

ニューラルネットによるプログラム学習のための手順生成と 日本語学習段階への適応

Procedure Generation for Program Learning by using Neural net and correspondence to Japanese Learning Stage

高橋 明義^{*1}, 椎名 広光^{*2}
Akiyoshi TAKAHASHI^{*1}, Hiromitsu SHIINA^{*2}

^{*1}岡山理科大学大学院総合情報研究科

^{*1}Graduate School of Informatics, Okayama University of Science

^{*2}岡山理科大学

^{*2} Okayama University of Science

Email: i18im03ta@ous.jp

あらまし: プログラムの手順を学習するシステムを構築するには、ソースコードから手順を自動的に生成することが必要となっている。本研究では、ソースコードから手順を生成するために、LSTM を用いてソースコードとそれに対応するコメントを学習することにより、新たなプログラムに対応するコメントを生成する。また、様々な難度の日本語文章を学習させることにより、使用者の日本語習熟度に応じたコメントを生成する。

キーワード: アルゴリズム学習, 手順生成, ニューラルネットワーク, LSTM

1. はじめに

近年、情報化社会が発展している中で小・中学校及び高等学校の授業でプログラミング関連の内容の授業を行うことが必修化される。この背景から高校生プログラミングの学習をより多く取り組むようになると考えられる。

プログラム学習のループリック⁽¹⁾が提案されているが、その中でもプログラムの理解と記述の段階が重要になってくると考えられる。特に、プログラムを学ぶ段階でのプログラムの理解を補助する仕組みが必要と考えられ、プログラムの各行の意味や、if 文, for 文, 複数文のブロック単位での意味を理解する必要があり、その補助としてコメントを自動で付けること等が必要である。

一方、ニューラルネットワークの機械学習を利用した日本語や英語などの翻訳の学習が可能である。RNN(Recurrent Neural Network)のモデルから長期依存を残しつつ学習を行うことを可能にしたLSTM(Long Short Term Memory, 以下 LSTM)⁽²⁾が開発されている。

本研究では、手順生成として、ソースプログラムのプログラム部分とコメントを分離した後、プログラム部分とコメント部分に対訳データとして LSTM を用いて学習を行い、次にコメントを学習した LSTM を用いて新規のコメント無しソースプログラムからコメントへ生成を行っている。また、留学生などの日本語の学習状況を考え、やさしい日本語への変換も行っており、手順生成とやさしい日本語変換の適用個所についても調査を行った。なお、本研究では、大学の情報系学科 1 年生での講義に使われた例や課題のプログラム 53 件を利用している。

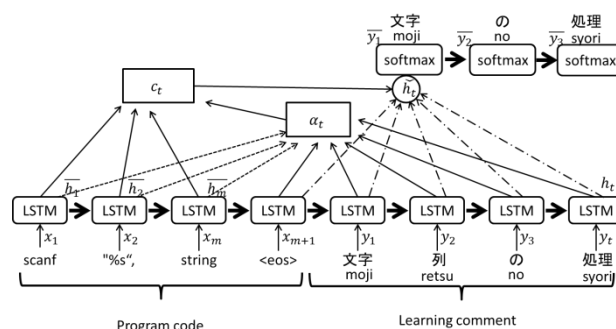


図 1 LSTM によるコメントの学習

2. 手順生成システム

プログラムコードの手順生成システムは、3つ変換処理を組み合わせている。

2.1 LSTM によるソースプログラムからの手順生成

(1) コメントの付されたプログラムからコメントを抽出しソースプログラムとコメントに分割し、それぞれを単語に分割する。

(2) ソースプログラムとコメントをそれぞれ単語単位に分割し、1行ごとに単語のつながりの関係性を学習する。単語に分割したソースプログラムを1単語ずつ LSTM に入力し、文脈情報を蓄積する。その後コメントを1単語ずつ LSTM に入力し、文脈情報を蓄積しながら単語を生成して学習を行う。入力単語 x_i に対しては、中間層の出力 $\tilde{h}_i$ を保持しておき、出力単語 y_t と x_i の類似度を正規化し α_t とする。この α_t と $\tilde{h}_i$ を使い文脈ベクトル c_t を作り、線形作用素 W で重みをつけて活性化関数 $\tanh$ により y_t に対する中間層の出力 $\tilde{h}_t$ を作る。中間層の出力 $\tilde{h}_t$ に線形作用素

Wで重みをつけ,softmax関数により $\hat{y}_t$ を作る.

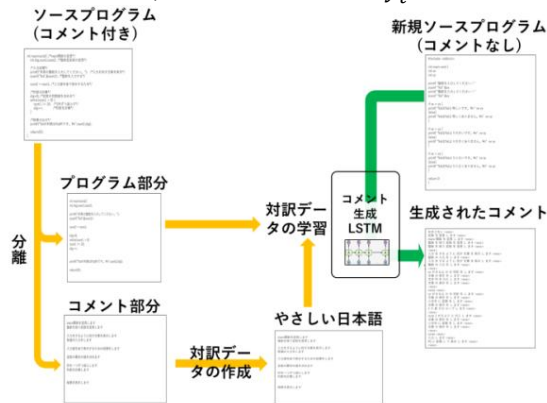


図2 コメント変換とやさしい日本語変換の組合せ

(3) (2)で行った学習をもとに新たなプログラムに対するコメントを生成する。プログラムを単語ごとに分割して LSTMに入力していき、最後まで入力を行った後に文脈情報をもとにコメントの1単語目を出力し、出力された単語と文脈情報をもとに次の単語を生成する。

2.2 やさしい日本語への変換

やさしい日本語への変換は、文献(3)に加え、プログラムのコメント変換用に合わせたやさしい日本語変換ルールに従って変換を行う。やさしい日本語への変換ルールを次に示す。

(1)難しい言葉をわかりやすく変換する。例えば、「最大値」は「いちばん大きい値」。専門用語や固有表現は変換しない。また、1文の情報を少なくする。場合によっては2文に分ける。

(2)学年別漢字配当表の難易度を利用して、難易度の高い漢字はひらがなに置き換える。

(3)文末を、ですます調に統一する。

2.3 手順の生成とやさしい日本語変換の組合せ

手順の生成とやさしい日本語変換の組合せについては、やさしい日本語を適用する箇所による組み合わせがある。本稿では、比較的结果が良い組み合わせについて述べる (図2)。

(1)学習時に、コメント付きソースプログラムをプログラム部分とコメント部分に分割し、プログラムの1行ごとのコメントの対訳データを生成する。

(2)コメント部分を、2.2のやさしい日本語への変換を利用してやさしい日本語のコメントに変換する。また、プログラムの1行ごととやさしい日本語のコメントの対訳データとする。

(3)プログラムの1行ごととやさしい日本語のコメントの対訳データに対して、2.1のLSTMによるソースプログラムからの手順生成の学習を行う。

3. 生成された手順の評価

コードとコメントの対訳を学習した場合とコードとやさしい日本語に変換したコメントの対訳を学習した場合の2つの手順の生成例を表1に示す。

表1 コメントの生成例(代表的な例を選択)

行	プログラムコード	コードとコメントの対訳を学習した場合	コードとやさしい日本語に変換したコメントの対訳を学習した場合
1	#define MAX 100	最大 値 を 定義 づける	一番 大きい 値 を 決めます
2	int main void {	main 関数 の 宣言	main 関数 を 宣言 します
3	char str1[MAX] = { ¥0 }	文字 列 型 変数 を 宣言 する	文字 列 を 使う 変数 を 宣言 します
4	scanf ("%s" str1)	文字 列 の 入力 を 行う	する の 文字 列 を 代入 します
5	scanf ("%d",&vx)	整数 の 入力	整数 の 入力 を します
6	printf ("整数を入力")	入力 を 促す 文章 を 表示	入力 を する よう に うながす 文章 を 表示 します
7	If (vx == vy){	0 以上 か どうか	vx が 0 以上 か の 判定 を します
8	printf ("%d は %d と 等しいです。 ¥n" vx vy)	文章 を 表示	文章 の 表示 を します
9	}else{	の 場合 の 結果 を 表示	文字 列 を 代入 します
10	printf ("%d は %d と 等しくありません。 ¥n" vx vy)	文章 を 表示	文章 の 表示 を します

表1の1行目は、生成したコメントが良好な例である。元コメントを学習した場合では、「最大値」と出力されていた部分が「一番大きい値」と分かりやすくなっている。しかし、9行目はコメント生成自体がうまくいっていない。これは、対訳データの条件分岐 else に対するコメントが多様なため収束していないと考えられる。

4. 今後の課題

通常のコメント生成とやさしい日本語のコメント生成の改善としては、対訳学習データの改善や構造情報の追加などの方法を考えている。また、attentionを導入した encoder decoder モデルを利用しておりモデルの改良を行いたい。

参考文献

- (1) 松永賢次, 萩谷昌己, 共通教科情報科ルーブリックにおける思考・判断・表現の位置づけ, 第10回全国高等学校情報研究会発表会全国大会(東京大会), (2017)
- (2) Greff, K. et al., "LSTM: A Search Space Odyssey", IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, Issue10, pp. 2222-2232(2017)
- (3) 弘前大学人文学部, 社会言語学研究室, <http://human.cc.hirosaki-u.ac.jp/kokugo/index.html>