Vol.2017-IFAT-126 No.2
2017/3/30

ニューラルネットを用いた引用リンクの解消

野本忠司[1,a)]

概要：学術論文では先行研究の引用が頻繁に行われるが，引用が原論文（被引用論文）のどの箇所を指しているのか明らかでないことが多い．この問題を本稿では引用リンクの解消と呼び，当問題の解決を図る．本稿では，CL-SciSumm プロジェクトで構築された３つの引用データセットを用いて，ニューラルネットと SVM について，それらがどの程度のリンク解消に有効なのかを検証した．予備実験の結果，MRR において，いずれも芳しい効果が確認できなかったことから，被引用箇所（通常は文）に一定の幅を与え，領域としてリンク解消を試みた．その結果，ニューラルネットが SVM に比べて安定的に優位な性能を示すことが確認された．

Finding Where Citations Come From: A Neural Approach

Tadashi Nomoto[1,a)]

Abstract: This work discusses how neural network models (NNs) may be put to use to resolve citation links in scientific literature, which requires locating passages in the source paper the author had intended when citing the paper. We develop neural machines of increasing complexity, largely drawing on the previous work by [1, 2]. Experiments are conducted using three datasets the CL-SciSumm project developed from a large repository of scientific papers at Association for Computational Linguistics (ACL). We argue based on the experiments that NNs are more susceptible to the amount of data they are trained on than conventional methods such as Ranking SVM. Prompted by a discovery that neither NNs nor Ranking SVM was able to deliver a practically meaningful performance in terms of mean reciprocal ranking (MRR), we turn to the idea of *extended target spans* (XTSs) where we aim at finding a region which likely contains a true target. We show under the XTS scheme that NNs consistently outperform Ranking SVM and TFIDF, an approach based on a standard term weighting metric.

1. はじめに

本稿では，学術文献に現れる引用が被引用論文のどこを参照しているのか，同定する手法について検討する．特に本稿では，引用文献のインデックスが現れる文を引用インスタンス，引用元をターゲットと呼ぶことにする．なお，引用インスタンス，ターゲットは１つ以上の文から構成されると仮定する．また，引用インスタンスの被引用箇所の同定問題を「引用リンクの解消」問題と呼ぶことにする．

本課題に関連した研究としては，ニュース記事間の意味的類似性のモデル化の向けた取り組みが挙げられる．これらは，主に単語頻度，名前付きエンティティ (named entity)，トピック構造，コロケーションなどテキスト内の情報を用いて文書間の類似度を捉えようとした [3–10]．しかし，TDT など多大な研究努力の末分かったのは，極めてシンプルな，TFIDF を用いた，コサイン類似度が一貫してより複雑な手法に勝るという事実であった．

もうひとつの流れとしては，文書間のハイパーリンク関係を用いた一連の研究である [11, 12]. 特に，[11] は「文脈ターム」なる概念を導入し，単語の曖昧性解消を行っている[*1]．

一方，[12] は，ウィキペディアのパイパーリンクを自動追加する仕組みを提案している．具体的には，文書とそれがリンクしている文書群で隣接行列を構成し，固有空間にマップしたのち，主成分分析で次元圧縮し，それぞれの文書の特徴を得る．文書間のリンクの有無をそれらの特徴の類似度を

1 国文学研究資料館
東京都立川市緑町 10-3
a) nomoto@acm.org

*1 ここでの曖昧性解消とは，テキスト中の語句をユニークなウィキペディアページに対応させることを意味する．文脈タームとは単一のウィキペディアページにリンクされている，注目語句の周辺に出現しているアンカーテキストである．

ⓒ 2017 Information Processing Society of Japan

表 1 訓練データ．Per-Topic Avg. はそれぞれのトピッククラスターにおけるデータの平均．Total はデータセット全体での総数．

	TAC2014	DSA2016	SRD2016
Total	53,477	83,293	123,103
Per-Topic Avg.	5,347	8,329	12,310

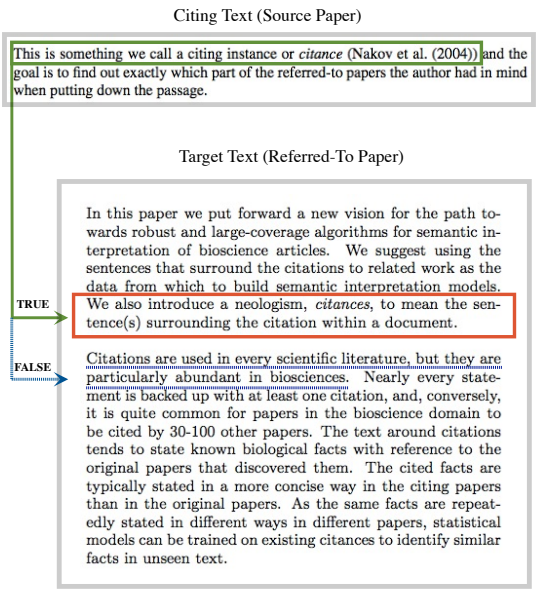

図 1 引用インスタンスとターゲット（文章は [13] から引用）．

使って判定しようと言うものである．単にハイパーリンクの共有度に基づいて文書間の関連度（実際は距離）を測っている [11] に比較して，技術的に洗練されていると言える．

2. アプローチ

図 1 のように，ある文献中で現れた引用を緑枠 (A)，その文が引く論文の被引用箇所を赤枠 (B)，引用とは関係ない箇所を青枠 (C) とする．この時，A と B の類似度を A と C の類似度より大きくすることを考える．引用箇所を d, 被引用箇所を r とし，両者の類似度 h を以下で与える．

$$h(d, r) = \mathbf{G}(d)^\top \mathbf{G}(r) \tag{1}$$

d に対する正の被引用箇所を r^+，それ以外の負の引用箇所を r^- と置くと，以下のような関係が成立することが望ましい．

$$h(d, r^+) > C + h(d, r^-) \tag{2}$$

C は正の定数とする（後述の実験では．[2] に従い $C = 1.0$ とした）．式 2 は以下のような損失関数に書き換えることができる．

$$\max(0, C - h(d, r^+) + h(d, r^-)) \tag{3}$$

本稿では，式 3 を最小化するニューラルネット (図 3) を考える．ただし，ψ は単語列 N から $N \times V$ への one-hot マッピングとする．V は語彙総数．また，隠れ層の重みは共有する．ここで，$S_i^- = \{e_i \mid e_i \neq r^+ \wedge e_i \in \mathbf{RP}\}$ （**RP** は被

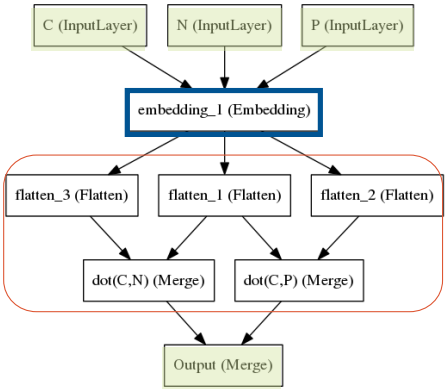

図 2 NN0 （KERAS によるプロット）

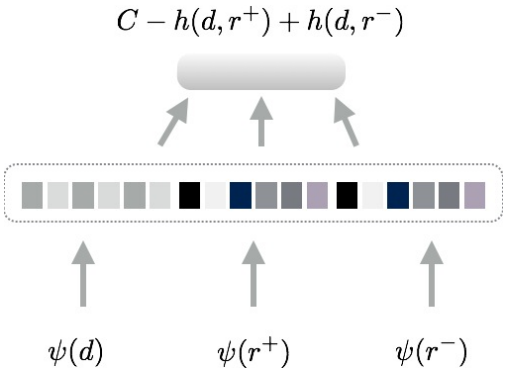

図 3 引用リンク解消モデル

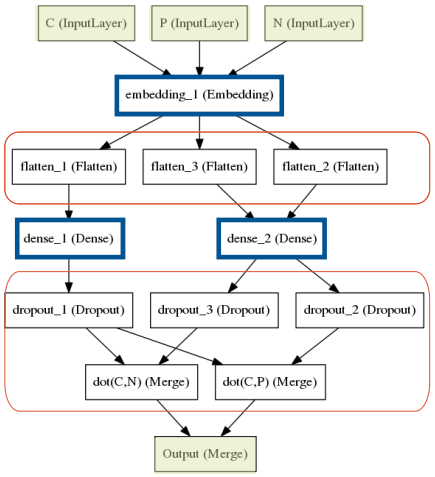

図 4 NN1 （KERAS によるプロット）

ⓒ 2017 Information Processing Society of Japan

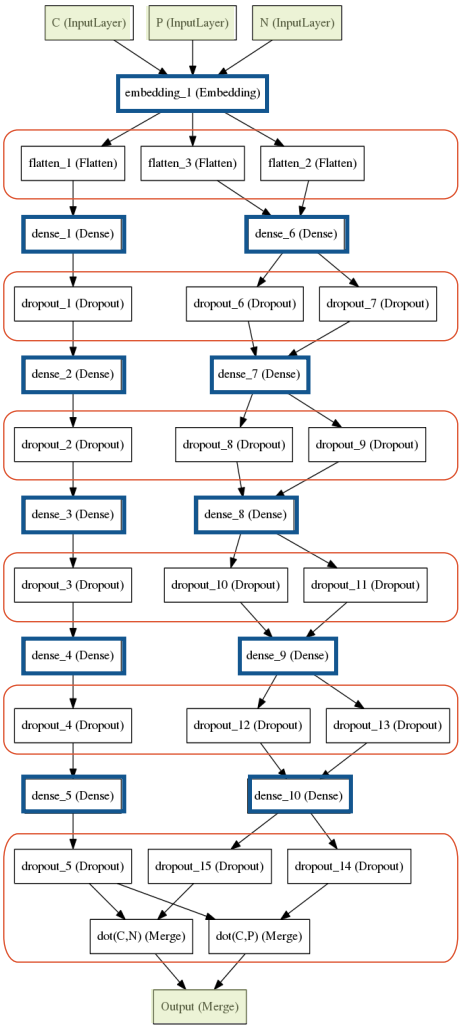

図 5 NN2（KERAS によるプロット）

引用文献中の文の集合），$D = \{(d_i, r_i^+, S_i^-)\}_{i=1...m}$ と置く．すなわち，各引用インスタンス d_i とその真（正）の被引用セグメント r_i^+ とそれに付随する非真（負）の被引用セグメント r_{ij}^- $(0 < j < n)$ を設定する．

学習は各 $(d_i, r_i^+, S_i^-) \in D$ について，SGD を用いて式 3 を最小化することにより行う．なお，本稿では深層学習ライブラリパッケージ KERAS[*2]を用いて，図 3 のモデルを実装した（図 2，以降 NN0 と呼ぶ）．ここで，青の枠は隠れ層，赤枠は行列操作関数を表す．隠れ層のユニット数を 60 とした．また，活性化関数は用いていない．

さらに本稿では，NN0 を改変した幾つかのモデル NN1, NN2 についても調べることにした（図 4, 5）．基本的に隠れ層の数と活性化関数の導入という点が異なる．表 3 にモデルパラメータの詳細を示す．隠れ層の数は，NN1 が 2，NN2 が 6 となっている．活性化関数は RECTIFIER を用いた．NN1, NN2 はさらにドロップアウトを導入した．SGD は ADAM を用いた．

*2 https://keras.io/

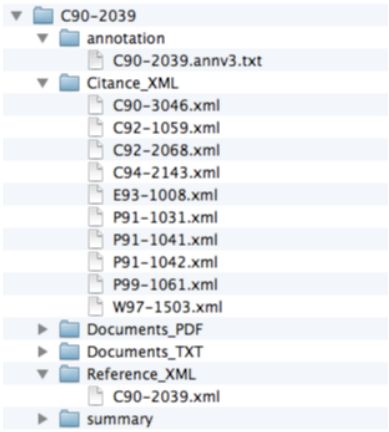

図 6 トピッククラスター C90-2039 の詳細

3. 実験

3.1 データ及び手順

データセットは CL-SciSumm シェアードタスクから提供された，3 つのコーパス，DSA2016 TAC2014, SRD2016 を用いる [14]．各データセットはトピッククラスターと呼ばれる 10 個のフォルダー群から成っている．それぞれのフォルダーには，1 つのレファレンスペーパー（被引用論文，つまり RP）と，RP を参照する論文 (CP) からなっている．さらに，CP の引用インスタンスとターゲットの位置情報を格納したアノテーションファイルが付随する．

フォルダーの一例を図 6 に示す．なお，データセットの論文は ACL Anthology から採取したものである[*3]．データの詳細は表 4 を参照されたい．

評価指標としては式 4 の MRR(Mean Reciprocal Ranking) を用いる．C は引用インスタンスの集合．$\mathrm{rank}_M(t[i])$ は引用インスタンス i に対してモデル M が返した候補ターゲットリストの中で，初めて正のターゲットが出現した順位を表す．つまり，正のターゲットの出現が早ければ早いほど，モデルのパフォーマンスは高いと判断する．

$$R(\mathcal{C}, M) = \frac{1}{|\mathcal{C}|} \sum_{t[i]:i \in \mathcal{C}} \frac{1}{\mathrm{rank}_M(t[i])} \qquad (4)$$

今回，ベースラインとして Ranking SVM (R-SVM) [15] と TFIDF を用いた．R-SVM については，以下の形式でデータをエンコードした．

$$1, \tau(\vec{d}) - \tau(\vec{r^+})$$
$$0, \tau(\vec{d}) - \tau(\vec{r^-})$$

ここで $\tau(\vec{d}) = (w_0, \ldots, w_i, \ldots, w_N)^\top$．$w_i$ は対応する単語

*3 http://aclweb.org/anthology

ⓒ 2017 Information Processing Society of Japan

表 2 $|D|$ の大きさ

	TAC2014	DSA2016	SRD2016
Total	53,477	83,293	123,103
Per-Topic Avg.	5,347	8,329	12,310

表 5 実験結果： TAC2014

n	±0	±1	±2	±3	±4	±5
NN0	0.050	0.077	0.123	0.158	0.181	0.190
NN1	0.086	0.106	0.140	0.164	0.171	0.206
NN2	0.097	0.072	0.073	0.089	0.099	0.180
R-SVM	**0.130**	**0.174**	**0.183**	**0.246**	**0.306**	**0.285**
TFIDF	0.049	0.108	0.147	0.188	0.206	0.227

の tfidf スコアとする．$\tau(\vec{r^+})$, $\tau(\vec{r^-})$ も同様．各要素について，引用インスタンスと正（負）のターゲットの差分を属性としてエンコードする．正データとの差分には 1, 負データとの差分には 0 のラベルを付与することにする．

TFIDF モデルでは，以下のスコア関数を用いた．

$$\cos(d, r) = \frac{\tau(\vec{d}) \cdot \tau(\vec{r})}{\|\tau(\vec{d})\| \, \|\tau(\vec{r})\|}. \tag{5}$$

3.2 拡大ターゲットスパン

さらに本稿では拡大ターゲットスパン (XTS) という概念を導入する．XTS は正のターゲットを含むある幅を持ったテキスト領域を指す．幅とは正のターゲットからの距離（文数）であり，以下では n で表すことにする．例えば，$\mathrm{XTS}_{n=\pm 5}$ は正のターゲットとその前後 5 文から成る領域を指す（図 7）．特に $\mathrm{XTS}_{n=\pm 0}$ を正ターゲットそのものとする．

このとき，モデルの予測ターゲットが XTS の領域に含まれれば，その予測を正解と見なす．すなわち，モデル予測が必ずも正のターゲットと一致しなくても，その近傍にあればよいと考える．

3.3 結果

表 5-7 に XTS の下での結果を示す（対応するグラフは図 8）．TAC2014 では XTS の幅の増減によらず R-SVM が NN0-2, TFIDF を圧倒している．一方, DSA2016, SRD2016 では NN0 が優位傾向にある．これは引用インスタンスの数の違いに依ることが大きい．つまり，学習データの大きさが性能に影響を与えたと考えられる[*4]．

表 2 を見ると，TAC2014 の $|D|$ が三つのデータセット中最も小さいことが分かる．このことは，R-SVM が他手法に比べてデータの欠乏に対して頑健であること，ニューラルネットがデータ量に影響され易いことを示している．逆に TAC2014 における $|D|$ が十分大きくなれば，ニューラルネットのパフォーマンスが向上する可能性があることを示唆している

[*4] これは別実験で確認済み

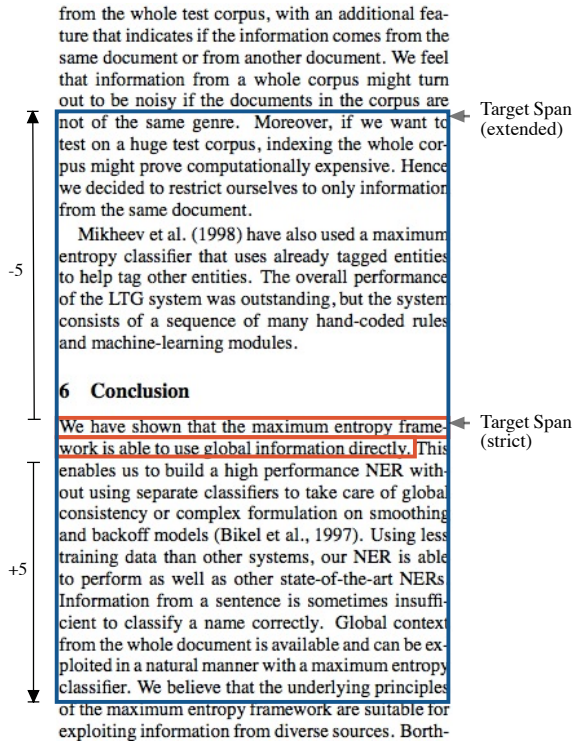

図 7 拡大ターゲットスパン $\mathrm{XTS}_{n=\pm 5}$． 赤枠は正のターゲット．青枠は拡大ターゲット．モデルの予測が青枠内にあれば正解とする．

表 6 実験結果： DSA2016

n	±0	±1	±2	±3	±4	±5
NN0	0.052	0.100	**0.154**	**0.189**	**0.242**	**0.252**
NN1	0.052	0.119	0.171	0.216	0.2259	0.210
NN2	**0.101**	0.126	0.152	0.192	0.2414	0.207
R-SVM	0.070	**0.127**	0.141	0.168	0.2072	0.213
TFIDF	0.032	0.074	0.092	0.140	0.157	0.174

表 7 実験結果： SRD2016

n	±0	±1	±2	±3	±4	±5
NN0	0.066	**0.102**	**0.164**	**0.188**	**0.230**	**0.255**
NN1	0.044	0.082	0.101	0.111	0.188	0.209
NN2	**0.076**	0.105	0.115	0.123	0.147	0.092
R-SVM	0.053	0.071	0.114	0.148	0.186	0.188
TFIDF	0.049	0.108	0.147	0.188	0.206	0.227

ⓒ 2017 Information Processing Society of Japan

表 3 ハイパーパラメータ

Model	#Hidden Layers	Activation	#Output Dimensions	Dropout Ratio
NN0	1	None	$\vec{I} \rightarrow$ 60 (1)	None
NN1	2	rectifier	$\vec{I} \rightarrow$60 (1) $\rightarrow$ 500 (2)	0.5
NN2	6	rectifier	$\vec{I} \rightarrow$60 (1) $\rightarrow$ 500 (2) $\rightarrow$300 (3) $\rightarrow$ 200 (4) $\rightarrow$100 (5) $\rightarrow$10 (6)	0.5

表 4 データセットの詳細．$|RP|$ は被引用文献に現れた文の総数，#CPs は引用インスタンスの総数．$|T|$ は正のターゲットの数，#Citances は引用インスタンスの総数．

TAC2014									
RP	|RP|	#CPs	$\|T\|$	#Citances	RP	|RP|	#CPs	$\|T\|$	#Citances
C90-2039	211	10	33	16	J00-3003	586	9	24	10
C94-2154	118	5	12	5	J98-2005	105	9	26	21
E03-1020	99	9	19	15	N01-1011	195	8	16	8
H05-1115	190	8	19	12	P98-1081	164	9	60	25
H89-2014	152	8	19	11	X96-1048	363	9	21	12
DSA2016									
RP	|RP|	#CPs	$\|T\|$	#Citances	RP	|RP|	#CPs	$\|T\|$	#Citances
C02-1025	205	18	31	23	N04-1038	258	20	44	24
C08-1098	226	22	37	29	P06-2124	247	12	38	18
C10-1045	321	13	42	33	W04-0213	161	13	28	18
D10-1083	248	11	21	18	W08-2222	165	9	13	9
E09-2008	63	10	8	8	W95-0104	338	25	68	39
SRD2016									
RP	|RP|	#CPs	$\|T\|$	#Citances	RP	|RP|	#CPs	$\|T\|$	#Citances
C00-2123	204	8	8	9	P05-1004	235	8	9	9
C04-1089	177	8	9	9	P05-1053	219	8	17	25
I05-5011	213	8	8	10	P98-1046	177	8	9	9
J96-3004	473	8	17	21	P98-2143	157	8	14	23
N06-2049	156	8	11	19	W03-0410	275	8	22	27

ⓒ 2017 Information Processing Society of Japan

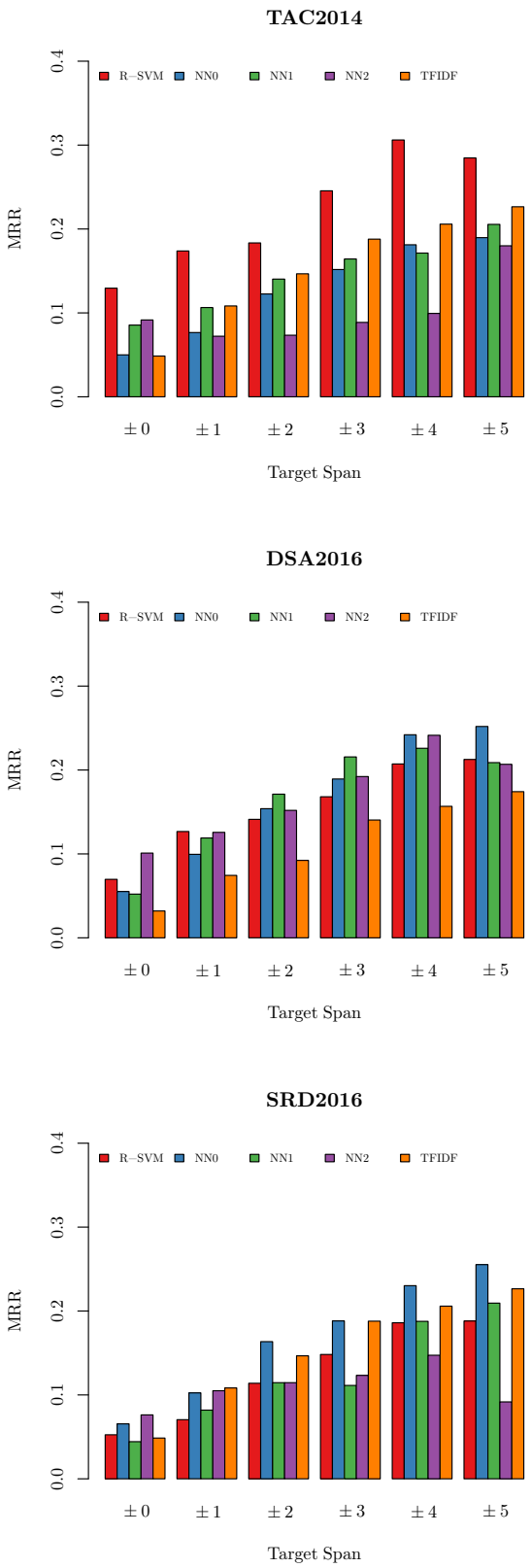

図 8 表 5-7 のグラフ

.

4. まとめ

以上，ニューラルネットを用いた引用リンクの解消法とその構成方法について論じた．また，TAC2014, DSA2016, SRD2016 の 3 つのデータセットを用いてその性能を評価した．$\mathrm{XTS}_{n=\pm 0}$ では，ベースラインを含めたいずれの手法においても芳しい結果が得られたかったが，n を大きくするにつれ，性能が向上することが確認された．特にニューラルモデルの精度がベースラインを顕著に上回ることが認められた．

TAC2014 での R-SVM と NN0-2 の振舞の違いは別の実験によって，学習データの乏しさによることが確認されているが，このことは，ニューラルネットが学習データの量に大きく影響されること示唆している．また，少なくとも，ニューラルネットが DSA2016, SRD2016 で優勢であると言う事実は，単語重要度に関する性質を TFIDF のような情報を与えなくとも学習できることを暗に示している．

参考文献

[1] Weston, J., Bengio, S. and Usunier, N.: Large scale image annotation: Learning to rank with joint word-image embeddings, *Machine Learning*, Vol. 81, pp. 21–35 (online), DOI: 10.1007/s10994-010-5198-3 (2010).

[2] Weston, J., Bordes, A., Yakhnenko, O. and Usunier, N.: Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction, *Empirical Methods in Natural Language Processing*, No. October, pp. 1366–1371 (online), available from ⟨http://aclweb.org/anthology/D/D13/D13-1136.pdf⟩ (2013).

[3] Lavrenko, V., DeGuzman, J. A. E., LaFallme, D., Pollard, V. and Thomas, S.: Relevance Models for Topic Detection and Tracking, *Proceedings of the Conference on Human Language Technology*, pp. 102–110 (2001).

[4] Brown, R.: Dynamic Stopwording for Story Link Detection, *Proceedings of HLT 2002: Second International Conference on Human Language Technology*, pp. 190–193 (2002).

[5] Chen, F., Farahat, A. and Brants, T.: Story Link Detection and New Event Detection are Asymmetric, *Proceedings of HLT-NACCL 2003*, pp. 13–15 (2003).

[6] Chen, Y.-J. and Chen, H.-H.: NLP and IR Approaches to Monolingual and Multilingual Link Dectection, *The 19th International Conference on Computational Linguistics (COLING-2002)* (2002).

[7] Nallapati, R.: Semantic Language Models for Topic Detection and Tracking, *Proceedings of the HLT-NAACL 2003 Student Research Workshop*, pp. 1–6 (2003).

[8] Larkey, L. S., Feng, F., Connell, M. and Lavrenko, V.: Language-specific models in multilingual topic tracking, *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, pp. 402–409 (online), DOI: http://doi.acm.org/10.1145/1008992.1009061 (2004).

[9] Lee, K.-S. and Kageura, K.: Korean-Japanese story link detection based on distributional and contrastive properties of event terms, *Information Processing and Man-*

agement, Vol. 42, pp. 538–550 (2006).

[10] Zhang, X., Wang, T. and Chen, H.: Story Link Detection based on Dynamic Information Extending, *Proceedings of the Third International Join Conference on Natural Language Processing*, pp. 40–47 (2008).

[11] Milne, D. and Witten, I. H.: Learning to Link with Wikipedia, *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, New York, NY, USA, ACM, pp. 509–518 (online), DOI: 10.1145/1458082.1458150 (2008).

[12] West, R., Precup, D. and Pineau, J.: Completing Wikipedia's Hyperlink Structure Through Dimensionality Reduction, *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, New York, NY, USA, ACM, pp. 1097–1106 (online), DOI: 10.1145/1645953.1646093 (2009).

[13] Nakov, P. I., Schwartz, A. S. and Hearst, M. A.: Citances: Citation Sentences for Semantic Analysis of Bioscience Text, In Proceedings of the SIGIR ' 04 workshop on Search and Discovery in Bioinformatics (2004).

[14] Jaidka, K., Chandrasekaran, M. K., Rustagi, S. and Kan, M.-Y.: Overview of the 2nd Computational Linguistics Scientific Document Summarization Shared Task (CL-SciSumm 2016), *The Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)*, Newark, New Jersey, USA (2016).

[15] Joachims, T.: Training Linear SVMs in Linear Time, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)* (2006).

ⓒ 2017 Information Processing Society of Japan