

文ベクトルを用いた文脈考慮型人工無能

理工学部 数理情報学科

T100052 中野 友貴

指導教員 佐野 彰

概要

人工無能とは、ユーザーの入力文に対して、人工無能のもつ辞書データやアルゴリズムによって文を生成し、返答するコンピュータプログラムのことである。チャットボットとも呼ばれる。大きく分けて辞書型とマルコフ連鎖型の2種類がある。本研究では比較的簡単に作成でき、また返答文が予想できない楽しさを重視し、文脈考慮型のマルコフ連鎖型人工無能を作成した。生成される文を日本語文として成立させやすくするため、連鎖の最小単位は形態素とした。また、形態素解析にはフリーソフトの MeCab を用いた。

まず、形態素を単位とする単純なマルコフ連鎖型的人工無能を作成し、様々な文章から形態素間の遷移確率を求め、それをもとに文を生成させた。文語の小説をもとに生成した結果、文語をうまく形態素解析できていないため、意味の通らない文が生成された。一方、ニュース記事のような口語文を用いることで、意味の通る文を作ることができた。この単純な人工無能の作成によりマルコフ連鎖で文を生成する際には、もととなる文章の性質が重要であることを確認した。

既存の人工無能では、文脈に基づいて会話ができるものは少ない。そこで、文脈を考慮して返答ができる人工無能の作成を行った。文脈を捉えるための最小単位は文であるが、文の意味を直接数値化することは難しい。本研究では文に含まれる語がその文の意味を代表していると考え、文に出現する語による文のベクトル表現を行った。さらに、このように作成された文ベクトルを過去の n 文に対して重ね合わせることによって、 n 次の文ベクトルを生成した。このような n 次の文ベクトルは過去 n 文での文脈を表現すると考えることができる。この n 次の文ベクトルから、最も寄与率の高い語を抜き出し、この語をもとに返答文を生成することで、文脈を考慮した人工無能を実現した。

マルコフ連鎖で文を生成するときに、重要となるのが連鎖の開始文字である。文ベクトルをもとに決定した重要な語を開始文字に指定してマルコフ連鎖を開始すると、画一的な文になり、面白みに欠ける。そこで重要な語から前後に連鎖させることで、開始文字を固定することなく文生成ができるようにした。これにより、重要な語を任意の位置に入れることができ、返答文の幅が広がった。

ニュース記事をもとに作成した人工無能で文を生成すると、小説をもとにしたものに比べて会話らしい短い文が生成できる。しかし、ニュースには様々な話題があるために、1 文の中に複数の話題が含まれてしまう。そこで会話文をもとに文を生成することで、さらに短く意味の通った返答文を生成することができた。さらに会話らしくさせるためには、確率データの充実や、未知語の学習機能やネット検索機能などをつけるといったことが課題として挙げられる。