Transformer による含意生成とその評価

樫原佑哉 [1]　Bisser Raytchev[1]　金田和文 [1]　檜垣徹 [1]
[1] 広島大学大学院　先進理工系科学研究科
ビジュアル情報学研究室
{yuya-kashihara,bisser,kin,higaki}@hiroshima-u.ac.jp

概要

最近の自然言語処理のモデルは特殊なタスクにおいて人間を超えるスコアを記録している. しかしこれらのモデルが言語表現の意味を理解しているかどうかは不明な点が多い. ここでは,「文章の意味を理解することは, その含意の関係にある事柄を推測することが必要である」という考えに基づき, 自然言語処理のモデルを用いて前提となる文章から含意の関係にある文章を生成する実験を行い, その評価を行う. その結果, Transformer ベースのモデル は人間に引けを取らない精度で含意を生成することができ, また含意に対する人間による評価との相関から, 含意に対する評価指標に適した自動評価指標を発見した.

1　はじめに

人間が扱う言葉は意味を持ち, 人間は状況に応じて言葉の意味を理解した上で会話などを行っている. また, 自然言語処理の分野では, 我々が普段話したり, 書いたりする自然言語をコンピュータで処理する様々なモデルが開発され, 最近では特殊なタスクにおいて人間の精度を超えるスコアを記録している. しかし, これらのモデルは特殊なタスクに対して非常に強力であるが,「モデルが言語の意味をどれくらい理解しているのか」, また「言語の意味を理解するためにモデルに必要なことは何か」については不明な点が多い. 意味の理解について様々な仮説が提唱されている中で, 本研究では Brandom [1] の「文章の意味を理解することは, その含意の関係にある事柄を推測することが必要である」という考えに基づき, 含意を生成する実験を行う. 含意とは, ある表現に対し, 必然な結果として導かれる事柄である. すなわち,「P → Q」の関係である. 例として前提文「A black dog crosses a puddle of water with a ball in its mouth.」 の含意文として,「The dog is crossing the water.」や「The dog is wet.」などが考えられる. 本研究の大きな目的は, 含意の生成を行う実験によって, モデルが含意を生成する規則を明らかにすることであり, 本実験ではモデルが含意をどの程度生成できるのか, どのモデル形態がこのタスクに最も適切か, そして生成した含意をどのように評価するかについて検証する.

2　関連研究

機械学習を用いた含意生成の研究は少ないものの, 私たちの知る限り 1 つ公表されている. Peter Blouw ら [2] は, 木構造のニューラルネットワークである Recursive Neural Network を用いて単一の前提文から単一の含意文を生成し た. このモデルでは, Encoder によって文章の木構造が生成されると, それを逆木構造ネットワークの Decoder を通して予測した文章を生成する. 本研究でも用いた SNLI データセット [3] で含意を生成し, その評価としてデータセット の含意文の単語との合致の割合, 人間による評価を行なっている.

3　扱うタスク

本研究では, 単一の前提文から単一の含意文を予測する含意生成のタスクを扱う. 一例を挙げると「A black dog crosses a puddle of water with a ball in its mouth.」から「The dog is crossing the water.」を生成する. このタスクはソース文からターゲット文を予測するという点で翻訳のタスクに似ているため, 使用するモデルとして, 機械翻訳におけるベースラインのモデルである RNN(Recurrent Neural Network) と, 最近の有力な翻訳モデルである Transformer [4], 更に Transformer の Decoder のみで構成されたモデルである GPT モデル [5, 6] を採用する. これらのモデルが含意をどの程度生成できるのか実験し, またその評価として人間の評価と自動的な評価指標を用意しその相関を調べどの指標が適切であるかを検討する.

4 使用したモデル

本実験では, ベースラインのモデルとして Encoder-Decoder モデルである Attention 付きの RNN と近年の翻訳タスクにおいて有力な同じく Encoder-Decoder モデルである Vaswani ら [4] の Transformer, そして大規模なテキストで事前学習を行った Decoder のみのモデルである OpenAI の [5, 6] を採用した.

4.1 Recurrent Neural Network

RNN は翻訳タスクにおいてベースラインとなる Seq2Seq で用いられるモデルである. 今回はアテンション機構のある GRU 層のシンプルなモデルでどの程度含意生成ができるのか評価する.

4.2 Transformer

Transformer は翻訳タスクにおいて当時の最先端の BLEU スコアを記録したモデルである. また自然言語処理の 11 のタスクで最先端のスコアを達成した BERT モデル [7] の基本構造となるモデルである. Transformer は Encoder と Decoder で構成される. Encoder は入力文に対する各単語の埋め込みベクトルに対し, Query, Key, Value のベクトルを生成したのち, 独自のアテンション機構 Self-Attention を用いて各単語が同一文中のどの単語に注目しているかを表現したベクトルを Decoder へ伝搬する. Decoder はそれを受けて予測した単語を順に出力する.

4.3 OpenAI GPT

OpenAI によって公開された GPT モデルは, Transformer の Decoder block のみで構成されるモデルである. 入力単語が与えられるとその位置情報を組み込んだ埋め込みベクトルを用いて入力文に対し文脈情報を計算し, その 値に基づき次の単語を予測する. このモデルは事前学習（Pre-Training）と微調整（Fine-Tuning）による学習方法を採用している. この方法はまず大量のデータセットでモデルを事前学習することでテキストの文脈を学習する. そしてその事前学習済みのモデルを用いて Fine-Tuning を行うことで各タスクに適応するモデルの学習が完了する. この方法により大規模なデータセットで事前学習されたモデルを誰で も利用することができ, Fine-Tuning を行うだけで様々なタスクに応用することができる. 現在も BERT をはじめ多く の事前学習済みモデルが公開されている. また, OpenAI は数百万の Web ページからなる更に大規模なデータセットで事前学習させた GPT2 も公開している. この実験では GPT と GPT2 をそれぞれ用いた. GPT2 は異なる大きさのモデルがあるが, 今回は small モデル (単語 の埋め込みが 768 次元) のものを用いる.

5 評価指標

含意生成のタスクでは生成された文章が含意文であるかどうか, その含意文がどのくらい正しいのかといった評価が難しいことが予想される. そして都度言語の専門家が見て判断することは非常にコストが高いためここでは様々な自動的評価指標を検討する. まず, 翻訳タスクで一般的に用いられる評価指標である BLEU[8], METEOR[9], TER[10] を使用した. また前提文から含意文を生成するタスクは前提文の重要な情報を抽出し簡単な文章に置き換える要約の要素を含んでいる. したがって要約で用いられる指標も検討する. 今回は要約タスクで用いられる ROUGE[11] を加えた. そして含意の評価として有力であると考えられる二つの指標 (Sentence-BERT[12], BERTScore[13]) を提案する.

5.1 Sentence-BERT

図 1 に示す Sentence-BERT は, BERT を用いたシャムネットワークで, 2 つの入力文に対し BERT で生成した固定長の埋め込みベクトルから類似度を計算する. 本実験ではこの事前学習されたモデルで前提文とその含意文の固定長埋め込みベクトルを作成し, そのコサイン類似度を計算する. SBERT は -1 から 1 の値をとる. 使用したモデルは 2 種類あり, 基本となる bert-base-nli-mean-tokens モデルを"SBERT_1", タスク STS のスコアが最も高い bert-large-nli-stsb-mean-tokens モデルを"SBERT_2" とする.

5.2 BERTScore

BERTScore は図 2 に示すように入力文に対し BERT による各単語の埋め込みベクトルから類似度を計算する. BERTScore は 0 から 1 の値をとる.

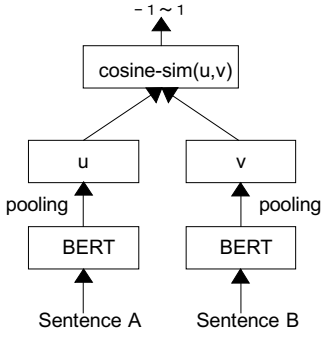

図 1　Sentence-BERT

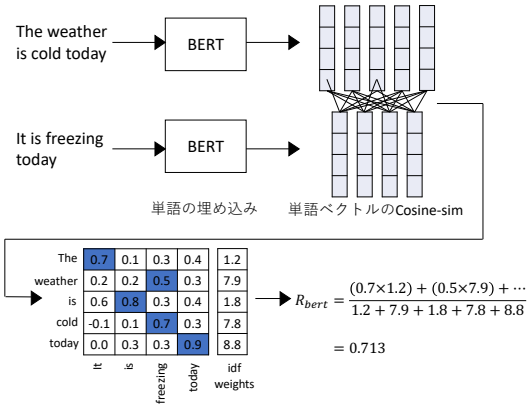

図 2　BERTScore

6　実験

実験では SNLI データセット [3] を用いる. このデータセットは, 前提文と仮説文の文ペアが 57 万セットあり, それぞれの文ペアがそのペアの関係を表す「含意」,「中立」,「矛盾」でラベル付けされているデータセットである. 本来は自然言語推論タスクと呼ばれる上記三種類のラベルを分類する分類タスクで用いられるが, 今回は含意生成タスクを扱うためにラベルが「含意」であるもののみを用いる.

6.1　実験 1

実験 1 では SNLI データセットのうちラベルが含意であるものを, トレーニングデータとして 150000 文ペア, バリデーションデータとして 20000 文, テストデータとして 20000 文を RNN と Transformer の学習に対して用いた. 各モデルの学習後, モデルを用いて生成した含意文に対し, 人間による評価と定量的評価を行った. また, 人間の評価と自動的な各指標の結果の相関を可視化してどの指標が含意に対して有効であるかを検証した.

表 1　生成した含意文の例

	sentence
Premise	A large group of bicycle riders riding down the road.
Reference	People riding their bikes on a road.
RNN	People are riding bikes.
Transformer	The bike riders are outside.
Premise	A girl in orange is hitting the ball in field hockey.
Reference	The girl is outdoors playing field hockey.
RNN	A girl is outside.
Transformer	Someone is playing sports.

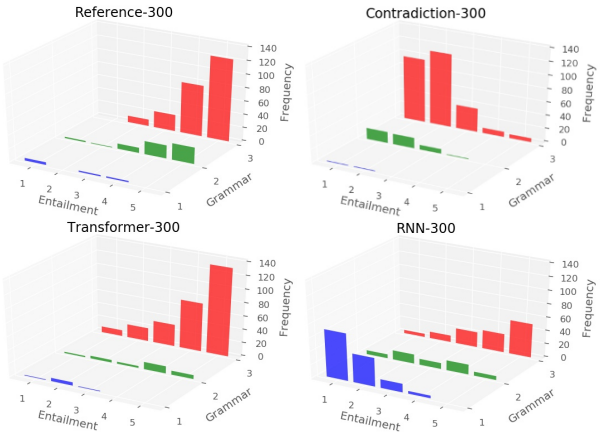

図 3　人間による評価の結果（被験者 A）

6.1.1　生成結果の例

表 1 に前提文と生成した含意文の例を示す. Premise が前提文, Reference がデータセットの含意文である.

6.1.2　人間による評価

RNN と Transformer の生成した含意の精度を評価するため, テストデータの前提文 300 文に対し, 前提文とデータセットの含意文 (Reference), 前提文と RNN の出力, 前提文と Transformer の出力の各ペアを評価する. また, 含意の評価の基準が難しいと予想されるため, 前提文とデータセットの矛盾文 (Contradiction) の 300 文ペアも使用する. 上記の 4 種類の文ペアをラベルを見せずに, ランダムに並び替えた 1200 文ペアに対して, 含意の正しさを 5 段階, 文法の正しさを 3 段階とした評価を英語に詳しい被験者 2 名に実施した. 被験者 A の結果を図 3 に示す. 図 3 より Transformer は人間に近い形状に分布していることがわかる. 被験者 B も同様の結果を示した.

6.1.3　人間による評価と自動的な評価指標の相関

最後に, 自動的な評価指標のうちどれが含意に対して適切な指標であるかを検討するため, 人間の評

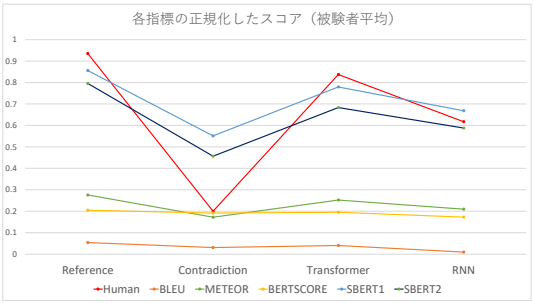

図 4　正規化した各指標と人間による評価の平均スコア

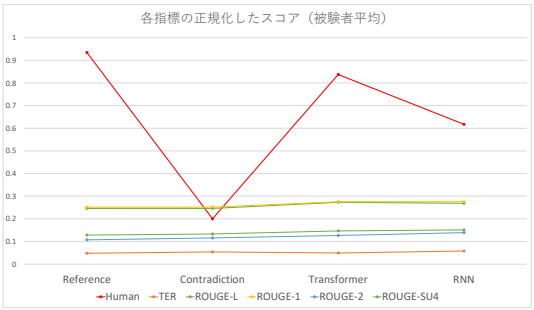

図 5　正規化した各指標と人間による評価の平均スコア 2

価との相関を可視化した. 図 4 は, 人間による評価に用いた 1200 文ペアの最大値と最小値で正規化した各指標と人間の評価を種類ごとにプロットしたものである. これより, SBERT が人間に近い形状をしていることがわかる. 同様に図 5 は要約の指標をプロットしたものである. いずれも相関が低い事がわかる.

6.2　実験 2

実験 2 では強力な事前学習済みのモデルを試すため, OpenAI の GPT モデルを採用する. このモデルは Transformer の Decoder block を基とした Decoder のみのモデルである. この実験の目的は, 最近あらゆるタスクで優れた成果を出している事前学習によって汎用性を獲得した大規模なモデルの含意文の生成能力を検証することと, Decoder のみのモデルでの生成能力を検証することである. 実験 2 では Hugging Face[14] の事前学習済みのモデルとデータセットを用いる. Hugging Face の SNLI データセットのうち, ラベルが含意であるものをトレーニングデータとして 183416 文ペア, バリデーションデータとして 3329 文, テストデータとして 3368 文を用いる. この実験では GPT と GPT2 それぞれを用いて学習と生成を行う.

Fine-Tuning の学習方法として GPT2 の実験で用いられた翻訳タスクの生成手法を学習方法として採用する. これは入力を「前提文=含意文」のまとめた形

表 2　GPT と GPT2（1 epoch, 最大長 20）が生成した含意文の比較例

	sentence
Premise	Two black dogs are running in the grass, one has a green ball in his mouth.
GPT 出力	two dogs playing in field.
GPT2 出力	two dogs run in the outdoors.
Premise	People crowded around a table with plates in front of them.
GPT 出力	people are gathered.
GPT2 出力	people at a dinner table eating and watching their plate.
Premise	in front of an Asian market store one man rests in a chair while another man cleans.
GPT 出力	the men are relaxing.
GPT2 出力	a man is sitting while someone cleans the store.
Premise	three men in a boat on a narrow river team together to carry a large pink sack onto their boat.
GPT 出力	there are three humans.
GPT2 出力	three men are carrying a sack on their boats.

でモデルに渡して言語モデリングで学習させる方法である. 含意生成を行う際はプロンプト「前提文=」の入力を与えてそれに続く含意文を生成させる. これはソース文からターゲット文を生成するというタスクにおいて, 文ペアを連結させて学習させたモデルに特定の条件付きサンプルを生成させるものであり, 貪欲な方法と言える.

6.2.1　生成結果の例（GPT, GPT2）

GPT と GPT2 で Fine-Tuning を 1 エポック行い含意文を生成させた. GPT の性能と直感的な比較するため, 無作為に選んだ前提文より得られた GPT と GPT2（両方 1 エポックの Fine-Tuining）の出力を比較した結果を表 2 に示す.

7　結論

今回の実験から Self-Attention 機構を持つ Transformer は人間に引けを取らない精度で含意を生成できることがわかった. Encoder Block の無い GPT モデルは 1 epoch のみ Fine-Tuning を行うだけで, Transformer と同様の生成結果を得られた. また, 含意の評価指標として BLEU や METEOR, ROUGE, TER 等の主に単語の一致具合や編集操作の量を見る指標よりも, Sentence-BERT の文の埋め込みベクトルによる類似度の方が適していると言える. 今後は単なる言い換えではなく常識的な推測を含んだ含意生成を目標にした更なるモデルでの実験や, 文の埋め込みベクトルを活用した新たな評価指標の検討を考えている.

参考文献

[1] Robert B. Brandom. **Making it explicit: Reasoning, representing, and discursive commitment**. Harvard university press, 1998.

[2] Peter Blouw and Chris Eliasmith. Inferential role semantics for natural language. In **CogSci2017**, 2017.

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **EMNLP Association for Computational Linguistics**, 2015.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmara, Jakob Uszkoreit, LlionJones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in neural information processing systems**, 2017.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[6] Alec Radford, Jeffrey Wua, Rewon Childa, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In **OpenAI blog 1.8 : 9**, 2019.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **arXiv preprint arXiv:1810.04805**, 2018.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. Bleu: A method for automatic evaluation of machine translation. **40th Annual Meeting on Association for Computational Linguistics**, pp. 311–318, 2002.

[9] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. **Second Workshop on Statistical Machine Translation**, pp. 228–231, 2007.

[10] MatthewSnover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. **7th Conference of the Association for Machine Translation in the Americas: Technical Papers**, 2006.

[11] Kavita Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. **arXiv preprint arXiv:1803.01937**, 2018.

[12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

[13] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. **International Conference on Learning Representations**, 2020.

[14] Hugging face - the ai community building the future., （2022-1 閲覧）. https://huggingface.co.