

オートコンプリートによる翻訳支援

大倉清司、富士秀、長瀬友樹
(株) 富士通研究所

okura.seiji@jp.fujitsu.com

1. はじめに

ブラウジングソフト、表計算ソフトなどで使われているオートコンプリート機能を翻訳支援システムに取り込み、翻訳効率化を図った。

オートコンプリートとは過去に入力された履歴などを記憶しておき現在入力中の文字列を予測する機構であるが、翻訳にそのまま適用するとオートコンプリートの対象が膨大になり、現実的でない。今回、オートコンプリートする対象語彙を翻訳対象範囲に絞り込むことなどにより問題を解決した。

2. 翻訳支援システム Cliché の問題点

Cliché は機械翻訳技術と訳例検索技術を統合したシステムである[1,2,3,4,5,6,7]。従来の Cliché は、翻訳結果や訳例検索結果をユーザに提示し、ユーザが書き込みボタンを押すなどして結果を訳文フィールドに取り込む仕組みになっていた。ユーザは取り込んだ結果を編集して訳文を完成させていた。しかし、機械翻訳結果や訳例検索結果のうち、ユーザが利用したいのはごく一部の表現である場合もある。そのとき、システム結果を取り込んで編集するよりもユーザが直接入力する方がはやくことが多い。システム結果が翻訳の参考になるとしても、それをコピーしたり、見ながら打ち込むのでは入力に時間がかかってしまうし、フルスペルで入力するときはスペルを間違える可能性がある。

3. 入力インターフェースの改良

ブラウジングソフトや表計算ソフトなどでは、オートコンプリートという機能により、入力の効率化が図られている。これは、ユーザが以前入力した文字列を覚えておき、ユーザがこれから入力しようとする文字列を予測して補完するものである。

原文を見ながら訳文を手入力する場面を考えてみよう。翻訳のために最も単純に考えられるオートコンプリート機構としては、ユーザが文字を入力すると、その文字ではじまる単語を辞書から検索し候補を表示するというものであろう。ところがこれを実装しようとするとき、例えば日英翻訳でユーザが "a" と入力したとき、システムは "a" から始まる英単

語を全て表示しなければならず、現実的でない。1文字では制限がゆるすぎるので2文字にしたとしても、例えば "ab", "ap", "as", "at" などからはじまる単語もかなりの数があり、ユーザは補完候補を提示されても選択するのに時間がかかってしまう。また、例えば3文字以上入力して補完候補がかなり絞り込まれたとしても、それら全てが入力したい単語に近いかどうかはわからない。

翻訳支援用オートコンプリート機能に必要なのは、ユーザが入力したい単語を補完候補として表示することである。ユーザが訳文を入力するときには、その原文は決まっているので、原文情報からユーザが入力すると思われる単語を予想することにした。単純には以下2段階の手順で補完候補を計算する：

1. **翻訳対象となる原文の処理**：原文の語句などをキーにして、訳文語句を出力する構成のもの（リレーショナルデータベースや、電子辞書、自動翻訳の辞書などを含む）を使い、原文に対する単語対応情報を計算する。
2. **ユーザ入力に対する補完候補の表示**：ユーザが打ち込んだ文字列をキーとして、1. で計算した単語対応情報の訳文語句側を検索し、その結果の単語リストを表示する。ユーザは入力したい単語があれば表示された候補を選択する。入力したい単語がなければ文字の入力、削除など作業を続ける。

日英翻訳の例をまじえてこの2段階の処理を詳しく説明する。「アルカリ溶液を攪拌する」という原文があるとしよう。処理1では、原文を解析して辞書引きを行う。形態素解析をする場合、解析結果が「アルカリ溶液 / を / 攪拌 / する」とすると、「アルカリ溶液」「攪拌」を辞書引きして、図1のような表を得る。

原文語句	訳文語句
アルカリ溶液	alkali solution
アルカリ溶液	alkaline solution
攪拌	stir
攪拌	mix

図1 単語対応情報の計算

続いて処理 2 で、ユーザが例えば「アルカリ溶液」の訳語である”alkali solution”を入れたいと思い、先頭の”a”を入力する。そのとき、上の「訳文語句」にある 4 つの語句から、”a”ではじまる語句を検索する。ここでは、”alkali solution”, ”alkaline solution”を得る。そこで、画面上に、この 2 語句を表示して、ユーザの選択を待つ。ユーザは”alkali solution”を即座に選択することができ、”alkali solution”が入力される。

しかし、この単純な仕組みだけでは十分な効率化にならない。既存辞書などを検索するだけでは、ユーザが苦勞して調べた単語が何度も出てくる場合に調べなおす必要がある。機械翻訳の辞書に登録する方法もあるが、手間がかかってしまうのが現状だ。また、辞書を検索するだけでは、多義語など候補が絞りきれないときが多く、表示されても使えないものがある。

調べた訳語を簡単にオートコンプリートの対象にでき、補完候補をさらに絞り込むことが必要である。Cliché の Divide and Conquer[1]の機構を利用し、Cliché を使うと自然に用語がたまっていく GUI を開発した。これは、分割した部分の訳を確定すると、それが自動的に補完候補に追加されるという仕組みである。部分訳の確定は翻訳の際に必要な作業であるため、従来の作業と同じ手間で補完候補が登録できる。これにより蓄積した用語も処理 1 の対象とする。

また図 1 のような単語対応情報に加え、ユーザ定義の語句かどうかの情報、辞書の訳語の場合はその辞書名情報、単語の信頼度情報[7]なども表に入れる(図 2)。原文語句と部分一致する語句を取り込んでよい。

原文語句	訳文語句	種別	辞書名	完全一	単語信
アルカリ溶液	alkali solution	自動翻訳	辞書 2		高
アルカリ溶液	alkaline	自動翻訳	辞書 2、辞		高
攪拌	stir	自動翻訳	辞書 1、辞		高
攪拌	mix	自動翻訳	辞書 1		低
アルカリ	alkali	ユーザ定	-		高
アルカリ溶液	alkali solution	ユーザ定	-		高
アルカリ金属	alkali metal	ユーザ定	-	x	高

図 2 様々な情報を含む対訳語句テーブル

これらの情報は、ユーザが原文を選択したときに自動的に計算し、表として画面上に出力すると効果的である(図 3)。オートコンプリートはこの表に基づいてユーザに提示される。

これらの情報をもとに、処理 2 においてユーザに候補を提示する際、ただ単に補完候補を表示するだけでなく、その出展や、信頼度が高い単語かどうかなどによりランキング表示することが可能となる。例えばユーザ定義語句を一番上に表示したり、信頼度の高い訳語のみ補完候補として表示すれば、ユーザ

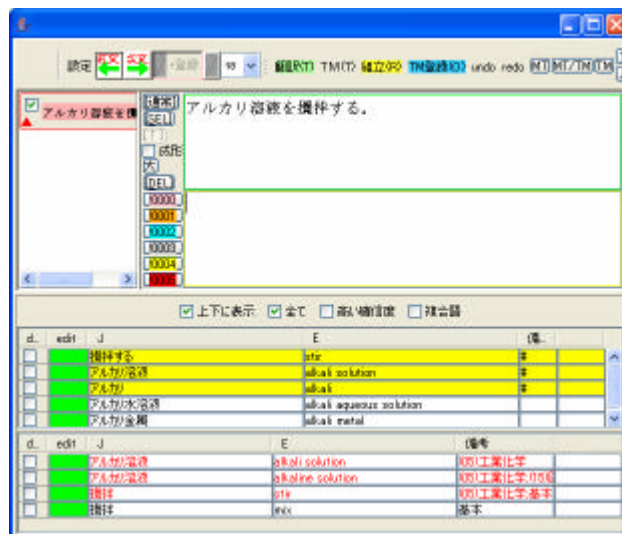


図 3 単語対応情報などの表示

が選択の手間を省ける(図 4)。今後は Cliché と更に統合し、入力支援をより効果的にしていく。

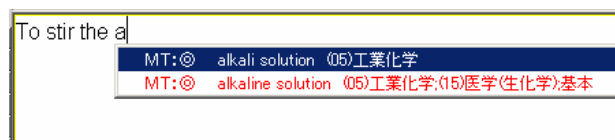


図 4 オートコンプリートによる入力支援

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] 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.
- [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] 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.
- [5] Seiji Okura, Tatsuo Yamashita, Masaru Fuji, Guowei Xu, Akira Ushioda. Integrating MT and TM: Cliché. 20th International Conference on Computational Linguistics(COLING), Exhibition.
- [6] 大倉清司, 富士秀, 徐国偉, 長瀬友樹, 潮田明. Cliché: さらに翻訳効率化のための翻訳支援インターフェース. 言語処理学会第 12 回年次大会予稿集, 2006.
- [7] 富士秀, 大倉清司, 長瀬友樹. 翻訳支援における訳語信頼度. 言語処理学会第 13 回年次大会予稿集, 2007.