「自動翻訳エンジンの金融専用モデル構築に関する委託研究」

2022年3月

国立研究開発法人情報通信研究機構（NICT）

要約

- 2021年11月までに**金融庁より翻訳バンクに提供された金融分野の和文・英文の文書データを用いて、 NICTは金融分野向けに特化した専用モデルについて試作・評価した** 。
  ① 金融庁より提供された**和文・英文の文書データ（約600件）**から自動の処理で金融分野の**日英対訳の文（約18万文）**を抽出し、更に自動および半自動の処理で精製して**高品質の日英対訳の文（約8万文）**を得た。
  ② NICTは、この**高品質の日英対訳の文（約8万文）**を用いて金融専用モデルを構築し、同モデルに基づく自動翻訳システムで得られた訳文100文について、第3者（翻訳者）が文毎に翻訳品質を5段階評価し、**金融専業翻訳者に匹敵する最上位品質の訳文が全訳文の5割弱を占めることを確認した。**
- NICTは、2021年12月末まで得られた上記成果と2022年1月までに金融庁以外の各機関から翻訳バンクに追加された全データを併せて再構築した**金融専用高精度モデルの技術移転を2022年3月1日に開始している。**

目次

１．研究開発の流れ

２．受入処理

３．アラインメント

４．クレンジング

５．翻訳精度の自動評価

６．翻訳精度の人手評価

７．結語

１．研究開発の流れ

特定分野における自動翻訳の精度を向上させるためには、汎用の翻訳モデルを当該分野の翻訳文書に合わせて変更する（これを**適応**と呼ぶ）。適応は文単位の対訳に基づく処理なので、入手した翻訳文書を高精度の文単位の対訳に効率的に変換しなくてはならない。これが本研究開発の前半にあたる。後半は前半の各処理後得られる適応モデルの翻訳精度の計測である。1章では前半を構成する3ステップの概要を説明する。

A) 先ず、翻訳例を取捨選択する。今回の翻訳例の四分の三は、以降の処理に好適な文書単位、すなわち、原文書のdoc ファイルに対応する翻訳文書のdocファイルであった（これをdocファイル対と呼ぶ。図1）。docファイルとはWORDのフォーマットであり、ボールド指定など翻訳で活用されない情報が含まれている。これらを削除したフォーマット（txtファイルと呼ばれる）に変換する（これを**受入処理**と呼ぶ）。

Introduction

The American Chamber of Commerce in Japan (ACCJ) fully supports the recently re-energized efforts, both by the national and Tokyo metropolitan governments, to reestablish Japan's position as a preeminent global and regional financial center. We believe it is imperative that Japan's standing in global finance more fully reflect the tremendous scale of its economy and investment capital. We further believe that adopting policies to achieve success in this effort is an important step for Japan to return to sustained and strong economic growth.

序文

在日米国商工会議所（ACCJ）は、日本政府と東京都が日本を世界有数の国際・地域金融センターとして再構築するために再び活発化させている昨今の取組みを全面的に支持する。ACCJは、国際金融における日本の地位が、その極めて大きな経済規模や投資資本規模をより十分に反映したものになることが必要だと考えている。また、この取組みを成功させる政策の導入が、日本が持続的で力強い経済成長に回帰するための重要な一歩になると確信している。

図1 docファイル対の例

～別々の文書ファイルのテキストに対訳関係がある～

B）文書単位で対応したデータから文単位で対応したデータ（文単位で対応したデータを、以後対訳候補と呼ぶ）に自動的に変換する必要がある（これを**アラインメント**と呼ぶ）。

C）自動で行うアラインメントは費用も時間も非常に小さくて済むが、原文と翻訳文の対応が間違っていたり、訳質に問題があることがある。この不良な文対応は適応による精度向上の障害になることがある。この障害を減少させるため不良な文対応を除外する（これを**クレンジング**と呼ぶ。クレンジングは自動と半自動処理の2段階に分かれている）。

今回の翻訳例の残り四分の一は、対訳候補が一つのdocファイルまとめられたもの（これを併記docファイルと呼ぶ。図２）である。併記docファイルは文単位であるものの、文書構造にかかわる情報、様々な人間向き注釈等の追記や意訳などがあり、適応の障害になるものを含むので、これらを除外した。

次ページの図３に処理のフローをまとめた。

①【日 付】
この日付は書類作成日ではなく、申請書提出日を記載する。

①【Date】
This date should be the submission date of written application for registration under Article 29, not the date of documentation.

②【宛 名】
記載例のとおり、本店等の所在地を管轄する財務局等が関東財務局であれば、「関東財務局長」と記載する（以下、「関東財務局長」宛に提出されることを例に記載）。

②【To】
As stated in the sample entry, describe “Director-General of the Kanto Local Finance Bureau” in cases where it is the Kanto Local Finance Bureau that has jurisdiction over the location of the head office, etc. of the applicant (hereinafter, on the assumption that the written application for registration is to be submitted to “Director-General of the Kanto Local Finance Bureau.”)

図2 併記docファイルの例

～一つのファイルの中に対訳関係がある～

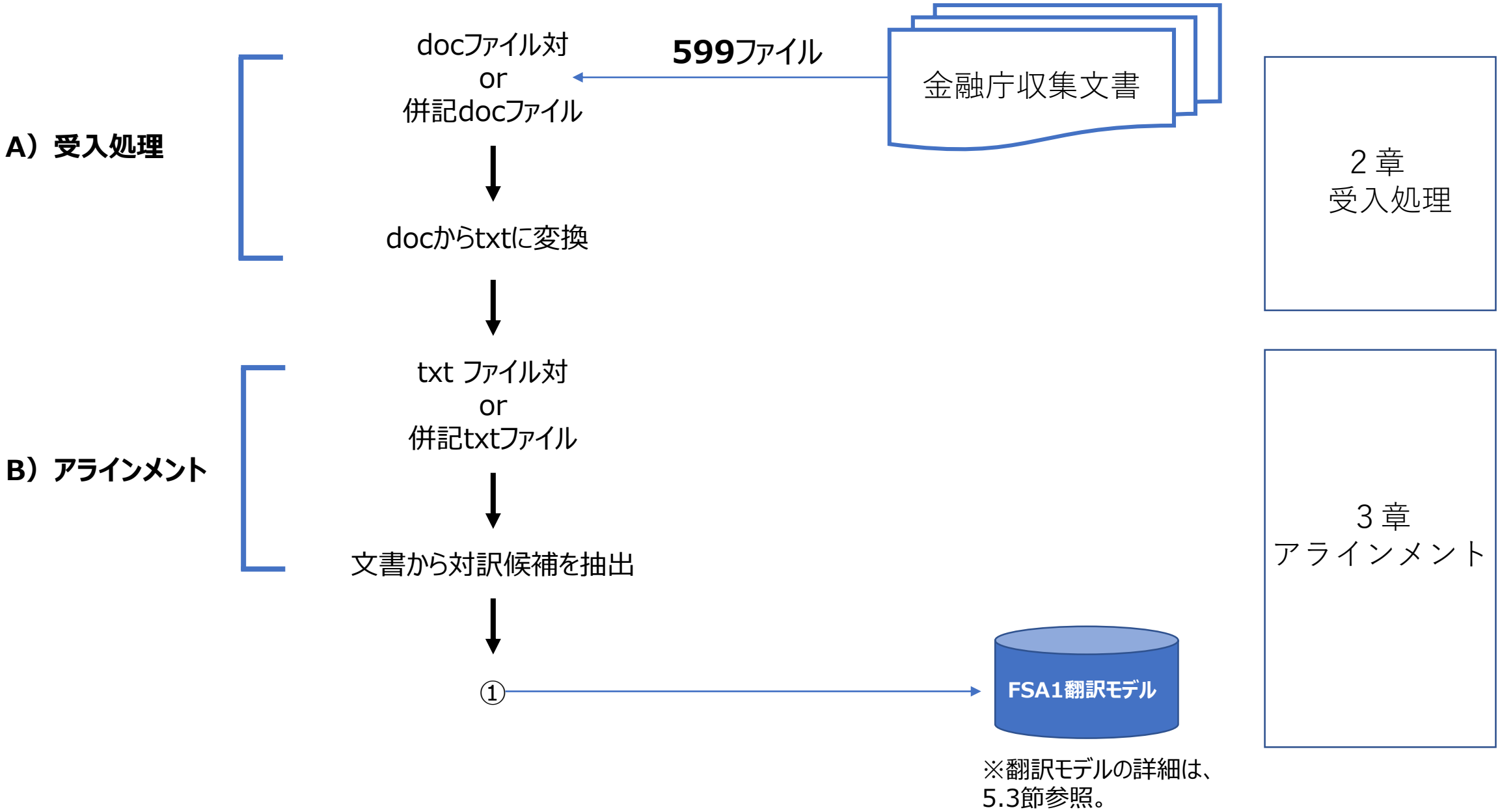


図3(a)　受入処理・アラインメント

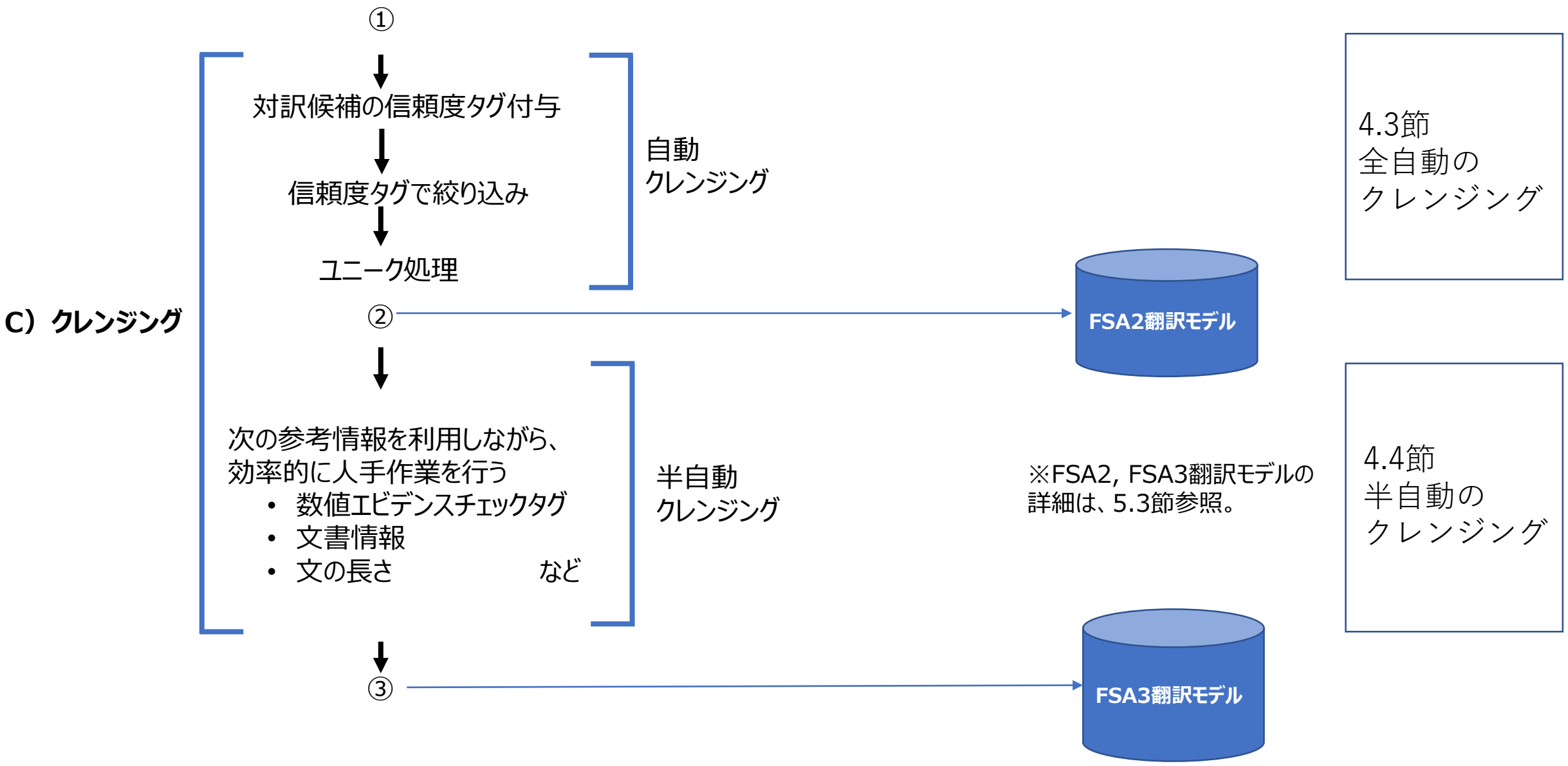


図3(b) クレンジングの流れ

2．受入処理

2.1 対象文書

金融庁より提供された日英対訳文書を精査し効率的に利用可能な文書を選別した（PPT等は効率的な作業が不可能であることから対象としなかった。）。

選別された文書は599あり、docファイル対（日英対で１と計数）と併記docファイルの個数を表1にまとめた。

表1 提供を受けたファイルの個数

ファイル種別	ファイル数（個）
docファイル対	453
併記docファイル	146
合　計	599

2.2 docファイルからtxtファイルへの変換

docファイルからフォントなど不要な情報を削除しtxtファイルに変換した。さらに、空行、全角スペース、空白（タブ含む）のみからなる等の無効なデータを削除した。

docファイル対から変換されたものをtxtファイル対、併記docファイルから変換されたものを併記txtファイルと呼ぶ。

3．アラインメント

txtファイル対や併記txtファイルから文単位で対応したデータである対訳候補を取り出すアラインメント処理を行った。

3.1 利用プログラム

●txtファイル対

対応関係にあるtxtファイル対からアラインメント・プログラム※を利用して、対訳候補を取り出した。

※ アルゴリズムの詳細は次を参照
https://www.aclweb.org/anthology/P03-1010.pdf

●併記txtファイル

日本語文と英語文の位置関係を利用するプログラムを新たに作成し対訳候補を取り出した。

表2 受入毎の対訳候補数

対訳候補数
179,856

４．クレンジング

3.1の自動抽出した対訳候補には次の障害が含まれている。

- 関係のない対訳
- 部分的にあっている対訳
- 余分な記号が付与された対訳
- 日付などの事実の対応が間違っている対訳
- 空白

自動翻訳の質を向上させるため、これらの障害を削除するクレンジングを実施した。

クレンジングには、人の手を介さないプログラムによる自動クレンジングと、プログラムにより補助情報を生成して人手で行う半自動クレンジングがある。
最初は、自動クレンジングを行い、次に半自動クレンジングを行う。

４.１　クレンジング作業の進行

クレンジング作業を進めるにしたがって、翻訳モデルの構築に利用する訓練用の対訳候補数が徐々に絞り込まれていく。図4に対訳候補数の推移を示す。クレンジング処理は処理１から処理５までの５つの処理で構成されている。処理１から処理５まで処理のレベルが進むにしたがって、前の処理の結果を受けて、次の処理を行う。障害のある対訳候補が除かれていくので、対訳候補数は徐々に削減されていく。

クレンジング作業の効果を確認するために、クレンジングの成果である対訳候補を訓練データとする３つの翻訳モデルを構築した。その翻訳モデルをFSA1、FSA2、FSA3と呼ぶことにする。図4では、FSA1翻訳モデルは処理1，FSA2翻訳モデルは処理3、FSA3翻訳モデルは処理5の対訳候補をそれぞれ利用していることを示している。翻訳モデルの詳細は5章で説明する。

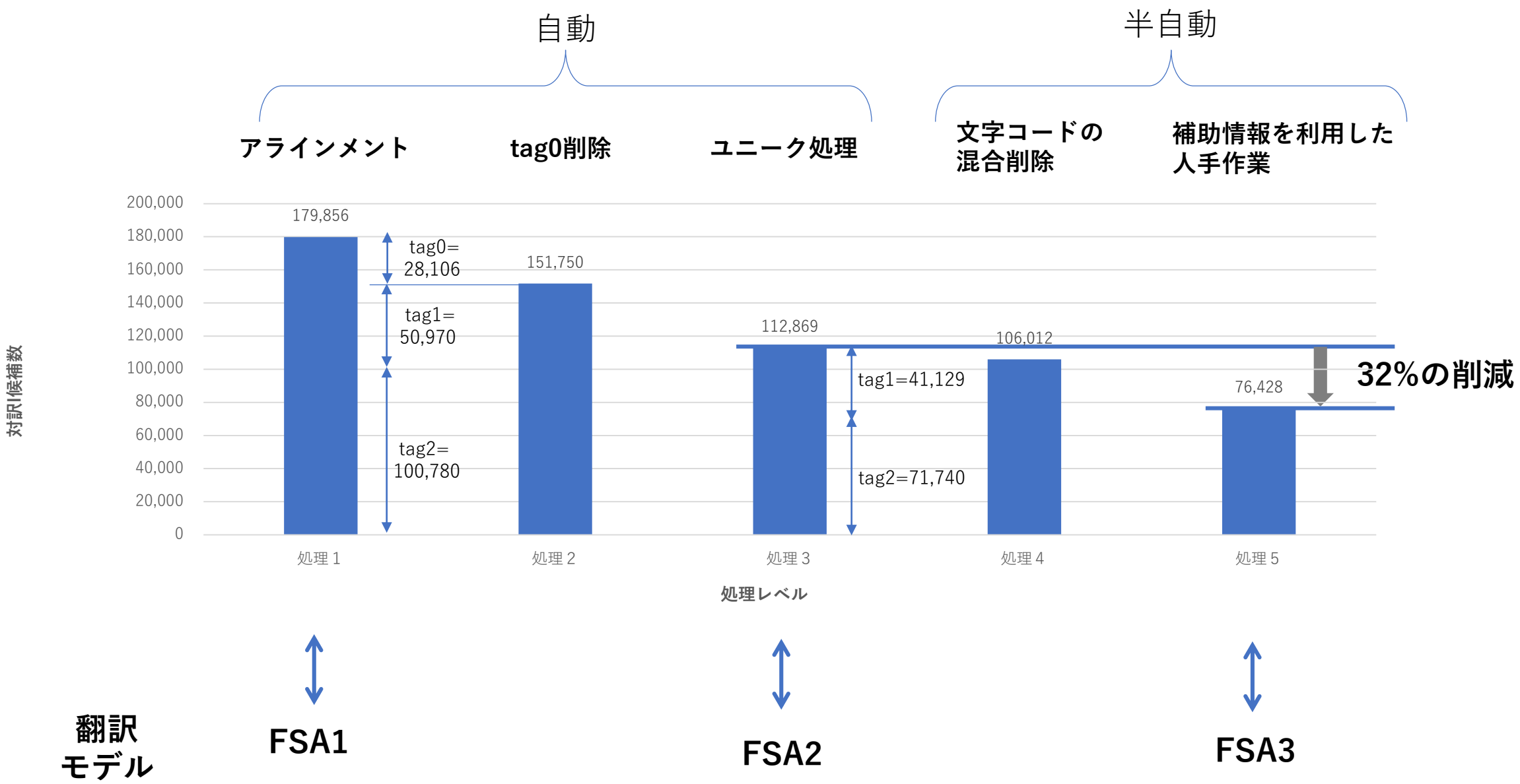


図4 クレンジング処理の進行と対訳候補数の推移及び翻訳モデルと処理レベルの関係

4.2　自動クレンジングと半自動クレンジング

クレンジングの処理１から処理３までは、プログラムによる処理であり、自動クレンジングである。

一方、処理４と処理５は、半自動クレンジングである。

処理1では、併記ファイルと別ファイルがあったので、それぞれに対応しながら、第３章で説明したアラインメントにより対訳を抽出した。図4では、ここを対訳候補数の起点としている。

処理2では、逆翻訳に基づく信頼度タグを付与して、適応に寄与できる対訳候補を選択した。信頼度タグには、tag=0,1,2のタグがある。

処理3では、複数の箇所に存在する同じ対訳候補を重複してクレンジング作業しないように、対訳候補をユニークにする処理を行った。

処理4では、日本語と英語の混合度合いを尺度を利用して、不適格な対訳候補を削除した。

処理5では、クレンジング分析タグなどの補助情報を利用しながら人手作業により、不適格な対訳候補を削除した。

処理3の結果に対して、処理4と５で不良な対訳候補32%を削除して、最終的には68%の良質な対訳を選び出すことができた。

4.3 自動クレンジング
処理１、処理２、そして処理３について説明する。

4.3.1 処理１（アラインメント）

３章で説明したアラインメントにより対訳候補を抽出する処理をクレンジングの起点として処理１とした。

4.3.2 処理2（信頼度タグによる対訳候補の選択）

最初に、信頼度スコアを対訳候補に対して計算し、この信頼度スコアを利用して対訳候補に信頼度タグを付与する。最後に、信頼度タグに基づいた対訳候補の絞り込みを行う。

4.3.2.1 信頼度スコアの付与

クレンジング作業を効率的に行うために、対訳候補にまず次に説明する信頼度スコアを計算する。

<u>信頼度スコア</u>

- 日英信頼度スコアとは、日英対訳文において、英語文を日本語文に自動翻訳し、その結果と元の日本語文との距離を０～１の値で表すものである。
- 英日信頼度スコアとは、その逆で、日本語文を英語文に自動翻訳し、その結果と元の英語文との距離である。

距離にはレーベンシュタイン距離※を用いた。

※https://en.wikipedia.org/wiki/Levenshtein_distance

4.3.2.2 信頼度タグの付与

日英信頼度スコアと英日信頼度スコアを利用して、各対訳候補に対して信頼度タグを付与する。本タグは、信頼度の高い方から順に、２、１、０の３つの値とする。値の意味は次のとおりである。

信頼度タグの意味
２： 日英信頼度スコア >= 0.5 且つ 英日信頼度スコア >=0.5
１： 日英信頼度スコア、英日信頼度スコアのいずれか一つ >=0.5
０： 日英信頼度スコア<0.5 且つ 英日信頼度スコア<0.5

信頼度タグが１あるいは２の対訳候補は、次のクレンジング作業の対象とした。

4.3.2.3 信頼度タグ＝0の対訳候補の削除

表3に示すように、処理1が終了した時点で、179,856の対訳候補があった。この対訳候補に信頼度タグを付与すると、その内訳は表3に示す通りである。

信頼度タグ＝0の対訳候補は、障害のある対訳候補であるので、これを削除した。

表3 信頼度タグ毎の対訳候補数

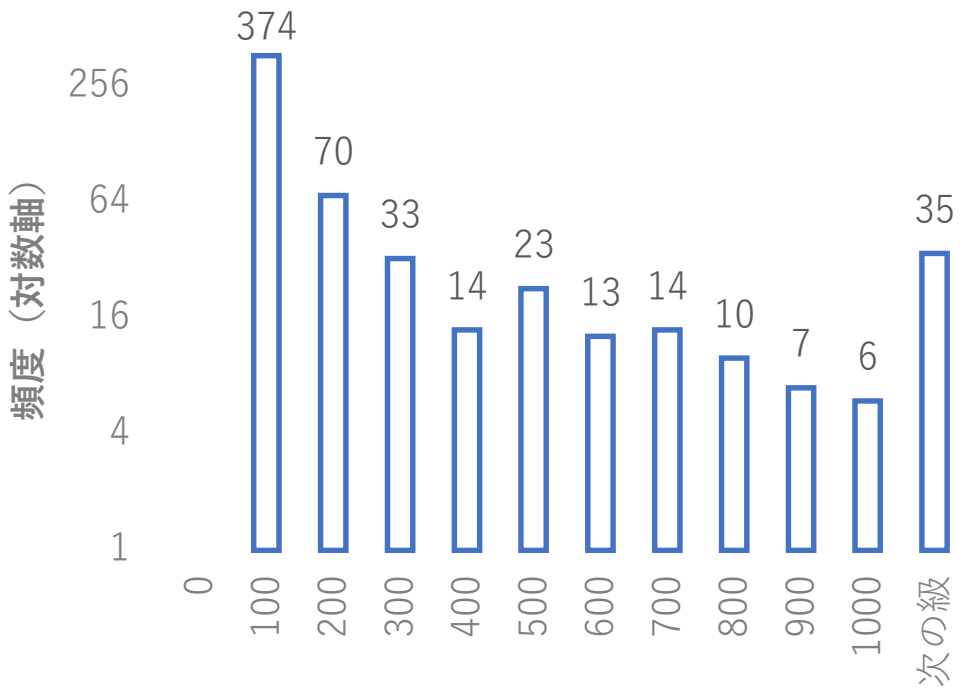
説明	対訳数
アラインメントで取り出された対訳候補数	179,856
信頼度タグ＝2の対訳候補数	100,780
信頼度タグ＝1の対訳候補数	50,970
信頼度タグ＝0の対訳候補数	**28,106**

4.3.2.4 ファイルに含まれる対訳候補の分析

受入のファイルごとに対訳候補の数にばらつきがある。

ファイル当たりの対訳数のヒストグラムを図5に示す。図5の縦軸は、対数としてしている。図5によると、 100個から200個未満の対訳を得ることができたファイルが374ファイル（62%)ある。また、200以上の対訳候補が取り出されたファイルもあることがわかる。

受入処理した599個のファイルは各ファイルそれぞれから有効な対訳が抽出できたといえる。



（注）横軸の次の級とは、1100以上の対訳候補の数をまとめて数えていることを表している。

図5 各ファイルに含まれる対訳候補の数

4.3.3 処理3（ユニーク処理）

対訳候補の中には、まったく同じ対訳が含まれている。そこで、（日本語、英語）の対訳対が完全に同じものを一つに絞りこむ処理を行った。

処理3の段階で、112,869対訳候補である。

処理3が終わった段階で、クレンジングの効果を測定するために、適用翻訳モデルの訓練用対訳を準備した。訓練用対訳は信頼度タグの情報で重みづけした。すなわち、信頼度タグ＝2は信頼度が高いので２度利用し、信頼度タグ＝１は１度利用した。 5章の翻訳精度の自動評価で利用した。

4.3.5 評価セット

ここでは、翻訳モデルの性能を評価するための対訳（以後、評価セット）を決定した。

- 先ず、ランダムに選んだ1,000文からなる評価セット①を準備した。
- 次に、評価に最適ではない20文字以下の文を削除し、更に、人手で確認しつつ不完全な対訳を削除し、**620対訳からなる評価セット**②を選択した。
- 人手評価を行うために、更に、ランダムに減数して**100対訳からなる評価セット**③を選択した。

以後、評価セット②を評価セット620、評価セット③を評価セット100と呼ぶ。

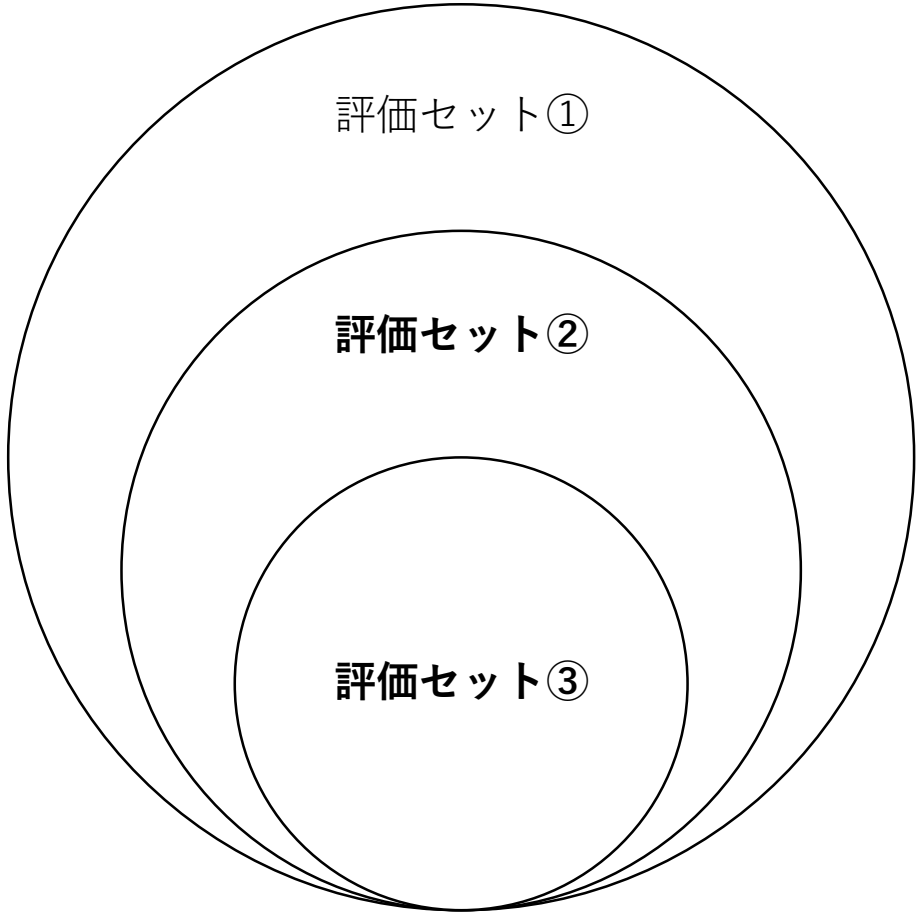


図6 評価セットの説明

4.4 半自動クレンジング

4.4節では、人手と自動を合わせた半自動クレンジングの処理4と処理５について説明する。

4.4.1　処理４（１バイトコードの含有率が多い対訳の処理）

次の例のように日本語テキストに１バイトのアルファベットが含まれている場合には、対訳候補として利用できないケースが多い。

日本語	スピードAgility	英語	Agility

一般に、１バイトコードと２バイトコードの割合を尺度として、**目視により不適格な対訳候補を排除した。**

４．４．２ 処理５（補助情報を利用した人手処理）

4.4.2.1に例示した補助情報を利用しながら、人手でクレンジング作業を行った。

4.4.2.1 数値エビデンスチェックタグの付与と本タグ利用の作業

日本語と英語に含まれる数値の内容が一致しているかプログラムによりチェック※をし、次の通りタグを付与して、作業プロセスの中で利用した。

数値エビデンスチェックタグ
１：　一致
０：　不一致

数値エビデンスチェックタグは、人手チェックの際に利用した。

最終的なクレンジング結果で本タグの割合を評価すると、
１：　86.96%
０：　13.04%
となる。

数値エビデンスチェックタグが0の対訳候補の例を図7に示す。数値を中心に対訳を確認することにより、対訳が誤っていることを効率的に確認できる。このように、数字エビデンスチェックタグは、数値の誤りにフォーカスしてチェック可能であり、人手だけだと見過ごされる誤りを発見できるので有効である。

不一致の例

a. **詳細は事業スキーム例②(ｖ)（26頁）をご参照ください。**
For more details, please refer to Business scheme case 2)

V,26頁に対応した英語表現がないので、削除した。

b. **健康増進特別企画として「事業所対抗 徒歩で伊勢神宮（197km）を目指せ！」を実施しました。**

to Ise Shrine! (197 km)" as part of its efforts to achieve "Goal 3: Ensure healthy lives and promote well-being for all at all ages.”

ゴール３に対応した日本語表現がないので、削除した。

図7 数値エビデンスチェックタグ＝０の例

5. 翻訳精度の自動評価

5.1. 目的と手法

本章は、金融庁文書の有用性とNICTで行った各処理の有用性を翻訳精度の観点で調べることを目的とする。

比較対象は、以下の４種類の翻訳モデルである。
①NICTから広く公開されている汎用モデル
②金融庁文書を用いることによって得られた３つの適応モデル

多数の翻訳結果の比較が可能になるように、プログラムによる自動評価法BLEUを採用した。

次節以降で、BLEUの概要とBLEUによって得られた評価結果を説明する。

5.2　BLEU

BLEUは訳質を示すスコアで、値が大きいほど高精度である。

BLEUは、①原文に対する人間による模範訳と②自動翻訳の訳文の類似度をコンピュータで計算する※。

日本語の原文、英語の参照訳①　、金融専用モデルの訳②の３つを表5に記載した#　。この表の　①と②はかなり類似度が高いので金融専用モデルは一定の翻訳品質と言える。

実際には、BLEUは人手評価との相関を高くする工夫として、1文単位で計算するのではなく、数百文まとめて計算する。

※　計算方法の詳細は以下の論文とプログラムを参照。
https://aclanthology.org/P02-1040.pdf
https://github.com/odashi/mteval

表5 BLEUは原文①に対する見本訳と②自動翻訳の訳の類似度を示す尺度

原文	自社の実態と相違するなど、虚偽記載の場合は、行政処分の対象となります。
①　参照訳	False statements, such as those that differ from the company's actual situation, will be subject to administrative punishment.
②　自動翻訳	False statements, such as those that differ from the actual situation of the company, are subject to administrative sanctions.

5.3. BLEUによる日英モデル比較

翻訳精度向上を簡易に確認するため、モデルによる翻訳精度の差を、プログラムで自動的に評価した結果を図12に示す。 また、評価に利用した評価文を

- 下記条件で評価。
  - **BLEU**（自動翻訳の訳文と人間の模範文の類似度）
  - 評価文数 評価セット620
- 4モデルを比較。
  - **汎用**
  - 今回追加した３モデル
    - **FSA1**（自動アラインメント後の適応モデル ）
    - **FSA2**（自動クレンジング後の適応モデル）
    - **FSA3**（半自動クレンジング後の適応モデル）

5.4. BLEUによる日英モデル比較

実験の結果、全体で18ポイントの改善。

- FSA1が汎用に対してスコアを大きく改善。
- FSA2はFSA1に対してスコアで2.4ポイント改善。
- FSA3はFSA2に対してスコアで2.2ポイント改善。

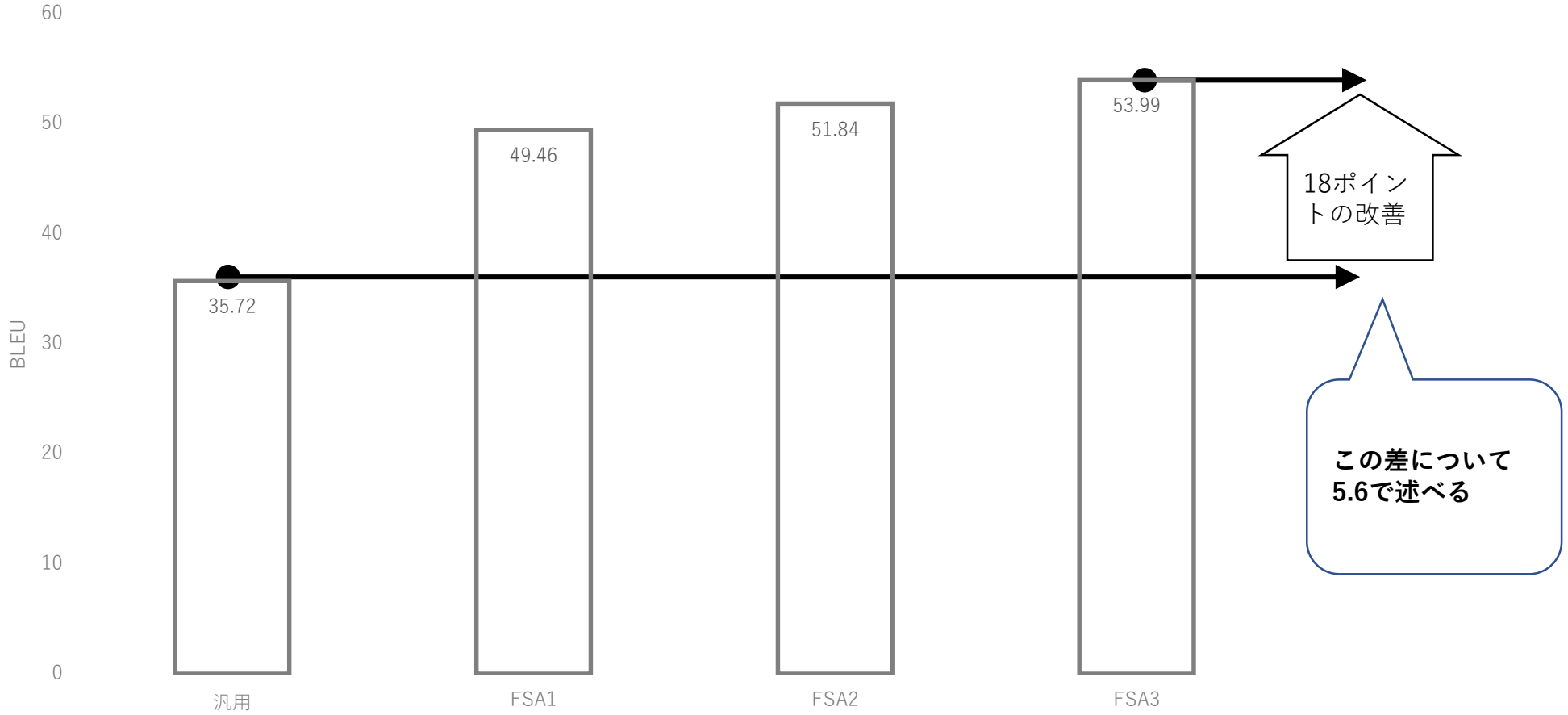


図12 自動評価による日英モデルの翻訳精度比較

5.5. BLEUによる英日モデル比較

実験の結果、全体で15ポイントの改善。

- 汎用からFSA1、FSA1からFSA2、FSA2からFSA3への改善の傾向は日英と同様であった。

金融庁文書からのアラインメント・クレンジングによって、BLEUで、**日英双方向で翻訳精度を改善することを確認できた。**

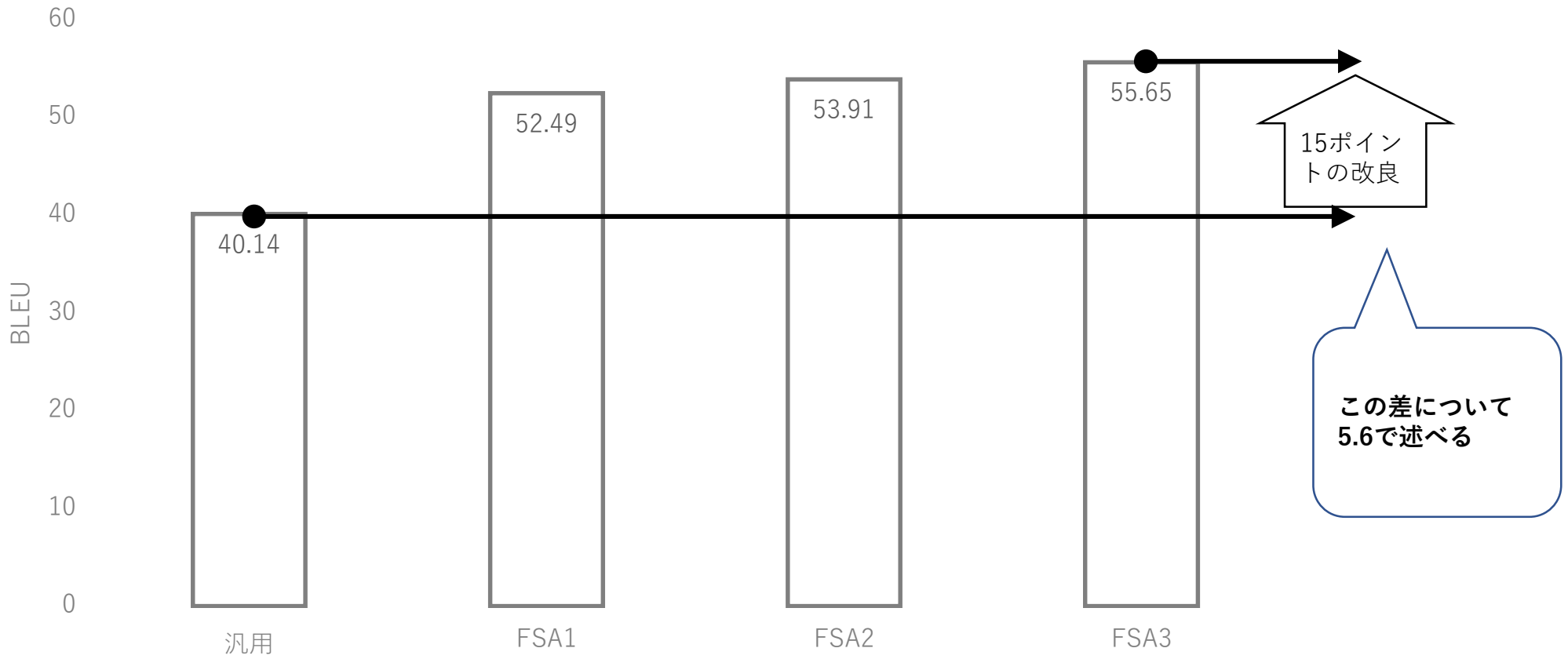


図13　自動評価による英日モデルの翻訳精度比較

5.6 BLEUの解釈

大まかにはBLEUは図14のように解釈できると言われている㊟ 。

この解釈に従うと5.4と5.5のBLEUの絶対値に照らして、**日英双方向で翻訳モデルを変えたことで平均的に「理解できる、適度な品質の翻訳」から「非常に高品質で、適切かつ流暢な翻訳」に改善できた**と言える。

次節で汎用と金融専用の２つの翻訳モデルの翻訳精度の差を人手評価により検証する。

BLEU	解釈
60<	人が翻訳した場合よりも高品質であることが多い
50～60	非常に高品質で、適切かつ流暢な翻訳
40～50	高品質な翻訳
30～40	理解できる、適度な品質の翻訳

本研究の精度改善

図14 スコアBLEUの『直観的な』解釈

㊟ https://cloud.google.com/translate/automl/docs/evaluate

６． 翻訳精度の人手評価

5. 章で、汎用とFSA3（ここではFSA専用モデルと呼ぶ）の翻訳精度の差が大きいこと分かったので、この２つの翻訳モデルについて人手で評価した。（前記620文からランダムに選んだモデルにとって未知の）原文100文に対応する翻訳結果を評価の対象にして、金融専門現業翻訳者1名が翻訳品質を 5段階、高い方からP, S, A, B, Xで判定した。

６.１ 評価結果

FSAモデルは、汎用モデルよりも総体的に高品質。

- FSAモデルは、Pランクが約5割、 P, S, A, Bの合計が約９割。
- 汎用モデルは、Pを始め各ランク約２割で、 P, S, A, Bの合計が約８割。

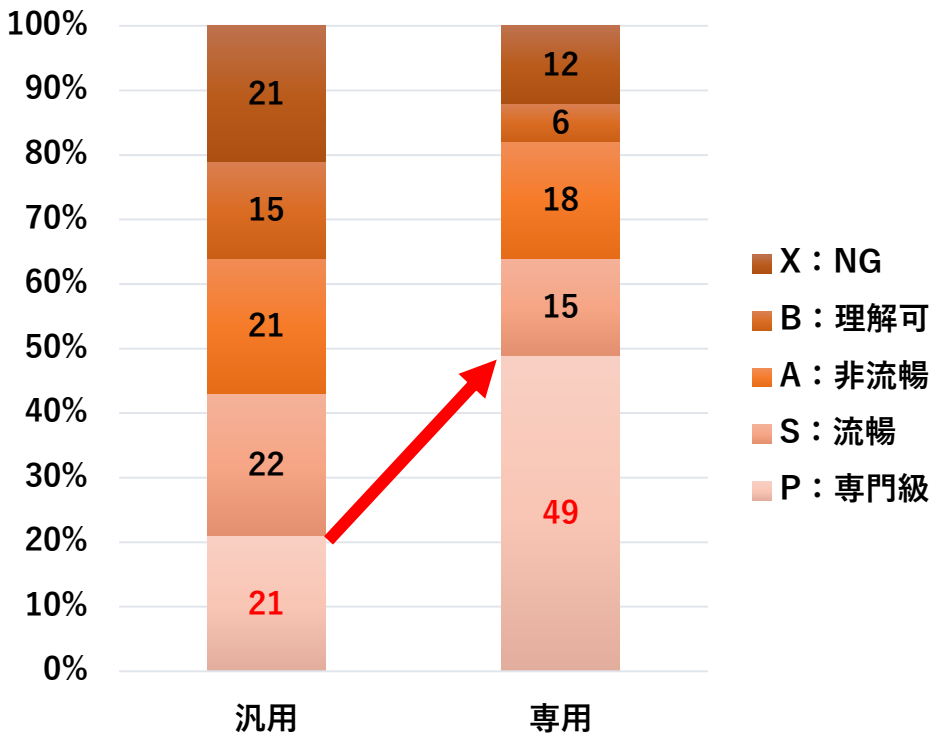


図15 ランク評価による翻訳精度のモデル比較

6.2　高精度化事例～訳文前後比較～

自動翻訳（汎用）の訳文がランクBで、自動翻訳（FSA）の訳文ランクPに改善した２例を示す。

日本語原文	自動翻訳（汎用）の訳文	自動翻訳（FSA）の訳文
保険契約に解約価額がある場合、契約を解約する保険契約者に損失を負担させることを可能にするための解約時の契約価額の削減	where insurance contracts have a surrender value, reducing the value of contracts upon surrender to enable losses to be imposed on policyholders that surrender their contracts.	A reduction in the value of policies at surrender to enable policyholders who surrender policies to bear losses if they have a surrender value ;
なりすましに対抗するため、各参加者に固有の生体認証番号（BIN）が作成されている。	A unique Biometric Identification Number (BIN) is created for each participant to counter identity fraud.	To counter identity theft, a unique Biometric Identification Number (BIN) is created for each participant.

7.　結語

本研究以前は、金融関係の対訳データの収集可能な規模および金融分野向け翻訳における収集対訳による改善については未知であったが、次のポジティブな結果が得られた。

A）金融庁の集中的努力で収集できた約600件の金融関係の文書から、**約8万件の高品質な対訳が取得できた。**

B）日英・英日の双方向で**翻訳精度改善への収集した対訳の有効性を確認できた。**

a. 日英・英日の**双方向で**自動評価BLEUを実施し、その直観的解釈において「理解できる品質」から「高品質で流暢な品質」に改善できた。

b. 第3者による日英方向の翻訳品質評価で、**金融専業のプロ翻訳者に相当する品質の訳文がほぼ5割に達し翻訳精度の大幅な改善が確認できた。**

C）未実施の高精度化手法がいくつかあり、一層の高精度化について期待できる。