日中対訳コーパスにおける単語・句の翻訳対応関係の付与

張 玉潔[*1]　王 主龍[*2]　内元　清貴[*1]　馬 青[*1,*3]　井佐原 均[*1]
[*1]情報通信研究機構
[*2]Fujitsu R&D Center LTD.　[*3]龍谷大学
[*1]{yujie.zhang, uchimoto, isahara}@nict.go.jp
[*2]wangzhulong@cn.fujitsu.com [*3]qma@math.ryukoku.ac.jp

１．はじめに

対訳コーパスは、その文、句、さらに単語レベルでの翻訳対応関係を利用することによりさまざまな構文単位の翻訳知識を抽出することができ、用例ベース翻訳と統計ベース翻訳の学習や評価にも必要不可欠である。

ＮＩＣＴは２００２年度からアジア言語を中心に多言語対訳コーパス（ＮＩＣＴ多言語コーパスと呼ぶ）を構築するプロジェクトを開始している[1]。このプロジェクトは構文情報や単語・句レベルでの対応付け（アライメント）など、詳細情報の付与に重点を置いている。ＮＩＣＴはまた、２００６年度から科学技術文献の日中・中日機械翻訳システムの開発にも着手した。機械翻訳システムのアプローチとして用例ベース手法を主眼としているため、大規模な日中対訳コーパスを構築し、それに対し単語と句レベルでのアライメントを自動的に行い、そこから翻訳知識としての対訳関係を抽出する技術の開発が必要となる。自動アライメント技術の開発及び評価を行うためにはまず、標準となる単語・句レベルで対応付けされたデータを準備しておく必要がある。

本稿に述べる研究の目的はＮＩＣＴ多言語コーパス構築の一環とする日中対訳コーパスにおいて、単語・句レベルでのアライメントを人手で付与する作業を行い、標準となる単語・句レベルでアライメントされたデータを構築することである。その第一ステップとして、対応付けの付与基準を定め、人手作業を補助するツールを開発した。

２．ＮＩＣＴ日中対訳コーパス

ＮＩＣＴ日中対訳コーパスは日本語原文と中国語訳文から構成される。日本語原文は新聞記事と雑誌から選んだものである。中国語訳文はプロの翻訳者により日本語原文から中国語に訳したものである。日本語文を一つずつ中国語文に訳したため、得られた対訳コーパスはすでに文レベルで対応付けされている。日本語文については、京大コーパスを使用しているため、形態素解析情報と構文構造情報はすでに付与されている[2]。中国語文には、単語分割と品詞情報を北京大学の基準[3]を用いて付与した[4]。コーパスの詳細情報を表１に示す。

表 1. ＮＩＣＴ日中対訳コーパスの詳細

	日本語	中国語
文	38,383	
単語	947,066	877,859
語彙	36,657	33,425
一回出現の語彙	15,036	13,238
平均文長（文字）	24.7	22.9

３．アライメント付与の補助ツール

人手による単語・句レベルでのアライメントの付与作業を補助するために、補助ツールを開発した。Melamed[5]が開発したツールなどを参考して主に以下の機能を備えたツールを開発した。

（１）自動アライメント技術がすでに実用のレベルに達しているため、人手作業は自動で得られたアライメントを修正することになる。補助ツールは自動的に得られたアライメント結果を表示

することができ、修正することもできる。本研究では多数の自動アライメントシステムの出力結果を統合した手法[6]を用い自動的なアライメントを行っている。

（2）ツールは作業者にとって操作しやすい可視化インターフェースを提供する。例えば、原文と訳文を横方向で表示することや、長い文の場合、スクロールで見ることができる。

（3）ツールは単語より大きい文法単位、すなわち句を選択することができる。これにより、もっと大きい単位あるいは構文構造の間にアライメントを付与することができる。

開発したツールのインターフェース画面を図１と図２に示す。操作画面の左側には文対の番号が表示されている。右上と右下のエリアにはそれぞれ中国語と日本語の構文構造が表示されている。中央のエリアにはアライメントの結果が表示されて、修正作業ができる。アライメントされた単語がグレー色になり、アライメントされていない単語が黄色になる。

単語間のアライメントを付与するには、マウスの左ボタンで日本語形態素と対応している中国語単語を選択すればよい。それによって日本語形態素と対応の中国語単語の間に線が引かれることになる。大きい単位（すなわち、複数の日本語形態素と複数の中国語単語）でのアライメントを付与する場合、原文の単位内のそれぞれの形態素から中央の点に線が引かれ、その中央の点から訳文の単位内のそれぞれの単語に線が引かれることになる。図１はその表示方法を示している。一方、Melamed の表示方法では、大きい単位でのアライメントができず、原文のそれぞれの単語が訳文のそれぞれの単語と線でつながっているため、線が乱雑で見にくい。

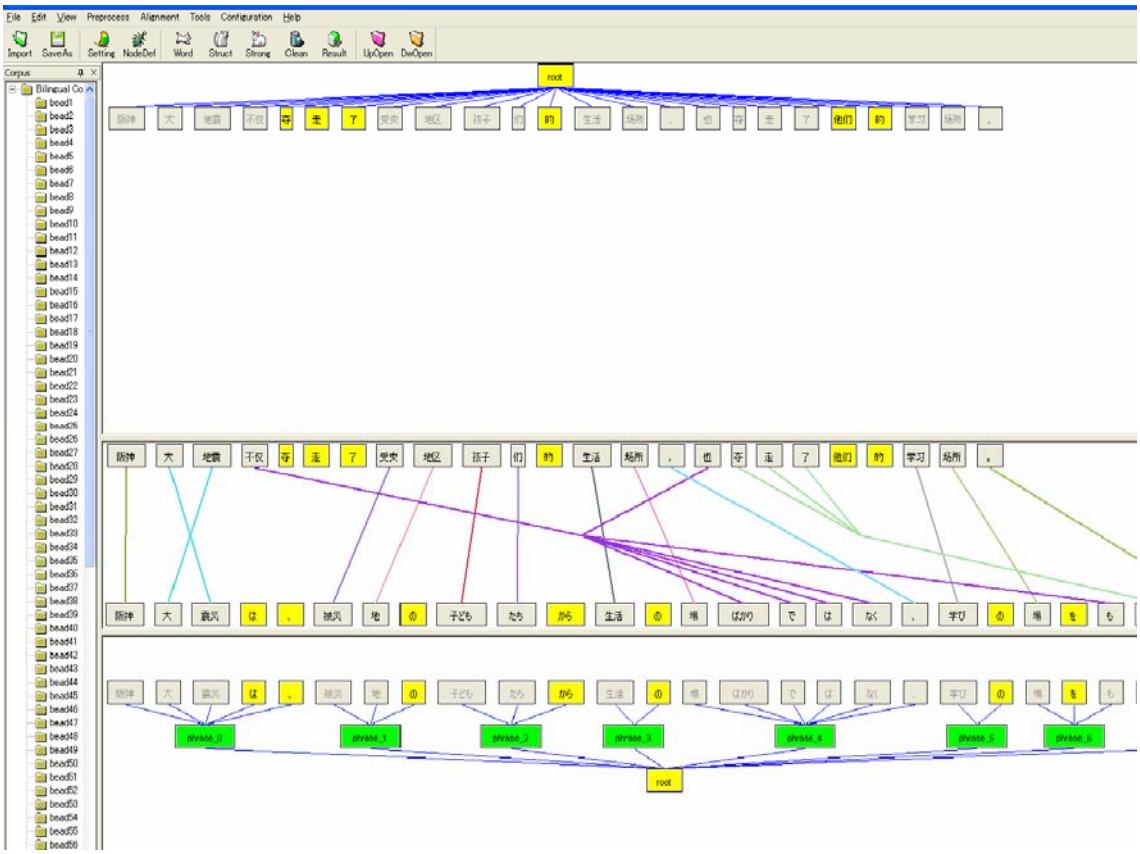

図１．アライメント付与作業補助ツールのインターフェース（単語レベル）

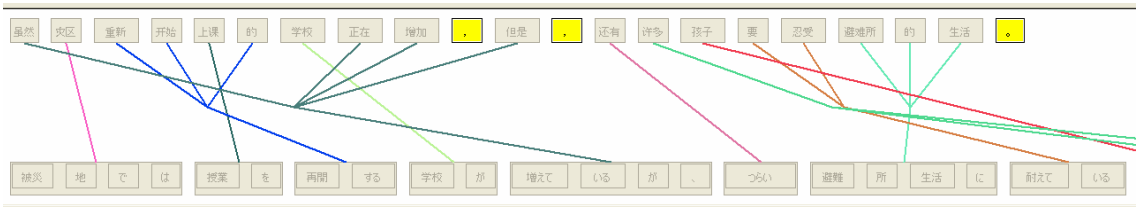

図２．アライメント付与作業補助ツールのインターフェース（句レベル）

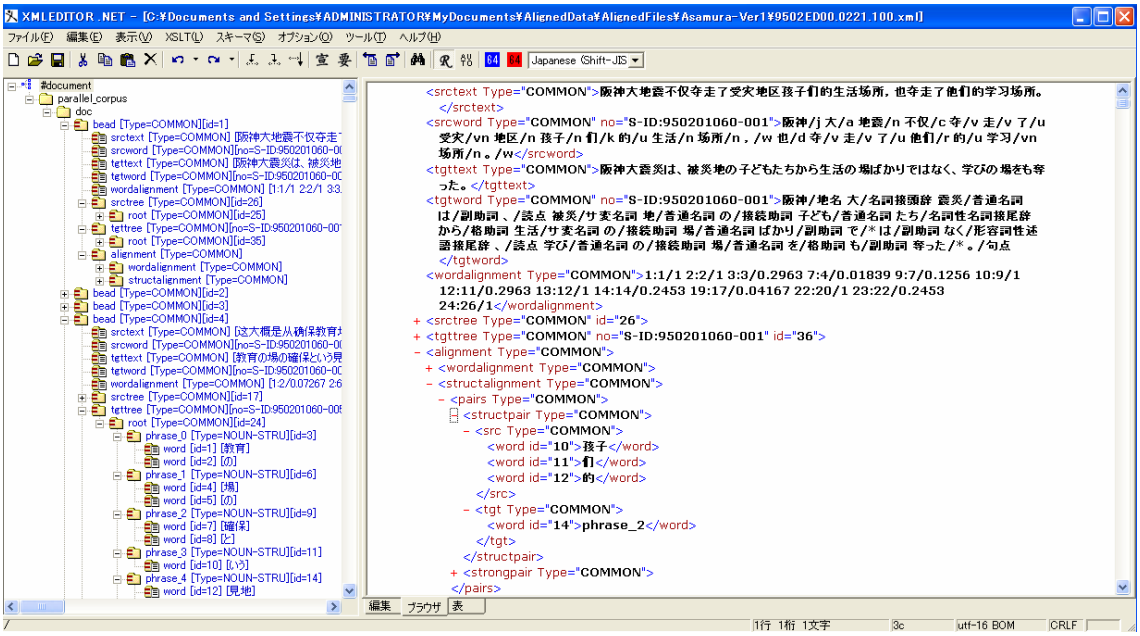

図３．アライメント付与結果を格納するファイルのフォマート

句レベルでのアライメント付与モードを選択すると、図２に示されているように、（日本語側のみ構文構造が付与されているので）日本語の句の列（文節の並び）が中央の修正エリアに表示される。句を構成する形態素列はより大きめな四角ボタンで表示されている。四角ボタンをクリックすると、句が選択されることになる。これにより大きい単位でのアライメント作業が可能となり、作業効率が向上する。

アライメントの付与結果はXMLフォマートのファイルに格納され、ユニーコードでエンコードされる。文、形態素情報、構文構造情報、単語アライメント、句アライメントを格納するために、いろいろなタグを設計し、使用した。図１に示されている文対のアライメントの付与結果を図３に示す。

４．アライメントの付与基準

① まず内容語に着目し、アライメントを修正・付与する。すべての内容語のアライメントが終わったあと、ほかの単語に対しアライメントを修正・付与する。

② 翻訳対応関係は事前に選定した複数の日中翻訳辞書により検証する必要がある。

③ 翻訳知識のカバー率を上げるために、最小単位でのアライメントが原則である。すなわち、アライメントされた日本語形態素列と中国語単語列はさらに二つ以上のアライメントに分割できないようにする。

④ 熟語および慣用表現の場合、文法上と意味上での対応関係を成り立たせるために、日本語側の熟語および慣用表現に対応する複数の形態素と中国語側のそれらをそれぞれグルーピングし、アライメントする。

⑤ 日本語の前置詞は中国語の「介詞＋方位詞」に対応している。しかし、介詞と方位詞は位置的に離れている。例えば、日本語の前置詞の文節「机の上**で**」は中国語の「**在** 桌子 **上**」に対応している。このような場合、中国語介詞「**在**」と方位詞「**上**」をまずグルーピングし、日本語の前置詞「**で**」に対応付ける。

⑥ 日本語の接続詞は中国語の二つの「連語」に対応している。しかし、二つの連語は位置的に離れている。例えば、日本語の接続節「遅い**が**」は中国語の表現「**虽然**晚了，**但是**」に対応している。このような場合、中国語連語“**虽然**”と“**但是**”をまずグルーピングし、日本語の接続詞“**が**”に対応付ける。

⑦ 日本語と中国語との間の一つの大きな違いは前者に形態素の変形があり、後者にないことである。後者の中国語は、「時態助詞」の「**了**」、「**过**」、「**着**」で時制を、「介詞」の「**被**」と「兼語動詞」の「**使**」で受動と使役を表現する。一方、日本語の場合、活用語尾が活用語幹と分離されている場合、その活用語尾と中国語文の「時態助詞」や「介詞」などと対応させる。一方、活用語尾が活用語幹と分離されていない場合、まず中国語文の「時態助詞」や「介詞」と関連動詞とをグルーピングして日本語文の活用語（活用語幹＋活用語尾）と対応させる。

５．おわりに

本稿では、ＮＩＣＴの日中対訳コーパスに単語・句レベルのアラインメントを付与する作業について、その付与基準及び作業補助ツールについて述べた。この作業は、平成１８年度からスタートし、現段階では、単語レベルの１万文対と句レベルの１万文対のアラインメント付与作業を完了した。平成２０年度の末までには全部完成し、公開する予定である。

参考文献

[1] Uchimoto, K., Zhang, Y., Sudo, K., Murata, M., Sekine, S. and Isahara, H. 2004. Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications. In Proc. of the MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources,pp.63-70.

[2] Maekawa, K., Koiso, H., Furui, F., Isahara, H. 2000. *Spontaneous Speech Corpus of Japanese.* In Proc. of LREC2000, pages 947—952.

[3] Yu, S. 1997. Grammatical Knowledge Base of Contemporary Chinese. Tsinghua Publishing Company.

[4] Zhang, Y., Uchimoto, K., Ma, Q. and Isahara H..2005-a. Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multi lingual Corpora. In the Tenth Machine Translation Summit Proceedings, pp.71-78.

[5] I. Dan Melamed. 2001. Empirical Methods for Exploiting Parallel Texts. The MIT Press. Linguistic Data Consortium. 2006. Guidelines for Chinese Word Alignment Annotation.

[6] Zhang, Y., Liu, Q., Ma, Q. and Isahara, H. 2005-b. A Multi-aligner for Japanese-Chinese Parallel Corpora. In The Tenth Machine Translation Summit Proceedings, pp.133-140.