

TR-I-0031

音声・言語インタフェースの予備検討

Preliminary Research on the
Interface between Speech Recognition and
Natural Language Processing

森元 逞 柿ヶ原康二 北 研二

Tsuyoshi MORIMOTO Kouzi KAKIGAHARA Kenzi KITA

1988.6

概要

音声と言語をどのような方式で結合するかは、自動翻訳電話システムを実現するために確立すべき最も重要な技術の1つである。特に問題となるのは、音声認識の結果として残存するあいまいさを、どのようにして絞り込むかである。

本レポートでは、音韻ラティス、文節ラティスから最も確からしい候補を求めるために、ラティスと辞書とのマッチングを行う方法、言語データベースから求めた統計量を用いる方法など、いくつかの方式案についての検討結果や問題点について述べている。

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

－ 目次 －

1. はじめに
2. 検討の概要
3. 辞書とのマッチングによる自立語の認定
4. 自立語と付属語、付属語同士の接続確率
5. 慣用表現
6. 格関係による述語の決定
7. 今後の検討課題

1. はじめに

音声と言語をどのような方式で結合するかは、自動翻訳電話システムを実現するために確立すべき最も重要な技術の1つである。我々は63年2月央～4月央の約2ヶ月間に、その予備的な検討を進めてきた。本資料はその内容について報告するものである。しかし比較的短期間であったことや、また関連技術の研究がまだ充分進んでないこともあり、我々が検討を行った内容は一部分に過ぎない。まだ多くの問題の解決や新たな技術の研究などが必要である。これらの残存問題については本文中でも言及しているが、最後のまとめで簡単に今後の研究方向を述べる。

2. 検討の概要

音韻ラティス、または文節ラティスから、もっともらしい候補を絞りこむには、言語情報、分野知識などの利用が必須である。このため、表1のような項目について検討を行うこととした。なお、本資料で検討を行った項目は下線のもののみである。

表1 検討項目・

レベル	(1)辞書	(2)構文	(3)慣用表現	(4)係り受け(格以外)	(5)格関係	(6)主題	(7)分野知識	(8)対話構造
(a)文節内(注1)	<u>1-(a)</u>	<u>2-(a)</u>						
(b)文内		2-(b)	<u>3-(b)</u>	4-(b)	<u>5-(b)</u>			
(c)文間(発話内)								
(d)発話間						6-(d)	7-(d)	8-(d)

(注1)自立語を1つのみ含むもの

以下、各項目毎にその狙いを述べる。(各項目の項番は表1の項番に対応している。)

(1)辞書

(a)音声認識の結果が音韻ラティスで与えられた場合、各音韻の組合せと辞書とをマッチングさせることにより、正しい単語を求める方法が考えられる。特に単語数が少ない場合、自立語の推定にはかなり有効であろうと思われる。

(2)構文

(a)自立語と付属語、付属語同士の接続条件

これらの接続関係には統計的差異が存在し、これを利用することにより付属語の認識率を高めることができると考えられる。

(b)日本語は語順がかなり自由であり、また会話文では省略が多いことから、あまり予測には使えないと思われる。従って今回検討では対象外とした。

(3)慣用表現(注1)

(b)会話文では、「有り難う/ございました」、「お願い/します」などのような慣用表現が多い。この場合、前の文節から後ろの文節の自立語を容易に認識できる。

(注1)2つ以上の文節にまたがったものとする

(4)係り受け(格関係以外)

(b)語と語の共起関係を用いることにより、「AのB」において、AからBを推定(またはその逆を)することが考えられる。しかしこれらの共起関係を現在の言語コーパスから求めるにはデータ量が少なすぎるため、今回の検討からは除外した。

(5)格関係

(b)述語とその格要素となる自立語が認定できている場合、各々の自立語がどの格に対応するのか、すなわちその格助詞が何であるかを、述語の内容と各自立語の意味素性の組合せ(結合価)から求めようというものである。逆に格要素が認定できている場合、それらから述語自身を推定することも考えられる。

ただし、文末に現れる意図表現などには対処できない。また応答文では以下の例のように、述語自体も省略されることが多いため、適用できない場合が多い。

ex. 「はい、そうです」、「いいえ、まだです」

(6)主題

(d)一般的に主題は以降の会話では省略されることが多いため、主題となる語自身を推定することはあまり意味がない(注1)。しかし、主題を管理しておくことにより、その副主題など、主題に関連する語を推定することは重要と思われる。ただし、これは分野知識による推定と同じであり、今回は検討を行っていない。

(注1)ただし、電話会話における質問/応答では主題自身のを反復する傾向があり、この部分では有効そうであるが、この問題は他のものに比較すると小さい問題である。

(7)分野知識

(d)各単語間(厳密には各単語の表す概念間)の関係を分野知識として定義しておく(注1)、この分野知識を利用することにより、文の命題内容を述べている部分の単語を予測しようとするものである。しかし、省略の多いような応答文には適用不可である。また、文末の意図表現にも対処できない。

(注1)これを「辞書」と呼ぶ方法もあるが、これらの関係は対象とする分野にも依存する部分も大きい。従って本資料ではこれをあえて「分野知識」と呼ぶことにする。

(8)対話構造

(d)我々の対象とする対話では、「内容(WH)質問→内容の回答」、「Yes-No質問→肯非応答」など、その流れがある程度パターン化されている場合が多い。この対話構造の性質を利用することにより、文末の意図表現(例えば質問しているのか、または自己の希望を述べているのか、など)を推定することが考えられる。ただし、これ単独ではなく、Litman&Allenのように上記の分野知識を利用した命題内容部分の推定と組み合わせて利用する必要がある。

以下では、これまで検討を行った項目について、その結果の概要を述べる。

3. 辞書とのマッチングによる自立語の認定(北)

(1)方法

- (a)音声収録された会話のスペクトログラムから、音声グループの研究者が目視により文節毎の音韻ラティスを作成
- (b)上記音韻ラティスの全ての組合せで辞書引きを行い、単語(自立語および付属語)を決定
- (c)辞書にはこれまで収集したキーボード会話に現れた単語を登録(単語数約2000語)
- (d)解析は以下の点を除き通常の状態素解析と同様のアルゴリズムを採用
- (e)候補が複数現れた場合は、文節を構成する単語数が最も少ないものを採用
- (f)1つの文節に2つ以上の自立語は現れない(従って複合語は1つの単語として登録しておく)ものとする

(2)結果

- A. 文節の総数: 129
- B. 解析不可能: 17
(単語が辞書に登録されていなかった、などの単純なミスが原因)
- C. 自立語を一意に決定できたもの: 92
- D. 自立語を一意に決定できなかったもの: 20
(形態素解析ができなかったものを含む)

$$\begin{aligned}\text{自立語を一意に決定できた割合} &= C/(A-B) \times 100 \\ &= 82\%\end{aligned}$$

一意に決定できなかったもの(項目D)のうち、以下にいくつかの例をしめす。

同音、類音: 私/渡し、良い/用意、原稿/銀行、予定/余計
形態素解析不可: お送りします(「多く」と一致)

また、付属語についての結果は以下の通りであり、決定できた割合はかなり低い。

- A. 付属語を含む文節数: 114
- B. 解析不可能: 15

C. 付属語を一意に決定できたもの : 28

ただし品詞の細分までは見ないものとする

例) 「と」における格助詞/並列助詞/準体助詞の違い

D. 付属語を一意に決定できなかったもの: 66

$$\begin{aligned}\text{付属語を一意に決定できた割合} &= C/(A-B) \times 100 \\ &= 28\%\end{aligned}$$

(3)評価

(a)今回対象としたデータは、音声グループの研究者がスペクトログラムから目視により作成した音韻ラティスであるため、音韻認識率自体がかなり良い値になっている。実際の音韻認識率はもう少し低くなると予想される。従って自立語の決定率もさらに低下する。

(b)今回の実験においても、同音語、類音語についてはほとんど決定できていない。これらを決定するには、さらに主題や分野知識などの上位レベルの情報を利用する必要がある。

(c)付属語の決定率は予想通りかなり低い。これらについては、例えば次節に述べるような方法など、別の対処が必要である。

4. 自立語と付属語、付属語同士の接続確率(柿ヶ原)

(1)方法

これまで収集した言語データベースから、以下の接続確率を求める。

- ・ 自立語(動詞、形容詞、副詞、名詞)→{付属語(助詞、助動詞)or文節区切り}
- ・ {助動詞or助詞}→{助動詞or助詞or文節区切り}

(2)結果

(a)以下に例を示すように、かなり統計的な有意差が得られた。これを付属語の認識率向上のために利用することができる。

接続される語			接続する語			接続確率

本動詞	5段	未然	助動詞	せる	連用:せる	8/47
				.		
				.		
		連用	助動詞	たい	連用:たい	2/630
					終止:たい	5/630
					連体:たく	18/630
		連用	助動詞	ます	未然:ませ	32/630
					連用:ます	106/630
					終止:ます	147/630
				.		
				.		

(b)以上の結果の評価のため、例題についてその文節生成能力を求めた。その一部を以下に示す。

かき	ます	か	↑	確率大
かく	そう	です	か	
かく	の	です	か	
かき	ましょう	か	↓	確率小
	.			
	.			

(3)評価

例題ではほぼ妥当な文節が出力されている。ただし、生成能力が高すぎる部分もあり、例えば以下のようなループを生成する場合もある。

かく	の	です	から	です	か
かき	ます	から	です	から	です

これを避けるにはどういう制約を付加すればよいかが今後の課題である。1つの案として、上記モデルがbigramに相当することから、trigramに拡張することなどが考えられる。

5. 慣用表現

(1) 方法

(a) これまで収集した言語データベースから、すべての単語の列について以下の評価関数を求める。

$$K(a) = (|a| - 1) \times n(a)$$

$|a|$: 単語列 a の長さ

$n(a)$: 単語列 a の出現頻度

ここで $K(a)$ は、 a を慣用表現として登録したことによる処理の削減量を意味する。

(b) $K(a)$ の大きいものを、慣用表現とする。

(2) 結果

(a) 以下に例を示すように、かなり妥当と思われるものが抽出された。

抽出されたイディオム	$K(a)$
でし・う・か	130
です・か	109
し・たい・の・です・が	88
の・です・が	84
分かり・ました	72
の・でし・う・か	72
.	
.	

(なお、本検討の詳細については[1]を参照されたい。)

(3) 評価

かなり妥当と思われるものが抽出できた。ただし、これを用いるとどの程度認識率の向上が図れるかについては、今後検討を行う必要がある。

6. 格関係による述語の決定(柿ヶ原)

(1) 方法

従来から提案している名詞と動詞の共起関係を利用し、文の述語(動詞、形容詞など)を推定する[2]、すなわち文の格要素である名詞は認定できたとして、こ

これらの名詞から述語を決定できないか、という問題について検討を進めた。具体的には各名詞を意味素性で表し、その意味素性の組みを結合価として持つ述語を抽出した。

(2)結果

以下に例を示すように、限定されたタスクでは、述語の候補をかなり絞れるという結果を得た。

名詞の意味素性	推定される述語
-----	-----
人間、場所/施設	利用する
行為、数/費用	要る、かかる、必要
具象、時間/変数(いつ)	出す、出る
行為、抽象/言語	運営する、行う、話す
.	.

(注)A/Bは意味素性がAで、かつそのサブカテゴリがBであることをしめす。

(3)評価

(a)1文内の名詞数が少ない場合や一般的な名詞である場合、必ずしも十分に述語の候補を絞りきれない場合がある。

名詞の意味素性	推定される述語
-----	-----
具体/文書	ある、受け付ける、送る、書く 出す、発行する、含む、持つ、 読む、無料

(b)対象とするタスクが少し広がると、候補となる述語数がかなり多くなることが予想される。

(c)上記の問題を解決するために、結合化パターンに統計量を取り入れる方法も考えられる。しかし、このような静的な情報のみの利用には限界があると予想される。話題や対話構造なども利用する必要があるだろう。

7. まとめと今後の検討予定

(1) まとめ

以上報告したように、各項目ともさらに検討を詳細化する必要を残しているが、全体として以下のようにまとめることができる。

(a)音声認識結果から正しいもの(単語など)を認定する方法として、言語データベースを基にした言語の統計モデルを利用することが、一部のものの認定には有効である。

(b)ただし、最終的な効果は音声認識の認識率に大きく左右されるため、実際の音声認識との連動をなるべく早く行い、評価する必要がある。

(c)また、現在の検討は各項目ごとに独立して行っているが、実際のシステムを実現するには、それらの相互関係を整理し、統合する必要がある。

(d)さらに、本報告でしめた各方式では対処できないもの(文末表現や省略語の多い応答文など)については本文中でも述べたが、分野知識や対話構造を利用することを考える必要がある。

(2) 今後の予定

(a)今後は上記(b)~(d)の研究を重点的に行うことが必要である。このため、本報告の検討結果や、さらに分野知識、対話構造なども積極的に利用するシステムを具体的に作成することにより、これらの研究を実証的に進めて行くこととしたい。

(b)また、今回報告した各方式の精度をさらに向上させるためには、今後ともさらに多くの言語データベースを収集してゆく必要がある。

【参考文献】

- [1] 北、森元 「テキスト・データベースからの慣用表現の自動抽出」、
ATR Technical Report, TR-1-0027
- [2] 柿ヶ原、相沢 「自立語からの対話文生成」、第36回情報処理全国大会