潜在的ディリクレ配分法を用いたテキスト含意認識のモデル

風間 健太郎 †　　藤田 桂英 ‡

† 東京農工大学　工学部　情報工学科　　‡ 東京農工大学大学院　工学研究院　先端情報科学部門

1　はじめに

テキスト含意認識とは [1]，英語では Textual Entailment と呼ばれ，ある 2 つのテキストが含意関係にあるか矛盾関係にあるかを推定するタスクとして定義される．テキスト含意認識の文やでは，多くの応用がなされている．Tatar らの研究では [1, 2]，文章中に現れた含意関係を 1 つの arc とみなし，排除することで冗長性を取り除いた．一方，質問応答の分野では，質問文と回答文には含意関係があるという仮説をもとにして，テキスト含意認識を用いた質問応答のシステムを構築した Harabagiu らの研究 [1, 3] がある．

以上で述べた通り，自然言語処理の分野におけるテキスト含意認識の応用先は多々あり，テキスト含意認識は有効性のある研究分野であることがわかる．本論文では，潜在的ディリクレ配分法を組み込んだ LSTM により，テキスト含意認識のために用意されたデータセットの分類精度を向上することを目的とする．

2　潜在的ディリクレ配分法

潜在的ディリクレ配分法 (Latent Dirichlet Allocation; LDA) とは，トピックモデルにおいて，トピック分布にディリクレ分布を仮定し，ベイズ推定する手法のことをいう [4]．トピックモデルでは，1 つの文書が複数のトピックを持つと仮定する．トピックモデルの生成過程を Algorithm 1 に示す．

3　テキスト含意認識に用いるコーパス

深層学習の発展に伴い，SNLI コーパス [5] のようなコーパスが登場してきた．SNLI コーパスに対して多大なリスペクトを与えたコーパスが MNLI コーパス [6] である．SNLI コーパスは，各文の「ジャンル」というものがなかった．例えば，どの話題についての文なのかが明示的に指定されていない．一方，MNLI コーパスは各文にジャンルが与えられている．例えば，9.11 に関する文であれば，その文には「9.11」とラベル付けされ，政治に関する文であれば，政治とラベル付けされる．

Algorithm 1 トピックモデルの生成過程

for トピック $k = 1, \cdots, K$ **do**
　単語分布を生成 $\boldsymbol{\phi}_k \sim \mathrm{Dirichlet}(\beta)$
end for
for 文書 $d = 1, \cdots, D$ **do**
　トピック分布を生成 $\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\alpha)$
　for 単語 $n = 1, \cdots, N_d$ **do**
　　トピックを生成 $z_{dn} \sim \mathrm{Categorical}(\boldsymbol{\theta}_d)$
　　単語を生成 $w_{dn} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{z_{dn}})$
　end for
end for

本論文では，LDA を用いたテキスト含意認識のモデルを提案する．LDA は，あるテキストをトピックごとに分類するモデルであった．つまり，膨大なコーパスがあれば，そのコーパス内に含まれるテキストを一定のトピックごとに教師なし学習して分けることができる．これより，MNLI コーパスを用いて LDA を学習させ，MNLI コーパス内に含まれるテキストをトピックごとに分ければ，そのトピック分布には MNLI コーパス内にもとから含まれるジャンルの情報が含まれており，学習時に有利に働くのではないかと考えられる．このことから，本論文では提案するテキスト含意認識のモデルに使うコーパスに MNLI コーパスを使用する．

4　潜在的ディリクレ配分法を用いたテキスト含意認識のモデル

本論文では，LDA トピック分布を用いないシンプルな LSTM モデルと，LDA トピック分布を用いた LSTM モデルの 2 つを提案した．LDA トピック分布を用いた LSTM モデルについて図 1 に示す．

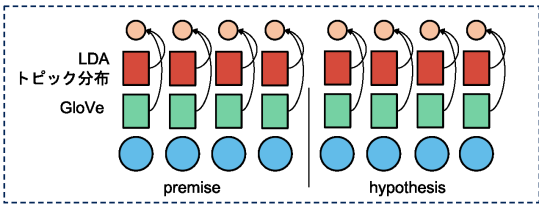

図 1: LDA トピック分布を用いた LSTM モデル

A model of textual entailment utilizing latent dirichlet allocation
†Department of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology
‡Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology

このモデルではまず，元の文章となる premise と対する文章 hypothesis をそれぞれ GloVe 埋め込み表現により単語を埋め込む．埋め込んだ単語列は一旦保持しておく．次に，LDA トピック分布を獲得する段階に移る．LDA トピック分布の生成は，LDA よりあるテキストのトピック分布を生成したあとに，各単語ごとのトピック分布の値を要素としたベクトルを作る．これを各文章ごとに行なって出来上がった行列が LDA トピック分布行列である．作成された LDA トピック分布行列と埋め込み表現を結合し，BiLSTM に入力したあとに，premise と hypothesis それぞれの隠れ層を結合し，最後に softmax で出力する．

LDA トピック分布を用いたモデルでは，MNLI コーパスに与えられたジャンルと LDA により推定されたトピックがそれぞれ同じものを指すと仮定して学習を行う．つまり，完全に一致していれば Accuracy は高く出るが，全く一致していなかった場合 Accuracy は低く出る．以上から，本論文の実験では LDA のトピック数を変化させて行う．

5 実験

5.1 実験設定

モデルを Keras を用いて実装し，トピック数 $K \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$ として変化させ実験を行なった．最も F 値の高いトピック数でシンプルな LSTM との比較を行なった．また，ランダムサーチを行い，最も最適なハイパパラメータでの実験下で行なった．

5.2 実験結果

トピック数を変化させたときの F 値の変化を図 2 に示す．

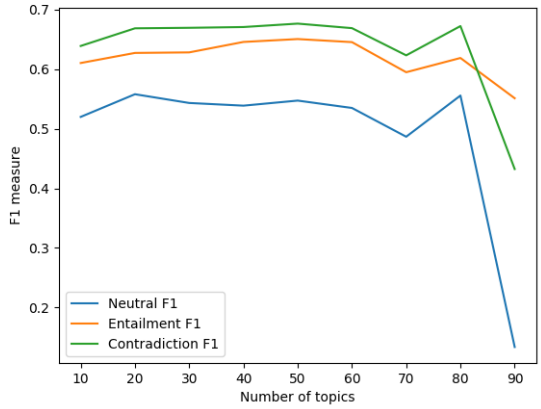

図 2: トピック数を変化させたときの F 値の変化

表 1: トピック数 30 のときの F 値

手法	適合率	再現率	F 値
BiLSTM (+GloVe)	0.48	0.49	0.48
BiLSTM (+GloVe+LDA)	0.63	0.64	0.63

図 2 より，トピック数 30 のときに最も良い F 値を達成できている．トピック数 30 のときの F 値を表 1 に示す．

表 1 より，LDA トピック分布を用いたモデルの方が用いないモデルよりも高い F 値を達成していることがわかる．提案手法では，トピックごとに分類することにより MNLI コーパスのジャンルをうまく利活用できたと考えられる．また，トピックの分布情報を LSTM に投入することが有効であることも示された．

6 まとめ

本論文では，LDA トピック分布を用いてテキスト含意認識を行うモデルを提案した．実験の結果，LDA トピック分布を用いるモデルが既存のモデルと比較して高い F 値を達成した．

参考文献

[1] 横手健一, 石塚満, et al. テキスト含意認識に有効な意味類似度変換及びその獲得法. 人工知能学会論文誌, 28(2):220–229, 2013.

[2] Doina Tatar, Emma Tamaianu-Morita, Andreea Mihis, and Dana Lupsa. Summarization by logic segmentation and text entailment. *Advances in Natural Language Processing and Applications*, 15:26.

[3] Sanda Harabagiu and Andrew Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912. Association for Computational Linguistics, 2006.

[4] 岩田 具治. トピックモデル. 機械学習プロフェッショナルシリーズ. 講談社, 2018.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[6] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.