4T-03　ニュースの同時通訳ための短文分割手法について

張玉潔 * 丸山岳彦 * 柏岡秀紀 * 浦谷則好 * 江原暉将 **

* ATR 音声言語通信研究所　** NHK 放送技術研究所

1 はじめに

ニュース文の機械翻訳において,長文を短文に分割することは翻訳精度を確保するために必要である[1]. また文章の推敲支援と係り受け解析のための短文分割についても研究されている[2][3]. 従来の研究は,長文の入力が終わった段階で短文分割の処理が始まる. しかし同時通訳の場合は,長文の全部を認定する前に分割を始めることが要求される. ここでは,ニュース文の入力に対し,リアルタイムに短文分割の処理を行い,適切な短文を生成する手法を提案する. また,その実験結果についても報告する.

2 同時短文分割

2.1 ニュース文の特徴

ニュース文は1文が長いという性質がある[4]. ニュース文の例を以下に示す.

[例1] A は ... について「... できたが, ... である。... しており, ... したい」と述べたのに対し,B は「... とともに, ... すると思う」と述べ, ... しました。

ニュース文が長文化する原因の一つは,直接引用を含むことにある. さらに,直接引用の中には複数の文が含まれていることもある. また,例1のように二つの直接引用を含む文も少なくない.

2.2 短文分割

従来の研究[5]ではまず主文と直接引用を分割する. 次に「が」,「ており」,「もので」,「のに対し」で終わる連用節を短文に分割して,その接続関係を次の短文の文頭で補う. 例えば例2のパターンを満たす長文は,例3に示す短文に分割される.

[例2] ... は ... ついて「Q_1。Q_2」と述べました。

[例3] [短文1] ... は ... ついて次のように述べました。
[短文2] 「Q_1。」
[短文3] 「Q_2」

2.3 同時短文分割

ニュース文の同時通訳において,短文を早い段階で翻訳処理に渡すためには,短文分割をニュース文の入力と同時に行うべきである. 例2では,“「”,“Q1。”,“Q2”,“」”が現れても,後ろの動詞“述べる”が揃わないと分割が始まらない. そこで,“「”が来た時点で,それまでの文脈から後ろには“述べる”のような意味を持つ動詞が来ると予測し,主文と直接引用を分割する. そして直接引用の内部で短文になるものが来た時点で分割を行う. その結果,“「”が入力されたあと[短文1]が出力され,句点が入力されたあと,[短文2]が出力され,“」”が入力されたあと,[短文3]が出力されるようになり,入力に対しリアルタイムに短文が分割できる.

そこで,次の点が問題になる. (1) 分割点の判定には分割点より後ろの文脈がどのくらい必要であるか. (2) 分割点より後ろの文脈は分割点より前の文脈と共起するか.

3 同時短文分割の手順

形態素解析と文節の区切りのあと,形態素の品詞,表記などの情報が記述された文節列が得られる. 同時分割にはこのような文節列がリアルタイムに入力されると仮定する.

$i\,(i\geq 1)$ 番目の文節 P_i が入力された時刻を単に i で示す. 分割点候補は任意文節 P_i の直後とし,C_i で表す. C_i の前の文節列を P_1^i で表す. 後ろの文節列は時刻とともに更新するので,時刻 $j\,(j\geq i)$ のときの文節列を $B_i(j)$ で表し,$B_i(i)=\emptyset, B_i(j)=B_i(j-1)\cup P_j$ になる. そして分割点候補 C_i に関して時刻 j で得られたデータを次のように定義する.

$$C_i(j)=(P_1^i, B_i(j)) \qquad (1\leq i\leq j)$$

ここの目的は時刻 j において $C_i(j)$ が分割点であるかどうかを判定することである. C_i の状態は“判定待ち”,“分割”,“非分割”という三つの値のいずれかを持っている. 分割手順は以下になる.

(1) 初期化　　未分割部分の始点:=1;
　　次の短文への情報:=$\emptyset$;

(2) 文末の入力が終わるまで次のことを繰り返す.

Automatic Segmentation of News Sentences for Simultaneous Interpretation

Yujie Zhang*, Takehiko Maruyama*, Hideki Kashioka*, Noriyoshi Uratani*, Terumasa Ehara**

*ATR Spoken Language Translation Research Lab.

**NHK Science and Technical Research Lab.

(2.1) 時刻 $j(\geq 1)$ でデータの作成と更新

$B_j(j) := \emptyset$; $C_j(j) := (P_1^j, B_j(j))$;

C_j の状態:= “判定待ち”;

$B_i(j) := B_i(j-1) \cup P_j$; $(1 \leq i < j)$

$C_i(j) := (P_1^i, B_i(j))$;

(2.2) 時刻 $j(\geq 1)$ で分割点の判定

for i:= 未分割部分の始点 to j do

if C_i の状態 = “判定待ち” then

begin

判定結果:=分割点の判定$(C_i(j))$;

if 判定結果 = “分割” then

begin

C_i の状態 := “分割”; 短文の生成;

次の短文への情報を設定する;

未分割部分の始点:=$i+1$

end

else if 判定結果 = “非分割” then

C_i の状態: = “非分割”

end

分割点の判定 と 短文の生成 という手続きは実装システムにより入れ替えることができる. 例えばルールベースによる方法と決定木による方法が考えられる. この手順により, C_i が分割点と判定される情報は後ろの n 個文節にしか依存しないなら, C_i が時刻 $i+n$ まで “判定待ち” で, 時刻 $i+n$ なら “分割” と判定できるようになる. $i+n$ は C_i の分割点ができた時点を示し, $T(C_i)$ で表す.

4 実験と結果

分割点の判定と短文の生成をルールベースによる方法で実装し, 44 個の分割規則を人手で作成した. そして直接引用を含む 210 文に対し同時短文分割の実験を行った. この 210 文は人手で 670 個の分割点のラベルを付け正解とした. 分割結果の一例を以下に示す. 例 4 の入力文が例 5 の 7 つの短文に分割された. 例 4 の添付数字は短文の番号と対応し, 入力においてその短文の分割点が判定できた時点を示す.

[例4] 会談では海部総理大臣が「日本はフィリピンの経済再建のためにできるだけの協力を[1]しており,[2] ことし2月に開かれた対フィリピン多国間援助構想でも15億7000万ドルの援助を約束している」と[3]述べたのに対し, ラモス国防相は「日本からの援助はフィリピン経済の安定に役立っている。[4,5] フィリピン経済は再建に向けて着実に前進しており,[6] 日本からの投資を期待している」と[7]述べました。

[例5] (1) 会談では海部総理大臣が次のように述べました。(2) 「日本はフィリピンの経済再建のためにできるだけの協力をしている。」(3) 「そして, ことし2月に開かれた対フィリピン多国間援助構想でも15億7000万ドルの援助を約束している」(4) それに対し, ラモス国防相は次のように述べました。(5) 「日本からの援助はフィリピン経済の安定に役立っている。」(6) 「フィリピン経済は再建に向けて着実に前進している。」(7) 「そして, 日本からの投資を期待している」

文末時点で短文分割を行う方法の結果と比べて, 同時短文分割の方がどれだけ早く得られたかを評価するため, 正しい分割点集合 S={u,v|文u の C_v が正しく “分割” と判定された} に対し次の尺度を定義する.

$$\bar{F} = \frac{\sum_{u,v \in S}(\text{文}u\text{の文節数} - T(C_v))/\text{文}u\text{の文節数}}{|S|}$$

669 個の正しい分割点から $\bar{F}$=40.37% が得られた. これは正しい分割点が平均して文長の 4 割くらいの早さで, あるいは文長の 6 割が入力された時点で得られたことを意味する. さらに分割点の判定には, 平均して後ろの 1.44 個の文節の文脈が必要であることが分かった. また 99.9% の再現率と 88.8% の適合率が得られた.

5 おわりに

本論文では, ニュース文の同時通訳の前処理として, ニュース文の入力と同時に短文分割を行う手法を提案した. 今後, 分割規則の整備と分割規則の自動抽出について研究する予定である.

なお, 本研究は以下の分担で行った. ATR ではアルゴリズムと処理プログラムの作成を行い, NHK では分割実験を行った.

参考文献

[1] 金淵培, 江原暉将: 日英機械翻訳のための日本語長文自動短文分割と主語の補完, 情報処理学会論文誌, 35(6) (1994).

[2] 武石英二, 林良彦: 接続構造解析に基づく日本語複文の分割, 情報処理学会論文誌, 33(5) (1992).

[3] 張玉潔, 尾関和彦: 決定木による日本語長文の短文分割, 自然言語処理, 7(1) (2000).

[4] 江原暉将, 沢村英治, 若尾孝博, 阿部芳春, 白井克彦: 聴覚障害者のための字幕つきテレビ放送制作への自然言語処理の応用, 言語処理学会第3回年次大会発表論文集 (1997).

[5] 江原暉将, 福島孝博, 和田裕二, 白井克彦: 聴覚障害者向け字幕放送のためのニュース文自動短文分割, 情報処理学会研究報告, NL-138-3 (2000).