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

山下 達雄　富士 秀　大倉 清司　潮田 明
{yto, fuji.masaru, okura.seiji, ushioda}@jp.fujitsu.com
富士通研究所
ドキュメント研究部

1　はじめに

コンピュータによる翻訳者を支援するシステムにおいて、機械翻訳と並び、訳例検索 (翻訳メモリ) は翻訳支援の核となる機能である。

ユーザが翻訳したい文と類似する文を訳例データベースから探し出し、その文の訳例を得ることができれば、ユーザはそれを修正することにより、少ない作業量で一貫性のある翻訳を行うことができる。

訳例検索機能を備えた翻訳支援システムとして、主にシステムの改版にともなうマニュアルの再翻訳などの産業翻訳の分野を対象に、既に商品化されたものもある [1]。

本論文ではこの訳例検索において、効率的な類似度計算方式と検索結果提示方式を提案し、実験によりこれらの方法が翻訳支援に有効であることを示す。

本論文の構成を説明する。第 2 章では、訳例検索に用いられる類似度をおおまかに「スコア」と「一致率」の二つに分類し、それぞれの特徴と、本研究での類似度計算方式の方針について述べる。第 3 章では、翻訳支援において、訳例データの部分一致箇所の再利用を容易にする新たな表示方法について述べる。第 4 章では、我々の提案する類似度計算方式と検索結果提示方式に基づき開発した訳例検索システムの実装と、データの作成方法について述べる。第 5 章では、提案手法の有用性を示すために行った、我々の開発した訳例検索システムによる実験について述べる。

本論文では、説明を容易にするために、翻訳者が翻訳したい英文をキーにし、英日訳例データベースを曖昧検索し、検索結果の訳例 (日本語文) を利用して訳文を作成するという英日翻訳支援の流れを念頭に置くことにするが、理論的にはいかなる言語対にも適用できる。また、訳例検索のキーとして使う翻訳対象の文を「検索キー文」、訳例データに含まれる対訳文を「訳例原文」「訳例訳文」と呼ぶことにする。訳例原文は検索キー文と同じ言語であり、訳例訳文は翻訳先言語である。

2　類似度計算方式

ここでは、訳例検索に用いられる類似度をおおまかに「スコア」と「一致率」の二つに分類し、それぞれの特徴について述べる。

一致率は、検索キー文と訳例原文の文全体での一致度を重視するもので、百分率などで表される。完全に一致する場合は 100% となり、これがこの類似度の上限となる。

計算方法としてさまざまなものが提案されているが、ここでは一番単純なものとして式 (1) を挙げる。B は検索キー文の単語数、C は訳例原文の単語数、A は二つの文で一致した単語数である。

$$\text{一致率} = \frac{A \times 2}{B + C} \tag{1}$$

以下の例では一致率は $\frac{3\times2}{8+4} = 0.5\ (50\%)$ となる。“[ ]” は一致した単語を表す。

検索キー文	This [is] [a] cool [cat] which I love
訳例原文	She [is] [a] [cat]

一方、スコアは、検索キー文と訳例原文との部分一致を重視するものである。単語が連続で一致している場合に得点を付加するなどの方法で、部分一致箇所が長い・多い訳例に高い類似度を与える。一致率のように類似度に上限は無い。ダイナミックプログラミングによるマッチング手法での出力値もこれにあたる。

こちらも計算方法として多くのものが提案されているが、簡単な例で説明する。例えば、単語が連続して一致していた場合、その連続数の二乗を足し合わせるという計算方法を考える。この場合、先の例文では、“[is] [a]” が 2 単語連続一致なので $2^2 = 4$ 点、“[cat]” は 1 単語一致ゆえ 1 点で、足し合わせて類似度は 5 となる。

実際のシステムでは、一致率を用いているものが多い。現状では訳例検索はマニュアルの改版などに用いられることが多く、ほとんど同じ文を訳例から見つけてきて最低限の変更で効率良く翻訳を行うという作業の流れが一般的になっている。これらのシステムでは、一般に、一致率がある程度 (例えば 80%) に満たない訳例は、訳例訳文との対応箇所が分かりづらく、活用できないケースが多い。一致率がある程度を越えれば、対応箇所が異なり箇所よりも多くなり、ユーザが単語同士の対応を認識しやすくなる。つまり、現状では類似度は足切りの基準として用いられているため、正規化されている「一致率」が採用されていると言える。

我々は、これまで活用が難しかった、一致率の低い訳例を、部分一致箇所単位で利用できるようにすることで、翻訳効率を向上できると考えた。そのためには、やはりスコアによる類似度計算が必要であるため、これを採用した。そして、対応箇所が分かりづらいという欠点を補うため検索結果提示方式を工夫した。これにより、ユーザが訳例の部分一致箇所の対応を容易に認識できるようなる。これについては次章で述べる。

3 検索結果提示方式

本章では、翻訳支援において、訳例データの部分一致箇所の再利用を容易にする新たな表示方法について述べる。

訳例検索の結果は類似度順に並べられた各訳例の羅列と仮定する。

従来の訳例データ検索結果の表示方法として、検索キー文と訳例原文で一致する単語、または、訳例原文と訳例訳文で意味的に対応する単語のみのハイライトというものがある。これを「二つ組」表示と呼ぶことにする。

本論文では、それに加え、検索キー文、訳例原文、訳例訳文の全ての文中の対応する単語をハイライトする「三つ組」表示方式を提案する。

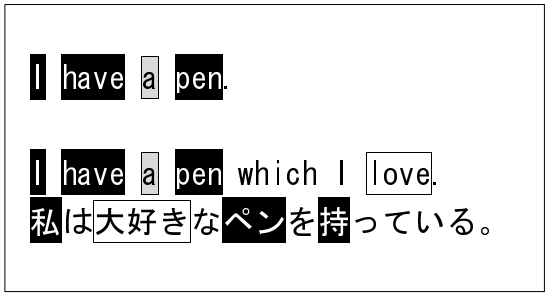

図 1: 訳例データの表示

図 1 に三つ組表示の例を示す。検索結果の各訳例は、検索キー文、訳例原文、訳例訳文の 3 文を一単位としてボックス表示している。上から、検索キー文、訳例原文、訳例訳文となっている。「I」-「I」-「私」、「pen」-「pen」-「ペン」などが三つ組に当たる。三つ組表示では、静的なハイライト表示だけでなく、どれかの文中のハイライトされた単語にマウスを合わせると残りの二つの文中の対応する単語がもう一段階ハイライトされるという動的なハイライトも行う。例えば、図中の検索キー文の「pen」にマウスを合わせると、訳例原文の「pen」、訳例訳文の「ペン」がもう一段階ハイライトされ、ユーザは対応箇所を容易に認識することができる。

三つ組表示により、検索キー文中の単語について、この訳例中での訳が一目で分かり、ユーザによる効率的な訳語選択が可能となる。また、三つ組単語対応をヒントに使うことで、ユーザによるフレーズの把握が容易になり、訳例の部分利用が促進される。訳例が長文であっても、目的の訳が容易に得られ、これまで再利用が困難であった訳例 (例えば、特許文の訳例データベース) も活用できるようになる。つまり、三つ組表示によって、スコアによる類似度の高い訳例を、信頼のおけるコンテキストを持つ訳例として利用できるようになるのである。

検索キー文と訳例原文の単語対応 (検索時にマッチングする) については 4.1 節で、訳例原文と訳例訳文の単語対応付け (事前に処理する) については 4.2 節で述べる。

4　実装

ここでは、訳例検索システムの実装と、対訳データの作成について説明する。

4.1　訳例データの検索

我々の開発したシステムでは、訳例データの検索は「絞り込み」と「マッチング」の 2 段階の処理を行っている。

絞り込み検索では、高速な文字列検索を可能にするデータ構造である suffix array [2][3] により、指定された絞り込み数分の結果を得る。この後、絞り込まれた検索結果の一つ一つを検索キー文とつきあわせ、ダイナミックプログラミング法で最適な照合結果を取りだすアルゴリズムにより単語同士の対応を得る (マッチング)。

検索結果は、複数の対訳データが類似度順に表示される。類似度計算方式は、「スコア」や「一致率」などが選択でき、その場で即座にランキング表示に反映させることができる。

4.2　訳例データの作成

我々のシステムで用いている訳例データは、訳例原文・訳例訳文のペアだけでなく、それぞれの文を構成する単語同士の対応情報も持っている。アライメントされた対訳文データに対して単語対応付けを行う方法には単語の共起情報を用いるものが多いが [4]、我々は図 2 のように意味情報 (図 2 中の I, BUY, APPLE) を用いて、単語対応付けを行っている。単語対応付けのための単語辞書 (意味タグ情報付き) は、英日・日英翻訳ソフト ATLAS[5] のものを用いた。それぞれの言語の文を構成する単語列を、両言語共通の意味タグ列に変換し、意味タグ同士のマッチングをとることにより実現した。

これらの訳例データは、言語情報・対応情報とともに、XML に準拠したフォーマットの訳例データベースファイルへ格納している。訳例データベースファイルは、検索に利用されるだけでなく、リアルタイムでの訳例データの追加・削除が可能である。

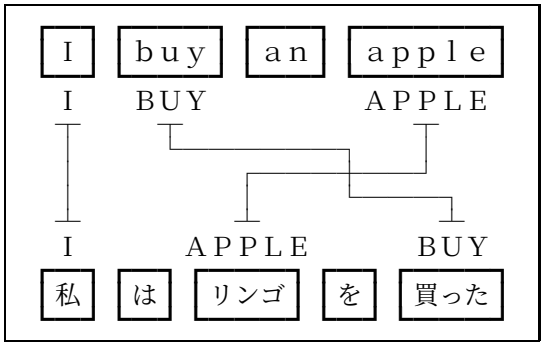

図 2: 英日単語対応付け

5　実験

提案手法の有用性を示すために、我々の開発した訳例検索システムで、「スコア」と「一致率」の比較実験を行った。

実験対象として契約書のテキストを選択した。契約書の訳例データベースには約 65000 対の英日対訳が含まれている。評価は翻訳業務の経験のある翻訳者 1 名が行った。評価用テキストは、13 文のみだが、その文だけでなく、翻訳者が文中の部分文字列 (実際の翻訳作業で訳例検索する単位) を選択して検索を行ったので、実際の検索回数 (評価対象数) は 103 である。評価の手順を以下に示す。

1. 絞り込み数 (4.1 節を参照) を 50 とし、50 個の検索結果を取り出す。
2. その 50 個の検索結果に対して、「スコア」と「一致率」のそれぞれの類似度計算方式でソートする。
3. それぞれの上位 N 個を取り出し、どちらの訳例が実際の翻訳作業に有効か判定する。

判定は、翻訳作業に役立つ訳文 (フレーズ) がその上位 N (= 1 or 3 or 5 or 10) 個にどれだけ含まれているかという観点で、「スコア (によるランキングの方が良い)」「ややスコア」「Draw」「やや一致率」「一致率」の 5 段階で行った。

図 3 に実際の判定の例を示す。1〜10 の数字はランキングを表す。「○」「△」「×」はそれぞれ、「そのまま利用できる訳例」「加工すれば利用できる訳例」「利用できない訳例」であることを表す。

図 4 に実験結果を示す。全般にスコアによる計算方式によるランキングの方が翻訳作業により有用な検索結果を提示しているということが分かる。

1. *shall be construed and determined in accordance with the laws*
   判定: Top1 & Top3 & Top5 & Top10=Draw

	1	2	3	4	5	6	7	8	9	10
スコア	○	○	○	○	○	○	○	○	○	○
一致率	○	○	○	○	○	○	○	○	○	○

2. *breach is cured*
   判定: Top1 & Top3=ややスコア, Top5=Draw, Top10=やや一致率

	1	2	3	4	5	6	7	8	9	10
スコア	○	○	×	×	○	×	×	×	○	○
一致率	△	△	○	○	△	○	○	△	○	○

3. *all persons*
   判定: Top1=ややスコア, Top3 & Top5 & Top10=スコア

	1	2	3	4	5	6	7	8	9	10
スコア	○	○	○	○	○	○	○	○	○	○
一致率	△	△	×	×	△	△	△	△	○	○

図 3: 判定例

6 おわりに

本研究では、今まで活用が難しかった「スコア」による類似度計算方式による訳例検索において、三つ組表示による提示方式を採用することにより、これまでの「一致率」による類似度計算方式よりも有用な検索結果を得ることができることを実験により示した。

我々は訳例検索機能のみでなく機械翻訳機能やその他翻訳支援に有用な機能を統合した翻訳支援システム Cliché[6][7] を開発しその有用性を示している [8]。本研究は、そのシステムの核の 1 つである訳例検索に着目し、翻訳作業効率向上のため、効率的な類似度計算方式と検索結果提示方式を提案したものである。本論文で提案した手法は現在の Cliché に導入されており、実際の翻訳作業に活用されている。

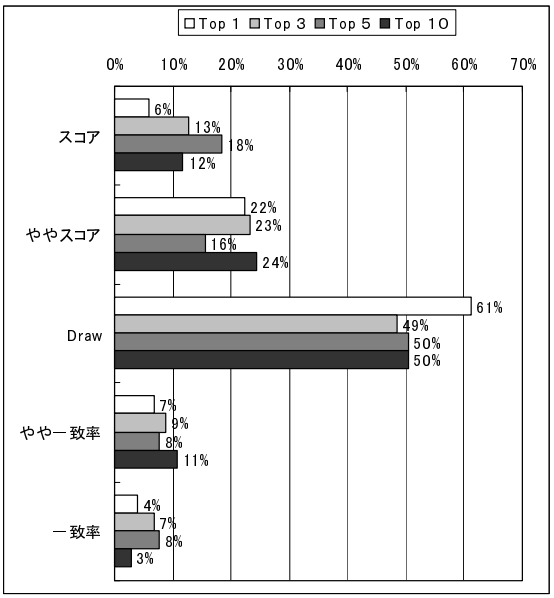

図 4: 実験結果

参考文献

[1] TRADOS. Translator's Workbench. http://www.trados.com/.

[2] Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. In *1st ACM-SIAM Symposium on Discrete Algorithms*, pp. 319–327, 1990.

[3] 山下達雄. 用語解説: Suffix Array. 人工知能学会誌, Vol. 15, No. 6, p. 1142, November 2000.

[4] I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, Vol. 26, No. 2, pp. 221–249, 2000.

[5] 富士通. 英日・日英翻訳ソフト ATLAS. http://software.fujitsu.com/jp/atlas/.

[6] 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.

[7] 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.

[8] 富士秀, 潮田明, 大倉清司, 山下達雄. 翻訳支援システム導入による効率化の評価. 言語処理学会第 9 回年次大会予稿集, 2003.