RNN言語モデルを用いた平叙文完成問題の自動解法

喜多智也　　平 博順
大阪工業大学 情報科学部
e1b14036@st.oit.ac.jp, hirotoshi.taira@oit.ac.jp

1 はじめに

我々は，大学入試センター試験で高得点を取り、東大入試に合格する人工知能を開発することを目標とする「ロボットは東大に入れるか」(東ロボ) プロジェクト [1] に参加している．本稿では，大学入試センター試験の英語科目における短文問題のうち，平叙文完成問題についての自動解法手法について述べる．

平叙文完成問題とは，以下のように与えられた文中の空欄に最もふさわしい選択肢を選ぶ問題である．

問題文: I had a severe toothache, so I made ☐ with the dentist.
選択肢: (1) a promise (2) a reservation (3) an appointment (4) an arrangement

この例では，歯医者の「予約」として最もふさわしい (3) の選択肢が正解となる．

近年の大学入試センター試験の英語科目では，大問が 6 問あり，前半 3 問が短文問題，後半 3 問が長文問題となっている．平叙文完成問題は前半の短文問題の 1 つであり，大問 2 において 10 問程度の小問が毎回出題され，文法・語法に関する問題であり，取りこぼしをなるべく避けたい問題となっている．

これまで，東ロボプロジェクトにおいて平叙文完成問題の自動解法手法として試みた手法としては，N-gram 言語モデルを用いる方法や VPOS 法による方法などがある [7]．N-gram 言語モデルを用いる方法は，問題の空欄に各選択肢を当てはめて文を完成させた場合のそれぞれの単語 N-gram の生成確率が最も高くなるような選択肢を正解として出力する方法である．Verb Part-of-Speech (VPOS) 法は，N-gram 言語モデルを改良した方法で，単語の N-gram に加え動詞の have, 完了形を作る have などを区別するため，動詞の詳細な品詞タグも擬似単語として単語列に含め N-gram を考慮する手法である [7].

N-gram 言語モデルや VPOS 法では，N-gram の N の値を大きくすればするほど，適切な選択肢を選択する精度が向上することが期待されるが，同時に計算量が急速に増加する．我々の実験では，5-gram での言語モデルの場合，約 7 割の正解精度が得られているが，それ以上の長距離の文脈を考慮しないと解けない問題も多く存在している．

本研究では，より長距離の文脈を考慮するため，Recurrent Neural Network (RNN) に基づく言語モデルにより単語列の生成確率を求めて正解の選択肢を出力する方法について提案する．実際のセンター試験の過去問を用いて評価したところ，RNN 言語モデルの利用により，N-gram 言語モデルを用いた場合に比べて計算量が大きく増大することなく大幅に正解率が向上した．

2 RNNを用いた平叙文完成問題の自動解法

文 S が $|S|$ 個の単語列 $w_1, w_2, \cdots, w_{|S|-1}, w_{|S|}$ であるとき，文 S 生成確率 $P(S)$ は，単語 w_i の生成確率の積として

$$P(S) = \prod_{i=1}^{|S|+1} P(w_i|w_0, \cdots, w_{i-1})$$

と表すことができる．ここで w_0 は文頭記号を，$w_{|S|+1}$ は文末記号と表すとする．このとき通常の N-gram 言語モデルは，単語 w_i の生成確率を

$$P(w_i|w_0, \cdots, w_{i-1}) \approx P(w_i|w_{i-N}, w_{i-N+1}, \cdots, w_{i-2}, w_{i-1})$$

と近似するモデルである．つまり，文章中の単語 w_i の生成確率を N 個前の単語から直前の単語までの生成確率の積として近似する．

一方 RNN 言語モデルは，各単語 w_i の生成確率 $P(w_i|w_{i-N}, w_{i-N+1}, \cdots, w_{i-2}, w_{i-1})$ を RNN を用いて求めるモデルである．複数の RNN 言語モデルを組み合わせることで通常の N-gram 言語モデルよりパープレキシティを大幅に減少させることが分かっている [4]．この RNN 言語モデルでは，原理的には単語 w_i より前に表れた全ての単語情報を使って生成確率を求めることができる．

我々は Chelba らの研究 [2] で高いパープレキシティを実現している Jozefowicz らのモデル [3] を用いて RNN 言語モデルを構築した．LSTM の内部構造を図 1 に，Jozefowicz らのモデル構造を図 2 に示す．2 層の LSTM の文字ベースの CNN を用いていることがこのモデルの特徴である．なお LSTM の隠れ層と Projection 層の次元は Chelba らの研究と同じくそれぞれ 8192 次元と 1024 次元とした．

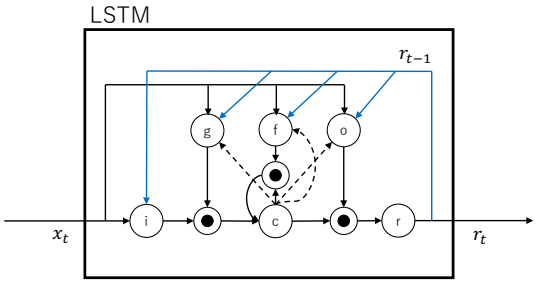

図 1: LSTM の内部構造

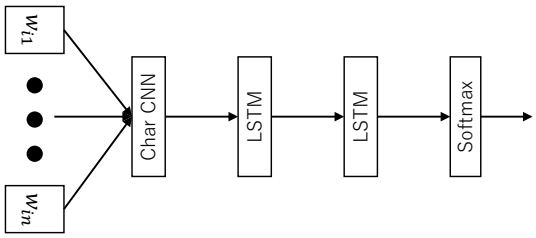

図 2: 言語モデル構築に用いた RNN の構造

3 評価実験

3.1 実験設定

実験では，2005 年から 2013 年までの奇数年のセンター試験の問題 5 年分，計 51 問を用いて評価した．各年の問題数は 2007 年のみが 11 問で他の年は 10 問であった．N-gram, RNN 言語モデルの学習には One Billion Word Benchmark コーパス [2] を用いた．コーパスの規模を表 1 に示す．また N-gram 言語モデルの構築には SRILM(ver 1.7.1)[5] を，VPOS 法での品詞タグ付けには Stanford Log-linear Part-Of-Speech Tagger[6](ver 3.7.0) を用いた．

表 1: コーパスの規模

単語数	793,471
文数	30,301,028

3.2 実験結果

N-gram 言語モデル，VPOS 法，RNN に基づく言語モデルそれぞれを使って評価用データ 51 問で評価した．マクネマー検定を行ったところ 1%の有意水準で RNN 言語モデルによる手法が VPOS 法より有意に優れていることがわかった．正解率を表 2 に示す．

表 2: 実験結果

手法	正解率
N-gram	67% (34/51)
VPOS	76% (39/51)
RNN	96% (49/51)

また RNN 言語モデルと VPOS 法における正解/不正解数の内訳を表 3 に示す．

表 3: 実験結果

	VPOS が正解	VPOS が不正解
RNN が正解	39	10
RNN が不正解	0	2

表 3 からわかるように，本評価実験では VPOS 法で正解だった問題はすべて RNN 言語モデルでも正解であった．

4 実験結果の分析

改善が見られた問題と本稿の手法でも正解できなかった問題について解析を行う．

4.1 改善が見られた問題

VPOS 法と RNN 共に正解だった 34 問と RNN のみが正解だった 10 問について傾向を調べた．それぞれ

の問題群の単語数の平均を表 4 に示す．これを見ると単語数が多い問題について改善が見られたと分かる．そこで単語数が多い問題についてさらに分析を行った．

表 4: 単語数の平均

	単語数
共に正解	16.9
RNN のみが正解	19.8

4.1.1 回答の根拠が選択肢の前方にある問題

2013 年度センター試験大問 2A 問 4 の問題文とシステム出力を表 5 に示す．この問題の単語数は空欄の単語も含め 22 であった．

表 5: 2013 年度センター試験 2A-4

問題文: Small children have teeth which usually fall out between the ages of five and twelve, after which they get their ☐ teeth.

選択肢		
	(1)	false
	(2)	forever
VPOS	(3)	general
RNN/正解	(4)	permanent

この問題の前半部分では「小さな子供は 5 歳から 12 歳の間に通常歯が抜け落ちる」と書かれている．後半部分で「その後に生えてくる歯」をどのような単語で表現するのが適切かが問われている．前半部分で述べられている歯は「乳歯」を指しているため,「その後に生えてくる歯」は「永久歯」を指しているため，(4) の permanent teeth が正解となる．

後半部分の”after which they get their ☐ teeth.”の空欄に各選択肢を当てはめた文の対数尤度をそれぞれの手法で計算した結果を表 6 に示す．

どちらの手法も後半部分の文だけでは選択肢 (1) の対数尤度が最も大きい結果となっている．しかし計算の対象を問題文全体とすると，RNN 言語モデルでは選択肢 (4) の対数尤度が最も大きい結果になる．これは RNN 言語モデルは VPOS 法より長期の文脈を考慮していると考えられる．

表 6: 2013 年度センター試験大問 2A 問 4

選択肢		RNN	VPOS
(1)	false	-45.6	-23.1
(2)	forever	-54.9	-28.5
(3)	general	-52.7	-26.1
(4)	permanent	-49.8	-24.4

4.2 エラー分析

本稿の手法で正解を用いてもきなかった 2 問について詳細に分析を行った．

4.2.1 細かな表現の区別が不十分なケース

2013 年度センター試験大問 2A 問 2 の問題文と，システム出力を表 7 に示す．

表 7: 2013 年度センター試験大問 2A 問 2

問題文: Of the seven people here now, one is from China, three are from the US, and ☐ from France.

選択肢		
	(1)	other
	(2)	others
RNN	(3)	the other
正解	(4)	the others

この問題では，ここにいる 7 人のうち中国出身が 1 人，アメリカ出身が 3 人，残りがフランス出身だと述べており,「残り」を英語でどのように表現するかが問われている．システムの出力を見ると，正解である「残り全員」を表す”the others”と選択肢 (3) の”the other”の区別ができていないことがわかる．”the others”と”the other”は表現も意味も非常に似ているが,「残り全部」の表すには”the others”を答える必要がある．正答するためには人数を数える必要があるが，現状そのようなモデリングはできていない．

4.2.2 可算/不可算の区別が不十分なケース

2009 年度センター試験大問 2A 問 5 の問題文とシステムの出力を表 8 に示す．この問題では,「ネクタイについた卵」をどのような英語で表現するかが問われ

ている．調理後の卵は不可算名詞として扱われるため選択肢 (3) が正解である．システムでは「調理済みかどうか」という概念を捉えていないため，正解できなかったと考えられる．これは文の流暢さを判断するだけの言語モデルでは正解できない問題であると考えられる．

表 8: 2009 年度センター試験大問 2A 問 5

問題文: You've got ☐ on your tie. Did you have fried eggs for breakfast?

	選択肢	
	(1)	a few eggs
RNN	(2)	an egg
正解	(3)	some egg
	(4)	some eggs

5 おわりに

本稿では，大学入試センター試験の英語科目における短文問題のうち，平叙文完成問題についての自動解法手法について述べた．N-gram 言語モデルを元にした手法では長距離の文脈を考慮するには計算量の増大が原因で問題があり，本研究では RNN 言語モデルによる手法を試みた．その結果，VPOS 法で正解だった問題はすべて RNN 言語モデルで正解できセンター試験の過去 5 年分の問題の正解率が大幅に向上した．回答の根拠と選択肢が離れている問題について分析を行ったところ，RNN 言語モデルによる手法は長距離の文脈を考慮していると考えられた．

また，エラー分析を行ったところ，細かな表現や可算/不可算の区別が不十分であることがわかった．今後は，RNN 言語モデルに加えて細かな表現の区別を行うために文で述べられている状況や可算/不可算の区別を行うために物体の状態などをモデルに取り入れることを検討したい．

謝辞

本研究を遂行するにあたり「『ロボットは東大に入れるか』大学入試センター試験関連オンラインタスクデータ」を利用しました．ご提供下さった「独立法人大学入試センター」および「株式会社ジェイシー教育研究所」に感謝いたします.「ロボットは東大に入れるか」を推進している新井紀子教授をはじめ，国立情報学研究所の方々に深く感謝いたします.
また，本研究の一部は以下の各氏（組織）との共同研究として行われました．東中竜一郎, 杉山弘晃, 成松 宏美 (以上 NTT)，菊井玄一郎， 磯崎秀樹（以上岡山県立大），堂坂浩二（秋田県立大），南泰浩（電気通信大）．熱心な議論に感謝いたします．

参考文献

[1] Noriko H Arai and Takuya Matsuzaki. The impact of ai on education–can a robot get into the university of tokyo? In *Proc. ICCE*, pp. 1034–1042, 2014.

[2] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[3] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[4] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, Vol. 2, p. 3, 2010.

[5] Andreas Stolcke, et al. Srilm-an extensible language modeling toolkit. In *Interspeech*, Vol. 2002, p. 2002, 2002.

[6] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180. Association for Computational Linguistics, 2003.

[7] 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩, 大阪工業大学. センター試験における英語問題の回答手法. 言語処理学会第 21 回年次大会 (NLP2015), 2015.