節境界を考慮した長文の単語アライメント

柏岡 秀紀
ATR 音声言語コミュニケーション研究所
hideki.kashioka@atr.jp

1 はじめに

講演などの独話では独りで発話しつづけるために、文の区切りが明確ではない。例えば、「新しいゲーム機が発売された」「若者の人気を集めている」という２文の内容を発話するとき、「新しいゲーム機が発売されて、若者の人気を集めている」となることが多い。さらに、この発話末も文末にならずに、「新しいゲーム機が発売されて、若者の人気を集めており、売上げが…」のように発話が続き、文が長くなることも考えられる。そのため、人手により文境界を付与しても、一文が長くなる傾向がある。さらに、文境界の判定も人による判定の揺れが大きい。また、長文は、テキストにおいても様々な対象分野であらわれる。特許文は、その一例であり、一文が複雑な構造を持つ。これら、長文を扱う処理はかなり改善されつつあるが、未だに多くの問題を含んでいる。係り受け解析の精度は、長文に関しては決して高いとはいえない。また、翻訳においても、対訳データの対応関係の推定は難しく、翻訳知識を長文から獲得することは困難である。

先の例で示したように、講演などの独話に現れる長文は、複数の文を節としてつなげたものも多い。また、特許文では、一つの用語を説明している節を埋め込んで、指示対象の曖昧さを無くすような文にするために、構造が複雑になっている。独話あるいは特許などにおいて、これら長文の翻訳や要約等の処理を効率的に行うためには、文より短い意味的なまとまりを持つ単位での処理が必要である。日本語では、講演の同時通訳データの分析から、節がその有効な単位であると考えている [5]。しかしながら、節にはその切れ目としての強さがあり [7]、単純に全ての節を分割し各々を処理単位とするより、適切な処理単位があると思われる。実際、日英の対訳データの対応関係を見ると、節の出現順序が異ることは多く、また、１割程度であるが節に対応する部分が２カ所以上に別れて現れる。

本稿では、節対応データの分析について述べ、節境界を考慮した長文の単語対応を行う手法を提案する。

2 節対応データの分析

NHK の解説番組である「あすを読む」２５０番組に対して節対応の取れた対訳データを構築している。[2] 以下にデータおよび分析について述べる。

2.1 節対応データの構築

本稿では、NHK の解説番組「あすを読む」[1]を講演データとして利用した。以下の手順で、２５０番組に対して節対応データの構築をおこなった。

1. 番組の書き起こしテキストの作成
2. 文対応を取った対訳の作成
3. 書き起こしテキストへの節境界情報付与 [4]
   （節境界判定処理プログラム (CBAP) を利用）
4. 日本語節境界情報を見ながら対訳テキストの分割、対応情報の付与

上記の手法で、英語側の分割では、特に統語構造的な制約を設けてはいない。表現内容が対応する部分を切り出しており、前置詞等については英語の慣習的な切れ目を尊重するとして、その判断を作業者にまかせている。日本語の節に含まれる表現内容が他の節に対応する表現で分割されている場合、対応する英語が２カ所以上に現れることになる。

2.2 節単位における対応関係

図１に節対応データのサンプルを示す。日英の表現を結ぶ線は、対応関係を示しており、太線は、その対応関係の線が交差していない切れ目を示している。

[1]月曜から金曜まで毎日 1 番組 10 分で放送されている解説番組

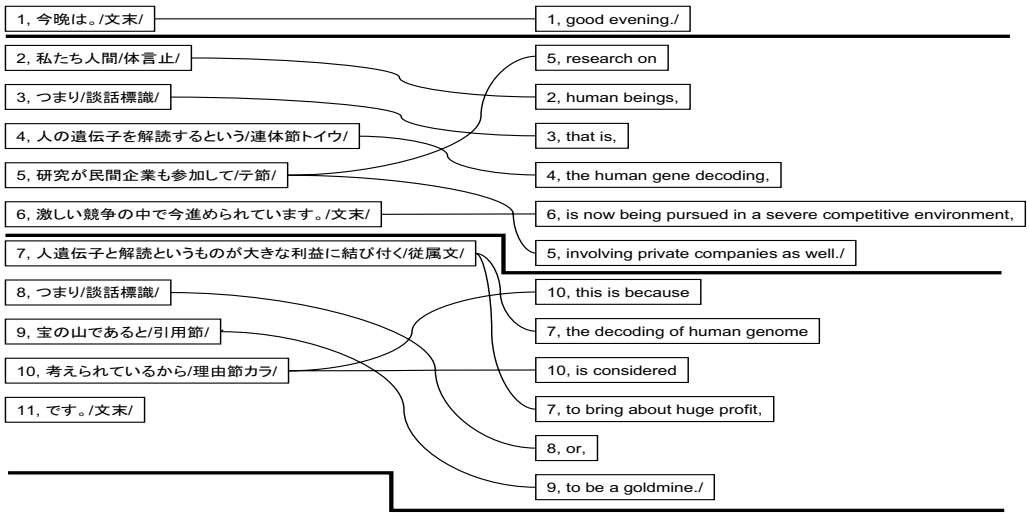

図 1: 節対応データサンプル

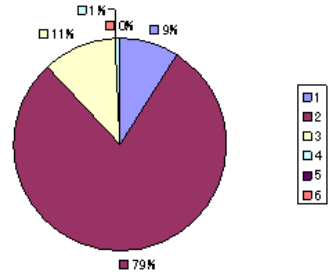

図 2: 日本語の節に対応する英語の部分数

図 2 に日本語の節一つに対応する英語部分の数を示す。

図 2 より、対訳部分における節毎に対応する部分は、９割近くがまとまっていることがわかる。そのため、大雑把なアライメントを行ったのちに、節対応を考慮して詳細なアライメントを行うことで、精度よくアライメントを取ることが期待できる。

3　長文のアライメント

講演などの独話データにあらわれる語彙は、その講演に応じて分野が決まり、各分野の専門用語が多く現れる。また、発話中に生じる略称等の出現も多く、表現の多様性に富んでいる。そのため、多くの講演データに対してアライメントを取ることを考えたとき、辞書を利用した手法では、辞書引き出来る語彙が限定されるため、ほとんど対応がつかないことが多い。また、特許においても専門用語が多く、同様の現症がおこる。そのため、統計的な情報によるアライメント手法が有効と考えられる。

3.1　GIZA によるアライメント

本稿では、統計的なアライメントのツールとして、GIZA を利用した [3]。文の対応が取れている（と思われる）データに対して処理を行った。日本語は、茶筌により単語に分割し品詞の情報も利用した [6]。英語は、スペースにより単語に分割し、一部、ピリオド、コンマの処理を行った。アライメントの段階では、英語の品詞情報や構文情報などは利用していない。特許文書でのアライメント結果を図 3 に示す。

図 3 から、ある程度の精度で対応付けが成功していることがわかる。しかしながら、特許において複数回あらわれる専門用語などは、適切な対応が取れていないことなどが見受けられる。また、非対応となっている部分も多く見られる。

3.2　節境界を利用したアライメント

図 3 は、日本語部分を節境界単位に分割した表示となっている。アライメントが取れている部分は、節ごとにある程度まとまりを持っていることがわかる。2 節で述べた「あすを読む」の節対応データでも節毎のまとまりがあることから、英語を上手く分割できれば、節

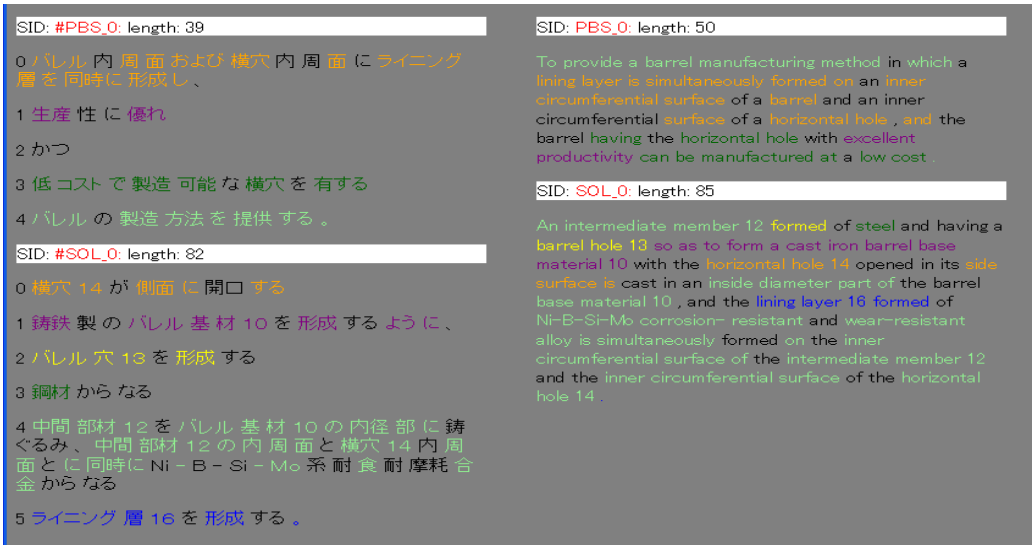

図 3: 特許文のアライメント結果

単位でアライメントを取り直すことにより、高精度な単語アライメントが実現できる。

以下に、本稿で提案する手法の手順を示す。

1. 文単位での単語アライメント処理を実行
   （本稿では GIZA を利用）
2. 日本語文を節境界単位に分割
   （本稿では CBAP を利用）
3. 英文を解析し一定のユニットに分割
   （本稿では Charniak Parser を利用 [1]）
4. 文単位でのアライメント結果を利用し、英文のユニットをある程度まとめあげ、日本語文の節境界との対応を取る
5. 節単位での単語アライメント処理を実行
6. 必要に応じ英文のユニットのまとめあげを、修正 5、6 を繰り返す。

上記手順のステップ 3 における英文のユニットの設定は、現在、Charniak Parser の出力を利用して、隣接する単語がパースツリーで深さが異ればユニットの境界としている[2]。また、ステップ 4 におけるまとめあげは、現在、次の手順にしたがって処理を行っている。

4-1 ステップ 1 のアライメント結果から、各ユニット内の単語に対応する語を最も多く含む節を各ユニットの対応する日本語の節とする。

4-2 隣接するユニットが同じ節と対応するなら、隣接するユニットをまとめる。

4-3 対応するユニットがない（ユニットに含まれる全ての単語が対応する節をもたない）ときは、後続するユニットとまとめる。

以下に、図 3 に示したデータから推定した節対応の結果の一部を示す。

- バレル の 製造 方法 を 提供 する 。
  To provide a barrel manufacturing method in which
- 生産 性 に 優れ
  with excellent productivity
- バレル 内 周 面 および 横穴 内 周 面 に ライニング 層 を 同時に 形成 し 、
  a lining layer is simultaneously formed on an inner circumferential surface of a barrel and an inner circumferential surface of a horizontal hole , and
- 補修 施工 する 場合 、
  effective repair when
- 高い 品質 を 維持 さ せる 。
  To maintain high quality

[2]英文のユニットは、いわゆるチャンクでよいと思われる。様々なツールが有るが、比較は行っていない。

3.3 考察

現在の手法では、前節で述べたステップ 1 のアライメントにおいて誤った対応となった部分に対して、ステップ 4 におけるまとめあげで、適切な節の対応関係を構築できない。そのため、ステップ 4-1 において、対応する節を最も多く含む節という基準で判断するのではなく、GIZA の出力する対応の尤度を考慮にいれ、同じ語が一文中に複数現れている場合は、それらの間での対応誤りが多くみられるので、その補正をした判断基準が必要である。実際に前節で示した対応の結果の中でも、対応する部分の長さの釣合が取れていないところも多い。補正するために、長さの釣合をとることも必要である。実際のデータでは、「あすを読む」の節対応データからもわかるように、ほとんど節に対応する部分は、まとまっており、本稿での提案手法を利用することで、節対応のついた長文の単語アライメントを実現することができる。さらに、節単位での対応情報を利用することにより、日英の表現の節レベルでの順序関係や長い修飾表現の対応関係の統計情報が得られるようになる。

4 まとめ

本稿では、節対応データの分析から、長文のアライメントに対して節境界単位の情報は有効であると判断し、節境界単位を利用した長文のアライメント手法を提案した。現在の試作システムでは、文全体でのアライメントの誤りの影響を強くうけているが、アライメントの尤度、対応する単位の長さ、同一語句の重複などを考慮し、目的言語の分割基準を修正することにより、精度よくアライメントを実現することが可能となると考えている。今後は、この目的言語の分割基準を改修したシステムの作成および実験を行い、提案手法の有効性を確認するとともに、日本語の節間の構造と目的言語である英語の構造、および節に対応するまとまりの分析を進める。また、日本語の節タイプおよび節間の関係から、英語での表現を節や句、文として表現するなどの判断基準を設定し、節単位での翻訳システムの構築を行いたい。

謝辞

本研究は総務省戦略的情報通信研究開発推進制度における研究委託により実施したものである。

参考文献

[1] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL 2000*, pp. 132–139, 2000.

[2] H. Kashioka, T. Maruyama, and H. Tanaka. Building a parallel corpus for monologue with clause alignment. In *Proceedings of MT Summit IX*, pp. 216–223, 2003.

[3] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19 – 51, 2003.

[4] 丸山, 柏岡, 熊野, 田中. 日本語節境界検出プログラム cbap の開発と評価. 自然言語処理, 11(3):39–68, 2004.

[5] 柏岡, 田中. 講演の同時通訳データの分析. 言語処理学会第 7 回年次大会発表論文集, pp. 433–436, 2001.

[6] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム 『茶筌』 version 2.3.3 使用説明書. 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座, 2003.

[7] 南. 現代日本語の構造. 大修館出版, 1974.