論述構造との同時予測による論述的な意見生成

阿部 智彦　　井之上 直也
株式会社 Nextremer　東北大学大学院情報科学研究科
tomohiko.abe@nextremer.com, naoya-i@ecei.tohoku.ac.jp

1 はじめに

近年、自然言語処理を用いて議論を解析する研究（議論マイニング; Argument Mining [1]）が盛んに行われている。こうした議論の解析に関連するタスクとして、**意見生成** (Argument Generation) がある。意見生成には、反論生成 [2] や対話形式の意見生成 [3] など様々な形式がある。本研究では特に、議論の的となる論題に対して自動で意見を生成する形式の意見生成を扱う。本タスクの入出力例を表 1 に示す。意見生成により、ユーザが感心のある様々な論題に対し、pros と cons をそれぞれ表す意見を提示することで、人の意思決定の支援などに応用できると考えられる。

生成した意見の質の評価軸の一つとして、論述構造の良さが挙げられる [4]。論述構造とは、主に、(i) “主張” や “根拠” などの論述単位と “支持” や “反論” などの論述単位間の関係性からなる**内容面**を表す構造と、(ii) 論述単位の並びなどの**レトリックの面**を表す構造で構成される。意見の論述構造を改善することで、主張や根拠、それらの関係性がより明確で、かつ、表現がより技巧的になり、より説得力のある意見となる。このため、意見生成においても、より良い論述構造を持つ意見を生成できるような機構を設計することは、重要な課題であると考えられる。意見生成に対する既存のアプローチは、大きく分けて**検索ベース** [5, 6]、**生成ベース** [2, 3] の二種類あり、検索ベース [5, 6] では、論述構造を持つような意見生成の手法が模索されている。しかしながら、生成ベース [2, 3] では、生成した意見がより良い論述構造を持つための明示的なモデル設計がされておらず、生成した意見の論述構造の評価も行われていない。

そこで本研究では、言語生成モデルに基づく意見生成モデルを拡張し、より良い論述構造を持つ意見を生成できるような改良を提案する。論述構造は単語よりも大きな単位（文、文章など）で決まるため、より良い論述構造を持つ意見生成を行うためには、長距離依存関係をより捉えられる機構が必要であり、こうした改良は自明でない。本稿の貢献は、次の通りである:

- 長距離依存関係をより捉えるための機構として、機械翻訳と統語構造との同時予測を用いた方法 [7] に着目し、意見生成とその論述構造を同時予測することで、より良い論述構造を持つ意見を生成する意見生成モデルを提案する。

入力 (論題)： Funding artists
出力 (意見) : (略) some countries depend on tourism as their main source of revenue, thus, the governments should help arts because their work crafts are very necessary to boost the tourist industry. (略)

表 1: 本タスクの入出力例 (Stab ら [9] の essay データから抜粋)

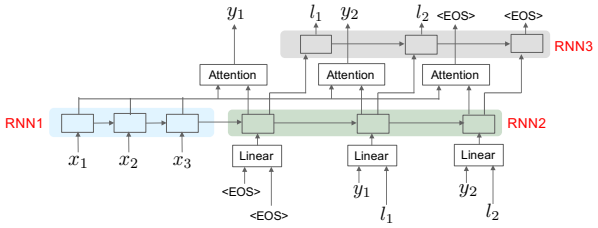
- 提案モデルは、任意の単語ベースの言語生成モデルを用いることができ、今後の言語生成モデルの進化をそのまま取り入れることができる。また、既存の論述構造のアノテーションをそのまま活用することができる。
- 論述構造との同時予測を行うことにより、BLEU [8] による自動評価と、論述構造に関する評価指標 [4] に基づく人手評価の両面から、より良い論述構造を持つ意見が生成されることを示す。

2 関連研究

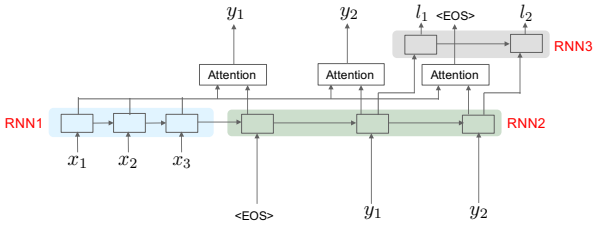
議論マイニングに対するアプローチとして、分割したサブタスクについて学習したモデルの出力結果を整数計画問題として統合する方法 [9] や、深層学習を用いて end-to-end で解く方法 [10] などがある。

意見生成に対する既存のアプローチは、大きく分けて、検索ベースの方法と生成ベースの方法がある。検索ベースの方法では、テキストデータから検索した文集合の並び替えによる方法 [5] や、ニューラルネットワークを用いて学習した分類器を使って、テキストデータから pros と cons に分類した文集合を検索する方法 [6] などがある。生成ベースの方法では、論点と意見をそれぞれ生成する 2 つの decoder を持つ seq2seq ベースのモデルを用いて反論を生成する方法 [2] や階層型 RNN を用いて対話の文脈を考慮した生成を行う対話形式の意見生成 [3] がある。生成ベースの方法は、関連する意見が検索対象のテキストデータや学習用データに含まれない論題に対しても、言語生成モデルの汎化能力により、新たな意見を生成できる可能性があるという利点から、本研究では生成ベースの方法に着目して、より良い論述構造を持つための改良を提案している。

構造を持つ文章を生成する方法として、統語構造との同時予測を用いた機械翻訳がある。その例として、翻訳と原言語側の単語系列の統語構造タグの予測とのマルチタスク学習 [11] や、目標言語側の単語系列にその supertag を挟んだ系列を予測する方法 [7] がある。



(a) ラベル系列生成とのマルチタスク学習を行うモデル



(b) 系列ラベリングとのマルチタスク学習を行うモデル

図 1: 論述構造と同時予測する 2 つのモデルの構造

3 論述構造との同時予測による意見生成

3.1 キーアイデア

1 節で述べたように、論述構造は単語よりも大きな単位（文、文章など）で決まるため、より良い論述構造を持つ意見生成を行うためには、長距離依存関係をより捉えられる機構が必要である。そこで本研究では、Recurrent Neural Networks (RNN) が長距離依存関係をより捉えられるような機構を導入することで、この問題を解決する。実際に、文献 [7] では、翻訳とその統語構造を同時予測する方法により、長距離依存関係をより捉えられるようになることで、より正しい統語構造を持つ翻訳結果が得られたことが示されている。これは、統語構造を表す supertag を用いることにより、RNN が時間方向に生成していく際、長期記憶に依存する必要が少なくなり、より局所情報に依存すればよくなるため、RNN の長短期記憶を補うことで、長距離依存関係をより捉えられるようになることを示している。

そこで、本研究では、統語構造と論述構造とのアナロジーを考え、意見生成を翻訳、論述構造を統語構造と対応付け、意見生成とその論述構造を同時予測することで、より良い論述構造を持つ意見生成モデルを提案する。これは、論述構造との同時予測により、論述タイプや論述単位間の関係性を持つ長距離依存関係に従って単語を生成することで、より良い論述構造を持つ意見を生成できることを期待している。論述構造と同時予測するモデルのバリエーションとして、(i) 論述構造を表すラベル系列生成とのマルチタスク学習と、(ii) 論述構造を表すラベルのラベリングタスクとのマルチタスク学習、を検証する。

3.2 ラベル系列生成とのマルチタスク学習

論題を表す単語系列 $X = \{x_1, \cdots, x_{|X|}\}$ を入力として、意見を表す単語系列 $Y = \{y_1, \cdots, y_{|Y|}\}$ の生成と、意見の論述構造を表すラベル系列 $L = \{l_1, \cdots, l_{|Y|}\}$ の生成とのマルチタスク学習を行う。論述構造を表すラベル系列の例として、“I think that we should study hard.” という文のラベル系列は、{“None”, “None”, “None”, “B-Claim”, “I-Claim”, “I-Claim”, “I-Claim”, “None”}などで表せる。モデルの構造を図 1(a) に示す。まず、エンコーダ側（RNN1）で入力系列 X をエンコードする。次に、デコーダ側の各タイムステップ t において、単語 y_{t-1} とラベル l_{t-1} を入力して得られる RNN2 の隠れ層に対して、Attention [12] を計算して得られる RNN2 の隠れ層を $\tilde{\boldsymbol{h}}_t^{(Y)} \in \mathbb{R}^d$、RNN3 へ入力して得られる RNN3 の隠れ層を $\boldsymbol{h}_t^{(L)} \in \mathbb{R}^d$ とすると、出力単語 $\hat{y}_t$ と出力ラベル $\hat{l}_t$ それぞれの条件付き確率は、

$$P_\theta(\hat{y}_t|y_{<t}, l_{<t}, X) = \mathrm{softmax}(\boldsymbol{W}_1^{(Y)}\tilde{\boldsymbol{h}}_t^{(Y)} + \boldsymbol{b}_1^{(Y)}) \quad (1)$$

$$P_\theta(\hat{l}_t|y_{<t}, l_{<t}, X) = \mathrm{softmax}(\boldsymbol{W}_1^{(L)}\boldsymbol{h}_t^{(L)} + \boldsymbol{b}_1^{(L)}) \quad (2)$$

と計算される。ここで、V_Y を語彙数、V_L をラベル数とすると、$\boldsymbol{W}_1^{(Y)} \in \mathbb{R}^{V_Y \times d}, \boldsymbol{W}_1^{(L)} \in \mathbb{R}^{V_L \times d}$ は重み行列、$\boldsymbol{b}_1^{(Y)} \in \mathbb{R}^{V_Y}, \boldsymbol{b}_1^{(L)} \in \mathbb{R}^{V_L}$ はバイアスである。単語とラベルの抽象度の違いを考慮するため、RNN3 を用いて異なる層から出力する場合を考慮している。

学習時には、下記の損失関数 $\mathcal{L}$ を最小化する:

$$\begin{aligned}\mathcal{L}(\theta) = &- \sum_{(X,Y,L)} \sum_{t=1}^{|Y|+1} \log P_\theta(y_t|y_{<t}, l_{<t}, X) \\ &- \eta \sum_{(X,Y,L)} \sum_{t=1}^{|Y|+1} \log P_\theta(l_t|y_{<t}, l_{<t}, X)\end{aligned} \quad (3)$$

ここで、θ は学習する全てのパラメータ、η は、ラベル予測の損失関数にかかる係数を表す。

3.3 系列ラベリングとのマルチタスク学習

論題を表す単語系列 $X = \{x_1, \cdots, x_{|X|}\}$ を入力として、意見を表す単語系列 $Y = \{y_1, \cdots, y_{|Y|}\}$ の生成と、意見の論述構造を表すラベル系列 $L = \{l_1, \cdots, l_{|Y|}\}$ のラベリングとのマルチタスク学習を行う。モデルの構造を図 1(b) に示す。まず、エンコーダ側で入力系列をエンコードする。次に、デコーダ側の各タイムステップ t において、単語 y_{t-1} を入力して得られる RNN2 の隠れ層に対して、Attention[12] を計算して得られる RNN2 の隠れ層を $\tilde{\boldsymbol{h}}_t^{(Y)} \in \mathbb{R}^d$、RNN3 へ入力して得られる RNN3 の隠れ層を $\boldsymbol{h}_{t-1}^{(L)} \in \mathbb{R}^d$ とすると、出力単語 $\hat{y}_t$ と出力ラベル $\hat{l}_{t-1}$ それぞれの条件付き確率は、

$$P_\theta(\hat{y}_t|y_{<t}, X) = \mathrm{softmax}\left(\boldsymbol{W}_2^{(Y)}\tilde{\boldsymbol{h}}_t^{(Y)} + \boldsymbol{b}_2^{(Y)}\right) \quad (4)$$

$$P_\theta(\hat{l}_{t-1}|y_{<t}, X) = \mathrm{softmax}\left(\boldsymbol{W}_2^{(L)}\boldsymbol{h}_{t-1}^{(L)} + \boldsymbol{b}_2^{(L)}\right) \quad (5)$$

と計算される。ここで、V_Y を語彙数、V_L をラベル数とすると、$\boldsymbol{W}_2^{(Y)} \in \mathbb{R}^{V_Y \times d}, \boldsymbol{W}_2^{(L)} \in \mathbb{R}^{V_L \times d}$ は重み行列、$\boldsymbol{b}_2^{(Y)} \in \mathbb{R}^{V_Y}, \boldsymbol{b}_2^{(L)} \in \mathbb{R}^{V_L}$ はバイアスである。単語とラベルの抽象度の違いを考慮するため、RNN3 を用いて異なる層から出力する場合を考慮している。

	圧縮前	圧縮後
t	69.3%	47.7%
r	61.7%	35.6%
d	46.0%	24.7%

表 2: 圧縮前と圧縮後の “None” 以外のラベル数の割合

学習時には、下記の損失関数 $\mathcal{L}$ を最小化する:

$$\mathcal{L}(\theta) = -\sum_{(X,Y,L)}\sum_{t=1}^{|Y|+1}\log P_\theta(y_t|y_{<t},X) - \eta\sum_{(X,Y,L)}\sum_{t=2}^{|Y|+1}\log P_\theta(l_{t-1}|y_{<t},X) \quad (6)$$

ここで、θ は学習する全てのパラメータ、η はラベル予測の損失関数にかかる係数を表す。

4 評価実験

4.1 データ

意見生成タスクの学習では、Stab ら [9] により構築された論述構造の付与された essay データセットを用いた。これらのデータを、学習用データ（292(論題, 意見) ペア)、開発用データ（30 ペア)、テスト用データ（80 ペア）に分割し、実験を行った。前処理として単語は全て小文字化した。また、各単語に対して論述構造を表す下記の 3 種類のラベルを付与した:

$$\{(t,r,d)|t\in\{\text{None, B-P, I-P, B-C, I-C, B-MC, I-MC}\},\ r\in\{\text{None, Support, Attack, For, Against}\},\ d\in\{\text{None}, -11,\cdots,-1,1,\cdots,9\}\}$$

ここで、t は論述タイプ、r は論述単位間の関係性、d は関係する論述単位との距離を表す。t に関して、B, I はそれぞれ議論単位の始めと内部を表し、P, C, MC はそれぞれ “Premise”、“Claim”, “MajorClaim” を表す。また、目標系列 (意見) について、“MajorClaim” が最初に現れるパラグラフまでを抽出することで圧縮した。圧縮の結果、目標系列の平均単語数は約 356 から約 114 に減少した。また、ラベル系列に含まれる “None” 以外のラベル数の割合の圧縮前後の変化を表 2 に示す。

今回、意見生成タスクの学習で用いるデータは小規模なものであり、言語生成モデルの汎化性能が悪い可能性があるため、RNN2 の言語モデルの事前学習を行った。RNN2 の事前学習では、Monolingual language model training data[*1]の中の “From the News Crawl Corpus(2011 only)” と意見生成タスクの学習に用いる学習用データとを合わせたものを学習用データ、“Development sets”[*1]と意見生成タスク学習に用いる開発用データとを合わせたものを開発用データ、“Test sets”[*1]をテスト用データとして用いた。前処理として、単語数が 3 未満の系列を除去し、単語は全て小文字化した。

[*1] http://www.statmt.org/wmt11/translation-task.html

モデル	BLEU
ベースライン	0.0141
MT-gen	**0.0194**
MT-gen w/o RNN3	**0.0179**
MT-tag	**0.0165**
MT-tag w/o RNN3	**0.0215**

表 3: BLEU による自動評価

4.2 学習設定

意見生成タスクの学習では、ラベル系列生成とのマルチタスク学習を行うモデル、系列ラベリングとのマルチタスク学習を行うモデル共に、RNN1、RNN2 として 3 層の LSTM、RNN3 として 1 層の LSTM を用いた。また、ハイパーパラメータの組合せとして、単語の埋め込み次元数 200、最適化手法 Adam、荷重減衰 (係数 5e-4)、ミニバッチサイズ 16、η を 1.0、dropout 確率 (0.1 と 0.6)、勾配クリッピング (閾値 5.0 と 10.0) を使って開発用データ上で調整を行った。

意見生成タスクの学習モデルについて、Attention つきの seq2seq をベースラインとして、(i) ラベル系列生成とのマルチタスク学習を行うモデル (MT-gen) と (ii) その RNN3 がないモデル (MT-gen w/o RNN3)、(iii) 系列ラベリングタスクとのマルチタスク学習を行うモデル (MT-tag) と (iv) その RNN3 がなくラベル予測に Attention を計算した後の隠れ層を用いたモデル (MT-tag w/o RNN3) を用いて比較を行った。

RNN2 の事前学習では、ハイパーパラメータの組合せとして、単語の埋め込み次元数 200、dropout 確率 (0.1 と 0.6)、最適化手法 Adam、荷重減衰 (係数 5e-4)、勾配クリッピング (閾値 5.0 と 10.0)、ミニバッチサイズ 128、語彙数 1 万を使って開発用データ上で調整を行った。

4.3 評価手法

テスト用データの意見を参照例としてモデルが生成した意見との一致度を評価するために、テスト用データの 80 個の論題に対して生成した意見に対して、BLEU [8] を求めることで自動評価を行った。また、生成された意見の論述構造の良さを評価するために、ランダムに選択した 30 個の意見に対し、AB テストによる人手評価を行った。AB テストによる人手評価の指標として、文献 [4] の指標のうち、論述構造に関する評価指標として、内容面に関しては “Cogency” と “Reasonableness”、レトリックの面に関しては “Arrangement” を用いた。“Cogency” とは、主張に対して適切な根拠を持つかどうか、“Reasonableness” とは、意見全体として論題の解決に貢献しているかどうか、“Arrangement” とは、根拠や主張が適切に配置されているかどうかを表す。AB テストでは、どちらのモデルから生成されたかのバイアスが入らないよう、ベースラインのモデルが生成した意見と、論述構造と同時予測するモデルが生成した意見を、左右ランダムに並べた。評価者は、各評価指標について、左右どちらが良いか、あるい

	Cogency			Reasonableness			Arrangement
	A	R	S	A	R	S	
=	100.0	76.7	96.7	100.0	76.7	100.0	26.7
<	0.0	13.3	0.0	0.0	13.3	0.0	20.0
>	0.0	10.0	3.3	0.0	10.0	0.0	53.3

表 4: 議論マイニングの専門家による人手評価の結果 (“A” は “acceptability”、“R” は “relevance”、“S” は “sufficiency” を表す。)

は同じかを選択する形式で行った。

4.4 結果

4.4.1 自動評価

テスト用データにおける BLEU による自動評価の結果を表 3 に示す。論述構造と同時予測するモデル全てにおいて、テスト用データの意見例との一致度が向上した。また、ラベリングタスクとのマルチタスク学習において出力する層を同じにすることで特に向上がみられた。

4.4.2 人手評価

ベースラインのモデルが生成した意見と、BLEU による自動評価において最も向上がみられたモデル (MT-tag w/o RNN3) が生成した意見に対する、議論マイニングの専門家による人手評価の結果を表 4 に示す。表中の各指標に対して、「ベースラインの方が良い (“<”)」、「MT-tag w/o RNN3 の方が良い (“>”)」、「同じ (“=”)」のそれぞれの結果となったサンプル数の割合 (% で表示) を示す。論述構造との同時予測により、内容面での向上はみられなかったが、レトリック面で向上がみられた。レトリック面での向上とは、例えば、“i think” など、主張が後続することを示唆する表現がより多くみられる点などを表す。

評価の一貫性を調べるために、もう一人の評価者が同一の評価を行った。評価者間の Kappa 値 [13] の結果を表 5 に示す。一致率の結果から、評価者間の評価がよく一致していることがわかる。また、実際に生成した意見の例を表 6 に示す。生成された意見例は、論述構造の内容面に関しては意味の通らない文章であるが、“in my opinion,” という、主張が後続することを示唆する表現が見られる点や、“academic” と類似な単語 “knowledge” を生成した後に、“academic” と対比にあたるような単語 “career” が生成されている点など、長距離にわたって依存関係を持つ単語が生成されている点が見受けられた。

5 おわりに

本研究では、議論の的となる論題に対して自動で意見を生成する意見生成において、論述構造との同時予測を行うことで、BLEU の自動評価から、テスト用データの意見例との一致度の向上がみられた一方、人手評価から、論述構造の内容面での向上はみられなかったが、レトリックの面での向上がみられた。今後は内容面の向上を目指すと共に、外部知識を用いながら論述構造を持つ意見を生成する機構の構築を目指したい。

Cogency	Reasonableness	Arrangement
0.598	0.592	0.591

表 5: 人手評価の一致率

入力 (論題)： non academic subjects should be removed from syllabus
出力 (意見)：nowadays , last individuals choose , set up one need to no one ways to acquire more knowledge or the whole society . in my opinion , it will be an important option . students in my interested in a huge subject for one courses ’ university subjects rather than career with a job .

表 6: 論述構造との同時予測モデルによって生成された意見例

参考文献

[1] Iryna Gurevych, Chris Reed, Noam Slonim, and Benno Stein. NLP approaches to computational argumentation. *ACL*, 2016.

[2] Xinyu Hua and Lu Wang. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of ACL*, pages 219–230, 2018.

[3] Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, 2018.

[4] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of EACL*, pages 176–187, 2017.

[5] Misa Sato, Kohsuke Yanai, Toshihiko Yanase, Toshinori Miyoshi, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP*, pages 109–114, 2015.

[6] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of NAACL-HLT*, pages 21–25, 2018.

[7] Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Predicting target language ccg supertags improves neural machine translation. In *Proceedings of WMT*, pages 68–79, 2017.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, 2002.

[9] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

[10] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *Proceedings of ACL*, pages 11–22, 2017.

[11] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *Proceedings of ICLR*, 2016.

[12] Minh-Thang Luong, Hieu Pham, and Christopher D.Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421, 2015.

[13] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.