

係り受け関係データから見たキーボード会話と電話会話の比較

6F-3

井ノ上 直己 江原 暉将 小倉 健太郎

ATR自動翻訳電話研究所

1. はじめに

現在我々は、自動翻訳電話の実現のための基礎研究を進めているが、そのためには会話文特有の言語現象を明らかにする必要がある。そこで、当研究所では会話文の持つ言語現象を明らかにする上で必要となる各種の基礎データを収集した大規模な言語データベースの構築を目指している[1][2][3][4]。

本報告では、ドメインを国際会議に限定し、模擬実験により収集した電話会話文およびキーボード会話文ののべ約4万語に対して得られた各種の統計データを示し、電話会話とキーボード会話の比較検討を行う。

2. 基礎データ

当研究所で作成している言語データベースには以下の3種類の情報が付加されており、これらのデータは一般性をもたせるために特殊な言語理論に依存しないという前提のもとで収集しており、さらに、これらのデータは均質に付与されるように品質の管理が行われている。[5]

(1) 単語に関する情報

(a) 会話中に現れた単語、(b) 単語の読み、(c) 標準表現、(d) 品詞、(e) 活用型、(f) 活用形、(g) 音便

(2) 単語間の関係

(a) 格、係り受け関係*、(b) 接続関係

(3) 言語間の関係

(a) 文、(b) 格構造、(c) 文節、(d) 単語、(e) その他

* ここで、係り受け関係とは「AのB」という表現で代表されるような名詞句内修飾関係、1文内における命題間の関係等を含んでいる。

また、上記の格、係り受け関係は、文脈を考慮しながら1文内で意味的に係り得る単語と単語との間で付与しており、これによりその文の意味構造を抽出できるように定義した。

3. 統計情報

本報告では、前節の各種情報の内単語間の関係に関するデータ、格、係り受け関係から得られた統計データを示し、電話会話とキーボード会話の比較を行う。

表1 形態素数および文数

メディア	のべ形態素数	のべ文数
キーボード会話	19,225(16,549)	1230
電話会話	21,837(16,345)	844

カッコ内の数字はシンボルを除いたものである

各種の統計情報を抽出するために利用した形態素数および文数は表1に示す通りである。また、1文中の平均形態素数は表2に示す通りである。

表2 1文中の平均形態素数

メディア	平均形態素数
キーボード会話	15.6(13.5)
電話会話	25.9(19.4)

カッコ内の数字はシンボルを除いた場合

表2より、電話会話文では1文内の語数(形態素数)がキーボード会話に比べて多くなっている。その理由としては電話会話では「あの」、「あー」や「えーと」といった間投表現が多くみられることもその理由として考えられるが、表3に示すよう1文内に含まれる動詞の数も多く、キーボード会話に比べて冗長な表現が頻繁に行われることも1つの理由と考えられる。なお、表3内には文数に対する累積分布も同時に示している。

一方、表4に1つの動詞に対して出現した係り受け関係数と動詞の数の関係を示す。

表4より電話会話では、キーボード会話に比べて意味関係が全くない動詞が頻繁に現れることが分かる。このことは、意味関係の定義から1文内の意味構造を表すのに不必要な動詞が、電話会話では頻繁に発話されていることを表している。このような動詞として、以下の例文に見られるような「…という…」という表現の「言う」、「…に関して…」の「関する」、「…に対して…」の「対する」という表現等が見られた。

例文

1人はたぶん今週末には、[あの]承諾できるというふうに聞いているんです。

また、表4には意味的に関係する語が現れなかった動詞を除いて平均した値を示した。これから、キーボード会話および電話会話ではほぼ等しく、約2個の意味的に関係する語が1つの動詞に対して出現するのが分かる。

表3 1文内の動詞の数と文数

1文内の動詞の数	キーボード会話	電話会話
1	427(53.7%)	274(42.8%)
2	224(81.9)	152(66.7)
3	106(95.2)	101(82.5)
4	26(98.5)	55(91.1)
5	8(99.5)	24(94.8)
6	2(99.7)	21(98.1)
7	2(100.0)	8(99.4)
8	-	1(99.5)
9	-	1(99.7)
10	-	1(99.8)
11	-	-
12	-	1(100.0)

カッコ内の数字は文数に対する累積分布

表4 係り受け関係数と動詞数

関係数	キーボード会話	電話会話
0	58	675
1	455	275
2	520	281
3	237	121
4	76	68
5	14	15
6	2	6
7	1	1
関係数0の場合を除いた平均値	1.99	2.07

以上より、電話会話のような全く文字化が行われない会話では、語調を整えるためあるいは発話

者が考えるための時間的な余裕を得るため(例えば「あの」、「えーと」等)に冗長な表現が行われるが、1つの動詞に対して実際に意味的に関係する語の数は、メディアに依存せず会話文ではほぼ同じであるといえる。従って、電話会話のような文字化の行われない会話に対して自然言語処理を行う場合、格、係り受けのような意味的な関係を利用すれば、冗長な表現の中から意味的に関係する部分だけをとり出して処理できる可能性がある。

4. おわりに

電話会話とキーボード会話との比較については既に、談話構造を明らかにする立場から比較が行われているが、今回は1文内に限った格、係り受け関係に注目し比較を行った。その結果、

(1)電話会話では1文内の形態素数がキーボード会話に比べて多く、かなり冗長な表現がなされている。

(2)電話会話では意味的に関係する語がない動詞の出現が非常に多い。

(3)意味的に関係する語がない動詞を除けば、1つの動詞に対して意味的に関係する語の数はメディアには依存せず、電話会話とキーボード会話とではほぼ同じであり、平均2個程度である。

ことが分かった。

このことは、電話会話のような文字化の行われない会話に対して自然言語処理を行う場合、表層の表現にとらわれるよりは、格、係り受け関係のような意味的な関係を利用すれば、キーボード会話の場合と同様に扱える可能性があることを示唆している。

謝辞

本研究の機会を与えてくださるとともに適切な助言を述べられたATR自動翻訳電話研究所 榎松明 社長、森元 逞データ処理研究室長に感謝します。また、熱心に議論して下さいデータ処理研究室諸氏に感謝する。

<参考文献>

[1]森元、小倉、飯田：「自動翻訳電話研究用データベースの収集について」、情処学会第36回全国大会4U-5(1988)

[2]江原、小倉、森元：「電話対話データベースの構築」、情処学会第40回全国大会(1989)

[3]篠崎、小倉、森元：「言語データベース作成のためのシミュレーション会話」、情処学会第37回全国大会5B-8(1988)

[4]井ノ上、小倉、森元「言語データベース用単語間の関係データ」、情処学会第37回全国大会5B-7(1988)

[5]篠崎、小倉、森元「言語データベース品質管理」、情処学会第36回全国大会4U-3(1988)

[6]有田、小暮、野垣内、前田、飯田「メディアに依存する会話の様式-電話会話とキーボード会話の比較-」、情処学会NL研究会61-5(1987)