6C−7

断片的な解析によるノイズを含んだ文の理解 *

劉学敏 西田豊明 堂下修司 †

京都大学 ‡

1 はじめに

我々は音声入力による統合的自然言語理解システムの作成を行なっている. このシステムは音声認識システムの出力を入力として利用する. 一方, 現在の音声認識の精度によって, その出力にはかなりのノイズが含まれている. このため, 自然言語理解システムにとって, ノイズによる入力の誤りに対処できる能力が必要である.

入力にはノイズが入っている場合, 単語を検出するために, 入力訂正を行なわなければならない. 一方, 入力訂正によって, 検出可能な単語の数が急に増える. これらの単語候補を絞り込むのは困難であり, また, 多数の単語候補を保持するのはその後の構文解析や意味解析にとって大きな負担となる. このような問題を解決するために, 本研究では, 断片的な解析という方法を提案する. この方法では, 我々はノイズを含んだ入力文をいくつかの断片に分けて, その中の誤りのない断片を優先して解析し, 次いで, その処理結果を利用して, 誤りのある断片の可能な構文構造と意味構造を予測し, その予測結果に基づいて, 入力誤りを訂正する. 本稿では, この方法の要旨について報告する.

2 断片の分けと管理

断片は単語検出の結果によって分けられる. 断片にはノイズのない断片とノイズのある断片と2種類がある. 単語を検出した場合, システムは単語の語境界情報を利用して, ノイズのない断片を生成する. 一方, 指定される位置から単語が検出できない場合, システムはこの部分をパスして, 新しい検出位置を見つける. このパスした部分をノイズのある断片として記録する. ここで, 新しい検出位置を見つけるために, 日本語に頻繁に出現する文字列 (例えば, 助詞, 動詞や形容詞の語尾の活用形など) を利用する. このように, 入力文を断片の系列に変換できる. ノイズのある断片には次のような入力誤りのどれかが含まれている.

(1) 置換誤り: 文字が誤った文字に置換された.
(2) 挿入誤り: 余分の文字が挿入された.
(3) 脱落誤り: 文字が脱落された.

一方, ノイズによって誤った単語を構成してしまうことがある. このため, 本研究では, ノイズのない断片に対する解析が失敗した場合, この断片をノイズのある断片に変更する.

断片の間には排他的なもの (例えば, 重なっている断片同士) がある. このため, 断片の管理が必要である. 断片の管理の目標として, 下の二つの機能を達成することである.

(1) 断片間の論理的な整合性の管理

断片間の排他性の保持, すなわち, 互いに排他的な断片を同時に採用しないこと, また, 既に処理が失敗した断片を再び処理しないことを保証する.

(2) 断片集合 (系列) の確からしさの管理

排他的な断片が存在するため, 与えられた入力文からいくつかの断片系列が生成できる. これらのうち, 最も確からしいものを取り出す.

上の目標を達成するために, 本研究では, 我々が提案した統合パーサ [1] の機能を利用する. この統合パーサは整合性維持エンジン CME と確からしさ維持エンジン PME の融合によって構成された統合的自然言語処理のための推論管理機構である.CME は言明の間の論理的な整合性を維持すると同時に, 可能な推論環境 (仮定の組合せ) を木の形で管理する. 一方,PME は仮定に与えられた確からしさを利用して, 現在の時点で最も確からしい推論環境を選出できる. このような機能を利用して, システムは排他的な断片が生成すると同時に, 各々の断片に応じて, 仮定を生成し, 統合パーサに与える. そして, 統合パーサによって, 確からしい断片系列を選出できる.

3 断片的解析

システムの処理を断片内の処理と断片の統合と二つの処理に分けて行なう.

断片内において, 単語の検出, また, 検出された単語の品詞性や意味概念の確定, そして, 単語候補集合の絞り込みなどの処理を行なう. 一方, 入力誤りによる単語候補の増加を抑えるために, 我々は入力誤りの特性に対して, 各々に異なった対処方法を採用する. 音声認識の場合, 母音の認識率が高いから, 脱落誤りは挿入誤りと置換誤りより発生しにくい. このため, 我々は単語検出を次の三つの段階に分けて行なう.

(1) 置換誤りだけ: 置換補正 (適当に文字を置換すること) によって単語の検出を行なう.
(2) 置換・挿入誤り: 削除補正 (文字を削除すること) によって,(1) の場合に変換する.
(3) 置換・挿入・脱落誤り: 断片間の処理によって得られた予測結果を利用して, 単語マッチング方法によって検出する.

---

*Understanding Erroneous Sentence with Fragmental Analysis

†Xuemin LIU, Toyoaki NISHIDA and Shuji DOSHITA

‡Kyoto University

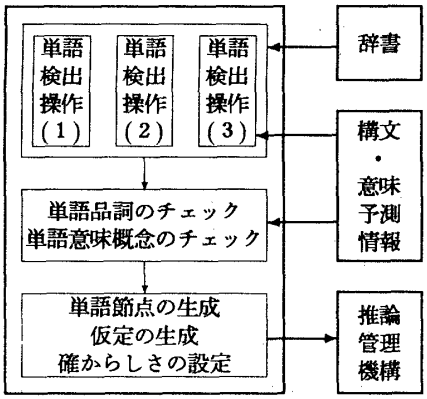

図 1: 単語処理過程の様子.

こうすると, 単語候補の数が大幅に減る. 本研究での実験によって, 登録した単語の数が約 4000 の場合, 脱落誤りも考えれば, 一つの断片 (平均 5 文字) から約 50 〜 200 の単語候補が検出できる. 一方, 置換・挿入誤りだけを考える場合, 可能な単語候補の数はその 15 分の 1 に減ることがわかった.

一つの断片から複数の単語が検出された場合, システムはまず, 各単語候補の品詞と意味概念をチェックし, 不整合な単語を除去する. 次に, 残った単語候補に対して, 仮定を生成し, 各仮定に対して, 確からしさを設定する. 最後, その処理結果を推論管理機構に与え, 管理させる. 単語検出処理の様子は図 1に示す. ここで, 単語検出操作 (1)(2)(3) は各々上の 3 つの段階に対応する.

一方, 断片の統合は断片の間において行なう. まず, 処理された断片に対して, 構文解析と意味解析を行なう. 構文解析と意味解析の結果を<構文構造 ・ 意味表現>というようなペアで表現する. 次いで, システムは構文解析の結果を利用して, ノイズを含んだ断片の可能な構文構造を予測し,<構文構造 ・ ϕ>のような構文的予測結果を生成する. 最後, この予測結果と意味解析の結果を統合して, ノイズのある断片に出現可能な単語集合を予測する. この予測結果を前述の単語検出 (3) の場合で利用する. この処理過程を一つの例によって説明しよう.

例 4. かいぎしつきてくたさ. (会議室 に 来て下さい)

この文には, 助詞 [に] が脱落され, また,[来て下さい] と言う部分にノイズが入っている. ここで,[会議室] が名詞であるから, その後ろに助詞が必要であることがわかる. その助詞を決めるために, システムはまず文の最後の部分を決める. この場合, 断片 [きてくたさ] に対する予測は次のように行なう.

(1) 構文的予測

A: <助動詞 ・ ϕ>
B: <動詞 ・ ϕ>
C: <形容詞 ・ ϕ>
D: その他

(2) 意味的予測

B1: (会議室に対する動作) 締める, 開ける, 行く, 来る, 予約する...
B2: (会議室の状態) 空いている, 締めている...
.....

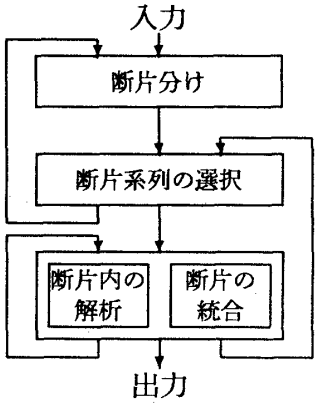

図 2: システムの処理過程の概要.

C1: (会議室の属性) 広い, 明るい, 暗い...
.....

システムは上の予測によって得られた単語集合を利用して, この断片から単語を検出する. ここで, 意味概念から関連する単語を取り出すために, 本研究では, 意味概念の関連を表すネットワーク [1] を利用する.

4 処理過程の概要

システムの処理過程の概要は図 2に示す.

与えられた入力文に対して, システムはまず断片分け処理を行ない, 次いで, 断片分けによって得られた断片系列の集合から一つ (確からしいもの) を選んで, これに対して, 断片内の解析と断片間の解析を行なう. 処理が失敗したら, また次の断片系列を選択し, 処理を続ける.

断片内の解析と断片間の解析の順序は処理過程で動的に決める. 解析の初期段階で, システムはボトムアップ的な解析方式を採用する. 即ち, 断片内の解析を断片間の解析より先に行なう. 一方, 処理の進行に伴って, 十分な予測情報があった場合, システムはその解析をトップダウン的な方式に変える.

5 おわりに

本稿では, ノイズによる入力誤りを効率的に訂正するために, 断片的な解析という方法を提案する. この方法では, ノイズのない断片を優先して解析することによって, 文の全体的な構造を早く把握することが可能, ノイズのある部分の訂正に有用な情報を提供できる. また, 入力誤りの特性に応じて, 各々に適当な対処方法を採用する. これによって, 単語候補の数の減少, 入力誤りを効率的に訂正できる一方, 無駄な処理を避けることができる.

参考文献

[1] 劉, 西田, 堂下: 統合パーサによる統合的自然言語解析, 情報処理学会論文誌, Vol. 31 No. 9, pp. 1293–1301, (1990).