自由対話実現のための自動文章生成モデルの提案

冨坂亮太† 鈴木崇史‡ 相澤彰子†‡
東京大学大学院情報理工学系研究科† 国立情報学研究所‡

1 はじめに

本研究では、ユーザーを楽しませるための対話システムの構築を目標とする。この目標を達成するために、本論文では、文書から知識を構築する方法を提案し、また、それを用いてユーザーの興味を引けるような用語を文章から抽出する方法を提案する。

ユーザーの興味を引く用語を抽出するために、本論文では、システムに特に興味がある分野を設定できるようにした。興味がある単語を、このシステムと会話するユーザーの興味と近いものに設定することにより、ユーザーが面白いと思える話題を提供することができ、ユーザーを楽しませることができると考える。また、Twitterからの最新の情報から学習することにより、システムが最新の話題の用語に強く反応できるようにした。これにより、システムが最近話題になっているような話題を提供できるようになり、よりユーザーの興味を引けるようなシステムになっている。

2 関連研究

自由対話を実現するシステムの構築の研究として、柴田ら[1]のような研究があげられる。これはWebからユーザーの発言の応答として合っている文を探しだし、そのままユーザーに返すことにより、ユーザーとの対話を実現するシステムである。また、意味ネットワークを用いて、自由な対話システムを評価しようと論文として、曽我部ら[2]のような研究がある。また、ニュース記事からユーザーが未知な情報を提供することによりユーザーが興味を持つ会話文を生成するシステムとして[3]のような研究もある。水野ら[3]の研究は特に本論文と関連が深いが、本論文では、ユーザーに未知な情報でユーザーの興味を引くのではなく、ユーザーの興味と近い話題により興味を引くという点で違う。

また、古くから人工無能(Chatbot)という会話システムが考えられている。これは、過去の会話文やニュース記事などからユーザーの発話への応答として適切なものを、マルコフ連鎖などを用いて作成する方法である。これらの研究では、いろいろな人間が書いた文章から文を生成するため、システムの発言が2転3転し人格が曖昧になり人間らしさが欠如すると言う問題がある。しかし、本研究の手法を組み入れれば、システムの興味のある分野を設定で切るため、会話に一貫性を持たせることが可能になると考えられる。

用語の抽出の研究も多数のものがある[4]が、本論文は、新語に対応し、用語の区切れを自動で判別し、ある特定の分野の用語に大きなスコアを割り当てることができるという点で新しいといえる。

3 提案手法

3.1 知識の構築法

まず、本論文での知識の構築法について説明する。Twitter(もしくはYahoo!検索結果のSummary)の文章を文字単位に分割し、ある文字が現れた後に現れた文字とその回数をカウントしていく。たとえば、「今日は晴れだ。」という文章が現れたとき、「今」の次に「日」、「日」の次に「は」といったように記録していく。また、5回以上カウントされたものについては、次回からはその二つを塊として扱う。例えば上記の例のような場合、「今日」という単語を過去に5回以上*見ているならば、「今日」の次に「は」が来ると記録する。

こうすることで、単語と単語のつながりが自然と学習され、また新語を自動的に学習していく事ができる。この方法だと「今日は」といったものもまとめて扱ってしまうが、文理解という観点において、これらをまとめて扱うのは問題がないと考えている。また、後述の方法により、「今日」と「は」を離して考えるべきなら、システムが離して考えると期待できる。

われわれは、このシステムでtwitterのpublic timelineの一ヶ月程度のデータを使って知識を構築し、また、Yahoo!検索で「コンピューター」という単語での検索結果1000件のSummary部で学習させ、コンピューター関連の単語をより強く記録させるようにした。

A model of automatic sentence making for free conversation with users
† Gradual School of Information Science and Technology,The University of Tokyo
‡ National Institute of Informatics

*トライアルエラーの結果、5回が最も直感的に優れた結果を与えたため

3.2 システムを用いた用語抽出法

次に上記のシステムを用いて用語を抜き出す方法を説明する。文字列 c1, c2, ..., c{t-1}, c{t}, c{t+1}に対して、c{t}の出現回数をn{t}としたとき、n{t}/n{t-1}をp{t}とする。さらに、p{t}/p{t+1}を計算し、それをスコアとする。

例えば「googleは便利だ。」といった文章から「google」という語を抜き出す場合を考えた場合、上記のシステムを見た場合「googl」の次ぎにくるのは「e」が多いが、「google」の次ぎにくるのは「の」や「が」などいろいろなものが考えられる。そのため、「googl」の後に「e」のカウント数の比(p{t})は1に近い値となる。しかし「google」の後の「は」のカウント数の比(p{t+1})は小さな値になると考えられる。そのため、p2{t}の値は大きくなり「google」という単語を抜き出すことができる。

ただし、「オンデマンド」といったものを考えた場合「オン」の後にる来るものは「オンライン」から「ラ」が多かったりする。これを防ぐために、「オン」のカウント数(c2)から、「オンラ」のカウント数を引いてやるといった処理が必要である。

4 実験

上記のシステムを用いてYahoo!NewsのRSSから実際に用語を抽出してみた。その例を表1に示す。

表を見ると、GoogleやAndroidなどコンピューター関連の用語に強く反応してくれていることが読み取れる。また、新語のNexus Oneにもきちんと対応できていることが興味深い。

抜き出された用語	カウント数の比(これが大きい用語に強く反応している)
Google	2482.12
自社	0.18
ブランド	4.68
Android	33461.83
携帯	2
Nexus One	188.79
」発売	2.31
ITmedia	4093.33
プライ	30.21
ズ	3.83E-04

表1. 「Google、自社ブランドAndroid携帯「Nexus One」発売（ITmedia エンタープライズ）」という文章から用語を抽出してみた結果

5 おわりに

本論文では、自由対話を実現するシステムを構築するための、新しい知識の構築法方を提案し、さらに、その方法により文章から用語を抽出する方法を提案した。
また、そのシステムでの用語抽出での実験で新語や、興味がある分野の用語をうまく抽出できることを確認した。

ただし、実験のなかでプライとズが分かれてしまったように、このシステムで文を綺麗に分割しようとするとまだ解決しなければいけない問題がある。ここでプライとズが分かれてしまったのは、プライズのカウント数が少ないため、サプライと言う単語のせいでシステムがプライで切れると勘違いしてしまったせいである。これはカタカナ語や英単語で多く見られる。ただし、ダウンロードなどがダウンで切れずにうまく抜き出せていたりするので、もっと多くのデータで学習させてさらにプライズのカウント数が増えればこういった問題は自然と解決されるのではないかと考えている。

この方法により、文中でシステムがより興味を持っているもの、また逆に興味がうすいもの知らないものと分けることができる。今後の研究ではこの方法でユーザーの発言を解析し、その中の用語を核として文を生成できるシステムの構築法を考えていきたい

参考文献

[1]柴田雅博，冨浦洋一，西口友美．雑談自由対話を実現するためのWWW上の文書からの妥当な候補文選択手法，人工知能学会論文誌 Vol.24, No.6, pp.507-519 [2009]
[2]曽我部将義，鳥海不二夫，石井健一郎．非タスク指向型対話システムの評価法(コーパス，学習，対話，要約)，IPSJ SIG Notes, 2005(117), pp.105-110 [2005]
[3]水野涼太，乾健太郎，松本裕治．ウェブニュースを利用した雑談対話システム言語・音声理解と対話処理研究 55, pp.1-6 [2009]
[4]北研二.確率的言語モデル,東京大学出版会[1999]