タイトル	Attention を用いたSequence-to-Sequence に基づく文脈ベクトルによる類似度
著者	越前谷, 博; Echizen ' ya, Hiroshi
引用	工学研究：北海学園大学大学院工学研究科紀要(21): 33-38
発行日	2021-12-24

研究論文

Attention を用いた Sequence-to-Sequence に基づく文脈ベクトルによる類似度

越前谷　　博*

Similarity by Context Vector based on Sequence-to-Sequence using Attention

Hiroshi Echizen'ya*

Abstract

本報告では文脈を考慮した文間類似度を求めるために，機械翻訳などの時系列データに広く利用されている Attention を用いた Sequence-to-Sequence の文脈ベクトルを利用した新たな手法を提案する．提案手法ではディープラーニングの技術である，Attention を用いた Sequence-to-Sequence によりモデルを学習する．そして，その学習したモデルの Encoder の出力である文脈ベクトルを利用する．具体的には類似度計算の対象となる 2 つの文それぞれの文脈ベクトル間のコサイン類似度を求めることで意味処理に基づく文間類似度を得る．本報告では，評価タスクデータを用いた小規模な性能評価実験を通して提案手法の可能性について述べる．

1　はじめに

時系列データを対象としたディープラーニングの研究は画像処理と同様に盛んに行われている．時系列データとは時間の経過に伴い変化する情報を持つデータであり，降雨量，株価，為替，交通量，売上など様々なものがあり，自然言語文もまた時系列データの一つである．自然言語文は単語を時系列に並べたものと考えることができ，文中の任意の単語はそれ以前の単語の並びに依存して文法と文脈によって決定される．つまり，自然言語文を扱うことの難しさは過去のデータについての情報を表現する仕組みが要求されることにある．

この自然言語文をディープラーニング，すなわち多層化されたニューラルネットワークで処理するための研究はディープラーニング技術の登場と共に活発に行われてきた．最も単純な考え方は過去の情報を過去の隠れ層として定義し，一般的なニューラルネットワークの隠れ層に伝えることである．それにより過去の隠れ層には再帰的に過去の状態が全て反映される．このように再帰的に過去の状態を反映したニューラルネットワークはリカレントニューラルネットワーク［1, 2, 3］と呼ばれる．

リカレントニューラルネットワークを自然言語処理分野の代表的なタスクである機械翻訳に適用するために改良されたものが Sequence-to-Sequence［4, 5］モデルである．機械翻訳では入力も出力も自然言語文となる．例えば，英語から日本語の翻訳を行う場合，Sequence-to-Sequence の入力は英語，出力は日本語となり共に自然言語文となる．このように入出力が自然言語文であってもそれらを扱えるように提案されたのが Sequence-to-Sequence モデルである．この Sequence-to-Sequence モデルは原言語文と目的言語文をそれぞれ処理する Encoder と Decoder の 2 つのリカレントニューラルネットワークを組み合わせたものとなっている．したがって，Sequence-to-Sequence モデルは Encoder-Decoder モデルとも呼ばれる．

その後，ディープラーニング技術のブレークスルーとして出現したのが Attention［6, 7, 8, 9］である．Sequence-to-Sequence モデルにこの

* 北海学園大学大学院工学研究科電子情報生命工学専攻
raduate School of Engineering (Electronics, Information and Life Science Eng.), Hokkai-Gakuen University

Attention を導入することで翻訳精度が大きく改善された．それまでの Sequence-to-Sequence モデルでは Encoder において入力の情報が最終的には一つのベクトルに集約され，かつ固定された大きさのベクトルとして Decoder に引き継がれることが問題であった．例えば，従来の Sequence-to-Sequence モデルが任意の英文を翻訳する際に先頭の単語から末尾の単語までをフラットに処理することで日本文に翻訳していた．それに対して，Attention を用いた Sequence-to-Sequence では英文中の任意の単語を注視したうえで訳文を得る．この Attention のメカニズムを全面に取り入れたモデルが Transformer［8］であり，現在の機械翻訳研究の主流となっている．

本報告では Attention を用いた Sequence-to-Sequence モデルを学習し，そのモデルの Encoder の出力を文脈ベクトルとして文間類似度を得る新たな手法を提案する．学習データには英語と日本語の対訳コーパスを用い，その対訳コーパスより学習モデルを構築する．そして，参照訳とシステム訳の 2 つの文を同一の学習モデルに通して文脈ベクトルにそれぞれ変換する．最後に文脈ベクトル間のコサイン類似度を求めることで類似度を定量化する．性能評価は提案手法により得られた類似度をシステム訳に対する評価値とみなし，人手により得られたシステム訳に対する評価値と比較し，相関係数を得ることで行う．性能評価実験の結果，提案手法の利用可能性が示唆された．

2　時系列データのためのディープラーニング

前節で述べた時系列データを対象としたディープラーニング技術の概要を述べる．

2.1　リカレントニューラルネットワーク

リカレントニューラルネットワークは過去の情報も再帰的に取り入れることで時系列データを処理するモデルである．過去からの影響を把握するためにリカレントニューラルネットワークでは「過去の」隠れ層を用いる．**図 1** にリカレントニューラルネットワークの概要を示す．

図 1 では入力層への入力データを $x(t)$，隠れ層の値を $h(t)$，出力層の値を $y(t)$，そして，過去の隠れ層の値を $h(t-1)$ としている．また，W は入力データ $x(t)$ に付与される重み，V は隠れ層の値 $h(t)$ に付与される重み，そして，U は過去の隠れ層の値 $h(t-1)$ に付与される重みである．リカレントニューラルネットワークの特徴は過去の隠れ層が過去の全入力データの情報を含んだものとなっている点である．すなわち，過去の隠れ層 $h(t-1)$ は $x(t-1)$ と $h(t-2)$ を入力とすることで得られる．したがって，隠れ層 $h(t)$ を得るための式は以下の式(1)となる．

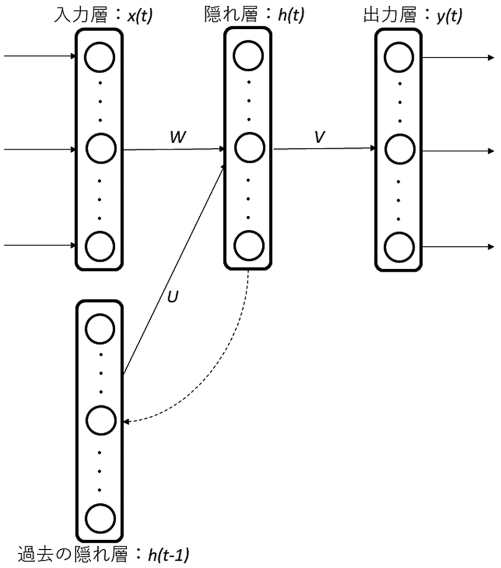

図 1 ：リカレントニューラルネットワークの概要図

$$h(t)=f(Wx(t)+Uh(t-1)+b) \tag{1}$$

また，出力層の値 $y(t)$ は以下の式(2)で表される．

$$h(t)=g(Vh(t)+c) \tag{2}$$

式(1)と(2)の b と c はバイアスを示す．また，$f(\cdot)$ と $g(\cdot)$ は活性化関数を示す．リカレントニューラルネットワークでは，パラメータは W, U, V, b, そして，c となり，これらを更新する必要がある．以下の式(3)から(7)にその更新式を示す．

$$W(t+1)=W(t)-\eta\sum_{z=0}^{\tau}e_h(t-z)x(t-z)^T \tag{3}$$

$$U(t+1)=U(t)-\eta\sum_{z=0}^{\tau}e_h(t-z)h(t-z-1)^T \tag{4}$$

$$V(t+1)=V(t)-\eta e_o(t)h(t)^T \tag{5}$$

$$b(t+1)=b(t)-\eta\sum_{z=0}^{\tau}e_h(t-z) \tag{6}$$

$$c(t+1)=c(t)-\eta e_o(t) \tag{7}$$

式(3)から(7)の e_h と e_o は誤差を示している．また，τ は 10 から 100 程度に設定するのが一般的である．誤差 e_h と e_o はそれぞれ以下の式(8)と(9)で与えられる．

$$e_h(t)=f'(p(t))\odot V^T e_o(t) \tag{8}$$

$$e_o(t)=g'(q(t))\odot(y(t)-t(t)) \tag{9}$$

式(8)の $p(t)$ は式(1)の活性化関数 $f(\cdot)$ の入力 $Wx(t)+Uh(t-1)+b$ である．式(9)の $q(t)$ は式(2)の活性化関数 $g(\cdot)$ の入力 $Vh(t)+c$ である．また，リカレントニューラルネットワークでは逆伝搬においても $t-1$ の誤差を考える必要がある．すなわち，逆伝搬の際には $e_h(t-1)$ を $e_h(t)$ の式で表すことになり，以下の式(10)より得られる．

$$e_h(t-Z-1)=e_h(t-z)\odot(Uf'(p(t-Z-1))) \tag{10}$$

2.2　Sequence-to-Sequence

Sequence-to-Sequence は機械翻訳や対話システムなどの言語処理分野で広く利用されているディープラーニングの代表的な技術の一つである．また，Sequence-to-Sequence は 2 つのリカレントニューラルネットワークを組み合わせた構成となっている．2 つのリカレントニューラルネットワークはそれぞれ Encoder と Decoder と呼ばれる．例えば，英日の機械翻訳であれば Encoder は英文を処理し，Decoder は Encoder による文脈ベクトルを用いて日本文を処理する．また，対話システムにおいては，Encoder が発話文を処理し，Decoder が Encoder による文脈ベクトルを用いて応答文を処理する．図 2 に機械翻訳をタスクとした Sequence-to-Sequence の概要図を示す．

図 2 は Sequence-to-Sequence モデルにより「this is a dog.」を「これは犬です．」に翻訳する処理の流れを示している．すなわち，「this is a dog.」というシーケンスに対して「これは犬です．＜eos＞」というシーケンスを出力している．「＜eos＞」は出力の末尾を示す記号であり，“end-of-sequence” の略である．また，Decoder は自身の出力単語を次のステップの入力としている．すなわち，Decoder の入力は「＜bos＞これは犬です．」というシーケンスとなる．ここで「＜bos＞」は文頭に付く記号であり，“beginning-of-sequence” の略である．文頭の単語においては前に単語が存在せず，直前の出力がないため，「＜bos＞」のような記号を入力する必要がある．

2.3　Attention を用いた Sequence-to-Sequence

2.2 節で述べたように，Sequence-to-Sequence では Encoder の出力は文脈ベクトルであり，結局は 1 つのベクトルとして表現されてしまう．しかし，本来は各時刻によって過去のどの時刻に注視すべきかは異なるはずである．そこで時刻の重みを考慮し，各時刻によって動的に変化するベクトルを生成した方がより良いモデルを生成できるとの考えから Attention が提案された．前節で述べた Sequence-to-Sequence の文脈ベクトルを c と

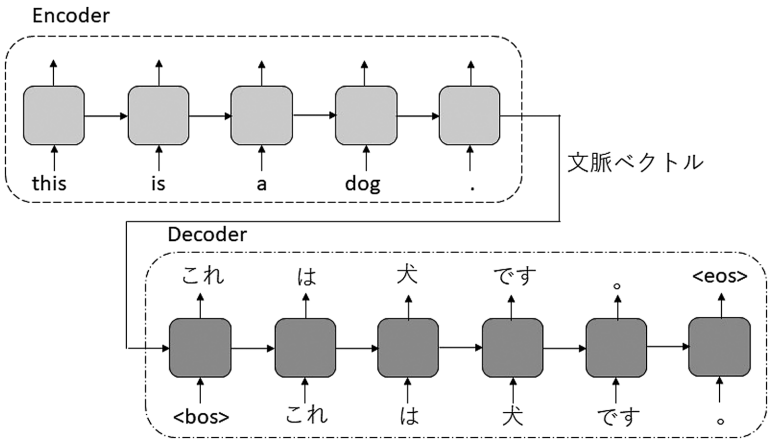

図 2 ：Sequence-to-Sequence の概要図

すると Decoder の最終状態 $h_t(t)$ は以下の式(11)となる．それに対して Attention を用いた Sequence-to-Sequence では時刻によって動的に変化する文脈ベクトルを $c(t)$ とすると Decoder の最終状態 $h_t(t)$ は式(12)として表される．

$$h_t(t)=f(h_t(t-1), y(t-1), c) \tag{11}$$

$$h_t(t)=f(h_t(t-1), y(t-1), c(t)) \tag{12}$$

このように各時刻 t によって変化する $c(t)$ を用いることが Attention の特徴である．この $c(t)$ をどのように定義するかについて述べる．$c(t)$ は Encoder の各時刻の値がそれぞれどれくらい Decoder に寄与しているかを表すものとなっていれば良いと考えられる．したがって，以下の式(13)に示すように各時刻の Encoder の値 $h_s(\tau)(\tau=1, ..., T)$ の加重平均を求め，それを $c(t)$ として Decoder に与えることで実現できる．

$$c(t)=\sum_{\tau=1}^{T} a(\tau,t)h_s(\tau) \tag{13}$$

式(13)の $a(\tau, t)$ は時間の重みとしての役割を担っている．すなわち，Encoder の各時刻の値がどれくらい $h_t(t)$ に影響するかを表す割合を意味する．次いで，重み $a(\tau, t)$ を定義する．ここで $a(\tau, t)$ は Encoder の各時刻における加重平均であるため，各時刻の重みの総和 $\sum_{\tau=1}^{T} a(\tau,t)$ は 1 になる．

また，$a(\tau, t)$ は $h_s(\tau)$ 及び $h_t(t-1)$ により決定される値として考えることができるため以下の式(14)の値をソフトマックス関数で正規化した式(15)より得られる値として定義できる．

$$w(\tau, t) := g(h_s(\tau), h_t(t-1)) \tag{14}$$

$$a(\tau,t)=\frac{\exp(w(\tau,t))}{\sum_{\tau=1}^{T}\exp(w(\tau,t))}=\mathrm{softmax}(w(\tau,t)) \tag{15}$$

ここで式(14)の関数 $g(\cdot)$ はスコア関数と呼ばれ，いくつか提案されている．以下の式(16)から式(18)に代表的なスコア関数を示す．

$$g(h_s, h_t) := \begin{cases} \nu^T \tanh(W_t h_t + W_s h_s) & (16) \\ h_t^T W_a h_s & (17) \\ h_t^T h_s & (18) \end{cases}$$

これらが Attention の定式化になるが，実装においては可読性及び可用性を高めるために Attention「層」の導入を行っている．図 3 に Attention 層を用いた Sequence-to-Sequence の概要図を示す．図 3 の $\tilde{h}_t$ が最終的に求める値であり，この値は Encoder の値 h_s を用いて新たに生成されるベクトルである．$\tilde{h}_t$ は以下の式(19)より得られる．

$$\tilde{h}_t(t)=\tanh\left(W_c\left[\frac{c(t)}{h_t(t)}\right]+b\right) \tag{19}$$

式(19)の $\left[\frac{c(t)}{h_t(t)}\right]$ は式(13)より得られる $c(t)$ 及び $h_t(t)$ を結合したベクトルを意味する．また，$c(t)$ に用いられる $a(\tau, t)$ は以下の式(20)より求める．

$$a(\tau,t)=\mathrm{softmax}(g(h_s(\tau), h_t(t))) \tag{20}$$

式(20)ではスコア関数の引数が $h_t(t-1)$ ではなく $h_t(t)$ になっている．すなわち，式(19)は h_t を $\tilde{h}_t$ に

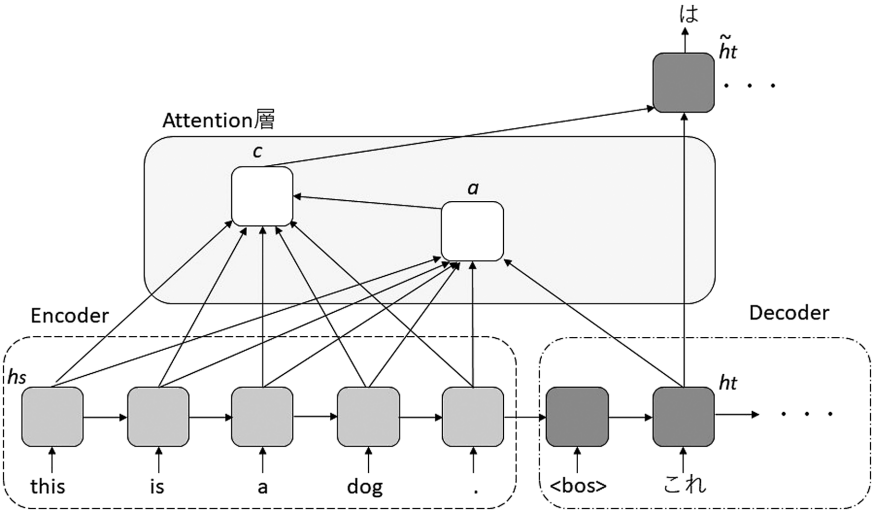

図 3 ：Attention を用いた Sequence-to-Sequence の概要図

活性化させる役割を担っていることを意味する．このような Attention を用いることで h_s のどの単語に注視すれば良いのかを決定することが可能となった．図 3 を用いて説明すると，スコア関数 $g(\cdot)$ として式(18)の内積を用いる場合，h_t の単語“この”と h_s の全単語との内積を求めることになる．その結果，式(15)より“この”に対応する単語“this”との間の $a(\tau, t)$ が最も大きな値になると考えられる．それに伴い，式(13)より“this”に注視した文脈ベクトルが生成され，Decoder はその文脈ベクトルを前提として日本文を出力していくことができる．本報告ではこの Encoder により出力される文脈ベクトルを用いて文間類似度を求める．

2.4　文脈ベクトルによる類似度計算

提案手法では Attention を用いた Sequence-to-Sequence のモデルより文脈ベクトルを得ることで文間類似度を求める．モデルの学習には学習データが不可欠であるが，今回は機械翻訳と同様に原言語文とそれに対する訳文のペアを学習データとして与えた．このような対応関係にある 2 言語の文ペアを用いた学習により，Attention を用いた Sequence-to-Sequence のモデルを生成する．

全ての学習データを用いてモデルを学習させた後，2 つの文を Encoder に与え，それぞれの文に対する文脈ベクトルを取得する．**図 4** に提案手法の概要図を示す．図 4 では 2 つの英文“this is a dog.”と“this is a puppy.”を同じモデルの Encoder に与え，その出力である文脈ベクトルのコサイン類似度により文間類似度を得ている．

3　性能評価実験

提案手法の有効性を確認するために小規模な評価実験を行った．データには WMT20［10］の評価タスクデータに含まれている英日の対訳文 1,000 を用いた．最初にこの英日の対訳データにより Attention を用いた Sequence-to-Sequence モデルを学習させた．その際に，学習データを増やすために全対訳文を 10 倍に増やし，10,000 の対訳文を用いて学習を行った．その際に用いた英文は評価タスクデータとして与えられている正解訳（英文）と原言語文（日本文）のペアを用いた．

このようにして生成した Attention を用いた Sequence-to-Sequence のモデルを利用して文脈ベクトルを取得した．類似度計算の対象となる文は学習に用いた正解訳及びシステム訳である．すなわち，原言語文に対する人手訳とシステム訳との類似度を求めた．

提案手法により得られた類似度に対する評価は評価タスクデータに含まれている人手による評価値と類似度の間の相関係数を求めることで行った．人手による評価値は 5 段階評価となっている．また，相関係数においてはシステム単位はピアソンの相関係数，文単位においてはケンドール τ を用いた．

実験の結果，提案手法の相関係数としてシステ

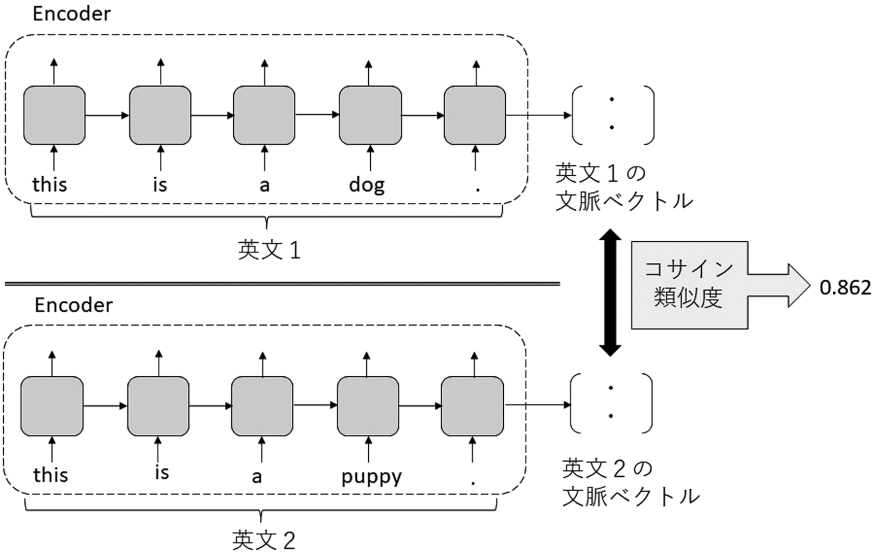

図 4 ：提案手法の概要図

ム単位は 0.785，文単位は 0.243 が得られた．一般的に評価タスクではシステム単位の相関係数は高く，文単位の相関係数は低い．提案手法においても同様の傾向が見られた．これらの結果から提案手法は文単位については不十分であったが，システム単位については今後に期待がもてる結果であった．

4　まとめ

本報告ではニューラルネットワークの代表的なアーキテクチャである Attention を用いた Sequence-to-Sequence に基づく文脈ベクトルを利用することにより，文の意味を考慮した新たな文間類似度の計算手法を提案した．小規模な性能評価実験の結果，精度としては改善の余地が多く残ってはいるが，それと同時に今後さらなる進展が期待できることもまた示唆される結果であった．

今後は精度向上のための改善に取り組む予定である．そして，その有効性を確認するためのより多くのデータを用いた性能評価実験を行う．

謝辞

本研究は，令和元年度北海学園大学学術研究助成費（総合研究）の助成を受けたものである．

References

［1］Andrew M. Saxe, James L. McClelland and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 2nd International Conference on Learning Representations (ICLR).

［2］S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation, Volume 9, Issue 8*. pp.1735–1780

［3］Felix A. Gers, Jürgen Schmidhuber, Fred Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation (2000) 12 (10)*. pp. 2451–2471.

［4］Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp.1724–1734.

［5］Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*.

［6］Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015*.

［7］Minh-Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation *arXiv preprint arXiv:1508.04025*.

［8］Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.

［9］Yankai Lin, Zhiyuan Liu and Maosong Sun. 2017. Neural Relation Extraction with Multi-lingual Attention. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL17*. pp. 34–43

［10］Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). Proceedings of the 5th Conference on Machine Translation (WMT). pp.1–55.