隣接単語系列の予測による汎用的な文の分散表現の構成

露木 浩章　　小川 哲司　　小林 哲則　　林 良彦
早稲田大学理工学術院
tsuyuki@pcl.cs.waseda.ac.jp

1 はじめに

本研究は，隣接単語系列の予測により文の分散表現を構成する手法を提案し，様々なタスクに適用することによって，その汎用性・有効性を評価する．

Word2Vec [1] に代表される単語の分散表現と同様に，文などのより長い言語単位に対しても適切な分散表現を構成するための文エンコーダの研究が盛んである．これらの研究では，大規模な文章コーパスを利用する隣接文の予測タスク等によって文エンコーダを事前学習し [2]，これを様々な目的タスクにおいてファインチューニングする [3] ことにより，汎用性を達成しようとしている．

本研究も同様のアプローチによるが，特に文エンコーダの学習において利用する言語表現の単位に着目する．すなわち，連続文からなるコーパスにおける文を単位にするのではなく，適当な個数の単語からなる単語系列を単位とし，その隣接関係を予測することを学習する．これにより，文の感情極性判定などの，多数の互いに独立な文からなる目的タスクのコーパス (以降，単独文コーパス) からも学習が可能となり，完結した文の形を成していない言語表現に適用することも可能となる．実際，標準的なタスクセットを利用した評価実験の結果から，(1) 教師ありタスクにおいて，提案手法で学習した文エンコーダは，既存の文エンコーダと同等の汎用性を持つこと，(2) 教師なしの意味類似度予測タスクにおいては，辞書中の定義・説明や SNS 上のコメントといった，必ずしも文を成さない言語表現に対して優位性を持つことを確認した．

2 関連研究

既存の汎用的な文の分散表現構成法は，学習データにおける正解ラベルの有無により分類できる．ラベルなしデータを用いる場合，様々なドメインで構成される大規模なコーパスを利用できるため，ドメイン変化に頑健なエンコーダの獲得が期待できる．

最も簡単な文エンコーダは，文中の単語ベクトルの平均をとる BOW アプローチにより構成できる．文のトピックなどのおおよその意味は，語順を考慮しない BOW アプローチによっても捉えることが可能である．

連続する文からなるコーパス (以降，連続文コーパス) における文の連続性に注目するアプローチとして QuickThought [2] がある．QuickThought は，所与の文に隣接する文を生成するのではなく，候補の中から選択することによって学習を行う．それまでの BOW アプローチでは考慮されてなかった語順がエンコーダの出力に反映される．

QuickThought に対し，本研究は単語系列の隣接関係を予測することを学習するため，単語系列の意味に焦点を合わせた手法である．本研究に類似したアプローチに ConsSent [4] がある．本研究と同様に単独文コーパスからも学習可能な汎用的な文の分散表現構成法の 6 種のバリエーションを提案しているが，本研究とは異なり，単語系列を文境界を超えて抽出することはない．このため，本研究では必ずしも必要ではない文の切り出しの前処理が必要となる．

3 提案モデル

3.1 アーキテクチャ

提案モデルは，入力の単語系列から隣接する単語系列を予測する．ただし，隣接する単語系列を実際に生成するのではなく，候補として与えられる単語系列群から正しいものを選択するように学習を行う．このため，提案するモデルは QuickThogut [2] と同様に，入力サンプルをエンコードする f，候補サンプルをエンコードする g という 2 つのエンコーダを用いる．これらのエンコーダは同じ構造を持つが，異なるパラメータを有する．

s_{ctxt} を学習データ中で入力サンプル s_i の前後に現れる正例とする．隣接サンプルの候補として使用する

表 1: 各モデルの学習サンプル例．L は系列長．学習テキスト: Summer vacation was over. And yet, he hadn't done his homework. Then, he called his teacher. "I have a cold."

Model	**Input sample(s_i)**	**Candidates of context sample(S_cand)**	
		Context sample(s_ctxt)	**Negative sample(s_ng)**
QT	And yet , he had n't done his homework .	Summer vacation was over . Then , he called his teacher .	" I have a cold . "
提案手法 (連続文コーパス)			
$L = 7$	yet , he had n't done	Summer vacation was over . And his homework . Then , he	called his teacher . " I have a cold . "
提案手法 (単独文コーパス)			
$L = 3$	he had n't	And yet , done his homework	Then , he [...]

負例を s_{ng} とし，s_{ctxt} と s_{ng} をあわせて候補サンプル群 S_{cand} とする．入力サンプル s_i が与えられたとき，候補群 S_{cand} から候補サンプル s_{cand} が選ばれる事後確率は以下のように表される．

$$p(s_{cand}|s_i, S_{cand}) = \frac{\exp[f(s_i) \cdot g(s_{cand})]}{\sum_{s_j \in S_{cand}} \exp[f(s_i) \cdot g(s_j)]} \quad (1)$$

入力サンプル s_i と候補サンプル $s_{cand} \in S_{cand}$ の共起確率を，各ベクトル同士の内積値を softmax 関数で正規化することによって計算する．負例との共起確率と比較して, 正例との共起確率が最大化されるように学習する．評価タスクに利用するベクトルは，2 つのエンコーダ f，g の出力を連結したものを使用した．

3.2 学習サンプルの作成

学習サンプルの単語系列は，学習に用いるコーパスの種類により 2 つに分類できる．学習サンプルの比較として，QuickThought と提案手法の簡単な学習サンプル例を表 1 に示す．

連続文コーパスから学習する場合，以下のようにして単語系列を作成した．系列長を L とする．文章 $T = \{w_1, w_2, \cdots, w_N\}$ が与えられたとき, w_1 から w_L までを 1 つ目の単語系列サンプルとする．同様にして w_{L+1} から順次単語系列に分割し，学習サンプルを収集する．最終サンプルは残った単語のみで単語系列サンプルとした．また，正例に類似したドメインの負例を候補に含めるため，入力サンプル周辺のサンプルを負例とした．

単独文コーパスから学習する場合，単語数 $2L$ 以上の文から連続する単語系列ペアを抽出し，学習サンプルとした．適当な単語系列を負例として使用する．

4 文の分散表現の汎用性評価

4.1 実験設定

大規模な文章コーパスである UMBC コーパス [5] を学習データとして使用した．これは 2007 年に 100M のウェブページからクローリングされたテキストを文章コーパスとして整えたものである．134.5M の文からなり, 総単語数 3.3B，1 文あたりの平均単語長は 24.9 である．UMBC コーパス中の文をシャッフルしたものを単独文コーパスとした．単独文コーパスを用いて学習サンプルを作成した場合，総単語数は $L = 25$ のとき 0.4B，$L = 30$ のとき 0.2B となった．

ハイパーパラメータを以下に述べる．実験で用いたエンコーダはすべて，次元数 1200 の BiGRU-Maxpooling である．バッチサイズは 400，負例をミニバッチデータからサンプルするため，候補サンプル数も 400 である．オプティマイザは Adam, 学習率は 5e-4 である．fastText [6] を学習済み単語ベクトルとして用いた．

標準的なタスクセットである SentEval [7] を使用して文の分散表現の汎用性を評価した．文エンコーダの汎用性を文の意味理解を必要とするタスクの文エンコーダとして適用することで評価する．単語系列長について検討した後，汎用性評価の結果を，教師あり評価タスクと教師なし評価タスクに大別して述べる．

4.2 学習に適した系列長の選定

提案する文エンコーダの学習に適した系列長 L を選定した．連続文コーパスを，学習データの総単語数を統一するために学習データとして使用した．教師ありタスクにおけるスコアの平均値を図 1 に示す．また，教師なし意味類似度タスク STS14 における，ピアソ

表 2: 教師あり評価タスクにおける文の分散表現の汎用性評価．L は系列長．SICK-R のみ 100 倍したピアソン相関係数．他は accuracy スコア．

Model	MR	CR	SUBJ	MPQA	TREC	MSRP	SICK-E	SICK-R
QT	81.5	**86.9**	94.5	**89.5**	87.8	76.2	**83.8**	**86.7**
提案手法 (連続文コーパス)								
$L = 25$	**82.1**	85.7	94.3	89.3	**88.6**	**76.4**	83.1	86.6
$L = 30$	82.0	85.3	**94.9**	**89.5**	88.4	76.1	83.5	86.5
提案手法 (単独文コーパス)								
$L = 25$	80.3	84.0	93.4	89.0	**88.6**	74.7	82.9	86.6
$L = 30$	80.6	84.3	93.8	88.8	86.8	73.6	82.8	86.4

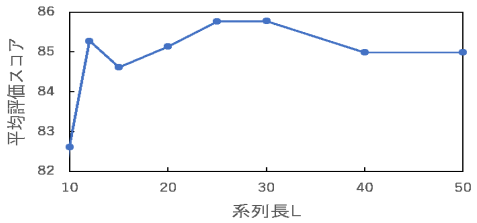

図 1: 系列長 L と教師ありタスクにおける平均スコア．

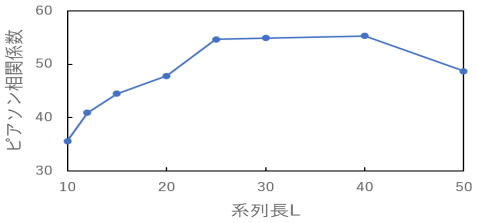

図 2: 系列長 L と教師なし意味類似度タスク (STS14) における平均スコア．スコアは 100 倍したピアソン相関係数．

ン相関係数を 100 倍したスコアの平均値を図 2 に示す．各タスクの詳細は後述する．また，図には示さないが，系列長が与える提案手法の汎用性への影響について調査するため，$L = 2$ の条件においても評価実験を行った．実験の結果，$L = 2$ における教師あり評価タスクの平均スコアは 65.9，STS14 のスコアは 18.8 であった．図より提案手法における単語系列サンプルの最適長は ($L = 25$) と ($L = 30$) であることが分かる．以降の実験項目では，最適系列長における提案手法とベースラインを比較する．

4.3 教師ありタスクにおける評価実験

文エンコーダの汎用性を，様々な教師ありタスクの文エンコーダとして適用することで評価した．文エンコーダの出力を入力として logistic 回帰モデルを学習，評価スコアを算出した．学習の際，文エンコーダのパラメータは更新しない．教師あり評価タスクは次の通りである：映画レビューの感情極性分類 (MR)，製品レビューの感情極性分類 (CR)，文の主観客観分類 (SUBJ)，意見の感情極性分類 (MPQA)，質問タイプ分類 (TREC)，言い換え表現の識別 (MSRP)，自然言語推論 (SICK-E), 文間類似度 (SICK-R).

教師ありタスクにおける評価実験の結果を表 2 に示す．連続文コーパスから学習した提案手法の文エンコーダは，教師ありタスクに転移する場合，QuickThought の文エンコーダとほぼ同等の性能を持つことが分かった．一方，単独文コーパスから学習する提案手法は，連続文コーパスから学習する提案手法よりも低いスコアとなった．単独文コーパスから学習することによって，文エンコーダの性能が劣化した原因として次の 2 つが挙げられる．負例をランダムにサンプリングしたこと，学習データが，単語数 $2L$ 以上の文のみから学習サンプルを収集したために，少なくなっていることである．したがって，単独文コーパスから学習した提案手法の文エンコーダの性能は，正例と類似したドメインの負例を収集し，学習データを増やすことで，向上が見込める．

4.4 教師なしタスクにおける評価実験

文エンコーダの汎用性を，様々なドメインの教師なし意味類似度タスク (STS14) の文エンコーダとして利用することで，評価した．STS14 は収集元ごとに以下の 6 つのデータセットに分かれている．各データセットの詳細を表 3 に示す．

教師なしタスク STS14 における評価実験の結果を表 4 に示す．提案手法の評価スコアは，学習に用いるコーパスに限らず，deft-news や images などの文を扱うデータセットにおいて QuickThought より低いことが分かる．すなわち，文に対する文エンコーダの表現力が，学習サンプルを文から単語系列に変えたことにより劣化した．一方で，提案手法で学習された

表 3: 教師なし評価タスク (**STS14**).

Name	**Type**	**Size**	**Examples**	**Score**
deft-news	News article	0.75k	"Tskhinvali is the capital of Georgia." "Paris is the capital of France."	0.4
deft-forum	Discussion forum	0.75k	"The problem is simpler than that." "The problem is simple."	3.8
OnWN	Word definition	0.75k	"create code or computer programs" "create code, write a computer program."	4.6
tweet-news	Tweet	0.75k	"the tricks of moving from wall street to tech # dealbook" "dealbook : moving from wall street to the tech sector proves tricky"	4.2
images	Image caption	0.75k	"A passenger train waiting in a station." "A passenger train sits in the station."	4.8
headlines	News headline	0.75k	"Stocks rise in early trading" "US stocks ease in choppy trading"	2.0

表 4: 教師なし評価タスク (**STS14**) における文の分散表現の汎用性評価. L は系列長. 表のスコアは 100 倍したピアソン相関係数.

Model	deft-news	deft-forum	OnWN	tweet-news	images	headlines	Overall
QT	**69.8**	**30.6**	55.2	63.1	**72.6**	**63.3**	**60.1**
提案手法 (連続文コーパス)							
$L = 25$	46.6	26.0	**58.7**	63.4	59.9	57.6	54.7
$L = 30$	47.2	26.0	56.3	**66.8**	59.5	57.5	54.9
提案手法 (単独文コーパス)							
$L = 25$	63.3	28.7	55.7	65.8	49.0	59.3	54.5
$L = 30$	65.7	29.2	58.3	65.6	48.8	58.2	55.0

文エンコーダは，辞書中の定義文を扱う OnWN と，SNS 上のコメントを扱う tweet-news など，必ずしも文の形を成さない言語表現を扱うデータセットにおいて QuickThought の文エンコーダを上回る性能を示した．提案手法により，単語系列の形態変化に頑健な文エンコーダが学習されたといえる．

5 おわりに

本稿では，隣接単語系列の予測により文の分散表現を構成する手法を提案し，様々なタスクに適用することによって，その汎用性を評価した．単語系列の隣接関係を予測することは，単独文からなる目的タスクのコーパスから文エンコーダの学習を可能とし，完結した文の形を成していない言語表現に適用することを可能とする．標準的なタスクセットを利用した評価実験の結果から，教師ありタスクにおいて，連続文コーパスから学習する提案手法の文エンコーダは，隣接文の予測タスクで訓練した既存の文エンコーダと同等の性能を持つことを確認した．一方，教師なし意味類似度予測タスクにおいて，提案手法の文エンコーダは，文に対する表現力が劣るものの，辞書中の定義・説明や SNS 上のコメントといった，必ずしも文を成さない言語表現に対しては優位性を持つことがわかった．今後は単独文から学習可能な類似手法 ConsSent [4] との比較実験を行う予定である．

参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 2013.

[2] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *ICLR*, 2018.

[3] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. *ACL*, 2018.

[4] Siddhartha Brahma. Unsupervised Learning of Sentence Representations Using Sequence Consistency. 2018.

[5] Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. Umbc_ebiquity-core: Semantic textual similarity systems. *ACL*, 2013.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 2017.

[7] Alexis Conneau and Douwe Kiela. SentEval: An Evaluation Toolkit for Universal Sentence Representations. *LREC*, 2018.