単語対応情報付き訳例検索システムの多言語対応

山下 達雄　大倉 清司　徐 国偉　富士 秀　Stefan Schmidlin　潮田 明
yto@jp.fujitsu.com (山下)
富士通研究所
知能システム研究部

1　はじめに

産業翻訳・特許翻訳などに求められる質の高い翻訳を行なうためには、現状の機械翻訳システムはその訳文の質の面で十分とは言えない。それを補うために訳文の修整などの人間の関与が必須である。そのために我々は、翻訳者を支援するシステム Cliché[1][2] を開発してきた。

翻訳支援システムにおいては、機械翻訳と並び、訳例検索 (翻訳メモリ, Translation Memory) が翻訳支援の核となる機能である。訳例検索とは過去の訳例の再利用を促進する機能である。これは、翻訳者を支援するという観点から、非常に有益な機能である [3]。

近年、インターネットの普及などにより様々なレベルでの国際コミュニケーションが活発化し、それに伴い、翻訳の需要が増大している。日本においては近隣アジア諸国、特に中国語を対象とした翻訳の重要性が高まっている。

本論文では、単語対応情報付き訳例データを様々な言語間で作成・利用する環境の設計方針とその実装について述べ、具体的に検証し、使用するリソースの質と付与できる対応情報の質の関係を示す。

本論文の構成を説明する。第 2 章では、翻訳支援における単語対応情報付き訳例データの検索ついて述べる。第 3 章では、自動単語対応付けに必要なリソースの分類を行ない、それらの利点、問題点について述べる。さらにその分類に基づいて、英日、日中、独英の訳例に対し、実際に単語対応付けを行ない、精度を測定し、分類と精度の関係について考察を行なう。

2　訳例の検索と結果提示方式

本章では、本研究の主旨である多言語間単語対応情報付き訳例データ構築について述べる前に、Cliché において、作成された訳例データがどのように使われるのか、また、単語対応情報がどのように役に立つのかについて述べる。

2.1　訳例データの格納と検索

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 (単語対応情報) も持っている。これらの訳例データは、XML に準拠したフォーマットで格納されている。また、リアルタイムでの訳例データの追加・削除が可能である。

データベースに格納する際、システムは、原文と訳文のペアを受け取り、両文に対して形態素解析を行ない基本形 (base form) を取り出し、それらに対して検索用のインデックスを作成する。また、同時に両文間の単語対応情報の付与を行なう (詳細は第 3 章で述べる)。検索用インデックスは、高速な文字列検索を可能にするデータ構造である suffix array [4] を採用している。

ユーザ (翻訳者) が訳例を検索する際、システムは、与えられた検索キーに対して形態素解析を行ない基本形を取り出し、それら基本形の列をキーとして訳例データベースを類似検索する。この検索は「絞り込み」と「マッチング」の 2 段階の処理を行っている。絞り込みでは、suffix array で高速検索を行ない指定された数に絞り込まれた結果を得る。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 (単語同士の対応) を得る (マッチング)。この際、類似度計算も行ない、これに基づき複数の検索結果をランキング表示する。ユーザはそれらを参照したり、エディタへ取り込むなどして翻訳に利用する。

2.2　検索結果の提示

我々のシステムでは、検索結果の各訳例の表示に、検索キー文、訳例原文、訳例訳文の全ての文中の対応す

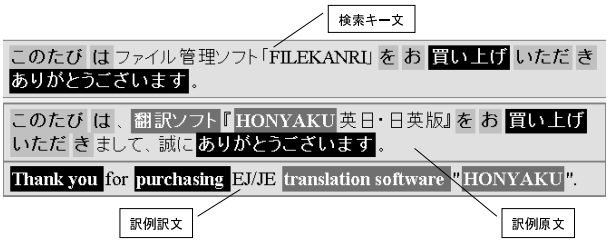

図 1: 訳例データの表示

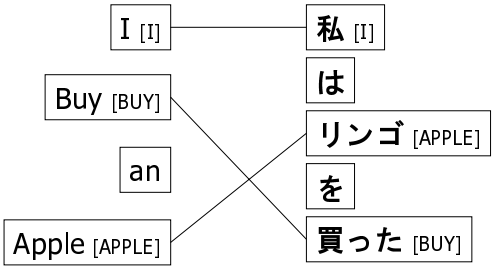

図 2: 英日単語対応付け

る単語をハイライトする「三つ組」表示方式を採用している。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。「買い上げ」-「買い上げ」-「purchasing」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「買い上げ」にマウスを合わせると、訳例原文の「買い上げ」と訳例訳文の「purchasing」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。もちろん「このたび」-「このたび」、「翻訳ソフト」-「translation software」間などの「二つ組」のハイライト表示も行なっている。

この技術により、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、特許文などの長文) も活用できるようになる。つまり、長さに関わらずあらゆる訳例を、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 [5]。この技術は多言語の訳例データベースを整備する際にもそのまま利用できる。

3 訳例への単語対応情報付与

本章では、本研究における単語対応付けの方針について述べ、実装と検証を行なう。

3.1 単語対応付けの手法

我々のシステムでの単語対応付けの手法を英日を例に具体的に説明する。英日の訳例に対する単語対応付けでは、英日それぞれの単語辞書に含まれている意味情報を用いている。

我々のシステムでは、原文・訳文ペアのそれぞれの文に対して、英日・日英翻訳ソフト ATLAS[1] の内部ルーチンを用い、形態素解析を行ない、さらに図 2 の I, BUY, APPLE のような意味情報 (概念 ID) を得て、それ利用している。単語対応付けは、各言語の単語を X, Y に配置したテーブルを用いて、概念 ID 同士のマッチングを行なうという単純なものである。概念 ID だけでなく、訳語候補などの表層文字列も補助情報として利用している。単語対応の曖昧性の解消は、距離などいくつかの要素を用いて行なっている。

対訳文データに対して単語対応付けを行う方法には、既存の整備された対訳辞書を用いず統計情報を利用するものが数多く提案されている。しかし、我々は統計情報による方法ではなく、整備された対訳辞書を用いる方法を採用した。これは、実際のユーザである翻訳者の要望に基づき[2]、対応付けのエラーを極力削減することを優先し、対訳辞書の整備に力を入れているためである。

3.2 単語対応付けに使用するリソース

我々は、単語対応付けの質の向上のためには、単に対訳辞書があれば良いというのではなく、相互の機械翻訳まで意識しなければならないと考える。しかし、そのためのリソースが整備されていない場合、既存のリソースを効果的に利用する必要がある。

ここでは、単語対応付けに使うリソースを以下のように分類した。我々は、言語対毎にそれぞれ異なる方針を取った。これについては次節で述べる。

- 形態素解析: 統一語彙、各言語独立
- 単語同定用情報: 概念 ID、訳語候補
- 解析する言語数: 両方 (2)、片方 (1)

まず、形態素解析について説明する。問題となるのは、対象とする両言語での形態素区切り方針の異なりである。適切な単語同定を行なうためには、両言語の

[1] http://software.fujitsu.com/jp/atlas/

[2] 翻訳者の方々の意見として、「単語対応付けに間違いがあるくらいなら対応が付いていない方が良い」というものが多い。

	英日	中日	独英
適合率 (%)	96.8	92.2	93.3
カバレージ (%)	32.2	17.1	39.2

図 3: 単語対応付けの精度

形態素解析が対訳辞書の語彙をベースとして行なわれる必要がある。双方向機械翻訳で利用する辞書以外は、基本的に各言語独立の辞書である。両言語で統一された語彙による形態素解析システムは、入手の難しいリソースと言える。

次に、単語同定のための情報について説明する。双方向の翻訳に対応した機械翻訳システムでは、それぞれの言語の辞書のエントリに、両言語で共通に用いる概念 ID を格納するという方法がある。この概念 ID を単語同定に用いる方法は、翻訳のためのコアな知識を利用できるという点で理想的である。また、翻訳元言語の辞書のエントリに翻訳先言語の訳の候補を複数格納する方法がある。この場合は、この訳語候補を翻訳先言語の文の単語の表層 (base form) 文字列と比較することにより同定を行なう。

最後に、解析する言語数について説明する。単語対応付けを行なうにあたって、対象の 2 言語について、単語の分離と base form を得る必要があり、最低でも形態素解析処理は必要である。これを前提に、さらに、各単語に前述の単語同定用情報が必要となり、これのための処理を両言語に行なうか一方の言語だけに行なうかがポイントとなる。単語同定用情報として概念 ID を用いる場合は、両言語を統一された概念 ID を持つシステムで解析する必要がある。しかし、単語同定用情報として訳語候補を用いる場合は、単語同定用情報を得るための処理はどちらか一方の言語に対して行なえばよい。もちろん両方の言語に対して行なえば理論的には単語対応付けの精度は向上する。

3.3 実装

我々は、英日、中日、独英の 3 つの言語ペアに対して、単語対応情報付き訳例データを作成した。以下、前節の分類に沿ってそれぞれの環境と精度を示していく。

適合率 (precision) は、システムが出力した全単語対応付けリンクのうち、正しく対応付けされたリンクの割合である。カバレージ (coverage) は、正しく対応付けされた単語の数 (=正解リンク数の 2 倍) を両言語の全単語数で割ったものである。対応付けの正しさは、各言語に精通した翻訳者が判断した。

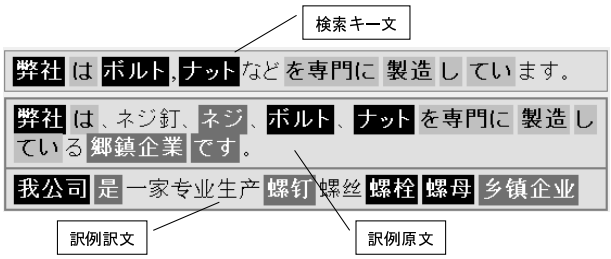

図 4: 中日単語対応付き訳例データの検索例

3.3.1 英日

英日の単語対応付けは、第 3.1 節の例で挙げた方法で行なっている。単語同定用情報として概念 ID を用いている。補助情報として訳語候補 (表層文字列) も用いている。英日双方向の翻訳が可能な翻訳システム ATLAS で、両言語を解析し、形態素解析 (base form の取得) と前記単語同定用情報を取得し、これらを用いて単語対応付けを行なった。特許庁のデータベースから入手した特許の対訳文 100 文を対象に精度を測定した。英語 2416 語、日本語 3032 語に対し、単語対応リンクは 907 個で、そのうち正解が 878 個であった。精度を図 3 に示す。

3.3.2 中日

中日の単語対応付けは、単語同定用情報として中日方向の訳語候補を用いた。既存の中日翻訳システムに手を加えて、中国語文に対して、形態素解析結果 (単語区切り) と各単語に対する日本語訳候補文字列を出力するようにした [6]。日本語文に対しては、前記中日翻訳システムとは別の形態素解析器で単語区切りと base form 取得を行なった。これら (中国語文の各単語の日本語訳候補と日本語文の各単語の base form) を用いて単語対応付けを行なった。中国語による中国企業の紹介文を日本語に翻訳した対訳コーパスの約 330 文で精度を測定した。中国語 4228 語、日本語 11396 語に対し、単語対応リンクは 1446 個で、そのうち正解が 1333 個であった。精度を図 3 に示す。また、図 4 にこの単語対応付き訳例データの検索例を挙げる。

3.3.3 独英

独英の単語対応付けは、単語同定用情報として独英方向の訳語候補を用いた。ドイツ語文、英語文それぞれに対し、それぞれの言語の形態素解析システムで base form を得る。独英辞書を加工し、ドイツ語の各単語に対して英語の訳語候補を挙げる。そして、これらの

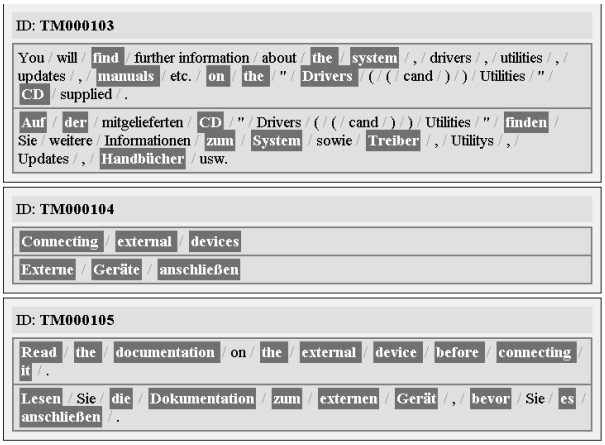

図 5: 独英単語対応付き訳例データの例

情報を用いて単語対応付けを行なった。富士通製品のドイツ語・英語のマニュアルから得た対訳文約 200 文で精度を測定した。英語 2398 語、ドイツ語 2261 語に対し、単語対応リンクは 979 個で、そのうち正解が 913 個であった。図 3 に精度を示す。また、図 5 にこの単語対応付き訳例データの例を挙げる。

3.3.4 考察

英日の単語対応では対象コーパスとして比較的長文である特許文を用いたため、両言語間での文章構造の異なり (フレーズ出現順など) が精度に影響した。

中日の単語対応では、形態素区切り方針が異なるシステムでそれぞれの文を解析したため、カバレージが低くなった。今後、カバレージを上げるためには、対訳辞書の語彙で両言語の形態素解析を行なう必要がある。また、エラーの多くは、機能語の対応付けの失敗による。内容語が形態素区切りの問題でマッチしない場合も、機能語だけが対応付けられることが多い。この際、曖昧性を効率的に解消する手段として、内容語の対応をヒントに使えないため、エラーが増える。これも前述の問題と同様、対訳辞書の語彙での形態素解析が必要となる。

独英の単語対応では、同系統の言語であるため、少ないリソースながら想定以上のカバレージが得られた。適合率については、中日と同様に機能語と形態素区切り (複合語) の問題により十分な値が得られなかった。

以下、関連研究について述べる。統計的な手法を用いた英仏間の単語対応付けでは、Cherry[7] らにより適合率 95.7%が報告されている。Melamed による研究 [8] では、カバレージ 36%に対して適合率 99.2%が報告されているが、本研究のように形態素区切りの問題を考慮した場合の適合率は 91.6%と報告されている。日英間の対訳単語対抽出では最大エントロピー法による方法で適合率 64.7%が報告されている [9]。

4 おわりに

本研究では、多言語間の翻訳支援のための単語対応情報付き訳例データの整備を目指し、実装と、使用するリソースの質と付与できる対応情報の質の関係の調査を行なった。我々はグローバル化により多言語間の翻訳はますます重要になっていくと考えており、新たな言語対に対応するシステムを既存のリソースを組み合わせて、いかに効率よく高精度に構築できるかを重視しながら、多言語翻訳支援の研究を進めている。

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 大倉清司, 徐国偉, 山下達雄, 富士秀, 潮田明. 多言語翻訳統合プラットフォーム cliché. 言語処理学会第 10 回年次大会予稿集, 2004.

[7] Colin Cherry and Dekang Lin. A probability model to improve word alignment. *Proceedings of ACL 2003*, 2003.

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

[9] 佐藤健吾, 斎藤博昭. 最大エントロピー法を用いた対訳単語対の抽出. 自然言語処理, Vol. 9, No. 1, pp. 101–115, 2002.