オフィス文書のための翻訳支援ツールの実装と評価

Implementation and Evaluation of Translation-Assistance Tool for Office Documents

伊川 洋平[*1] Yohei Ikawa
金山 博[*1] Hiroshi Kanayama
竹内 広宜[*1] Hironori Takeuchi
渡辺 日出雄[*1] Hideo Watanabe
三品 拓也[*1] Takuya Mishina
秋本 仁志[*1] Hitoshi Akimoto
清水 淳也[*1] Junya Shimizu

[*1]日本アイ・ビー・エム株式会社 東京基礎研究所
Tokyo Research Laboratory, IBM Japan

We developed a translation-assistance environment for office documents. This environment improves the productivity of translation work by providing tools for resource management, document quality checking, and manual translation. In order to efficiently assist the work for office documents, these tools must work directly with the applications. Therefore, we developed the Office Document Analysis Framework, which provides APIs for scanning and manipulating office documents, and implemented our tools in the translation-assistance environment using this framework. We also evaluated the efficiency of the translation-assistance tool by comparing translation times with the tool to manual translation times.

1. はじめに

ビジネス環境のグローバル化に伴い，オフショア開発の活発化や知識共有，生産性向上などの観点から，翻訳に対するニーズが高まっている．これまで長年にわたって機械翻訳に関する研究が行われてきている [1] が，機械翻訳は精度の点で問題があり，正確さが求められるビジネス文書の翻訳においては，人手による翻訳に頼らざるを得ないのが現状である．そこで我々は，人手による翻訳作業を支援し，翻訳にかかる時間を軽減しつつ，翻訳の品質を高めることを目的として，翻訳支援環境を構築した．

また，ビジネスにおいて頻繁に作成・利用される文書形式は，ワープロ文書，スプレッドシート，プレゼンテーションなどの，いわゆるオフィス文書である．アプリケーション上で文書の完成形を確認しながら作業できるツール群を提供することで，効果的な作業支援を行うことが可能となる．そこで我々は，オフィス文書における生産性向上のためのアプリケーションの基盤となる Office Document Analysis Framework (ODAF) を開発し，その上に翻訳支援環境におけるツール群を構築した．

オフィス文書翻訳支援環境の全体構成を図 1 に示す．この翻訳支援環境は，リソース管理，文書の品質チェック，人手による翻訳作業の 3 つのプロセスを反復して行うことで，翻訳のためのリソースを拡充しながら翻訳作業の効率を高めていくことを目的としている．また，辞書，翻訳メモリはサーバー上で管理されており，言語リソースの共有を行うことでチームで分担して翻訳作業を行う際に品質の平準化を図る．

本論文では，オフィス文書翻訳支援環境の基盤となっている ODAF と，人手による翻訳作業を支援する Translation Assist Tool (TAT) について述べる．また，TAT についての評価実験を行い，その有効性について検証する．

連絡先: 伊川 洋平，日本アイ・ビー・エム株式会社 東京基礎研究所，yikawa@jp.ibm.com

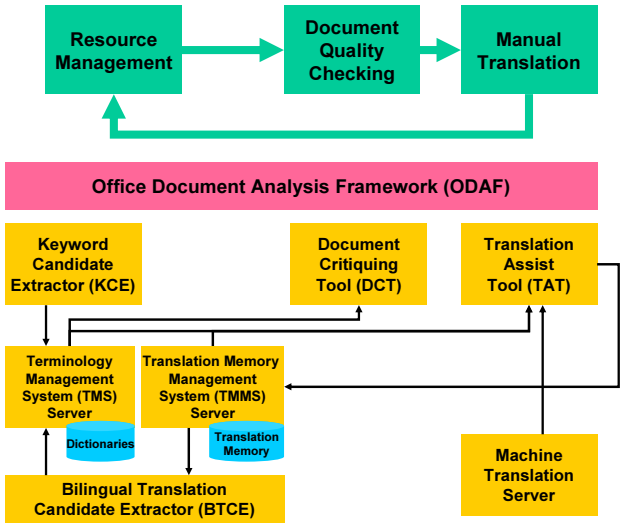

図 1: オフィス文書翻訳支援環境の全体構成

2. 関連研究

近年，言語グリッド [5, 6, 7] に関する研究が活発に行われている．言語グリッドは，インターネット上の言語リソースや機械翻訳などの言語処理機能を自由に組み合わせて新しいサービスを生み出すことを可能にする．一方で，本研究では手作業による翻訳の支援を効果的に行うことを目的としている．

市場には様々な翻訳メモリツールが存在する．例えば，Felix [3] は，オフィス文書上で翻訳支援を行う点で本研究における翻訳支援ツールと類似している．本研究で提案する翻訳支援環境は，翻訳メモリツールに加えて，翻訳対象の文書の品質を改善する文書校正ツールや，翻訳対象の文書から辞書登録する単語を抽出するツールを提供し，翻訳作業全体の効率化を目指すものである．

また本研究では，オフィス文書を操作するための共通のフレームワークを開発した．翻訳支援環境で提供されるクライ

アントツールは，このフレームワーク上に実装されている．オフィス文書に対してデータの読み出しや書き込みを行うためのインターフェイスとして，例えば Apache POI [2] などが知られている。本研究で開発したフレームワークは，データの読み出しや書き込みに加えて，オフィス文書上で指定箇所をハイライトしたり，現在選択しているオブジェクトのテキストなどの情報を取得することができ，ユーザーがインタラクティブにオフィス文書を操作できるアプリケーションを容易に構築することが可能である．

3. Office Document Analysis Framework

Office Document Analysis Framework (ODAF) は，オフィス文書から構造情報やテキストを取得したり，文書を操作するための API を提供する．ODAF は現在，Microsoft® Office (Word, Excel®, PowerPoint®) に対応しているが，ODAF の概念は特定のアプリケーションに限定されない．例えば，Open Document Format [4] や HTML などあらゆる形式の文書に対応することが可能である．

以下，ODAF において重要な概念である Location Path と，ODAF が提供する API を概観し，オフィス文書翻訳支援環境において ODAF 上で開発された，生産性向上のためのクライアントツールについて説明する．

3.1 Location Path

ODAF では，文書の構造情報を Location Path と呼ばれる形式で表現する．Location Path は，ドキュメント，ユニット，オブジェクト，テキスト範囲の 4 階層から成るパスの形式で表現される，テキストボックスや表のセルなどの文書を構成する要素を一意に特定するための識別子である．

Location Path の一般的な表記を以下に示す．

/DocType(FileName)/UnitType(UnitName)
/Obj_1Type(Obj_1Name)/...
/Obj_nType(Obj_nName)/TextRange(begin,end)

また，Location Path の例を以下に示す．

- /Document(C:\xxx.doc)/Paragraph(3)
- /Workbook(C:\yyy.xls)/Sheet(3)/Cell(5,10)
- /Presentation(C:\zzz.ppt)/Slide(5)/Group(1) /Shape(2)/TextRange(2,5)

必ずしも全ての要素を記述する必要はなく，必要に応じた詳細度のものを利用する．ただし，上位の要素を省略して下位の要素を記述することはできない．1 つ目の例では，ユニットまで指定されており，オブジェクトは指定されていない．2 つ目の例では，オブジェクトまで指定されており，3 つ目の例では，オブジェクト中のテキスト範囲まで指定されている．

ドキュメントの階層では，文書をファイル単位で指定する．

ユニットの階層では，文書形式ごとに定義された特定の単位を指定する．例えば，ワープロ文書では段落，スプレッドシートではワークシート，プレゼンテーションではスライドが主なユニットである．

オブジェクトの階層では，テキストボックスやセルなど，ユニットに含まれる要素を指定する．オブジェクトはユニット中の要素を階層的に指定することができ，グループ中のテキストボックスやグラフ中のデータラベルなど，階層構造を持つオブジェクトを指定できるようになっている．

テキスト範囲の階層では，オブジェクト中のテキストの範囲を指定する．例えば，後述する ODAF Manipulator が提供するハイライトの API は，Location Path で指定されたオブジェクト中のテキスト範囲をハイライトする．

3.2 ODAF APIs

ODAF は，ODAF Scanner, ODAF Manipulator の 2 種類の API で構成されている．

ODAF Scanner は，文書中のオブジェクトの Location Path を取得したり，Location Path で指定されるオブジェクトの情報を取得するための API 群である．例えば，文書の先頭から逐次的に Location Path を取得する API，アプリケーション上でユーザーが選択しているオブジェクトの Location Path を取得する API，Location Path で指定されたオブジェクト中のテキストを取得する API が提供されている．

また，ODAF Manipulator は，文書の内容を変更したり，アプリケーションのビューを操作するための API 群である．例えば，文書中のテキストを変更する API，Location Path で指定されたオブジェクトのテキストをアプリケーションのビューを切り替えてハイライトする API が提供されている．

ODAF Scanner, ODAF Manipulator の API を利用することで，オフィス文書上で動作する生産性向上のためのあらゆるアプリケーションを効率的に開発することが可能になる．

3.3 ODAF を利用したアプリケーション

図 1 に示したオフィス文書翻訳支援環境で提供されるクライアントツールは，ODAF 上に実装されている．ここでは，Keyword Candidate Extractor と Document Critiquing Tool の概要を示す．Translation Assist Tool については，第 4 節で詳細を述べる．

3.3.1 Keyword Candidate Extractor

Keyword Candidate Extractor (KCE) は，オフィス文書中に含まれる名詞句をキーワードとして抽出する．これから翻訳を行う文書に対して KCE を実行し，頻度の高い順番にキーワードの訳語を TMS に登録していくことで，翻訳対象に即した効率のよい辞書リソースの作成が可能となる．

KCE では，ODAF Scanner の API を利用してオフィス文書からテキストを取得し，キーワードを抽出する．

3.3.2 Document Critiquing Tool

Document Critiquing Tool (DCT) は，オフィス文書中に含まれる不適切な表現を検出し，対話的にユーザーに書換えを促す文書校正ツールである．DCT では，単純な文法チェックにとどまらず，様々な観点で文書の品質を低下させる要因となる不適切表現を検出する．

DCT では，ODAF Scanner の API で文書中のテキストを逐次的に取得し，テキスト分析を行う．不適切な表現が検出された場合は，ODAF Manipulator の API で検出結果をハイライトしてユーザーに提示し，適切な表現に書換えるためのダイアログボックスを表示する．検出結果が文書上でハイライトされていることで，ユーザーは検出箇所の周辺の情報を参照しながら文書校正を行うことができる．

4. Translation Assist Tool

Translation Assist Tool (TAT) は，手作業によるオフィス文書の翻訳作業を支援するツールである．TAT は ODAF 上に実装されており，アプリケーション上で翻訳を行うテキストを

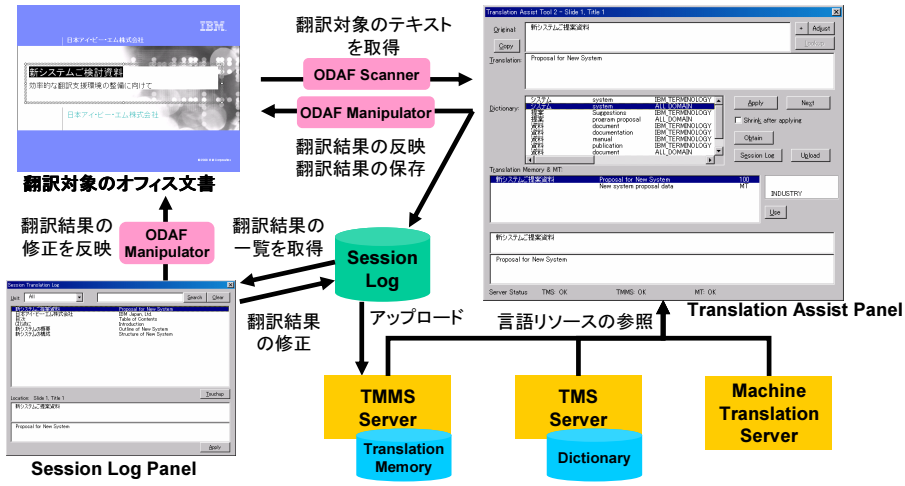

図 2: Translation Assist Tool を使用した翻訳作業の流れ

直接選択して翻訳作業を行うことができる．また，言語リソースを参照する際に，辞書を管理している TMS サーバー，翻訳メモリを管理している TMMS サーバー，機械翻訳サーバーに対して同時に問い合わせを行い，結果をまとめて参照することができる．TAT を使用して翻訳した結果は，TMMS サーバーにアップロードすることで，チーム内で翻訳結果を共有することができる．

4.1 TAT の使用方法

TAT を使用した翻訳作業の流れを図 2 に示す．

TAT で翻訳作業を行うには，翻訳を行うテキストをアプリケーション上で選択して Obtain ボタンを押す．すると，ODAF Scanner により選択したテキストが取得され，Translation Assist Panel (TAPanel) に辞書，翻訳メモリ，機械翻訳の参照結果が表示される．続いて，TAPanel で翻訳結果を作成して Apply ボタンを押すと，ODAF Manipulator により文書上に翻訳結果が反映され，その翻訳対と Location Path がセッションログに記録される．

翻訳結果を確認・修正する場合は，Session Log Panel (SLPanel) を開く．SLPanel のリストにはセッションログに記録された翻訳対が一覧表示されている．リストから翻訳対を選択すると，ODAF Manipulator によりアプリケーション上で該当箇所がハイライトされる．また，SLPanel で翻訳結果を修正することで，セッションログと文書上の翻訳結果が同時に修正される．このように，Location Path によってセッションログと文書上の翻訳結果が紐付けられているため，アプリケーション上で該当箇所をハイライトし，周辺情報を確認しながら修正作業を行うことができる．

また，TAT 上で行った翻訳結果は，セッションログに保存されている段階ではまだ共有されていない状態にある．翻訳結果を共有するには，アップロードを実行して，セッションログに保存されている翻訳結果を TMMS サーバーにアップロードする．

4.2 TAT 翻訳とマニュアル翻訳の比較

翻訳を行う際に言語リソースを参照する必要がある場合，マニュアル翻訳では一度の問い合わせに対して 1 つのリソースしか参照できないのに対し，TAT 翻訳では複数のリソースを同時に参照することができる．これにより，リソース参照にかかる時間を削減できるほか，意外な気づきが得られる可能性があり，訳質の向上にも寄与するものと考えられる．

また，言語リソースを参照するプロセスで TAT を使用した後，さらに TAT 上で翻訳作業を行うことにより，翻訳結果が自動的に記録され，チーム内で共有できるようになる．同一文書を複数人で分担して翻訳を行う際に，翻訳結果が共有されていることで，訳質の平準化が期待される．

その一方で，TAT 翻訳では TAT のインターフェイスを通して翻訳作業を行う必要があり，文書上に直接翻訳結果を入力していくマニュアル翻訳と比べて，時間のロスが発生する可能性がある．この問題を軽減するために，TAT ではオフィス文書上で直接操作可能なインターフェイスを提供し，時間のロスが少なくなるようにしている．

5. 評価実験

本研究では，TAT による翻訳支援の有効性を確認するために，TAT を使用して翻訳を行った場合と，TAT を使用せずにマニュアル翻訳を行った場合とで，翻訳時間の比較を行った．

5.1 評価の目的

TAT 翻訳では，TAT のインターフェイスを使用して翻訳作業を行うため，マニュアル翻訳と比べて時間のロスが発生している可能性がある．しかし，TAT を使用して翻訳を行っていくことで翻訳メモリが拡充され，翻訳時間が短縮されてくると考えられる．そこで，TAT 翻訳における翻訳メモリのカバー率と翻訳時間の関係を測定し，同じ文書をマニュアル翻訳した場合の翻訳時間と比較した．

また，TAT による翻訳支援が有効かどうかは，翻訳者のレベルによっても異なってくる．例えば，言語リソースの参照をほとんど必要としないようなレベルの高い翻訳者にとっては，TAT で言語リソースを参照できることにメリットはなく，TAT のインターフェイスを利用することによる時間のロスにより作業効率が落ちてしまう．逆に，言語リソースの参照を頻繁に行う翻訳者にとっては，TAT で言語リソースを参照することによって翻訳作業が効率化されるものと考えられる．そこで本実験では，翻訳者のレベル別に測定結果の分析を行った．

5.2 実験環境

本実験では，10 人の翻訳者がプレゼンテーションスライド 10 枚を日本語から英語に翻訳するタスクを実施し，スライド

表 1: 翻訳者のグループ分けと人数の内訳

		翻訳の方法	
		TAT 翻訳	Manual 翻訳
スキルレベル	Avg	3	3
	High	2	2

毎の翻訳時間を測定した．

測定に参加した翻訳者は，翻訳業務を専門に行っているスタッフで，通常業務で行っている翻訳作業の実績を元に，あらかじめスキルレベルが設定されている．これを元に，翻訳者のグループを Avg と High に分け，その中で TAT を使用するグループと，マニュアル翻訳を行うグループを編成した．この 4 グループの人数の内訳を表 1 に示す．

また，あらかじめ実験で使用するスライドに含まれるテキストの翻訳結果を翻訳メモリに登録しておき，スライド毎に翻訳メモリのカバー率が 8%〜100%の間で変動するようにした．この翻訳結果は，TAT 翻訳のみ参照可能である．なお，TAT 翻訳・マニュアル翻訳共に，通常業務で利用している言語リソースを自由に参照できるものとした．

5.3 実験結果と考察

スライド毎の翻訳メモリのカバー率と，文字あたりの平均削減時間の関係を図 3 に示す．このグラフでは，削減時間が正の値であれば，TAT 翻訳により翻訳時間が短縮されたことを表している．

始めに，翻訳メモリの蓄積による時間削減効果について確認する．スキルレベルが Avg, High のグループの両方で，翻訳メモリのカバー率と削減時間が正の相関を示していることから，翻訳者のレベルに関わらず，翻訳メモリの蓄積により翻訳時間の削減効果があることが分かる．

次に，TAT 翻訳による時間削減効果について確認する．

スキルレベルが Avg のグループは，翻訳メモリのカバー率が低い場合でも，TAT 翻訳により翻訳時間が削減されていることが確認された．これは，スキルレベルが Avg のグループは，翻訳作業を行う際に言語リソースの参照を頻繁に行っており，TAT 翻訳で TAT のインターフェイスを利用する手間と，マニュアル翻訳で言語リソースを参照する手間が，同じであるためと考えられる．また，翻訳メモリのカバー率が低い場合でも TAT のインターフェイスを使用することによる翻訳時間の増加が見られないことから，言語リソース参照を頻繁に行う翻訳者にとっては，TAT が提供するオフィス文書上で操作可能なインターフェイスにより，翻訳支援が有効に行われていることが確認された．

一方，スキルレベルが High のグループは，翻訳メモリのカバー率が十分高い場合でも，TAT 翻訳により翻訳時間が増加していることが確認された．これは，スキルレベルが High のグループは，翻訳作業を行う際に言語リソースをほとんど必要としておらず，TAT のインターフェイスにより本来必要のない言語リソースを参照している分だけ，翻訳時間が増加しているものと考えられる．

また，TAT 翻訳では共有された翻訳メモリを参照しているため，マニュアル翻訳と比べて訳質が向上している可能性があるが，現段階では訳質の評価を行っていない．これは今後の課題とする．

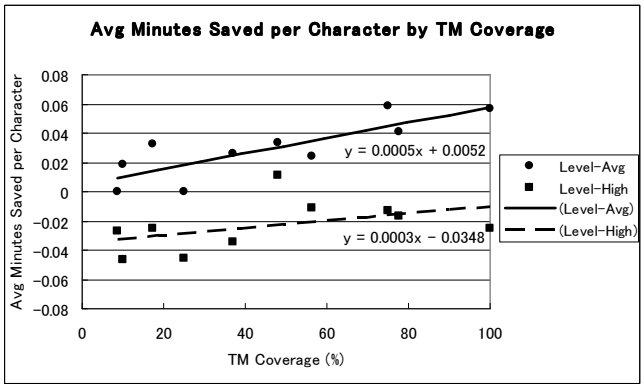

図 3: 翻訳メモリのカバー率と TAT 翻訳による文字あたりの平均削減時間の関係

6. おわりに

本研究では，オフィス文書の翻訳作業を支援するための環境を構築した．また，オフィス文書上で効率的な作業支援を行うためのツールを容易に作成するための Office Document Analysis Framework を開発し，その上に翻訳支援のためのツール群を構築した．また，Translation Assist Tool について評価実験を行い，スキルレベルが平均の翻訳者に対して，有効に翻訳支援が行われていることを確認した．

今後の課題としては，TAT 翻訳による訳質の向上について評価を行うことや，翻訳メモリだけでなく，辞書についても翻訳時間の削減や訳質の向上について評価を行うことなどが考えられる．

参考文献

[1] 701 Translator, IBM Press Release (http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html), 1954.

[2] The Apache POI Project, http://poi.apache.org/

[3] Felix, http://felix-cat.com/

[4] Open Document Format, http://www.oasis-open.org/home/index.php

[5] Language Grid, http://langrid.nict.go.jp/en/index.html

[6] T. Ishida. Communicating Culture. IEEE Intelligent Systems, Special Issue on the Future of AI, Vol. 21, No. 3, pp. 62-63,2006.

[7] S. Sakai, M. Gorou, R. Inaba, Y. Murakami, T. Yoshino, Y. Naya, Y. Hayashi, M. Tanaka, T. Takasaki, S. Matsubara, Y. Kitamura, T. Ishida. Supporting Multicultural Society with the Language Grid, International Conference on Informatics Education and Research for Knowledge-Circulating Society (ICKS'08), 2008.

＊ Microsoft, Excel および PowerPoint は Microsoft Corporation の米国およびその他の国における商標.