6B－5

自動アライメント機能を組み込んだ対訳コーパス構築環境 BACCS*

山崎 毅文　春野 雅彦
ＮＴＴコミュニケーション科学研究所

1 はじめに

対訳コーパスは、機械翻訳システム構築において必要な様々な言語知識を取り出す上で、重要な情報源である。実際、対訳コーパスから様々な言語知識を抽出する試みがなされている [1][5] が、精度の良い知識を抽出するためには、なるべく大量の文対応の取れた対訳コーパスがあることが望ましい。

対訳コーパスから文レベルの対応関係を特定する、いわゆる文対応 (アライメント) の研究に関しても、様々な試みがされているが [2][3][4][7]、構造の異なる言語族間 (例えば日本語と英語) で正しく文対応がとれた対訳コーパスを収集するための方法論やツールについて、あまり研究がなされていない。

本稿では、文対応の取れた対訳コーパス収集を容易にするための対訳コーパス構築環境 BACCS (Bilingual Aligned Corpus Construction System) について述べる。BACCS は、従来手法とは異なるロバストなアライメントプログラムを有し、プログラムの出力するアライメント結果をグラフィカルなインタフェースを通じて簡単に確認／修正が可能な対訳コーパス構築環境である。以下、BACCS が有するアライメントプログラムの詳細、及び BACCS の持つ機能について述べる。

2 統計情報と辞書情報を用いた自動アライメントプログラム

2.1 アライメントプログラムの概要

BACCS が有するアライメントプログラムは、統計情報と辞書情報を用いて得られる訳語対を基に、アンカー (対応関係が確定した文対応ペア) を見つける操作を繰り返すことによって、文対応を行なう。統計情報の利用は、文脈に応じた情報が利用できる点や、形態素解析の誤りに対処可能であるというロバスト性がある反面、複数回出現する単語しか利用できないという欠点がある。一方、辞書情報の利用は、一度しか出現しない単語に対しても対処可能である反面、訳語選択の多様性、形態素解析の誤りに対応不可という欠点を有する。本アライメントプログラムは、この両者の長所を兼ね備えたものである。

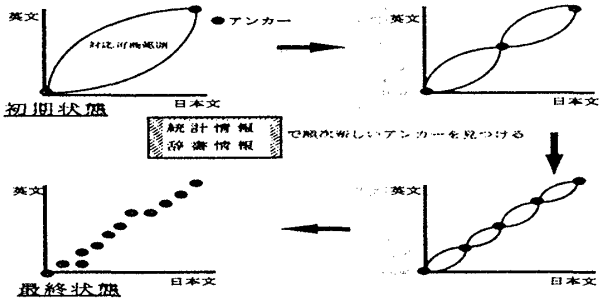

図 1: アライメントアルゴリズムの流れ

アルゴリズムの流れは、Kay のアルゴリズム [3] を基本としている。図 1 で示す通り、既に決定したアンカーを基に対応可能範囲を生成する。続いて、統計情報と辞書情報によって得られる訳語対から、新たなアンカーを見つけ出す。この操作を繰り返すことで、順次文対応ペアを見つける。

統計情報による訳語対の生成手順は、まず対応可能ペアに出現する日本語単語、英語単語の相互情報量と t-score[2] を計算し、次にそれらの値がある決められた閾値以上のもつものを訳語対とする。この閾値は、初期的には小さめに設定し、最終状態に近づくに従い大きくするというアニーリング手法により決定する。

2.2 アライメントプログラムの評価

"Scientific American" とその日本語訳版である日経サイエンスからの対訳テキストと、WWW を通じて読売新聞ホームページから得られる社説の対訳テキストを用いて、本アライメントプログラムの評価を行なった。適合率／網羅率による評価で、91～96 % の精度を得ている。

本プログラムの詳細、実験評価の詳細は、文献 [6] を参照されたい。

*BACCS: Bilingual Aligned Corpus Construction System with alignment program
Takefumi YAMAZAKI, Masahiko HARUNO
NTT Communication Science Laboratories

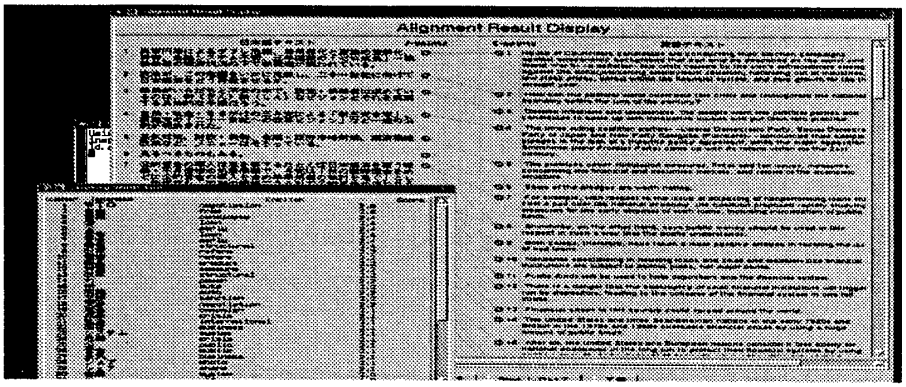

図 2: BACCS の起動画面 (左下: 単語登録, 中央: アライメント編集)

3 アライメント構築環境の必要性

先の実験結果で示す通り、提案したアライメントプログラムによって、ある程度の精度の文対応が可能であるが、いつも完全な対応関係が得られるとは限らない。よって、高品質な大量の対応付けされた対訳コーパスの蓄積は、人間による確認／修正作業が必要である。また、本作業によるユーザからのフィードバック情報を通じてアライメントプログラムの改良への手がかりを得ることができる。さらに、アライメントプログラムの副産物として得られる統計情報による訳語対の内、ユーザの確認作業により再利用可能な訳語対の保存により、アライメントプログラムの精度の向上も期待できる。

4 対訳コーパス構築環境:BACCS

対訳コーパス構築環境 BACCS は、前述したアライメントプログラムのアライメント結果をポインティングデバイスで容易に修正／確認ができるインタフェースを提供する。また、ユーザは、BACCS を通じてアライメントプログラムで得られる訳語対の中からユーザ辞書に登録する訳語対を選択することも可能である。

BACCS を用いたコーパス作成手順は次の通りである。まず、対象となる対訳テキストを選択し、アライメントプログラムを起動し、その結果をディスプレイに表示する。次に、マウス操作により、確認／修正作業を行なう。最後に、アライメントプログラムが出力する訳語対の中から正しい訳語対を選択し、ユーザ辞書に登録する。BACCS の起動例を図 2 に示す。

BACCS は、X window 上で Tcl/Tk を用いて実現されている。現在、BACCS を用いて大量対訳コーパスを作成中であり、構築支援ツールとしての効果も合わせて検証中である。

5 おわりに

本稿では、対訳コーパス構築環境 BACCS について述べた。ロバストなアライメントプログラムの組み込みと修正確認作業が容易なインタフェースの提供により、効率的な大量対訳コーパス構築が可能になった。今後の課題として、ユーザからの修正情報を利用したアライメントプログラムの改良、また蓄積された対訳コーパスからの対訳フレーズの自動抽出等が挙げられる。

謝辞

WWW Homepage 上の社説データ利用の許可を頂いた、読売新聞メディア企画局に感謝致します。

参考文献

[1] Ido Dagan and Ken Church. *Termight:* identifying and translating technical terminology. In *Proc. Fourth Conference on Applied Natural Language Processing(ANLP)*, pp. 34–40, 1994.

[2] Pascale Fung and K W Church. K-vec: A new approach for aligning parallel texts. In *Proc. 15th COLING*, pp. 1096–1102, 1994.

[3] Martin Kay and Martin Roscheisen. Text-translation alignment. *Computational Linguistics*, Vol. 19, No. 1, pp. 121–142, March 1993.

[4] J.C.Lai P.F.Brown and R.L. Mercer. Aligning sentences in parallel corpora. In *the 29th Annual Meeting of ACL*, pp. 169–176, 1991.

[5] Frank Smadja and Kathleen McKeown. Translating collocations for use in bilingual lexicons. In *ARPA Human Language Technology Workshop 94*, pp. 152–156, 1994.

[6] 春野雅彦, 山崎毅文. 辞書と統計を用いた対訳アライメント環境. 言語処理学会 第 2 回大会, 1996.

[7] 宇津呂武仁, 松本裕治. 対訳辞書及び統計情報を用いた二言語対訳テキスト照合. コンピュータソフトウェア, Vol. 12, No. 5, pp. 12–21, 1995.