4Q1-3in2

文書用ニューラルネットワークの半教師あり end-to-end 学習

Semi-supervised End-to-end Learning for Neural Networks for Documents

河東 孝[*1]
Katoh, Takashi

[*1]株式会社富士通研究所
Fujitsu Laboratories LTD.

We often use document vectors made by some unsupervised learning method such as Doc2Vec as features when we apply a supervised learning for documents. Unfortunately, the features may not have enough information for the supervised learning task. The purpose of this paper is to make good features for the supervised learning task by an end-to-end learning. In the end-to-end learning, we use the supervised loss from the supervised learning to train Doc2Vec.

1. はじめに

文書に対して分類等の教師あり学習を行う場合，何らかの方法で文書をベクトル化した後，これを特徴量として教師あり学習を行う方法が一般的である．素朴な文書のベクトル化は，例えば，文書に出現する単語集合 (Bag of Words) を one hot ベクトルに変換することで可能となる．より高度なベクトル化として，分散表現の教師なし学習を使う方法がある．

Word2Vec [3, 4] は文書内での単語の共起関係に基づいて単語をベクトル化する手法であり，単語間の意味的な関係をとらえた分散表現が学習されることから，近年，自然言語処理の分野で注目を集めている．Doc2Vec [2] は単語だけでなく文書に対する分散表現が学習できるように Word2Vec を拡張したものである．Doc2Vec を使うと，文書をベクトル化することができるので，これを特徴量として文書の分類等の教師あり学習を行うことが可能となる．一方，Doc2Vec のような教師なし学習によって作成された文書ベクトルには分類等の教師あり学習に必要な情報が保存されているとは限らず，教師あり学習の性能が十分に出ないという問題点がある．

本稿では，教師あり学習器をニューラルネットワークで構成し，教師あり学習の誤差を逆誤差伝播法で特徴生成器として働く Doc2Vec に伝えることで，教師あり学習に有用な特徴量を作成する半教師あり end-to-end 学習を試みる．実験では，小説から抽出した文書の 2 値分類を例に，学習の進行と特徴量の関係を観察する．

2. Doc2Vec に基づく半教師あり学習

2.1 Doc2Vec

Doc2Vec は Word2Vec を拡張して提案された文書に対するベクトル化手法であり，文書のみから学習が進むため教師なし学習に分類される．図 1 に Doc2Vec による学習の概略を示す[*1]．

Doc2Vec は文書内のある単語を周辺の単語とその文書の ID から予測する学習器と言える．入力は，ある単語とその周辺の単語を切り出したウィンドウと文書 ID である．周辺の単語は

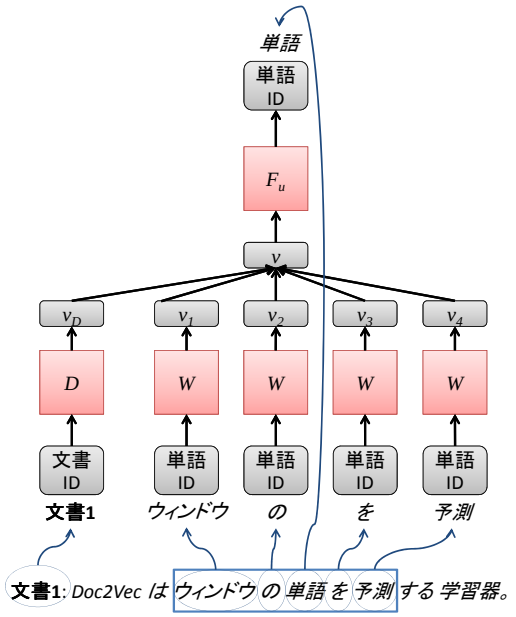

図 1: Doc2Vec のフレームワーク

コンテキストと呼ばれ，コンテキストの各々の単語は，各々の単語に対する単語ベクトルを保存している単語行列 W によって単語ベクトル $v_i (i = \{1, 2, 3, 4\})$ に変換される．文書 ID も同様に，各々の文書 ID に対する文書ベクトルを保存している文書行列 D によって文書ベクトル v_D に変換される．単語ベクトルと文書ベクトルから平均ベクトル v が計算され，これを特徴量に線形分類器 F_u がウィンドウ中央の単語を予測する．線形分類器の誤差が最小となるように行列 W と D が学習される．

2.2 半教師あり学習の方式

Doc2Vec で作成された文書ベクトルを教師あり学習の特徴量として使用する場合，Doc2Vec と何らかの教師あり学習器 F_s を組み合わせて，図 2 のような構成になる．教師なし学習と教師あり学習の組み合わせであり，全体では半教師あり学習となる．このような組み合わせでの学習方法にはいくつかの方式が考えられる．

連絡先: kato.takashi_01@jp.fujitsu.com

*1 Doc2Vec として提案された 2 つのモデル Distributed Memory Model (PV-DM) と Distributed Bag of Words (PV-DBOW) のうち PV-DM を扱う．PV-DBOW についても同様に半教師あり学習にすることができる．

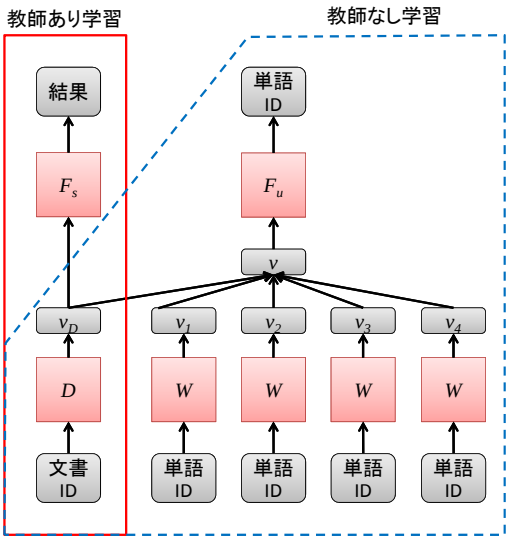

図 2: Doc2Vec をベースにした半教師あり学習のフレームワーク

2.2.1 preprocessing 方式

1 つ目は，従来からよく行われている方式である．すなわち，先ず Doc2Vec の学習を完了させ，その後，文書行列 D を固定して教師あり学習を行う．この方式では，教師あり学習の特徴量である文書ベクトルの学習が教師あり学習とは独立に行われる．そのため，教師あり学習に有用な情報が文書ベクトルに残らない可能性がある．

2.2.2 fine tuning 方式

教師あり学習に有用な情報を回復するために，2 つ目の方式として，学習後の文書行列 D の調整が考えられる．すなわち，Doc2Vec の学習を完了させたのち，教師あり学習を行うと同時に，逆誤差伝播法を使い教師あり学習の誤差で D を修正する end-to-end 学習を行う．逆誤差伝播法を使用するため，教師あり学習器はニューラルネットワークで構成する必要がある．残念ながら，Doc2Vec を使う場合，この方式は以下の理由で想定通りに機能しない．逆誤差伝播法で D に伝わる誤差は教師あり学習時に使用する文書に関するものだけである．Doc2Vec の D は各々の文書 ID に対してそれぞれ独立の文書ベクトルを割り当てることが可能なので，逆誤差伝播法で変化する特徴量は教師ありの文書だけであり，教師あり学習によって作成されたモデルを適用する文書を含む教師なしの文書の特徴量は変化しない．したがって，この方式では汎化性能が向上しない．

2.2.3 parallel training 方式

3 つ目は，Doc2Vec の学習と逆誤差伝播法による D の修正を含む end-to-end の教師あり学習を同時並行で進める方式である．fine tuning 方式 で見たように，教師あり学習による D の修正は，教師あり文書のベクトルにしか影響しない．しかし，Doc2Vec の学習が同時に進行している場合，教師あり文書のベクトルの変化はすべての単語ベクトルに影響を及ぼし，さらに，単語ベクトルの変化は教師なし文書を含むすべての文書のベクトルに影響を及ぼす．したがって，この方式により，教師あり学習に有用な特徴量が生成できる可能性がある．この方式も逆誤差伝播法を使用するため，教師あり学習器はニューラルネットワークで構成する必要がある．

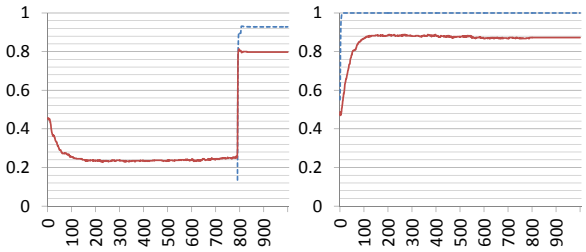

図 3: タスク 1 のエポック数に対する分類精度．preprocessing 方式（左） と parallel training 方式（右）．青破線: 訓練データ，赤実線: 評価データ．

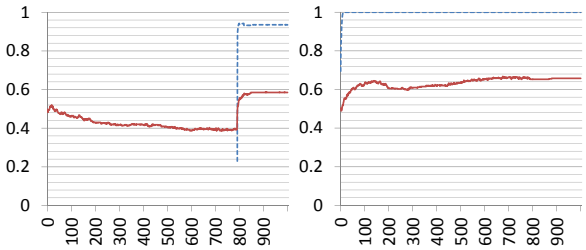

図 4: タスク 2 のエポック数に対する分類精度．並び順と線種は図 3 と同じ．

3. 文書分類への適用実験

実験として，Doc2Vec を使った半教師あり学習を青空文庫[*2]で公開されている小説から作成した文書の 2 値分類に適用し，学習の進行を観察した．

文書データは，夏目漱石の「坊ちゃん」および「吾輩は猫である」を用い，1 段落を 1 文書として抽出した．タスクは以下の 2 種類を用いた．

タスク 1 2 つの小説のどちらから抽出した文書かを分類するタスク．

タスク 2 小説「吾輩は猫である」の前半から抽出した文書か後半から抽出した文書かを分類するタスク．

2 つのタスクともに,「教師ありの正例」,「教師ありの負例」,「教師なしの正例」,「教師なしの負例」をそれぞれ 200 文書, 合計 800 文書をランダムに選択し，教師なし文書を正例/不負に分類する問題とした．すなわち，教師あり学習器に対しては，教師ありの文書が訓練データ，教師なしの文書が評価データとなる問題である．文書ベクトルは 2 次元とし，教師あり学習器には線形分類器（単層パーセプトロン）を使用した．

学習は preprocessing 方式 と parallel training 方式で行った．preprocessing 方式では，800 エポックの Doc2Vec 学習と 200 エポックの教師あり学習を行った．parallel training 方式では 800 エポックの学習を行った後，preprocessing 方式と比較しやすいように，文書行列 D を固定して 200 エポックの教師あり学習を行った．教師あり学習の回数は preprocessing 方式の 200 エポックに対して，parallel training 方式では合計 1000 エポック行うことになるが，教師あり学習器が単純なため 200 エポックでも十分学習可能と考えられるのでこの差異は無視できる．

図 3 と図 4 はそれぞれタスク 1 とタスク 2 におけるエポック数に対する分類精度の推移である．preprocessing 方式では

*2 http://www.aozora.gr.jp/

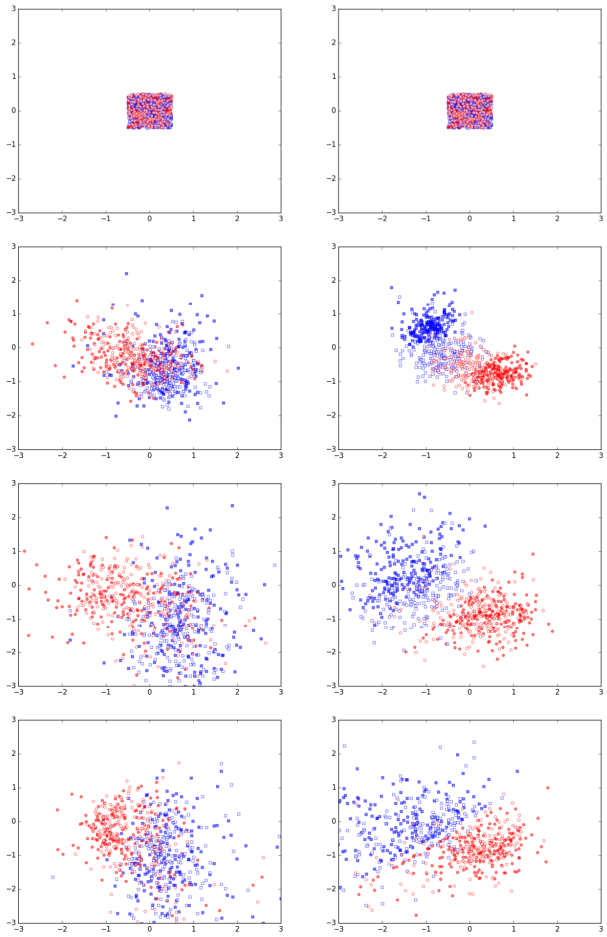

図 5: タスク 1 の文書ベクトル．上から 0, 50, 150, 800 エポック時．preprocessing 方式（左） と parallel training 方式（右）．赤●: 教師ありの正例，青■: 教師ありの負例，赤○: 教師なしの正例，青□: 教師なしの負例．

800 エポックまでは教師あり学習が進行しないため，それ以前の分類精度は学習器の初期値依存であり，意味のある数値ではない．preprocessing 方式において，教師あり学習が 1000 エポックで完了した時点での評価データの分類精度はタスク 1 が 0.80，タスク 2 が 0.59 となった．一方，parallel training 方式では 1 エポック から教師あり学習が進行し，徐々に分類精度が向上していることが確認できる．parallel training 方式において，1000 エポック時点での評価データの分類精度はタスク 1 が 0.87，タスク 2 が 0.66 となっており，どちらのタスクでも，preprocessing 方式に対して精度が向上していることが確認できる．

図 5 と図 6 はそれぞれタスク 1 とタスク 2 における文書ベクトルの分布である．preprocessing 方式では学習が進行しても，正例と不例が特徴空間で混ざり合った状態になっており，教師あり学習に適した特徴になっていないことが確認できる．parallel training 方式では，学習初期の段階で教師ありデータが分離し，その後，徐々に教師なしデータが同じラベルの教師ありデータに近づく形で移動していることが確認できる．これは，教師あり学習による教師ありデータのベクトルの変化が Doc2Vec の学習を通じて間接的に教師なしデータに影響を及ぼしているためと考えられる．これらの影響により，学習終了時点で教師あり学習に適した特徴が生成されたため精度が向上したと考えることができる．

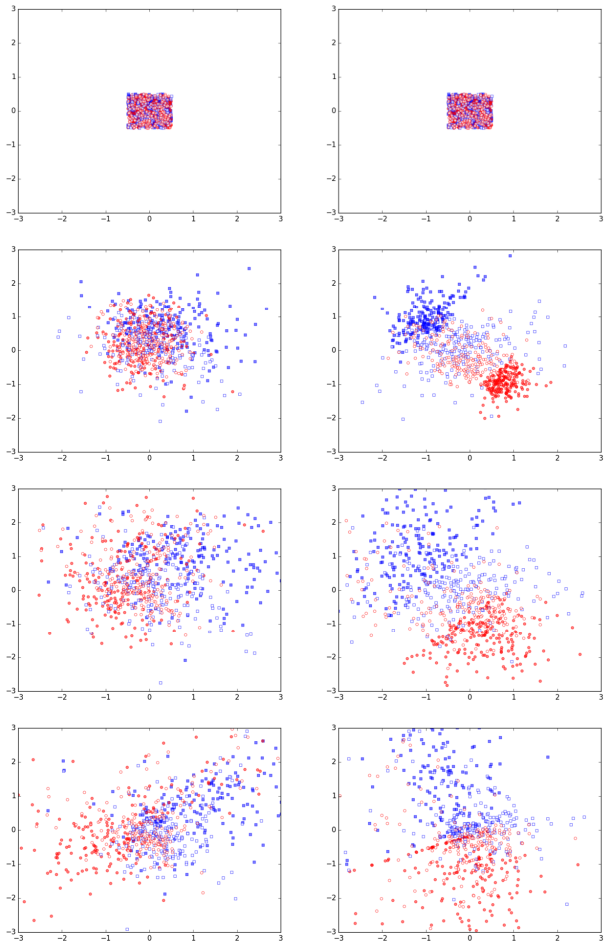

図 6: タスク 2 の文書ベクトル．並び順と記号の意味は図 5 と同じ．

4. まとめ

本稿では，ニューラルネットワークで構成された教師あり学習器と，特徴生成器として働く Doc2Vec を組み合わせて教師あり学習に有用な特徴量を作成する半教師あり end-to-end 学習を試みた．小説から抽出した文書の 2 値分類のタスクにおいて，Doc2Vec の学習とその文書ベクトルを使った教師あり学習を個別に行う方式に対して，Doc2Vec と end-to-end の教師あり学習を同時並行に行う方式により分類精度が向上することが確認できた．

実験で使用した教師あり学習器は単純なものであり，特徴量の次元も小さかった．教師あり学習器としてディープニューラルネットワークと組み合わせた場合の挙動は明らかではない．また，タスクは小説の 2 値分類という限定的なものであった．今後，学習器とタスクのさまざまな組み合わせでの評価を行わなければならない．

最後に，分散表現の学習には SSI [1] に代表される教師ありの方式も存在する．これらの方式との関係を明らかにし，比較を行うことは今後の課題である．

参考文献

[1] Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Chapelle, O., and Weinberger, K.: Supervised Semantic Indexing, In Proceedings of

the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp.187–196 (2009).

[2] Le, Q., and Mikolov, T.: Distributed Representations of Sentences and Documents, In Proceedings of the The 31st International Conference on Machine Learning, ICML 2014, pp.1188–1196 (2014).

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013 (2013).

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Phrases and their Compositionality, In Proceedings of the 26th Neural Information Processing Systems, NIPS 2013, pp.3111–3119 (2013).