機械学習による対立語を用いた漫才台本の自動生成

11371001　青木 哲（灘本研究室）

あらまし：本研究では，これまで研究してきた漫才の自動生成におけるつかみの改善を提案する．具体的には，ニュースのカテゴリの対立語を抽出し，その対立語に関連する人物を抽出する．これにより得られた人名を用いて新たなボケである取り違えボケの提案を行う．

1. はじめに

近年，動画サイトやマイクロブログなど，インターネットを利用した様々なコンテンツが増加し，メディアの多様化が進行している．それと対照的に，テレビやラジオなど，従来から存在するコンテンツは衰退の傾向がある．

一方，近年様々なロボットが人の生活に急速に浸透してきている．しかしながら，人とロボットとのコミュニケーションは未だ円滑とは言いがたい．神田ら[1]は，ロボット同士の対話観察を人が行うことで，ロボットに対して他の人間に対するように自然なコミュニケーションが可能になることを心理学的実験を通して実証している．そこで我々は，テレビなどが持つ受動的に情報が得られるという利点がロボットと人の円滑なコミュニケーションの補完になる点及び，娯楽性の高い漫才に着目し，漫才ロボットの研究をしてきた[2][3]．

これまで提案してきた漫才台本のつかみは漫才の本ネタと関係のない季節の挨拶のみを行っている．そのため，ニュース本文に基づく対話を生成している本ネタとの間で話が急激に変化し，ニュースの主題が明確になりにくいという問題が発生している．そこで本論文では，つかみの特徴である漫才の導入部であり最初のひと笑いである点に着目し，そのニュースが属するカテゴリ(以下主題カテゴリ)と，それに対する対立語を用いたオチによるつかみを提案する．本論文ではこのボケを「取り違えボケ」と呼ぶ．

対立語の抽出にはニューラルネットワークを用いた機械学習の手法の一つである Word2Vec[4] が生成する単語ベクトルを用いる．

2. 提案手法

2.1 取り違えボケの生成

本論文で新たに提案する取り違えボケは，まずボケが主題カテゴリについて述べる．主題カテゴリはニュースサイトのディレクトリ構造を用いて取得する．次にツッコミがそれに対して誰が有名かと質問をする．それに対しボケは取り違えた対立語の関連人物について述べる．本論文では有名な人物であればそのカテゴリに関する知識があまりない人でも理解できると考え，人名を用いる．

ボケ	最近テニスについて勉強しとんねん
ツッコミ	ほぉ～，どんな人が有名？
ボケ	せやな，例えば本田圭佑とかかな
ツッコミ	なんでやねん，それはテニスやなくてサッカーや
ツッコミ	テニスなら錦織圭とかが有名や
ボケ	そんなんどっちでもええやろ
ツッコミ	怒られるで
ツッコミ	ところで，テニスと言えば，地球でこんな話題あったの知っているか？

図 1 生成例

表 1 対立語の抽出例

基の語	先輩	ラーメン	夜
対立語	後輩	蕎麦	昼

例えば，主題カテゴリが「テニス」，対立語が「サッカー」の場合，サッカーに関連する人名として「本田圭佑」を取得する．そして主題カテゴリであるテニスに関連する人名として「錦織圭」を取得し，これらを用いる．生成例を図 1 に示す．

この時，主題カテゴリの人名はニュースの主題とする．ただし，抽出した主題語がフルネームではなかった場合，ニュース記事の本文からフルネームを取得する．

2.2 対立語の決定

Word2Vec の特徴の一つに「似た使われ方をする語は似たベクトルを生成する」というものがある．そこで本論文では，「似た意味を持つ単語は似た使われ方をする」と定義し，Word2Vec を用いて対立語の抽出を行う．

教師データには Twitter より無作為に取得したツイートを用いる．ここで Twitter を用いる理由として，まず，大量のテキストデータを得られる点が挙げられる．これは Word2Vec に代表されるニューラルネットワークによる機械学習における特性で，大量の教師データが必要となるためである．次にリアルタイムな話し言葉の情報を得られる点が挙げられる．これは，書籍や Wikipedia の記事などのデータを教師データ

表 2 実験結果

アンケートの選択肢	主題が明確か				対話は自然か				漫才が面白いか			
	既存		提案		既存		提案		既存		提案	
言える	4	67%	5	83%	1	17%	4	67%	0	0%	0	0%
どちらかといえば言える	1	17%	1	17%	3	50%	1	17%	2	33%	3	50%
どちらとも言えない	1	17%	0	0%	1	17%	1	17%	4	67%	3	50%
どちらかといえば言えない	0	0%	0	0%	1	17%	0	0%	0	0%	0	0%
言えない	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%

として用いた場合，書き言葉で書かれているため，最終的に話し言葉として生成を行う本研究に適合しない.

以上より教師データをツイート群とし，Word2Vec による学習を行い求めた単語ベクトルにおいて基となる語のベクトルに最も類似しているベクトルを持つ語を対立語として抽出する. ここで，教師データには 281,412,044 ツイートを用いた.抽出された対立語の例を表 1 に示す.

2.3 関連語の決定

対立語の人名の抽出手法には,Wikipedia のカテゴリを用いる．まず対立語と同じタイトルの Wikipedia の記事の下にある全ての記事を取得する．次に事前に作成した人名リストを用いて取得した Wikipedia の記事から人名の記事を抽出する．最後に，取得したページから認知度が最も高い人名を取得する．認知度は，Web の検索結果数を用いる.

3. 評価実験

提案手法の有用性を示すため，評価実験を行った．実験は 20 代の男女 6 人に対し行った．実験は既存手法の季節の挨拶と表情ボケによるつかみの漫才と，提案手法の取り違えボケを用いたつかみの漫才の 2 つを被験者に見せた.「主題が明確か」「対話は自然か」「漫才が面白いか」という 3 つの設問に対し，「言える」「どちらかといえば言える」「どちらとも言えない」「どちらかといえば言えない」「言えない」の 5 段階で評価を行った. 表 2 にその結果を示す. 表 2 は，それぞれの回答数と，その数が設問に対して占める割合を示す.

4. 考察

「主題は明確か」という設問に対しての回答は，既存手法は「どちらとも言えない」を選んだ被験者が 17%いたのに対し提案手法はこの項目を選んだ被験者がいなかった. また,「言える」と答えた被験者が 67%から 83%に増加したため，ある程度改善したと言える.「対話は自然か」という設問に対しての回答は，既存手法は半数が「どちらかといえば言える」という回答の被験者に対し，提案手法は半数以上が「言える」を回答した．さらに「どちらかといえば言えない」と「言えない」を選んだ被験者はいなかったため改善したといえる.「漫才は面白いか」という設問に対しての回答は,「どちらかといえば言える」の回答が 33%から 50%に増加したため，ある程度改善したと言える．このように，提案手法を用いることにより，主題が明確になったことがわかる.

5. まとめと今後の課題

本論文では，つかみに用いる新たなボケであり，主題を明確にする事ができるボケである取り違えボケの提案と，それに対する評価実験を行った．これにより漫才のつかみにより主題が明確になり，対話がより自然になったことを確認した.

今後の課題は，主題語が人名以外であった場合の生成手法の提案が挙げられる．提案手法ではニュースの主題語が人名であることが多い為，主題カテゴリの人名は主題語としている．しかし，主題語が人名ではない場合，提案手法のままでは生成が不可能となる為，この場合に適応可能な改善手法が必要となる.

また，対立語の抽出手法に対する評価実験を行うことも今後の課題である.

参考資料

[1] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu.，” Development and evaluation of an interactive humanoid robot robovie “. IEEE, ICRA 2002, pp. 1848-1855,2002.

[2] 真下遼, 梅谷智弘, 北村達也, 灘本明代, “Web ニュースからの漫才台本自動生成を用いたコミュニケーションロボット" Web DB Forum 2014.

[3] 真下遼, 灘本明代, “対立語抽出に基づく Web ニュースからの漫才ロボット台本自動生成手法の提案", DEIM Forum 2014C2-4

[4]Word2Vec, https://code.google.com/p/word2vec/