日英翻訳システムの改良とタグ付きコーパスの作成

6Q－6

畑山満美子　白井諭
NTTコミュニケーション科学研究所

平野志奈　野原ゆかり　阿部さつき
NTTアドバンステクノロジ（株）

1　はじめに

現在，様々な自然言語解析システムが研究されており，解析精度の向上も目覚しいものがあるが，さらなる精度向上が求められる．機械翻訳システムALT-J/Eにおいても，各処理の改良が行なわれているが，形態素解析，構文解析，意味解析などの各処理には一連の流れがあり，前処理における解析誤りが次の解析に悪影響を与えるため，各処理の独立した改良が難しい場合がある．

そこで，

1. 翻訳システムの中間の処理の入出力情報をコーパスとして定義する，これにより，各処理の相互依存性を相対的に低下させ，処理の改良を容易にする，
2. 各処理に補正処理を組み込み，修正情報を補正ルールとして取り込むことにより，人手修正のばらつきを防止する，

ことを狙いとして，詳細な情報を均一に付与したタグ付コーパスの作成方法を提案する．

タグ付コーパスにはEDRコーパス[1]，RWCコーパス[2]などが既に存在するが，人手修正を行なっているため，タグのばらつきが生じていると考えられる．京大コーパス[3]は，解析処理の改良と一体的に進めることにより品質の問題への対処を考えているが，高精度の意味解析を検討することには情報の種類が不足していると考えられる．

そこで本稿では，京大コーパスの考え方を発展させ，機械翻訳システムALT-J/Eの内部情報を利用することにより，より深い情報を付与することを考える．また，日英翻訳の検討に使用している対訳コーパスをソースデータとして使用することにより，大規模な対訳コーパスが作成できると考える．

付与する情報として，形態素情報，構文（係り受け）情報の他に，意味に関する情報として，単語の意味属性，複合語構成の解析情報，文型パターン，機械翻訳への利用を考えて，英語の対訳記事を付与する．

このような対訳タグ付コーパスの作成はタグ情報の誤りを大量に均一的に修正し，機械翻訳システムALT-J/Eの改良が可能である．また，正しい形態素情報の付与されたコーパスを与えることにより，構文解析システム処理の単独な改良が可能である．また，正しい形態素，構文解析情報から誤り補正ルールを獲得し自動的生成する，パターン対辞書の向上，など，様々な自然言語処理，機械翻訳処理への利用が考えられる．

解析システムを改良する場合，アドホックな改良にならないために，ある基準を設ける必要がある．このため，今回のコーパス作成では，誤りパターンをある程度獲得してから統計処理をして人手で重み付けを行ない，解析処理を改良する．一方，誤りデータ集から自動的にコーパスの補正をするルールを生成する．

2　対象とするソースデータ

対象とするソースデータは，日経新聞4紙のうち，記事対応のとれている日英記事とする．

対象記事は商用の日経テレコンデータベース（日本語記事；テレコンBIZ．英語記事；Japan News & Retrieval）から取得した．

現在，記事対応のとれているものは全体の3割だが，それらについては99%の記事対応の精度が保証されている[4]．

3　コーパス作成の概要

3.1　情報の付与

日本語コーパスに付与する情報は以下の形態素情報，構文情報，意味情報である．ここで，本研究では係り受けの特定をもって構文を解析したと定義する．同様に，絞り込んだ結果の意味属性を決定することで意味を解析したと定義する．本研究は機械翻訳を念頭においているため，日本語品詞の意味属性を決定することが訳語の語義決定につながるからである．

日本語コーパスに付与する情報

- 形態素情報
  文節，単語表記，読み，単語標準表記，品詞，
  活用形，自立語属性．
- 係り受け情報
  文節の係り受け，複合語内部の単語の係り受け．
- 意味情報
  複合語のまとまりとしての品詞・活用形，
  複合語内部の品詞・活用形，
  意味属性（一般名詞，固有名詞，用言），
  文型パターン．

英文コーパスには，日英記事対応のタグを付与する．

---

*Improving Japanese-to-English Machine Translation by Compiling a Tagged Corpus.

†Mamiko HATAYAMA, Satoshi SHIRAI, Shina HIRANO, Yukari NOHARA, Satsuki ABE, NTT Communication Science Laboratories, NTT Advanced Technology Corporation.

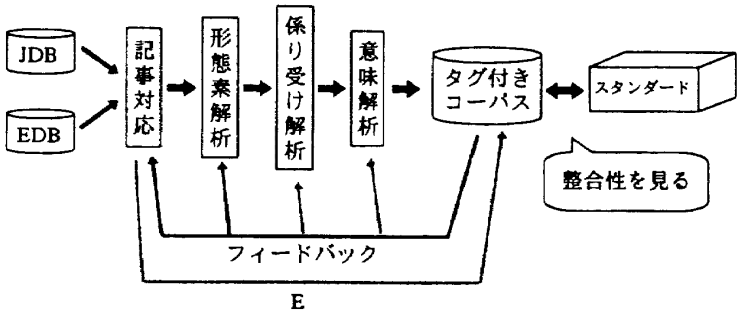

図 1: コーパス作成の手順

3.2 作成の手順

記事対応のとれている日本語記事に対し，形態素解析，係り受け解析，意味解析，を行なう（図1）．

まず，日英記事対応のとれている日本語文に対し形態素解析を行ない，形態素情報を付与する．この解析結果の自動付与には，機械翻訳システム ALT-J/E の形態素解析システム ALT-Jaws を用いる．得られたタグ付コーパスに修正を加える（3.3節）．これによって，正しい形態素情報の付与されたコーパスが作成できる．

次に，正しい形態素情報が付与されたコーパスをフィードバックし，構文解析を行なう．これによって，形態素解析誤りの悪影響を受けない，純粋な構文解析処理を行なうことができ，同様にして構文解析処理システムに特化した解析誤りを修正する．

同様にして，修正を行ない正しい情報（形態素，構文）の付加されたコーパスをフィードバックすることによって，他処理の影響を受けずに意味解析を行なう．

第1段階では，形態素情報と意味情報の一部の付与を行なう．

3.3 情報の修正

大量のタグ付コーパスを作成するためには，何らかの機械処理を行なって解析情報を付与するのが一般的である．しかし，現在の機械処理では，どんな高精度であっても 100% の精度を得ることが出来ない．そこで，機械処理による解析誤りを人手で修正する必要がある．人手での修正では，修正結果の均一性が問題となる．そこで，補正ルールの解析誤りの分析に基づいて自動生成を行ない，修正作業のコスト軽減と質の高さを保持する．また，補正ルールを現システムに反映させることにより，現システムの改良を行なう．

具体的には，ALT-Jaws による形態素情報を自動付与し，その結果を人手で修正する．修正はひとつひとつを修正するのではなく，誤り箇所とパターンを人手でチェックし，誤りパターン集を作る．誤りパターン集を統計処理し，（人手で）重み付けを行なう．次に，重みによってコーパスの補正ルールを自動生成する．

これによって正しいタグ付コーパスから ALT-J/E の解析システムの改良を行なう．

同時に，修正ツールの開発も行なう予定である．

4 今後の課題

第2段階として，3.1節の情報を備えたタグ付コーパスの完成を目指す予定である．

英語記事は記事単位でのマッチが示されているが，現在，文単位でのマッチングが行なわれているところである [4]．将来的には，英語コーパスに英文の解析情報も付与し，英文タグ付コーパスと日本語コーパスとのマッチングを付加することも考えられる．

今後，このタグ付コーパスの作成を通して，正しい形態素情報，構文解析情報から学習によってルールを獲得し，自動的生成する研究 [5]，正しい形態素情報，パターン対辞書を使って，係り受け解析の改良を行なう研究 [6]，などの加速とフィードバックが期待される．

5 おわりに

本稿では，日英機械翻訳システムの改良と一体的にタグ付コーパスを構築する方法を提案した．構築作業を開始し，現段階の記事・文対応の精度で，約１５万文の日英対訳コーパスが得られる予定である．今後の記事・文対応の精度の向上によっては，それ以上のコーパスが得られることが期待できる．

参考文献

[1] http://www.iijnet.or.jp/edr/ を参照．

[2] 新情報処理開発機構 (RWCP)．テキストデータベース報告書 (1996)．

[3] 黒橋禎夫，長尾眞．京都大学テキストコーパス・プロジェクト，言語処理学会，第3回年次大会 (1997)．

[4] 高橋大和，白井諭，大山芳史，渡辺いづみ，上田洋美．日英新聞記事の記事対応コーパス自動作成，言語処理学会，第3回年次大会 (1997)．

[5] 春野雅彦，白井諭，大山芳史．決定木を用いた日本語係り受け解析，自然言語処理シンポジウム“実用的な自然言語処理に向けて”(1997)．

[6] 松尾義博，白井諭．格フレーム解析を結合した日本語係り受け解析，言語処理学会，第4回年次大会 (1998)．