動詞語尾表現崩れ箇所を検出する。

提案手法を前記の企業内業務連絡ログに適用した結果、企業内業務連絡ログの日中翻訳品質 (Acceptability) は本手法の適用で 3.41 から 3.48 に向上することを確認した。

表 1: 社内の IM 会話ログの概要

ユーザ数	587
会話単位数	2000
総発話数	22715
1 会話当たりの平均発話数	11.3
発話の平均長さ	19.5 (文字)
発話間の平均応答時間	40.7 (秒)

2 IM を用いたリアルタイム業務連絡の会話ログの分析

本章では、我々が分析したIMを用いたリアルタイム業務連絡の日本語会話ログについて説明する。本研究では、インスタントメッセージングとして「Lync¹」を利用した。表 1 は業務連絡の会話ログの概要を示している。587 人のユーザが 3 カ月間に Lync を利用した日本語会話ログ、合計 22,715 件の発話を収集した。会話ログが社内の業務連絡のため、会話の開始・終了の境界が不明確である。本稿では、2 つの発話の間隔に基づき、発話から 10 分以上の間隔で会話を分断し、全会話ログは 2000 個の会話単位に分割された。1 つの会話単位における平均発話数は 11.3 であり、発話の平均長さは 19.5 文字である。業務連絡におけるインスタントメッセージの発話がニュース記事や書類のようなテキストより短いことを示している [3]。

次に、上記の会話ログの発話の特徴を観察し、発話の特徴が日英・日中機械翻訳の品質に対する影響を観測した。本稿では、機械翻訳の品質を判断する指標として「Acceptability」を採用しました [4]。また、本稿では「Acceptability」を最高位の「5」から最下位の「1」で 5 段階の数値で翻訳品質を評価した。機械翻訳器は、日中翻訳ソフト「J-北京 7²」と日英翻訳ソフト「ATLAS V14³」を採用した。

日英・日中翻訳を行うことで、会話ログの特徴が異なった言語の機械翻訳に対する影響する属性を観察できる。表 2 では分析結果を示している。会話ログで観測された特徴の数が多いため、表 2 では機械翻訳に影響がある代表的な特徴を示している。「特

¹ Microsoft™ Lync: <http://products.office.com/en/lync/>

² 株高電社 J-北京 7: <http://www.kodensha.jp/soft/jb/>

³ 株富士通 ATLAS V14: <http://software.fujitsu.com/jp/atlas/>

表 2: 日本語インスタントメッセージ発話特徴の出現確率と翻訳結果の Acceptability の分布

特徴	特徴の出現確率	日中翻訳の Acceptability	日英翻訳の Acceptability
顔文字	7%	4.29	4.33
動詞語尾崩れ表現	8%	2.65	2.41
その他の口語表現	23%	3.58	3.26
固有表現の使用	16%	3.50	3.62
途切れた発話	12%	3.36	3.10
主語、目的語省略	3%	3.30	3.0
URL、ソースコードなど	2%	4.16	3.81
話者の共通知識	16%	3.52	3.62
全発話平均		3.62	3.43

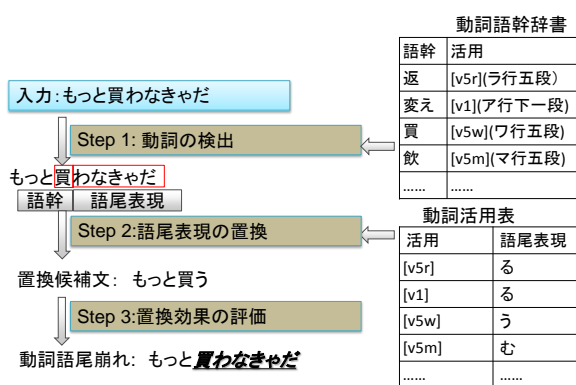


図 1: 提案手法の処理例

特徴の出現確率」が該当特徴種類を持つ発話が会話ログに出現する割合を示す。「日中（日英）翻訳の Acceptability」は会話ログの日中（日英）翻訳結果の平均翻訳品質を示す。

表 2 では、特徴「動詞語尾崩れ表現」の出現確率は 8% であるが、翻訳品質（日中: 2.65、日英: 2.41）が全発話平均（日中: 3.62、日英: 3.43）に比べて大幅に低くなっていること。他の特徴と比べても、「動詞語尾崩れ表現」を持つ文の翻訳品質が他の特徴を持つ文の翻訳品質より低いことがわかった。

特徴「動詞語尾崩れ表現」の出現確率は 8% であるため、業務連絡会話ログ（平均発話数 11.3）では、1 会話単位につき約 1 回の動詞語尾崩れ表現が発生すると考えられる。そのため、機械翻訳を用いた IM 会話では、動詞語尾崩れ表現の出現はユーザ間の意思疎通に障害が生じる原因と考えられる。

以下では、IM 会話の翻訳における日本語動詞語尾崩れ表現を検出するための提案方法を説明する。

3 提案手法：動詞語尾を一般表現で置換することで日本語動詞語尾崩れ表現を検出

3.1 日本語動詞語尾崩れ表現の検出

本稿の提案手法は、会話文中の日本語動詞語尾を一般表現で置換し、置換前後の翻訳品質の品質変化

を元に置換された表現は動詞語尾崩れ表現であるかどうかを判断する手法である。日本語動詞の語尾が崩れた表現の場合、動詞語尾部分が翻訳できないため翻訳品質が低下するが、動詞語尾部分を翻訳可能な一般語尾で置換することで翻訳品質が向上すると考えられる。そのため、動詞語尾部分を一般語尾で置換して翻訳品質の変化を評価することで動詞語尾崩れ表現を検出できる。例えば、機械翻訳ソフトが「買わねえよ」を翻訳できないが、「買わない」が翻訳できる。

図 1 では、本稿の提案手法の処理流れおよび実行例を示している。提案手法は以下の 3 ステップを含んでいる：

ステップ 1: 動詞(語幹と語尾表現)の検出

入力文に対し、動詞語幹辞書を用いて動詞の語幹部分を検出する。動詞語幹辞書は形態素解析などで用いた辞書であり、動詞の語幹およびその活用形を示すものである。検出された動詞語幹に続くひらがな文字列は動詞の語尾表現とみなす。IM を用いた業務連絡会話ログの分析により、IM 発話ログは短く、動詞が発話の最後にあるため、動詞の語幹以降、文末までの文字列は動詞の語尾表現とみなせ、動詞語尾崩れ表現の可能性があると考えられる。図 1 では、入力文「もっと買わなきゃだ」に対し、動詞とその語尾「買わなきゃだ」が検出された。

ステップ 2: 語尾表現の置換

ステップ 2 では、ステップ 1 で検出された動詞（語幹と語尾表現）に対し、語尾表現の部分を動詞語幹の活用形情報に基づき、一般表現で置換する。一般表現は動詞の活用形であるため、機械翻訳ソフトが処理できる。図 1 では、「買わなきゃだ」の動詞に対し、動詞語幹「買」の活用情報を用いて、一般表現「買う」で置換し、置換結果「もっと買う」を生成した。

ステップ 3: 置換効果の評価

ステップ 2 の置換結果とその原文（入力文）に対し、置換結果の翻訳品質が原文の翻訳品質より大幅に向

上したかどうかを評価する。翻訳品質の自動評価方法は 3.2 で説明する。置換結果の翻訳品質が原文の翻訳品質大幅に向上する場合、ステップ 2 の置換により、原文の翻訳できない部分を取り除いたと考えられ、置換された動詞語尾部分は翻訳できない語尾崩れであると判断できる。図 1 では、ステップ 2 で生成された置換結果「もっと買う」の翻訳品質を評価し、その評価値を原文「もっと買わなきゃだ」の翻訳品質と比較して、置換結果の翻訳品質が原文より大幅に向上したため、置換された語尾表現「わなきゃだ」が翻訳できない部分、すなわち語尾崩れであると判定する。この検出結果（動詞語尾崩れ：買わなきゃだ）をユーザに提示することで翻訳不可の表現を防ぐことが可能である。

3.2 原文と折り返し翻訳結果の類似度による翻訳品質の自動評価

提案手法のステップ 3 では、原文と折り返し翻訳結果の類似度を評価することで翻訳品質の自動評価を行う。「折り返し翻訳」は元言語から一度目標言語に翻訳し、更にその翻訳結果を元言語に翻訳する手順である[5]。翻訳目標言語を理解できないユーザが機械翻訳ソフトを利用する際、折り返し翻訳の結果が原文と一致するかどうかを確認することで翻訳品質を推測することがよく見られる現象である。折り返し翻訳が原文と一致する場合、目標言語への翻訳品質が高いと推測できる。

提案手法は、原文と折り返し翻訳結果の類似度を図 1 で示した概念構造を比較することで評価する。概念構造は文の意味構造を単語の概念および概念間の関係で表示する意味構造である。図 1 では、原文「もっと買う」とその（日英日）折り返し翻訳結果「それはそれをさらに買います」の概念構造を生成し、構造の要素（ノード、ノード間の関係など）に図例のように点数を付与し、図 1 で示した計算式で概念構造の類似度スコアを計算した。本事例では、原文と折り返し翻訳結果の類似度は 0.71 である。

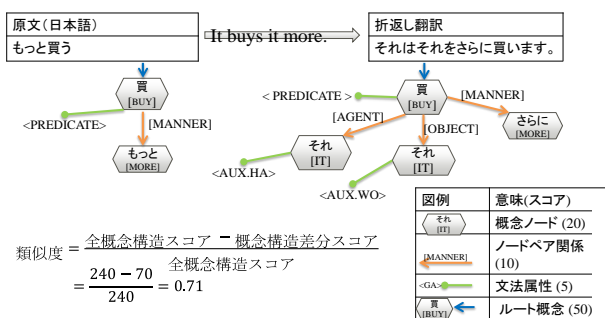


図 2: 原文と折返し翻訳結果の類似性の計算例

3.3 翻訳品質の変化による日本語動詞語尾崩れの判断

ステップ 3 では、原文および語尾を置換した置換結果の翻訳品質評価値を求めた後、その翻訳品質評価値の変化により、動詞語尾置換の効果を評価する。翻訳品質変化の閾値を設定し、語尾置換結果の翻訳品質の向上が閾値を超えた場合は置換された語尾表現が翻訳できない崩れた表現と判断する。その他の場合は語尾置換が文全体の翻訳品質の向上に貢献できないと判断する。

4 動詞語尾崩れ表現の検出実験

本章では、動詞語尾崩れ自動検出の提案手法効果を検証する実験を説明する。動詞語尾崩れ表現の出現原因が多様であり、崩れた表現が表す意味を推測することが困難である。そのため、本研究は動詞語尾崩れ表現の自動検出が目的であり、語尾崩れ部分の意味に相当する一般表現の自動サジェストを行わない。

本稿の実験で使われたデータは 2 章で紹介した IM を用いた（日本語ユーザ同士の）リアルタイム業務連絡の会話ログである。まず、我々は提案手法を会話ログに適用し、日中翻訳における翻訳できない動詞語尾崩れを検出した。その後、評価作業（ユーザ）が指摘箇所と同じ意味を持つ一般表現で置換し、手動置換前後の翻訳品質を評価した。提案手法の検出結果が正確であれば、ユーザが検出された箇所を一般表現で書き換えることで翻訳品質が向上し、会話の意思疎通の支障が軽減する。

表 3 では動詞語尾崩れ表現の自動検出結果を示している。会話ログは 22,715 件の発話があるが、重複発話を削除して 14,394 件の発話が残る（セル「会話ログの全発話」）。その中、提案手法が検出した動詞語尾崩れを持つ発話は 2,229 件がある（セル「検出された動詞語尾崩れの発話」）。

次に検出された 2,229 件の動詞語尾崩れ表現に対し、ユーザが手動で原文と同義の一般語尾表現で原文の語尾を修正した。語尾の手動修正は同義表現が望まれるため、同義表現が見つからない場合、該当箇所は「手動修正不可」と記された。その後、手動修正前・後の発話を日中翻訳ソフトで翻訳し、作業による翻訳品質評価（Acceptability）を行う。

表 4 では上記の手動修正前・後の日中翻訳品質を示す。「検出された動詞語尾崩れの発話」は自動検出された箇所の手動修正前・後の平均翻訳品質を示し、「全発話」では会話ログの全発話の平均翻訳品質を示す。その結果、自動検出された動詞語尾崩れの平均翻訳品質が手動修正前の 3.03 から手動修正後の 3.46 に増加し、全発話の平均翻訳品質が 3.41 から 3.48 に増加した。全発話の翻訳品質は劇的に増加

表 3: 総発話数と検出された動詞語尾崩れの数

	発話数
検出された動詞語尾崩れの発話	2229
会話ログの全発話	14394

表 4: 検出された動詞語尾崩れの修正前/後の日中翻訳結果の平均 Acceptability

説明	平均 Acceptability
検出された動詞語尾崩れの発話 (2229 件) の手動修正前	3.03
検出された動詞語尾崩れの発話 (2229 件) の手動修正後	3.46
全発話(14394 件) の手動修正前	3.41
全発話(14394 件) の手動修正後	3.48

表 5: 提案システムの処理結果に対する手動修正による日中翻訳品質の変化

Acceptability	発話数	%
増加	752	33.7%
同様	1163	52.2%
減少	60	2.7%
手動修正不可	254	11.4%
合計	2229	

しなかったが、提案手法が翻訳できない動詞語尾崩れ表現の検出が有効であると考えられる。ユーザが提案手法の処理結果に従い、指摘された動詞語尾を修正すれば、翻訳できない発話を軽減でき、IM 会話の効率が向上できると考えられる。

また、本提案手法で検出された動詞語尾崩れ表現を持つ発話の修正前の翻訳品質が 3.03 であり、全会話ログの平均翻訳品質 3.41 より大幅に低いことがわかった。本提案手法は正確に翻訳品質が低い発話を検出でき、翻訳品質の自動評価の利用にも可能であると考えられる。

表 5 では、検出された動詞語尾崩れの発話に対する手動修正による翻訳品質の変化を示す。2,229 件の動詞語尾崩れを持つ発話に対し、33.7% の発話が手動修正により翻訳品質が向上した。手動修正により翻訳品質が低下した件数は 2.7% のみである。一方、52.2% の発話は翻訳品質の変化がなかった。我々は、Acceptability を翻訳品質の評価指標としている。Acceptability での 5 段階の翻訳品質は明確な定義があり、定義を満たさないなら、ユーザの主観で「多少わかりやすくなった」と思っても Acceptability の段階を引き上げることができない。例えば、翻訳できない動詞語尾崩れ表現と未知語を同時に持つ発話 (Acceptability: 1) に対し、動詞語尾崩れ表現を修

正して動詞部分の意味が正確に翻訳できても、未知語部分が翻訳できないため、「重要な情報は全部翻訳した」の条件に満たさなく、Acceptability の向上は引き上げができない。そのため、表 5 の翻訳品質変化は Acceptability が変更なしの発話数が最も多かった。

5 まとめ

本稿では、機械翻訳を用いた IM 会話における意思疎通の障害を軽減するため、動詞語尾崩れ表現の自動検出手法を提案した。我々はまず、日本語の社内業務連絡の会話ログを分析することで動詞語尾崩れ表現が IM 会話の翻訳品質低下の重要な要因であることを明らかにした。一般動詞語尾を用いて会話文の動詞語尾表現を置換し、置換前後の翻訳品質を比較することで動詞語尾崩れ表現を自動に検出する手法を提案した。

提案手法の効果を検証するため、我々は提案手法を日本語業務連絡会話ログに適用させ、検出された動詞語尾崩れ箇所を手動修正してその翻訳品質の変化を検証した。実験結果は、検出された動詞語尾崩れ発話の平均翻訳品質が手動修正前の 3.03 から 3.46 に増加したことがわかった。ユーザが提案手法の検出結果に従って指摘箇所を修正すれば、機械翻訳を用いた IM 会話の意思疎通の障害が軽減すると考えられる。今後は、本手法の適用で意思疎通の障害の軽減効果を定量に検証する方式を検討し、機械翻訳を用いた IM 会話に対する提案手法の効果を検証する予定である。

参考文献

- [1] Inaba, R. *Usability of multilingual communication tools*. Springer-Verlag, City, 2007.
- [2] Aiken, M. Multilingual communication in electronic meetings. *SIGGROUP Bull.*, 23, 1 (2002), 18-19.
- [3] Komine, H., Kinukawa, H. and Nakagawa, H. *Sentence Extraction based on Document Frequency and Text Length*. City, 2002.
- [4] Goto, I., Chow, K. P., Lu, B., Sumita, E. and Tsou, B. K. *Overview of the patent machine translation task at the NTCIR-10 workshop*. City, 2013.
- [5] Miyabe, M., Yoshino, T. and Shigenobu, T. *Effects of undertaking translation repair using back translation*. ACM, City, 2009.