nwjc2vec の効果的な fine-tuning のためのパラメータ設定

熊谷 佳奈　古宮 嘉那子　新納 浩幸
茨城大学工学部情報工学科

{14t4027y, kanako.komiya.nlp, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1 はじめに

本論文では，日本語分散表現データ nwjc2vec を効果的に fine-tunig する際に，分散表現構築プログラムのパラメータをどのように設定すべきかを調査する．

nwjc2vec は国語研日本語ウェブコーパス（以下 NWJC）[1] から構築された大規模な分散表現データである [8]．NWJC は超大規模コーパスであるため，そこから構築された nwjc2vec は非常に高品質であると考えられる．実際，いくつかの報告でこの点が確認されている [9][6][7]．また NWJC は様々なコーパスを含んでいるために，広い範囲の領域で利用できると考えられる．ただし nwjc2vec であっても領域適応の問題が生じていることが示され利用領域に応じて fine-tuning することが提案されている [5]．

nwjc2vec を fine-tuning することは有益であるが，それを行うためにどの程度の規模のコーパスが必要なのかは明らかではない．理想的には小さなコーパスで効果的に fine-tuning できることが望まれる．ここでは小規模なコーパスを用いて分散表現構築プログラムのパラメータの調整だけで，どの程度の fine-tuning できるかを調査する．なおここでは分散表現構築プログラムとして word2vec を用いる．

分散表現の評価には，LSTM を用いて，得られた言語モデルにより行う．具体的には LSTM を用いて言語モデルを構築する際に，分散表現の学習も同時に行うが，実験では分散表現の学習は行わずに評価対象の分散表現を利用して言語モデルを学習する．利用した分散表現が高品質であれば，得られる言語モデルも高品質であると考えて評価を行う．

結果，nwjc2vec の効果的な fine-tuning には word2vec のパラメータのうちバッチサイズが最も影響していること，適切でないパラメータでは fine-tuning が逆効果になることが判明した．また fine-tuning にはコーパスの量が本質的であることも確認できた．

2 関連研究

一般に分散表現の優劣は評価法によって異なり，タスクに応じて分散表現をチューンニングすべきことが指摘されている [3]．

最も単純なチューンニングの方法は，本論文で行ったように学習済みの分散表現を初期値として，追加コーパスを用いて再度，分散表現を学習する fine-tuning である．この場合，大規模な追加コーパスを必要とするが，辞書などの外部知識を組み入れて分散表現を改善する試みもある．論文 [4] では分散表現を学習する目的関数の部分に，事前知識を利用した形に変更することで，分散表現を改善している．また論文 [2] では大量のコーパスから構築した分散表現を, 外部知識を使って再学習する retrofitting と呼ばれる手法により分散表現をチューニングしている．

3 提案手法

3.1 nwjc2vec の fine-tuning

分散表現を構築する際の word2vec のパラメータが複数存在する．まずそれらパラメータの基準値を設定する．その基準値のパラメータと追加コーパスにより nwjc2vec を fine-tuning する．次に基準値の中で，ウィンドウサイズのみを変更して，追加コーパスにより nwjc2vec を fine-tuning する．同様にして今度は基準値の中で，バッチサイズのみを変更して，追加コーパスにより nwjc2vec を fine-tuning する．同様にして最後に基準値の中で，エポック回数のみを変更して，追加コーパスにより nwjc2vec を fine-tuning する．

追加コーパスとしては，毎日新聞 ’93 年度版から’99 年度版の 7 年分の記事からランダムに抽出した 10 万文を用いる．

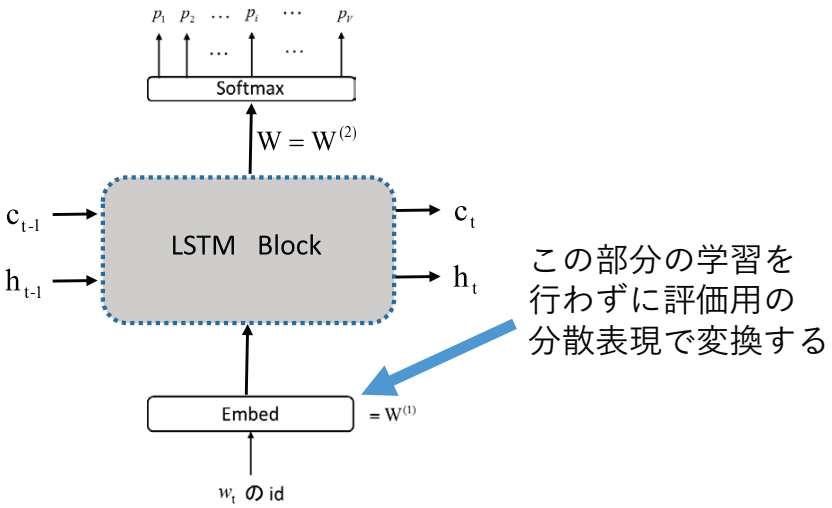

図 1: LSTM の時刻 t 時の入出力

3.2 分散表現の評価

fine-tuning により得られた分散表現を評価するには，類似度が付与された単語ペアのデータを利用するのが一般的であるが，ここでは論文 [7] で用いられた手法を用いる．論文 [7] では LSTM による言語モデルを用いて分散表現の評価を行っている．通常，LSTM により言語モデルを学習する場合，分散表現も同時に学習するが，ここではその学習を行わずに（図 1 参照），単語から分散表現への変換は評価対象の分散表現を用いて行う．LSTM で利用する訓練コーパスが同一の場合，得られた言語モデルの優劣により利用した分散表現の優劣を表すと考える．言語モデルの評価にはパープレキシティを用いる．パープレキシティはモデルの複雑さを表しており，値が小さいほどモデルの品質が良いと判断できる．

また言語モデルの訓練用コーパスと言語モデルの評価用コーパスは，ともに毎日新聞 ’93 年度版から’99 年度版の 7 年分の記事からランダムに抽出した 10 万文及び 1 万文である．またこれらは nwjc2vec を fine-tuning するため利用した追加コーパスとは重複していない．

4 実験

4.1 実験設定

fine-tuning する際の word2vec のパラメータの基準値を表 1 に示す．

表 1: word2vec のパラメータの基準値

ユニット数	200
ウィンドウサイズ	5
バッチサイズ	10
エポック回数	10
使用モデル	skip-gram

まず表 1 の基準値を用いて，nwjc2vec の fine-tuning を行い，分散表現を構築する．ここで構築できた分散表現を base_emb と名付ける．次に表 1 の値からウィンドウサイズのみを 8 に変更し，nwjc2vec の fine-tuning を行い，分散表現を構築する．ここで構築できた分散表現を win_emb と名付ける．同様にバッチサイズのみを 20 に変更し，nwjc2vec の fine-tuning を行い，分散表現を構築する．ここで構築できた分散表現を batch20_emb と名付ける．また同様にバッチサイズのみを 100 に変更し，nwjc2vec の fine-tuning を行い，分散表現を構築する．ここで構築できた分散表現を batch100_emb と名付ける．また同様にエポック回数のみを 20 に変更し，nwjc2vec の fine-tuning を行い，分散表現を構築する．ここで構築できた分散表現を epch_emb と名付ける．

4.2 実験結果

上記 5 つの分散表現（base_emb, win_emb, batch20_emb, batch100_emb, epch_emb）を用いて，10 万文からなる訓練用コーパスを用いて，LSTM に

表 2: パラメータ設定変更時のパープレキシティ

epoch	nwjc2vec	base_emb	win_emb	batch20_emb	batch100_emb	epch_emb
1	91.03	93.70	95.36	91.51	89.69	95.06
2	73.20	75.21	75.71	73.43	72.36	75.89
3	68.65	70.21	70.52	68.69	67.54	70.30
4	**67.43**	68.85	**69.33**	**67.56**	**66.23**	68.46
5	67.52	**68.84**	69.51	67.70	66.35	**68.17**
6	68.17	69.55	70.20	68.37	67.13	68.54
7	69.08	70.37	71.11	69.37	68.17	69.29
8	70.06	71.48	72.22	70.56	69.37	70.36
9	71.09	72.71	73.40	71.80	70.58	71.49
10	72.18	73.92	74.66	73.06	71.82	72.68

よる言語モデルの構築を行った．LSTM の学習での各 epoch 終了時に得られている言語モデルのパープレキシティを 1 万文からなる評価用コーパスを用いて測定した．この結果を表 2 と図 2 に示す．表 2 と図 2 には fine-tuning を行わず nwjc2vec を用いた場合のパープレキシティも示している．

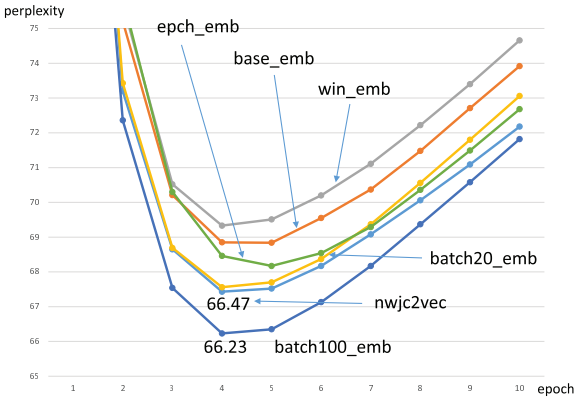

図 2: パラメータ設定変更結果

batch100_emb だけが fine-tuning した効果があった．他の分散表現はどれも fine-tuning の効果はなく，むしろ性能が下がっている．つまり不適切なパラメータを使ってしまったら，fine-tuning が逆効果になる危険性があることが示された．

5 考察

実験では最初に設定した基準値が適切ではなかったために効果的なパラメータを得ることはできなかった．バッチサイズの基準値を 100 に設定しなおして，再度，同様の実験を行えばよいと考えている．ただし表 2 と図 2 からウィンドウサイズやエポック回数よりもバッチサイズが最も fine-tuning の効果に影響していることが分かるので，batch100_emb の性能辺りが限度だと予想している．

また nwjc2vec の効果的な fine-tuning で最も重要な要因は追加コーパスのサイズであると考えられる．この点を確認するため，先の実験では追加コーパスとして 10 万文からなるコーパスを用いたところを，コーパスのサイズを変え，20 万文のコーパス，30 万文のコーパスを用いて，同様の実験を行った．なおこの際の word2vec のパラメータは表 1 の値のうちバッチサイズを 100 に設定したものを用いた．

この結果を表 3 と図 3 に示す．表 3 と図 3 から追加コーパスのサイズが大きいほど fine-tuning の効果が高いことが分かる．

表 3: 追加コーパスのサイズ

epoch	10 万文 (batch100_emb)	20 万文	30 万文
1	89.69	89.55	87.94
2	72.36	71.50	70.28
3	67.54	66.96	65.83
4	**66.23**	65.65	**64.61**
5	66.35	**65.62**	64.75
6	67.13	66.27	65.44
7	68.17	67.32	66.45
8	69.37	68.46	67.56
9	70.58	69.64	68.78
10	71.82	70.89	69.92

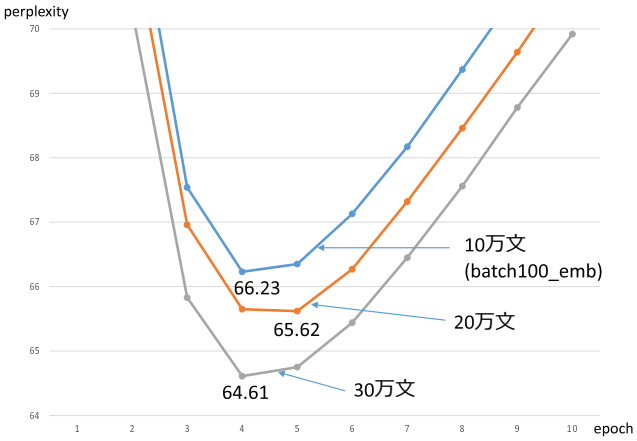

図 3: コーパス量変更結果

本論文で行ったチューニングの手法は，分散表現のチューニングとしては最もシンプルな方法である．学習には新たに大規模な追加コーパスを必要とする．追加コーパスを用いる代わりとして，辞書などの外部知識を融合する方式がある．その方式を利用して，nwjc2vec を改善する手法について今後調査していきたい．

6 おわりに

本論文では nwjc2vec に対して少量の追加コーパスを用いて fine-tuning を行う際の，word2vec の最適なパラメータを調査した．

その結果，バッチサイズの調整が最も fine-tuning には影響があることと不適切なパラメータの設定では fine-tuning が逆効果になることが判明した．また実験の結果から効果的な fine-tuning のためには，追加コーパスのサイズが本質的に重要であると予想し，同様の実験を行った．その結果，予想通りコーパスサイズが大きいほど fine-tuning の効果が高いことが確かめられた．

nwjc2vec を fine-tuning するには大規模な追加コーパスが必要と考えられる．大規模な追加コーパスの代用として外部知識を利用する方式について，今後調査を行おうと考えている．

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである．

参考文献

[1] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. *Alexandria: The Journal of National and International Library and Information Issues*, Vol. 25, No. 1-2, pp. 129–148, 2014.

[2] Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of NAACL*, 2015.

[3] Tobias Schnabel, Igor Labutov, David M Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *EMNLP 2015*, pp. 298–307, 2015.

[4] Mo Yu and Mark Dredze. Improving Lexical Embeddings with Semantic Knowledge. In *ACL (2)*, pp. 545–550, 2014.

[5] 新納浩幸，古宮嘉那子，佐々木稔． nwjc2vec の fine-tuning. 国語研言語資源活用ワークショップ, pp. PB–4, 2017.

[6] 新納浩幸, 古宮嘉那子, 佐々木稔. 順方向多層 LSTM と分散表現を用いた教師あり学習による語義曖昧性解消. 情報処理学会第 232 回自然言語処理研究会, pp. NL–232–4, 2017.

[7] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec : 国語研日本語ウェブコーパスから構築した単語の分散表現データ．自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[8] 浅原正幸, 岡照晃. nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ．言語処理学会第 23 回年次大会発表論文集, pp. 94–97, 2017.

[9] 山木翔馬, 新納浩幸, 古宮嘉那子, 佐々木稔. 教師データを用いた語義の分散表現の構築．言語処理学会第 23 回年次大会発表論文集, pp. 78–81, 2017.