ニューラルネットによる再ランキングを利用した文書トピックの導出

野本忠司
国文学研究資料館・総合研究大学院大学
nomoto@acm.org

要旨

本稿では，ニューラルネットによる再ランキングを利用して任意のテキストのトピックを検出する方法を紹介する．基本的に著者が過去に提案したウィキペディアを利用したトピック抽出法（ウィキラベル）にニューラルネットによる学習機能を付加し．低頻度トピックには非学習型，高頻度トピックには学習型と適宜モデルを使い分け，トピックの導出を行う．ニューヨークタイムズから採取した比較的大規模なデータを用いて SingleRank, TextRank などの最新手法と比較したところ，提案手法に顕著な優位性があることが確認された．

1 背景

ニューヨークタイムズの 2013 年 6 月〜12 月までのニュース記事に人手付与されたトピックの頻度分布を分析したところ，付与トピックのほとんどは 1 回しか出現しないこと分かった．具体的にはトピック全体の 42.3%が 1 回，2 回以下 62.0%，5 回以下 78.9%, 10 回以下が実に 87.4%であった．ニュース記事に自動でトピックを付与する問題を考えると，正例が圧倒的に少ないため通常の学習を使ったテキスト分類にはなじまない．しかし，その一方で 100 回を超える頻度で出現するトピックも存在する．筆者はニュース記事のトピック付与モデルとして非学習型のアプローチ（ウィキラベル）を提案しているが [4]，これは基本的に低頻度トピックを想定したアプローチである．

このような背景から，本稿ではウィキラベルにニューラルネットをベースにした学習機能を付加し，低頻度トピックには既存の非学習型ウィキラベル，高頻度トピックには学習型のウィキラベルと，トピックの性質によってモデルを切り替える折衷型モデルを提案する．

ウィキラベルとは入力テキストに対してそれと最も近いウィキペディアのページタイトルをそのテキストのトピックと考える手法である．形式的には，以下のような形をとる．

$$\ell_d^* = \arg\max_{\ell : p[\ell] \in \mathbb{U}} f(p[\ell], d)$$

d は入力テキスト，$p[\ell]$ はタイトル ℓ を持つページを表す．$\mathbb{U}$ はウィキペディアページの集合．ここで

$$f(p[\ell], d) = \lambda Sr(p[\ell], d) + (1 - \lambda) Lo(\ell, d) \quad (1)$$

とする．$Sr(p[\ell], d)$ は文書とページの類似度，$Lo(\ell, d)$ はペナルティー項で，ℓ 中に d に含まれない単語が存在すると減点される．さらに文圧縮によって ℓ から様々な変種を生成し，ラベルの候補として利用する．なお，Sr, Lo はいずれも 0 から 1 までの値をとる．

2 再ランクモデル

本稿では Weston et al (2010) の画像・テキストの同時学習モデル [5] の考え方を流用し，文書 d とその候補トピック ℓ の類似性を直接測るモデルを構成する（式 2）．ℓ はウィキラベルの出力とする．

$$\mathcal{M}(d, \ell) = \mathbf{G}(d)^\top \mathbf{F}(\ell) \quad (2)$$

$\mathbf{G}(d)$ は d，$\mathbf{F}(\ell)$ はラベル ℓ の隠れ層へのマッピングを表す（図 1 参照）． このとき以下の制約を考える．

$$\forall_{i,j} \;\; \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^+) > 0.1 + \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^-)$$

ℓ_j^+ は正例ラベル，ℓ_j^- は負例ラベルとする．上の制約は d と正例との類似度が負例より必ずある幅を持って上回ることを要請する．（正例，負例の細かい定義については第 4 章で説明する．）上式から以下が得られる．

$$\forall_{i,j} \; 0 > 0.1 - \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^+) + \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^-)$$

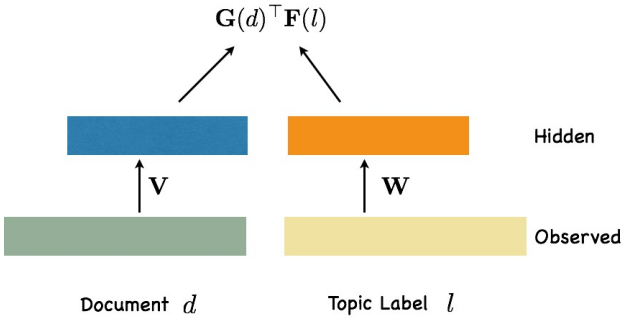

図 1: 再ランクモデル

これは右辺の値が 0 以上のとき，制約が満たされていないことを意味する．よって，以下を解くことを考える．

$$\text{minimize:}[0.1 - \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^+) + \mathbf{G}(d_i)^\top \mathbf{F}(\ell_j^-)]_+$$

$[x]_+ = \max(0, x)$ とする．次に以下を定義する．

$$\mathbf{G}(d) = \mathbf{V}^\top \psi(d)$$

さらに

$$\mathbf{F}(\ell) = \mathbf{W}^\top \phi(\ell)$$

$\psi(d)$, $\phi(\ell)$ は d 及び ℓ から二値表現への写像とする．ここで，ある文書 d_i に付与された正例ラベルを ℓ_i^+, 負例ラベルを ℓ_i^- と表記する．$\mathbf{V}$ は $N_v \times K$, $\mathbf{W}$ は，$N_e \times K$ の行列とする．K は隠れ層の大きさ．N_v, N_e は，それぞれ $\psi(d)$, $\phi(\ell)$ のベクトル長を表す．つまり，$\psi(d) \in \{0,1\}^{N_v}, \phi(\ell) \in \{0,1\}^{N_e}$. さらに，入力層 $\psi(d)$, $\phi(\ell)$ に対応した隠れ層をそれぞれ $\mathcal{H}_\text{V}$,$\mathcal{H}_\text{W}$ と表記する．**隠れ層の各ユニットに対して対応する入力層のユニットがすべて接続していることに注意．**

$\mathbf{W}$ と $\mathbf{V}$ の値は以下の手続きによって求める．

1. 各 ℓ_i^+ と ℓ_i^- について，確率的勾配降下法を実行して以下を最小化する（i は文書のインデックス）：

$$[0.1 - \mathbf{G}(d_i)^\top \mathbf{F}(\ell_i^+) + \mathbf{G}(d_i)^\top \mathbf{F}(\ell_i^-)]_+$$

2. $\mathbf{W}$ と $\mathbf{V}$ を列について，正規化する．つまり，$\forall_k \ \|w_{\cdot k}\|_2 = 1, \forall_j \ \|v_{\cdot j}\|_2 = 1$. $w_{\cdot k}$ は $\mathbf{W}$ の i 番目，$v_{\cdot j}$ は $\mathbf{V}$ の j 番目の列インデックスを表す．

$\mathbf{V}$ と $\mathbf{W}$ の初期値は $\mathcal{N}(-\frac{6}{\sqrt{n}}, \frac{6}{\sqrt{n}})$ に従ってランダムに生成する．n は N_v あるいは N_e を表す．

本稿では確率的勾配降下法として AdaGrad/RDA を用いた．各要素の更新は以下の式に従う．

$$x^{t+1,i} = -\operatorname{sgn}(\bar{g}_{t,i})\eta t \frac{[|\bar{g}_{t,i}| - \lambda]_+}{\sqrt{\sum_{\gamma=1}^t (g_{\gamma,i})^2}} \qquad (3)$$

特に η を 1，λ を 0.001 とする．$g_{t,i}$ はステップ t における要素 i の勾配を表す．また $\bar{g}_{t,i} = 1/t \sum_{\gamma=1}^t g_{\gamma,i}$.

v_{jh} を $\psi(d)$ の j 番目の要素から $\mathcal{H}_\text{V}$ の h 番目の要素へリンクの重みとする．特に t ステップ時の重みを $v_{jh}^{(t)}$ と書く．このとき，$t+1$ における重みは，式 3 の勾配 $g_{t,i}$ を以下の $v'_{jh}(t)$ に代入，$\bar{g}_{t,i}$ を再計算することで得られる．

$$\begin{aligned} v'_{jh}(t) &= \frac{\partial L}{\partial v_{jh}} \\ &= \psi(d)^{[j]} \sum_k^{N_e} w_{kh}^{(t-1)} (\phi(\ell^+)^{[k]} - \phi(\ell^-)^{[k]}) \end{aligned}$$

ただし，$L = 0.1 - \mathbf{G}(d)^\top \mathbf{F}(\ell^+) + \mathbf{G}(d_j)^\top \mathbf{F}(\ell^-)$. $w_{kh}^{(t-1)}$ は $\phi(\ell)$ の k 番目から $\mathcal{H}_\text{W}$ の k 番目のユニットを繋ぐリンクの $t-1$ 時の重みとする．$\phi(\ell)^{[k]}$ は $\phi(\ell)$ の k 番目のユニット，$\psi(d)^{[j]}$ は $\psi(d)$ の j 番目のユニットを指す．

一方，w_{ih} が $\phi(\ell)^{[i]}$ から $\mathcal{H}_\text{W}$ の h 番目のユニットを繋ぐリンクの重みを表すとすると，その $t+1$ 時の値 $w_{ih}^{(t+1)}$ は以下の勾配を式 3 の $g_{t,i}$ に代入，$\bar{g}_{t,i}$ を再計算することで得られる．

$$\begin{aligned} w'_{ih}(t) &= \frac{\partial L}{\partial w_{ih}} \\ &= (\phi(\ell^+)^{[i]} - \phi(\ell^-)^{[i]}) \sum_k^{N_v} v_{kh}^{(t-1)} \psi(d)^{[k]} \end{aligned}$$

3 実験

提案モデルの有効性を検証するため，オンラインで取得したニューヨークタイムズの 2013 年 6 月〜12 月の記事 19,952 本（以下，NYT2013）を対象に評価実験を行った．各記事には平均して 4 程度のトピックが人手付与されている，本稿ではこの付与トピックを正解として扱う．ウィキペディアは enwiki-20120902（2012 年 9 月ダンプバージョン）を用いた．ベースラインとして文献 [1] で紹介されている以下の 4 つのキーワード抽出手法を利用した．

TFIDF　テキストから {JJ}*{NN,NNS} にマッチする文字列を拾い，文字列の構成要素ごとに TFIDF を算出し合計したものを当該文字列の重要度スコアとする．

TEXT RANK (TRANK)　単語をノードと見立て近接した単語間でネットを構成してページランク（以下の式）を適用する．

$$\mathcal{S}(V_i) = (1-d) + d \sum_{V_j \in I(V_i)} \mathcal{S}(V_j) \frac{z\mathrm{w}_{ji}}{\sum_{V_k \in O(V_j)} \mathrm{w}_{jk}}$$

w_{ji} は単語 w_j から w_i に向かうエッジの重み．$I(V_i)$ は単語 V_i に直接接続している単語群，$O(V_j)$ は単度

表 1: ROUGE-W(s_1, s_2) の例

s_1	s_2	ROUGE-W
The United States of America	The United States of America	1
The United States	The United States of America	0.529
States	The United States of America	0.077

V_j が直接接続している単語群を表す．本稿では文献 [3] に従い，ダンピングファクターを 0.85 とした．

SINGLE RANK (SRANK)　SingleRank は基本的には TextRank に軽微な修正を加えたものである．単語の重要度の計算方法は同じであるが，TextRank と異なり，文字列を構成している名詞などの構成素に対して重要度を求め，それを合算するところに主たる違いがある．

EXPAND RANK (DRANK)　SingleRank は単一文書に対して PageRank を適用するが，ExpandRank は当該文書に近い文書を複数個選択し，PageRank を適用する．他の設定は同じ．なお，本稿では文献 [1] の著者が提供しているコードを利用した．[1]

ところで本稿では ROUGE-W と呼ばれる共通部分文字列の長さをベースにした尺度を用いて，出力トピックと正解ラベルの類似度を測った．例えば，正解トピックを ℓ，トップ k の候補ラベル集合を $\mathcal{C}|_k$ とする時，両者の合致度 $\mathcal{G}(\mathcal{C}|_k, \ell)$ を以下の式で測る．

$$\mathcal{G}(\mathcal{C}|_k, \ell) = \max_{c \in \mathcal{C}|_k} \text{ROUGE-W}(c, \ell)$$

トップ k のうちで正解トピックと類似度が最大になる候補のスコアが $\mathcal{G}$ の値となる．なお任意の ROUGE-W スコア v について $v \in [0, 1]$ が成立する（表 1 参照）．

4　結果と考察

実験結果を表 2 に示した．表の数値は文書毎に $\mathcal{G}$ を計測し平均したものである．MBT-X はウィキラベルの出力をそのまま使い，AMP は MBT-X の出力をニューラルネットで再ランクした．表では，AMP が最も優勢で，その後に MBT-X が続く，文献 [1] で「state of the art」と呼ばれた SRANK, DRANK, TFIDF はその後塵を拝している．

AMP は出現回数が 10 回以上の正例ラベルについてのみ学習を行い，そうでないものについては MBT-X の結果をそのまま利用する折衷型モデルである．ここで「正例ラベル」とは MBT-X が出力したラベルのうち，人手付与されたラベルとの ROUGE-W スコアが 0 を超えるものを指す．例えば，ある特定の文書について，人手付与されたラベル集合 $\mathbb{G}$ と MBT-X 出力ラベル $\mathbb{F} = \{a, b\}$ があるとき，$\exists_{y \in \mathbb{G}}$ ROUGE-W$(a, y) > 0$ なる a を「正例ラベル」，$\neg\exists_{y \in \mathbb{G}}$ ROUGE-W$(b, y) > 0$ なる b を「負例ラベル」と呼ぶ（以下では正例ラベルが MBT-X の出力に含まれる文書を「正例文書」，そうでないものを「負例文書」と呼ぶ）．AMP の目的は，MBT-X の出力ラベル群を，人手作成ラベルに近いものが上位に来るように並び替えることである．出現回数が 10 回以上の正例ラベルを持たない文書，つまり負例文書（のラベル）は再ランクの対象にしない．ちなみに NYT2013 では，再ランクの対象となった文書数は 1,811，全体の 9.1%である．月単位では約 258 となる．[2]　従って，**ほとんどの文書は再ランクの対象にならない．**

このように学習あり・なしモデルを混在させたトピック導出の方式を本稿ではアンフィビアンモデル (Amphibian Model (AMP)) と呼ぶ．上でも述べたが，AMP では正例を持たない文書については MBT-X の結果をそのまま利用し，正例を持つ場合はニューラルネットよる再ランクを実行する．

表 3 は正例文書を対象にした 10 回交差検定の結果を示している．再ランクモデルの $\psi(d)$ と $\phi(\ell)$ の長さ N_v, N_e はそれぞれ 15,164 と 3,247（図 1 のウグイス色と黄色の箱の長さ），隠れ層の長さ K は 60 とした (図 1 の青と橙の箱の長さ)．比較のためランキングサポートベクターマシン (SVM-R) を結果を併記している [2]．[3]　評価尺度は前と同様 $\mathcal{G}$ である．AMP では再ランクの効果が負例文書が圧倒的に多いため希薄化してしまうが，正例文書に限定するとその効果は歴然とする．

表 4 に各手法に出力ラベルの具体例を挙げる．この例は，**文書トピックと文書中のキーワードは必ずし**

[1] http://www.hlt.utdallas.edu/~saidul/code.html

[2] 本稿の実験では正例文書について，月単位で 10 回交差検定を用いて順位の入れ替えを行っている．

[3] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 2: NYT2013 の結果

k	TFIDF	SRANK	XRANK	DRANK	MBT-X	AMP
1	0.1157	0.1112	0.0702	0.1096	0.2101	**0.2279**
2	0.1889	0.1815	0.1289	0.1824	0.3146	**0.3382**
3	0.2448	0.2364	0.1773	0.2374	0.3810	**0.4068**
4	0.2882	0.2798	0.2176	0.2820	0.4284	**0.4548**
5	0.3252	0.3147	0.2523	0.3170	0.4632	**0.4905**

表 3: 10 回交差検定の結果

	NYT2013		
k	MBT-X	AMP	SVM-R
1	0.2606	**0.3432**	0.0441
2	0.4143	**0.4986**	0.0961
3	0.5098	**0.5817**	0.1368
4	0.5778	**0.6305**	0.1789
5	0.6240	**0.6677**	0.2202

表 4: NYT2013 の記事に自動付与されたトピックラベルの例．正解（人手付与トピック）は *Same-Sex Marriage*, *Civil Unions and Domestic Partnership*, *Macklemore (Rapper)*, *Ryan Lewis*, *Mary Lambert*.

k	TFIDF	SRANK	MBT-X	AMP
1	gay teenagers	gay slurs	Macklemore	Same sex marriage
2	gay marriage anthem	gay rights part	the United States	same sex marriage
3	gay rights part	gay godfather	Same sex marriage	Homosexuality
4	theme song	gay rights	Perry	Ex gay
5	gay marriage	gay marriage anthem	Homosextuality	Marriage Equality Act

も一致しないことを示す好例である．ここに TFIDF, SRANK を含む従来のキーワード手法の限界があると言える．なお今回は紙面の都合で割愛したが，NYT2013 と同様の結果が．フォックスニュース（18,163 記事）[4]，ガーディアン紙（23,438 記事）[5] を使った実験でも得られている．

5 おわりに

以上，非学習型モデルと学習型モデルを折衷したトピック導出方式について説明した．ニューヨークタイムズを使った比較的大規模なデータについて実験し，効果を確認した．特に現在スタンダードとされている幾つかのモデルを凌ぐ性能が得られたことは特筆に値する．ただし本方式を実際に導入するには，ウィキラベルが出力したトピックを低頻度型あるいは高頻度型に自動で分類する必要があるが，現段階では具体的な手続きが定かではない．将来的な課題としてはこの点に対応すること，Denoising Autoencoder などによる式 1 の Sr の改良などを考えている．

なお，以下サイトで非学習型のウィキラベル（英語，日本語）をデモ公開している．

► http://www.wikilabel.jp/ja

► http://www.wikilabel.jp/en

[4] http://www.foxnews.com
[5] http://www.theguardian.com

参考文献

[1] Kazi Saidul Hasan and Vincent Ng. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *COLING 2010*, pp. 365–373, 2010.

[2] T. Joachims. Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, pp. 217–226, 2006.

[3] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[4] Tadashi Nomoto. Mediameter: A global monitor for online news coverage. In *Proceedings of the First Workshop on Computing News Storylines*, pp. 30–34, Beijing, China, July 2015. Association for Computational Linguistics.

[5] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning*, Vol. 81, pp. 21–35, 2010.