

「講演など独話データの知的構造化に関する研究開発」(041307012)

A Study of Intellective Annotation Technology for Monologue

研究代表者

柏岡 秀紀 株式会社 国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
Hideki Kashioka ATR Spoken Language Communication Research Laboratories

研究分担者

松原 茂樹†
Shigeki Matsubara
名古屋大学 情報連携基盤センター†
Information Technology Center, Nagoya University

研究期間 平成 16 年度～平成 18 年度

本研究開発の概要

講演など独話に情報を付加し再利用可能な高度な知識源とし、容易にアクセスでき、要約や検索、翻訳などの応用技術が進めば、人の知的活動に有意義である。講演など独話の処理は、従来の対話などのインタラクティブな音声言語処理と異なり、処理単位、構造の把握が必要となる。講演等の音声言語データの収集により研究環境が整いつつある状況に鑑み、本研究課題では、まず、1)長く続いている発話の処理単位を明確にする。そのために意味的なまとまりを持つ意味処理単位を、同時通訳の訳出単位として分析する。2)WEB等によりネットワークからアクセスが容易な独話の構造を定式化する。また、その加工技術を開発する。

さらに、3)処理単位、構造の定式化により知的活動への働きかけを明確にするための応用技術の一つとして、要約機構の開発を行い、その有効性を検証する。

Abstract

In recent years, monologues are recorded into archives and have become research targets for speech processing and/or natural language processing. Monologues include lectures, TV news scripts, academic presentations, personal talks, and so on. We should develop the method to access these monologues data. However each utterance or sentence tends to be long and complicated in monologues. Then, it is necessary to define some short and suitable units instead of whole sentences for processing monologues.

1. まえがき

近年、研究会などで講演や解説など独話をデータとして、映像および利用されたスライドと共に蓄積されつつある。また、講演等を対象とした音声言語データの収集(開放的融合研究「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」など)が行われており、独話を対象とした研究環境が整いつつある。このように蓄積されたデータは、人間の知的資源として有効かつ効率的に活用できるようにすべきである。そのために、講演論文集などとの関連を記述できる表現形式を整え、有効な活用法を確認しつつ蓄積していく必要がある。収録された音声あるいは映像ファイルを単に蓄積しているだけでは、有効かつ効率的に講演データを活用することは困難である。現在、PDAなどの端末から高性能な処理能力を持つ端末まで様々な形態によりネットワークを介してデータにアクセスできるユビキタス環境が整いつつある。しかしながら、このようなユビキタス環境において、これらデータを各個人の利用環境に合わせて効率的にアクセスできるデータ加工技術の整備、アクセス法について十分な技術が確立しているとはいえない。蓄積されたデータを、ネットワークを通じて必要とときに容易に効率的にアクセスできるように加工するための技術開発は重要な課題である。

本研究課題では、そのような状況に鑑み、収録された講演等の独話の書き起こしテキストを対象に構造を付与し、関連する音声・映像・スライド等をひとまとまりのデータとして蓄積する技術を開発することにより、ネットワークを介した情報端末において容易にかつ効率的に活用するための基礎的枠組みを確立することを目的とする。その要

素技術として、

(1) 音声言語処理から見た基礎的な処理単位としての意味処理単位の確立と自動検出技術の開発、

(2) 独話内の意味処理単位間の統語的關係、意味処理単位に含まれる情報をより深く理解するための他の知的資源とのリンクを考慮した独話の構造情報の定式化と構造情報を付加するための機構の開発、

(3) 構造情報の利用効果を検証するため、独話の要約に着目し、圧縮率や注目事項などの個人の要求に応じた要約を実現する機構の開発、

を目標とする。

2. 研究内容及び成果

2.1 独話における意味処理単位の同定

独話に対して収録された同時通訳のデータを利用し、同時通訳者の発声の特性と原話者の発話内容との関係の分析を行うことにより、同時通訳者の処理単位と考えられる意味処理単位の分析を行った。その結果、節境界単位を意味処理単位が妥当であると判断した。節は、統語的な観点では、単文に相当する要素から構成されており、節単位での処理は、既存の自然言語処理ツールとの相性が良い。節境界単位は品詞の局所的な連鎖(節境界の直前の約3形態素及び直後の約2形態素の情報)パターンにより、節の種類、境界が判定可能であり、パターンマッチングを利用した節境界判定システムを構築した。各節は、その統語的な特徴から大分類5種(文末を含む)、小分類10種の階層的な139種の節タイプに整理した。さらに、節境界の強さが異なることが南らにより指摘されていることを考慮し、各

基準に節境界の強さを示す4段階の指標をつけた。

2.2 係り受け解析システムの開発

意味処理単位である節境界単位の情報を利用し、節内部での解析と節間の解析を分離することで、解析の効率を高めた高精度の処理機構を構築した(図1)。「あすを読む」の書き起こしデータにおいて、節境界単位を越えて節末文節以外が他の節境界単位に係るものは、10%未満であった。また、従来の係り受け解析では、主語が重複している場合に、遠い文節との係り受けを主として正解とした。たとえば、「私は、国際会議に投稿し、発表の中で議論を戦わせた。」という文では、「私は」は、「議論を戦わせた」に係ると解析される。実際には、「投稿し」とも係り受け関係があり、これらを考慮すると、ほとんどが節境界単位内部で係り受け関係が閉じているといえる。節境界単位内部の係り受け解析では、「節末の文節以外はその内部で係り受けが閉じている」という性質を利用した。

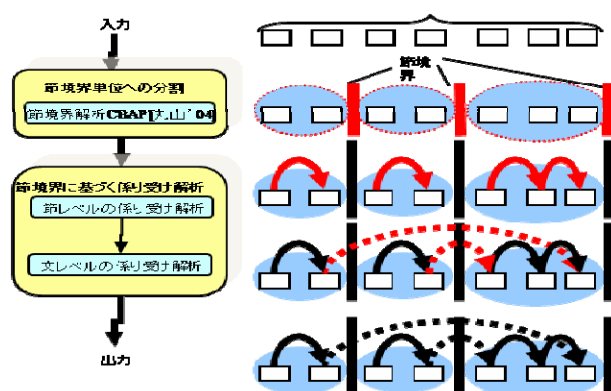


図1 漸進的解析機構の処理の流れ

係り受けの解析精度は、改善の余地があるが、基本となる解析手法は、統計的な最尤推定に基づくものであり学習データの整備、量の増加により精度向上を見込むことができる。さらに、上記の提案手法において、文末の係り先を決定することができれば、独話全体の構造を係り受け構造解析ができる。文末の係り先は、文末の表現をテンプレートに従った言い換えにより節末として処理できる。このように文を長くすることは、係り受け候補を増大させ、解析精度の低下および処理速度の遅延を招く恐れが強いが、本研究課題において提案した意味処理単位に分割して解析する手法では、係り受け候補の増大を防ぎ、漸進的な処理を考慮することで、文末の解析でも効率的に行うことが可能となる。処理速度について図2に示す。

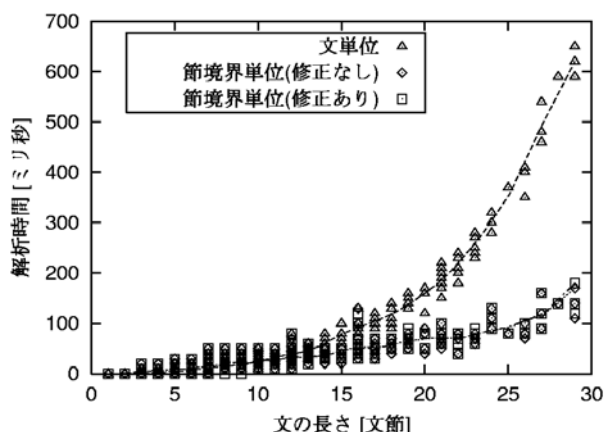


図2 提案手法及び文単位の処理における解析効率

2.3 独話の要約機構の開発

構造解析手法は、漸進的な解析機構となっている。この利点を生かし、順次解析される単位に応じて、部分的な文節を削除することによる要約手法を構築した。従来の用語の重要性による削除候補を決定するのではなく、意味処理単位の長さに応じて順次削除していく手法であり、構造的な情報を有効に活用した手法となっている。さらに、局所的な構造情報を利用する手法は、漸進的な処理が可能となることから、リアルタイム性を保持することが可能となり、リアルタイムな字幕生成の実現が期待されることから、漸進的処理における対象となる発話の時間長に対して出力文字数(要約結果)の比率を一定以下とする字幕生成の制約を用いたシステムを開発した。

3. むすび

本研究課題では、講演などの独話を有機的なネットワーク構造としてデータベース化し知識源として捉えること、および、「文」より短い処理単位を明確にすることに新規性・独創性がある。単に、収録したデータの蓄積ではなく、処理単位を明確化し、構造化することで、要約や翻訳等の技術を高度化するとともに、統合した処理も可能となるといった革新性を持つ。講演の処理技術の研究開発は、自然言語処理技術、音声処理技術を融合した研究分野を活性化させるという先導性を持つ。また、講演を録画した映像データを MPEG 形式で利用し、言語情報やその構造、他の知識源との関係を付加情報として与えることも考慮した。この点で国際水準における新規性がある。自然言語処理、音声処理を融合した研究分野である音声言語処理の分野に独話処理の分野を開くことができる。意味処理単位は翻訳の単位として位置づけることができ、単位判定ツールは独話の同時通訳システムの開発における基礎技術となる点でも、国際的に類を見ない研究開発といえる。

【誌上发表リスト】

- [1] 大野 誠寛、松原 茂樹、柏岡 秀紀、稲垣 康善、"節境界に基づく独話の漸進的係り受け解析"、電子情報通信学会誌, vol. J90-D, no. 2, pp. 556-566, (2007.2.)
- [2] Tomohiro Ohno, Shigeki Matsubara, Hideki Kashioka, Takehiko Maruyama, and Yasuyoshi Inagaki, "Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries", Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL2006), pp. 169-176, Jul. (2006.7.17).
- [3] Hideki Kashioka, "ANALYSIS FOR STRUCTURE OF MONOLOGUE", 4th Symposium on "Intelligent Media Integration for Social Information Infrastructure" (Nagoya) (2006.12.8)

【申請特許リスト】

- [1] 大野、松原、柏岡、加藤、係り受け構造解析装置およびコンピュータプログラム、日本、2005年1月25日
- [2] 大野、松原、柏岡、加藤、リアルタイム字幕生成装置およびコンピュータプログラム、日本、2006年3月2日
- [3] 柏岡、講演構造解析装置およびコンピュータプログラム、日本、2005年2月24日