6T-6

音声対話システムのための 簡略表現の生成規則と出現頻度の調査

秋田谷 樹 † 駒谷 和範 † 佐藤 理史 † 中野 幹生 ‡

† 名古屋大学大学院 工学研究科 ‡ ホンダ・リサーチ・インスティチュート・ジャパン (HRI-JP)

1. はじめに

音声対話システムでは，ユーザが音声認識辞書の単語を異なる表現で発話した場合，正しい音声認識結果が得られない．特にユーザは長い名称から一部を省略して発話することが多く，この現象はシステムに不慣れなユーザが利用する場合に顕著に現れる．この問題に対して，発話されうる簡略語を予想し，音声認識辞書を拡張する必要がある．

本研究では，簡略語を自動的に獲得し辞書を拡張する．手順を図 1 に示した．まず，音声認識辞書中の単語から，規則に従って文字列を抽出し，簡略語の候補を生成する．ここでは，カバー率を重視して生成する．ここで生成される候補数は，意味のない文字列を多数含むため，語彙サイズの増大にともなう音声認識率の低下が問題となる．したがって，次に，生成した候補のうち，意味のない文字列をフィルタリングにより除去し，残った候補を簡略語として辞書中へ追加する．

これまでに，簡略語を生成する規則を適用し，Web 上での出現頻度や生起確率を用いて候補を絞り込む研究 [1] や，辞書中の単語から簡略語の候補を自動生成し，音の類似度を利用したスコアリングにより適切な確率を付与する研究 [2] などが行われてきた．文献 [2] では，生成する文字列は単語の組み合わせに限定しているため，ドメインによっては十分なカバー率を得られないことがある．本研究では，54 名の被験者から新たに収集したデータを用いて生成規則を整理し，大規模コーパス中の出現頻度を利用して候補を絞り込む．

本稿では，まず人手で収集した，簡略化される前の単語 (以降,「元の単語」と表記) と簡略後の組の例に対し，簡略語の生成規則を列挙して，規則毎に分類する．次に，大規模コーパス中の出現を調査することによる候補の絞り込みを提案し，実験により簡略語の候補数や収集したデータに対するカバー率を評価する．

2. 簡略語の生成規則

ここでは，文献 [3] に基づき，簡略語が生成されるときの規則性について述べる．簡略語は長い名詞が存在したとき，発話者が時間の短縮や労力の軽減を目的として用いるものである．したがって音の長さが短くなるほど理想的だと考えられるが，聞き手が元の単語を推測できなければ意思疎通が図れない．そのため，ほとんどの簡略語は 2 モーラ以上の長さを持つ．ここでモーラとは音の長さの単位であり，仮名 1 文字が 1 モーラに対応するが，”ふぁ”などの小文字を含む場合も 1 モーラと数える．日本語における多くの簡略語は 2〜4 モーラとなる傾向がある．

元の単語からモーラを選ぶ方法にも規則性がある．元の単語が単一名詞の場合には，語頭や語尾から連続する n モーラを選択する規則が多い．元の単語が複合名詞の場合には，構成する単語のうちの一つを選んで用いる場合や，単語を 2 つ選んで語頭 2 モーラずつを選択するものがよく見られる．また，”フルオーケストラ”に対する”フルオケ”など，特殊なモーラ (っ，ん，ー) が出現したとき，それらのを省略して次の 1 モーラを抽出する例や，”世帯分離”に対する”世分”など，漢字の切れ目を優先的に選択する例が存在する．

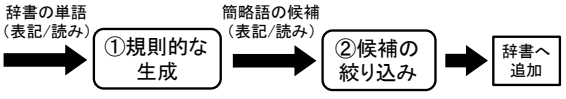

図 1: 簡略語の獲得手順

Investigating in Generation Rule and Occurrence Frequency of the Abbreviated Words for Spoken Dialogue Systems: *Tatsuki Akitaya, Kazunori Komatani, Satoshi Sato (Nagoya Univ.) and Mikio Nakano (Honda Research Institute Japan Co., Ltd.)*

3. 簡略語の生成規則の調査

本研究では収集した元の単語と簡略語の組に対して，どの規則によって簡略語が生成されたかを分類し，調査した．

我々は 54 人の被験者から，施設名・店名・場所に関する 403 種類の元の単語と簡略語の例を収集した．データ収集時には「マック」や「ファミマ」のような，よく使われるもの以外を挙げるよう教示した．調査では，元の単語が名詞である 339 例を選択し，以下の手順で生成規則を分類した．

1. 形態素解析器により元の単語が単一名詞か複合名詞かを判定.
2. 単一名詞に対しては，モーラの抽出箇所とモーラ数で分類.
3. 複合名詞に対しては，簡略語生成時に選択される単語の数で分類．その後，選択された単語数が 1 と 2 の場合に対して，モーラの抽出箇所とモーラ数で分類.

形態素解析器には KyTea を使用した．KyTea では，単語境界確率を用いた確率的な単語分割と，各単語の発音と品詞の推定を行う．集計結果を表 1，2 に示す．表 1 は元の単語が単一名詞である 37 例，表 2 は複合名詞

表 1: 単一名詞の分類結果

抽出箇所	モーラ数	出現数	割合	例		
語頭	2	3	8 %	ドナドナ	⇒	ドナ
	3	12	32 %	サイゼリア	⇒	サイゼ
	4	3	8 %	ラクアルオダサガ	⇒	ラクアル
語末	3	9	24 %	コトブキヤ	⇒	ブキヤ
	4	3	8 %	トランペット	⇒	ペット
	5	1	3 %	ハウステンボス	⇒	テンボス
その他	4	1	3 %	クアラルン プール	⇒	クアラン
複合名詞的規則		5	14 %	イタリアントマト	⇒	イタトマ
合計		37	100 %			

表 2: 複合名詞の分類結果

選択単語数	抽出箇所	モーラ数	出現数	割合	例
1単語	単語全て	-	84	28 %	モス/バーガー ⇒ モス
	語頭	4	9	3 %	キリマンジャロ/山 ⇒ キリマン
	語末	2	3	1 %	高田馬場/駅 ⇒ 馬場
2単語	単語全て	-	18	6 %	多田/小/学校 ⇒ 多田小
	語頭+語頭	2+1	17	6 %	ロイヤル/ホスト ⇒ ロイホ
		2+2	139	46 %	ゲーム/センター ⇒ ゲーセン
		その他	6	2 %	東武/動物/公園 ⇒ 東武公
	その他	-	16	5 %	自動/車/学校 ⇒ 車校
3単語以上			10	3 %	東京/工業/大学 ⇒ 東工大
合計			302	100 %	

表 3: 使用される頻度の高い生成規則

単一名詞	語頭 n モーラの抽出 $(n = 2, 3, 4)$
	語末 n モーラの抽出 $(n = 2, 3, 4)$
複合名詞	1 単語の選択
	2 単語の選択
	語頭 2 モーラと語頭 1 モーラの抽出
	語頭 2 モーラと語頭 2 モーラの抽出

表 4: 候補生成の結果

生成した候補数	正解数	カバー率 [%]
3205	279	82

である 302 例に対する分類結果である．表 1 の複合名詞的規則に分類したのは，KyTea では単一名詞として判定されたが，本来複合名詞であるものである．

なお，特殊モーラや漢字の切れ目による生成が起こりうる場合には，出現数がより多い規則に分類した．例えば”イメージフォーラム”に対する”イメフォ”の例では，単純には語頭 2 モーラと語頭 1 モーラの組み合わせだが，後半の語頭 2 モーラに長音の省略が起きているとすれば，語頭 2 モーラずつの組み合わせに分類できる．表 1 の複合名詞的規則に分類されたものに対しては，他の形態素解析器との組み合わせや，KyTea が出力する各文字間における単語分割の信頼度の利用に基づく，単語分割精度の向上を今後検討する．表 2 の，3 単語以上選択されているものは，10 例中 7 例が学校名であり，高等学校を高校，女子という単語が出現する場合は女のみに短縮する例が多い．このように，学校名とそれ以外の名詞で，生成規則の傾向に違いが見られる．

表 3 の規則を実装した．この規則は，表 1，2 において，使用される頻度が高かった 10 種類の規則である．収集した元の単語を使って候補生成を行った結果を表 4 に示す．カバー率は (1) 式のように定義し，正否の判定は，元の単語の表記が一致し，かつ簡略語の表記あるいは読みが一致する場合に正解とした．

$$\text{カバー率} = \frac{\text{正解単語数}}{\text{収集した簡略語の数}} \quad (1)$$

表 4 より，使用頻度の高い規則を適用することで，339 例の元の単語に対し，3205 例の簡略語の候補を生成した結果，約 82 %のカバー率を実現できることを確認した．

4. 大規模コーパスを用いた候補の絞り込み

大規模コーパス中に出現しない単語は，使用される頻度が低い単語であるとし，コーパス中の単語の出現を調べることで，意味のない簡略語候補の除外を試みる．絞り込みの手順を以下に示す．

1. 簡略語の表記に一致する文字列を含む文を抽出．ただし，簡略語が元の単語の部分文字系列である場合には，元の単語の表記を含む文には，常にその簡略語が含まれるため，その文は除外．
2. 簡略語の表記と，抽出した文をそれぞれ形態素解析し，変換後の文字列が合致する文章を検出．
3. 簡略語の候補から，検出されなかったものを除外．

表 5: 絞り込みの結果

絞り込み後の候補数	正解数	カバー率 [%]
2142	235	69

この手法では，簡略語の表記のみを判定している．したがって同じ表記で別の単語を指す場合でも，使用されていると判定する．”エクセルシオールカフェ”を”エクセル”と簡略化する例がこれに相当する．この例では，”エクセル”という文字列はソフトウェア名として一般的な文字列であるが，出現する分野が異なるためどちらも利用されている．以上のことから，同じ単語であるかを調査せずに，使用されない候補を判定することが可能である．このように，実際には意味の異なる単語の場合もあるが，出現しない場合には使われない表現である，という推定は可能である．

実験により，候補の絞り込みの有用性を検証した．実験で使用したコーパスは，Yahoo 知恵袋データ (第 2 版) の全データである質問のデータ 1600 万件と回答のデータ 5000 万件データである．

絞り込みの結果を表 5 に示す．表 4 と比較すると，候補の絞り込みにより候補数が約 2/3 に減少している．一方，カバー率は 13 ポイント低下した．このカバー率の減少分には，あまり一般的でないと考えられる簡略語，例えば「学ホ (元の単語:学生ホール)」や「北マ (元の単語:北里マート)」が含まれていた．以上より，コーパスを用いた候補の絞り込みによって，実際には使用されない簡略語を除外できる可能性があると言える．

5. おわりに

本稿では，収集した単一名詞，複合名詞に対し，8 割以上を網羅する規則を明らかにした．また，大規模コーパス中の単語の出現を調査することで，コーパスを用いた候補の絞り込みによって，使用されない単語を棄却できる可能性を示した．今後は，KyTea で分割できなかった複合名詞を判定する方法の検討や，音声データをに対する音声認識による評価を行う．

参考文献

[1] 榎将功，皇甫美華，大田健紘，柳田益造．日本語における略語自動生成法の検討とその音声インタフェースへの応用. 情報処理学会研究報告. 2007-SLP-69, pp. 313—318 (2007)
[2] 森信介, 駒谷和範, 勝丸真樹, 尾形哲也, 奥乃博. 音声対話システムにおける簡略表現認識のための自動語彙拡張. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3398—3407 (2011)
[3] 窪薗晴夫．新語はこうして作られる (もっと知りたい!日本語) 岩波書店 (2002)