Sequence to Sequence Model と Attention Mechanism による

自然言語からの意味抽出

1. 概要

　本研究は、自然言語の意味理解がコンピューターにとって困難であることの解決を目的としている。自然言語は曖昧な意味や語順を含んでおり、それがコンピューターによる自然言語の意味理解を困難にしている。そこでこの研究では、Recurrent Neural Network による Encoder Decoder Modelの一種であるSequence to Sequence Model およびAttention Mechanismの適用によって自然言語からの大まかな意味の抽出がどれほどの精度で可能なのかどうかを調査した。具体的には、自然言語（英語）で書かれた文章と、それに対応する形式言語で書かれた意味のリストを学習させ、未知の文章に対する出力の精度を調査した。その結果、ある程度の文法規則を学習させることは可能となったが、データセットの大きさに精度が強く依存することもわかった。

2. 研究の動機

　近年、ニュースなどで”人工知能・AI”というワードをよく聞くようになった。

実際、2012年のDeep Learningによる大きなブレークスルー以降は、画像認識・音声認識などの分野においてかなり大きな発展が見られる。顔認識や音声入力などがわかりやすい例である。

　しかし自然言語処理・理解の分野においては、この分野を揺るがすような、一般人でも実感できるような成果はあまりない。そこで私は、近年機械翻訳などの研究に使われているSequence-to-Sequence Model(以後Seq2Seq) と、それの欠点を補う Attention Mechanism(以後Attention) を用いて自然言語からのプログラム、すなわち形式言語を生成する方法を提案する。簡単に言えば、自然言語のもつ曖昧性を消して、シンプルに意味の部分だけを取り出すことを可能にする、ということである。これにより、Apple社のSiriやGoogle社のGoogle Assistantに代表されるAIアシスタントの質問に対する回答の精度を上げることが可能となる。また先に述べたように、既存の成果があまりない理由について考察することもこの研究の目的の一つとする。

3. 研究の方法

　文章（英語）のリストとそれに対応する一定の規則のある言語で書かれた意味のリストを、Attentionを適応させたSeq2Seq に学習させ、未知の入力に対してどれほどの精度で意味を取り出せるかを調べる。

学習に使用した文章に含まれる単語は以下の通りである。また、それらの単語や文章はすべて小文字アルファベット・数字・空白・ピリオドのみで構成されている。

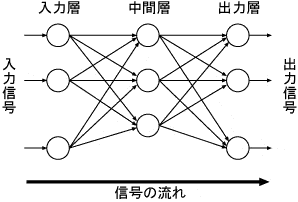
　曜日（sundayなど), 数字, time, is, today, tomorrow, tell, the, day, can, days, if, after, like, sunny, cloudy, rainy, later, be, how, weather, now, will, it, then, me, what, adorable, friend, nothing, schedule, look, everything, good, great, clear, am, go, shopping, hello

　人間同様、コンピューターも未知の単語については扱うことができないので、入力する文章はすべて上記の単語のみで構成されている。

4. 用語解説:

ニューラルネットワーク:

　脳はニューロンがシナプスで接続されたネットワーク構造をしており、学習によるシナプスの結合強度の変化により問題解決能力を持つ。それらの脳機能のいくつかの特性をコンピューター上で模した数学モデルである。



この図では、それぞれの○がニューロンを表しており、それらは自分以外のニューロンとつながっている。

それぞれのニューロンは、他のニューロンからの信号の強さにより興奮、すなわち他のニューロンへの信号の伝搬をする。

図に示した入力層、中間層、出力層という構造をしているニューラルネットワークは、階層型ニューラルネットワークと呼ばれる。これは後述のディープラーニングのベースとなる技術である。

\*1:ニューラルネットワークの模式図

ディープラーニング:

　階層型ニューラルネットワークの中間層が複数個存在するものを指す。これにより性能の向上が期待されることが昔から言われていたが、当時はコンピュータの性能不足や局所解への収束が問題となっていた。それが近年の研究成果やコンピューターの発達、インターネットの普及による訓練データを準備することが容易となり、2012年ごろから広く使われるようになった。

リカレントニューラルネットワーク:

　ニューラルネットワークで時系列データの予測を可能にしたモデル。あるニューラルネットワークの出力が次のニューラルネットワークの中間層として使用されることで過去の情報の保持を可能としている。

Sequence to Sequenceモデル:

　リカレントニューラルネットワークモデルの一種。任意個の入力からの任意個の出力を可能としているため、機械翻訳などに用いられる。

Attention Mechanism:

　Sequence to Sequence モデルではどんな長さの入力も固定長ベクトルに圧縮されてしまうため、大きな入力に対する学習の精度が下がりがちである。そのためAttentionでは、入力と出力の部分ごとの対応を学習させることを可能とし、大きな入力に対する学習精度を単なるSequence to Sequence モデルよりも向上させている。

6. 結果

　学習させたデータセット内の文章からある程度の法則性を見出し、未知の入力に対して反応することが可能となった。

以下の表に、

* いくつかの認識させた文章
* それが正しく認識されたかどうか
* データセットに含まれるか

を示す。

|  |  |  |
| --- | --- | --- |
| 文章 | 正しく認識された？ | データセットに含まれる？ |
| how will the weather be | Yes | Yes |
| what time is it | Yes | Yes |
| 2 days later | Yes | Yes |
| 8 days later | Yes | No |
| 7 days after tomorrow | Yes | Yes |
| after 7 days | Yes | No |
| If it is clear tell me plan today | Yes | No |
| If it is cloudy tomorrow | Yes | No |
| Add to the schedule that | Yes | No |
| tell me how the weather will be | Yes | No |
| Can you tell me how the weather is | No | No |
| If it is rainy the day after tomorrow then add to my schedule that ~~ | No | No |

7. 考察

　結果より、データセットに存在するような文章には完璧に回答できていた。また、データセット内に含まれる特定のテーマの文章なら、少々誤った文法を使った口語体のような入力に対しても、高い確率で正しい出力が可能となった。

　正しく認識されなかった入力のほとんどは、正しく認識された文章を部分的に含んでいた。このことから、コンピューターが文章の意味が、文章のどの部分に集中しているかを十分に学習できていないことがわかる。

　例として、

‘Can you tell me how the weather is’ と ‘tell me how the weather will be’ は ‘tell me how the weather’ が共通しており、意味の大半がこの部分に集中している。

しかし今回、コンピューターはこのことを学習することはできなかった。

原因としては、データセット内の文章と形式言語で書かれた意味のリストの数が少なすぎたということがある。文章と形式言語で書かれた意味のリストを、手作業で作りあげなければならなかったことが主な原因である。この方法では大量の文章を学習させるには、かなりの労力が必要である。また、人間は意識的に意味のリストを与えてやらなくとも、言語の学習と理解が可能であり、そのことは、根本的な研究の方法の転換および見直必要性を示唆している。

8. 参考文献

\*1: 村上・泉田研究室 ニューラルネットワーク

<http://ipr20.cs.ehime-u.ac.jp/column/neural/image/layer.gif>