翻訳支援システム導入による効率化の評価

富士 秀、潮田 明、大倉 清司、山下 達雄

富士通研究所

fuji.masaru@jp.fujitsu.com

1. はじめに

翻訳支援システムの導入による効率化の度合いを定量的に評価するための手法を考案し、この手法をもとに我々の作成した翻訳支援システム Cliché を対象とした実験を行なった。この結果、特許文書を対象とした日英翻訳作業において約 4 倍の効率化を達成した。評価手法は、単なるシステム性能ではなく、実際的な作業環境における人手作業を含めたトータルな作業としての効率化測定手法となるように設計した。

2. 従来評価の問題点と解決手段

システム開発では、一般的に、作成したシステム自体の性能的な優位性を定量評価することが多い。例えば、処理速度の定量測定や解析精度の定量評価等がこれにあたる。しかし本研究は、人間が支援システムを利用した場合の作業全体を対象としており、そこでは、作業環境までも考慮した評価手法の設計が必要となる。特に、翻訳作業に特化した検討要素としては以下が挙げられるが、それぞれについて、従来評価実験における問題点と本実験における解決手段について述べる。

速度と訳質

翻訳作業効率の測定において、単に作業速度だけを測定しても有意な差が出ない場合が多い。これは、実際の人間の作業では、訳質を犠牲にしてまで高速に作業したり、また逆に訳質ばかりに拘って必要以上に作業が遅くなるような場合が頻繁に発生するからである。このことから、作業速度と訳質の両方を測定することによって初めて有効な測定ができる。なお訳質を評価するには、安定した訳の正解セットがあった方が望ましく、本実験では対訳の特許明細書をこの用途に用いることによって、この点をクリアした。

対象文書への慣れ

支援あり翻訳と支援なし翻訳のような 2 条件間の測定を比較する際には、条件の適用順が結果に影響を与える。これは、被験者が対象文書を扱うに連れて対象文書に慣れて処理速度が上がっていくためである。例えば、最初に条件 1 での測定を一通り行なってから条件 2 での測定に移ると、条件 2 の方が不当に高速な測定値となってしまう。この問題を解決するために、本手法では、各条件での測定を小さな単位に分け (小さな文書を複数用意し)、条件 1 の文書と条件 2 の文書を交互に評価対象とすることによって公平な測定となるようにした。

文書間の差

支援あり翻訳と支援なし翻訳のように 2 種類の条件を比較する際に、条件 1 と条件 2 で全く同一の対象文書とすると、各文書を 2 回評価することになり、2 回目のほうが 1 回目よりも慣れた状態になり公平な測定とならない。そこで、文書としては異なっていながらも、内容的には同等の文書のセットを各条件用に用意する必要が出てくる。翻訳の実験における内容の同等性は、同じ分野の文書を選ぶ、文書量を揃えることによってある程度実現できる。また予備実験から、長文を含む文書ほど難易度が上がる傾向があることがわかったので、この補正も行なった。

分野特化と再現性

支援システムでは一般的に、作業に先立って作業対象文書の分野にシステムをチューニングしておく必要がある場合が多い。チューニングのためには、対象分野の学習用文書を事前に大量に入手できる必要がある。今回の実験では、文書データ量も多く、また分類コード等によって分野が明示的に分けられている特許文書を使うことによって、十分にチューニングした状態での実験を行なうことができた。

システムの安定性

作業効率を評価するような実験では、システムの反応やインタフェース等が結果に影響を与える。システムが理論的に優れているだけでは不十分で実用に耐えられるレベルまで作りこまれている必要がある。本実験で用いた翻訳支援システムは、評価実験に先立ってユーザビリティー評価および改良を繰返すことにより実用的に使えるシステムに作りこんだ。[5]

3. 本実験の構成

3.1. 言語対

本論文では、日本語の特許明細書を英語に翻訳するという、日英方向の翻訳実験について述べている。しかしながら、これはあくまでも本評価手法が扱える言語対の一例であり、論理上はいかなる言語対にでも適用できる。

3.2. 対象文書

特許庁から公開されている特許の明細書データで、日本語明細書と英語明細書で対応のとれるものを対象とした。以下で述べる実験では、ある特定の 1 分野を対象に対訳データを作成した。約 240 対の対訳明細書から、文単位で対訳関係にある約 3 万 8000 文対を人手で作成した。この対訳文の集合を、およ

そ 9 対 1 の割合で学習用とテスト用（兼、評価用正解対訳文書）に分けた。

学習文書
支援システムを対象分野に対してチューニングするために用いる。

テスト文書（兼、評価用正解対訳文書）
被験者がテスト時に翻訳対象とするための文書として用いる。テスト用として分けた明細書から 8 件を無作為抽出し、抽出した各明細書の中で、文単位で対応の付きやすい項目（今回は、「実施例」部分を利用）の先頭から 10 文を取り出した。この合計 80 文の日本語のみを取り出したものを被験者用のテスト文書とした。また、日本語と英語の対そのままを評価者用の正解対訳とした。

3.3. 実験条件

本実験では、翻訳者ワークベンチによる支援なしの翻訳作業と支援ありの作業の間の効率の比較を行なう。この二つの条件を表 1 にまとめる。

表 1.：実験条件

支援なし翻訳	支援あり翻訳
基本的に手で翻訳するが、支援システム以外はどんな情報源でも参照してよい	支援システムを含むあらゆる情報源を参照してよい
実際に多用された情報源は、オンライン辞書、ウェブ検索エンジン、等	実際には、最初に支援システムを参照し、情報が見つからない場合に、手翻訳で使われるようなその他の情報源を参照することが多い

用意したテスト用文書は、半分が支援あり条件で評価され、残り半分が支援なし条件で評価される。具体的には、上記で用意した 8 件（各 10 文）の文書対は二つに分けられ、最初の 4 件が支援あり条件、残りの 4 件が支援なし条件で評価される。

3.4. 支援システム

上述の学習用対訳文データを用いて、我々の開発した、機械翻訳と訳例検索[4]の統合による翻訳支援システム[3]のチューニングを行なった。主なチューニング内容は、学習用対訳データから抽出した対訳語句（表 2）の機械翻訳システムへの登録、および対訳文データの訳例データベースへの登録等である。

表 2：登録語句の例

日本語語句	英語語句
周期装置	periodic device
配置シーケンス	placement sequence
変調信号	modulation signal
フィードバック経路	feedback path

3.5. 被験者

今回の実験では、被験者として、特許翻訳の専門知識はないが英語力の高い日本人翻訳者 2 名が翻訳を行なった。被験者は、今回の対象とは全く別の文書を使い、支援システムが十分に使いこなせる状態になってから、本実験に取り掛かることとした。なお、対象文種（特許）や対象分野に関する事前の学習は行わない状態で実験を行なった。

4. 実験手順

従来評価手法の問題点でも述べたように、本実験は、2 条件（支援あり、支援なし）を公平に比較することが求められる。そこで、両条件を交互に適用するような手順とした。また、翻訳速度の推移と同時に訳質の推移も観察する必要がある。そこで、各文書に対する翻訳および翻訳の見直しを何回か繰返し、その度に速度と訳質の両方を記録していくという手順を採用した。

4.1. 条件適用の順番

テスト文書の半数は支援なし条件で、残り半数は支援あり条件での実験に割り当てられるが、文書に対する慣れの効果を公平にするために、2 条件が交互に適用されるようにした。今回は、処理の煩雑さを軽減するため 2 文書ずつの交互とした。この処理順を表 3 に示す。また、全 8 文書を一通り処理したあとには、評価者による評価およびフィードバックの処理が入る。これに引き続き、フィードバック内容をもとに、翻訳者は全文書に対して 2 回目の翻訳処理を行なう。2 巡目以降は主にそれまでに翻訳した内容の見直しをすることになる。

表 3 ：処理順と条件適用

処理順	繰返し	作業者	文書	作業内容	条件
1	1 巡め	被験者	1	翻訳	支援なし
2			2		
3			3		支援あり
4			4		
5			5		支援なし
6			6		
7			7		支援あり
8			8		
9		評価者	1～8	評価	
10			1～8	ﾌｨｰﾄﾞﾊﾞｯｸ	
11	2 巡め	被験者	1	翻訳	支援なし
12			2		
13			3		支援あり
14			4		
15			5		支援なし
16			6		
17			7		支援あり
18			8		
19		評価者	1～8	評価	
20			1～8	ﾌｨｰﾄﾞﾊﾞｯｸ	
21	3 巡め	被験者	1	翻訳	支援なし
22			2		
…	…	…	…	…	…

4.2. 速度測定

表３中の各翻訳処理では、被験者は、対象文書を一通り翻訳もしくは翻訳見直しをするが、これにかかる所要時間を測定して記録する。
翻訳業界で頻繁に用いられる「翻訳速度」は、対象文書の単語数を翻訳時間で除算したものである。なお、業界の慣例として、正規化の基準は英語ワード（単語）数で行なうことが多いので、本評価でも原文の和文の単語数ではなく、原文の正解対訳である英語文書のワード数を計算に用いた。
さらに、後述の分析では、単位ワードあたりの翻訳時間を指標として使っている。これは、翻訳速度の逆数であり、翻訳所要時間を英文ワード数で割ったものである。

4.3. 訳質評価

全テスト文書を一通り翻訳した時点で、評価者は訳文の訳質評価を行なう。訳質評価は、被験者の作成した訳文の英語文書を、正解英語文書と比較することによって行なう。訳質評価は、次のような方針で行なった。

文書中の評価対象
本評価では、文種（特許）固有表現、分野（今回対象にしている特定の特許分野）固有表現を評価の対象とし、また、意味を正確に伝えるために最低限必要な文法構造（係り受けの妥当性等）も評価の対象とした。
これ以外の、細かい言い回しや、意味の正しい伝達とは直接関係ない修飾表現や語順等は評価の対象外とした。また、被験者の作成した訳文に関する自然さや理解容易性も評価の対象外とした。

正否の判定
被験者訳と正解訳が一致した場合には、そのまま被験者訳を正解とみなす。一致しなかった場合は、それが当該分野で許容される範囲の訳語の揺れであるかどうかを確認し、範囲内であれば被験者訳を正解とみなし、範囲外の場合は不正解とみなす。

評価の単位
本実験では、評価結果は文単位で集計する。文中に正否判定で不正解となった要素がまったくない文を「正解文」とし、文中に正否判定で不正解となった要素を一つ以上含む文を「不正解文」とする。ある時点でのある文書の「訳質」の評価値は、文書中の全文に対する正解文数の割合である。

4.4. 評価結果のフィードバック

全文書の翻訳（または見直し）が一通り完了した時点で、全文書に対する訳質評価を実施し、その評価結果が被験者にフィードバックされる。フィードバックされるのは、各文書についての、その時点での文単位の正解・不正解の状況である。言い換えれば、被験者は、文中に誤りが含まれるか否かのみを知ることになる。

4.5. 時間設定

今回の実験では、条件毎に訳質を揃えた実験を行なうために、十分に作業時間を与えた状態での翻訳作業を行なった。
実務場面での翻訳作業では納期等の関係から限られた時間内で翻訳作業を行なうことになる。しかし、本実験で時間制限を設けると、いずれかの条件（特に支援なし条件）で、訳質が十分に上がらないうちに制限時間を超えてしまう場合が出てくる可能性があり、そうすると、条件間で訳質が揃わなくなる。このような状況を回避するために、実務よりも十分に長い時間をかけても翻訳（および見直し）を続行するという方針とした。

5. 結果と分析

5.1. 訳質・翻訳時間の推移

前項の構成で実験を行い、翻訳時間に対して訳質がどのように変化するかを観察した。図 1 および図 2 は、1 名の被験者の翻訳時間（単位ワードで正規化したもの）と訳質の推移を表した図である。図 1 は、支援なし条件、図 2 は支援あり条件での測定結果である。

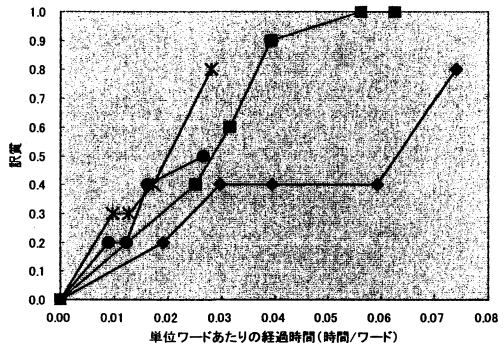

図 1. 支援なし条件での文書毎の時間・訳質の推移

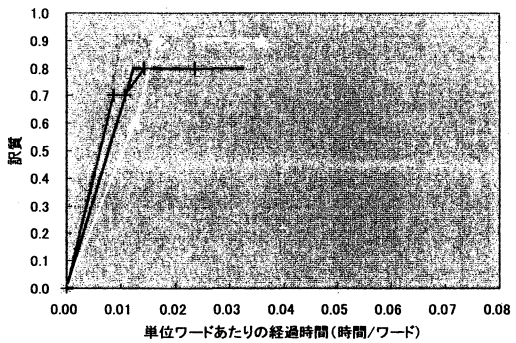

図 2. 支援あり条件での文書毎の時間・訳質の推移

図 2 から、支援あり条件では、各文書とも概ね早い時点で訳質が向上していくことがわかる。これに対して、図 1 の支援条件では、支援ありと同じ訳質に到達するのに大幅に時間がかかることがわかる。

5.2. 条件による訳質・速度差の見積

条件間での訳質と翻訳時間の差を見積もるために、訳質推移の結果から、条件毎の傾きを表す線形近似を行なった。この際に、一度ある訳質まで到達して、その後時間をかけても訳質が上がっていかないような測定点は外してから、近似を行なった。なお、いずれの条件でも、特にそのようにしなくても原点近くを通る近似直線となったが、ここでは明示的に原点を通る近似直線を使った。

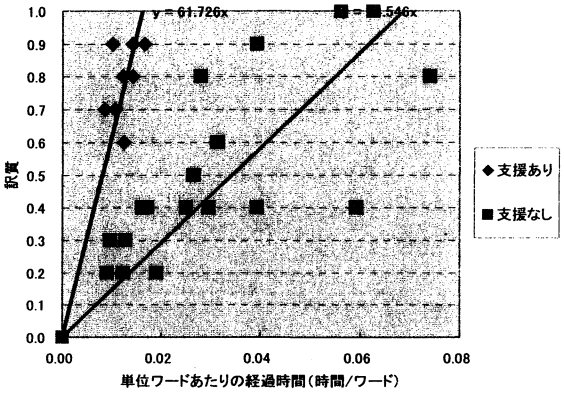

図 3. 翻訳時間と訳質の傾きの差の見積り

図中、条件同士で傾きの差を計算することによって、条件間の速度差が計算できる。ここでは、約 4 倍の速度差がでていることがわかる。なお、もう 1 名の翻訳者による同様の見積でも、ほぼ同等の速度差が観測された。

6. 考察

以上の結果から、本評価手法を使うことによって、支援あり条件と支援なし条件の間の速度差を見積もることができた。しかし、本実験には以下のような問題が生じていると考えられる。

条件間の相互作用

2 条件を交互に適用することによって実験の公平性は保証されるが、支援あり条件が支援なし条件に影響を与えている可能性がある。

実際の翻訳作業では、支援システムがなければ支援なし条件で作業し続け、あれば支援あり条件で作業続けるはずであり、今回のように交互に適用することによって、両条件がお互いに影響を及ぼしあっていた可能性は否定できない。

訳質評価の問題

今回の実験では、被験者の作成した訳文の自然さや理解容易度は評価の対象外としている。しかし、実際の翻訳現場では、訳文を推敲することによって納品レベルの文書を完成させる。このため、この点に関しては、実際の場面から求められる評価と異なる場合が出てくる可能性がある。

時間区切り

今回の実験では、被験者に時間の区切りを任せ、かかった時間を記録してもらう形態とした。しかし、この方法だと、特に支援あり条件の測定で訳質が急激に向上しているような箇所でも、測定点が十分に取れていないという問題が起こった。今後の測定では、測定点の時間間隔を陽に指定することにより、注目したい箇所で十分な測定点が確保できるような実験とする。

7. 結論

実用場面を想定した人手作業における翻訳支援システム導入の効果を見積もるための評価実験を設計した。評価実験は、単なる翻訳速度のみならず、同時に訳質の推移を測定することによって、作業経過の全体像を把握することができるものとなった。この評価手法を用いて、特許明細書の日英翻訳作業を対象に我々の開発した翻訳支援システムを使った実験を行なったところ、支援なし条件と支援あり条件の間で速度差を観察することができた。翻訳時間と訳質の推移から傾きの線形近似を行なうことよって、支援システム導入が約 4 倍の効果があることを示すことができた。

謝辞

Cliché 開発の段階から Cliché の評価に全面的にご協力いただいた㈱十印に感謝いたします。

参考文献

[1] Hitoshi Isahara, et al., (1995) JEIDA's Test-Sets for Quality Evaluation of MT Systems. In proceedings of MT-Summit V.

[2]富士秀,畠中伸敏,伊藤悦雄,亀井真一郎,隈井裕之,介弘達也,吉見毅彦,井佐原均. 機械翻訳システムの有効性の評価～どのような人にとってMTは役立つか～.言語処理学会第 8 回年次大会予稿集,2002.

[3]潮田明,富士秀,大倉清司,山下達雄.機械翻訳と訳例検索を統合した翻訳支援システム.言語処理学会第 9 回年次大会予稿集,2003.

[4]山下達雄,富士秀,大倉清司,潮田明.翻訳支援に有効な訳例検索の類似度計算方式と検索結果提示方式.言語処理学会第 9 回年次大会予稿集,2003.

[5] 大倉清司、山下達雄、富士秀、潮田明. 機械翻訳と訳例検索を統合した翻訳支援システムのインターフェース.言語処理学会第 9 回年次大会予稿集,2003.