D-003

単語と文の分散表現素性に着目したニュースストリームに対する文単位の新規性判定

Sentence-level Novelty Detection of News Streams focused on Feature extracted from Distributed Representation for Words and Sentences

田村 壮慶† 立間 淳司‡ 青野 雅樹‡
Masamichi Tamura Atsushi Tatsuma Masaki Aono

1. はじめに

近年，インターネットの浸透に伴い，様々なニュース記事に大量にアクセスできるようになってきた．膨大な記事の中でユーザは本当に新鮮な記事だけにアクセスしたい．そこで，本研究では英語のニュースストリームを入力とし，文単位で新規性があるかどうかを判定する方法を提案する．

2. 関連研究

これまでにニュースストリームに対する文単位での新規性判定に関して様々な手法が考案されている．Blott ら[1]はニュースストリーム中の以前の文に出現していない単語が閾値以上ある場合に新規性があるとしている．Gamon [2]はグラフ表現を用いて文を表現し，Support Vector Machine (SVM)で分類を行う手法を提案している．Karkali ら[3]は tf-idf 値を改良した独自のノベルティスコアが閾値以上の場合に新規性があるとしている．Lee [4]は単語分散表現の平均ベクトルで文を表現し，以前の文との類似度を素性の一つとして深層学習を行い，新規性判定を行っている．

3. 提案手法

まず，前処理としてニュースストリームを構成する文に対して単語単位の分割を行う．NLTK[5]の English stopwords に含まれる単語，及び記号等をストップワードとして除外する．得られた単語と文そのものを用いて以下に述べる素性を抽出し，学習器によって回帰を行う．

本研究で用いる素性とその抽出方法について述べる．

- **Unique Words (UW) :** Blott ら[1]の手法で用いられた，ニュースストリーム中の以前の文に出現していない単語の数を素性とする．

- **Bag-of-Words similarity (BoWsim) :** Bag-of-Words (BoW)素性では，単語の出現回数 (TF : Term Frequency) を重みとして文のベクトル化を行った．n 番目の文Sに対する TF ベクトルをS_n，同様に表現したそれ以前の文の集合を$C\{c_1,\cdots,c_{n-1}\}$とする．このとき次式のようにS_nとC の要素全てとのコサイン類似度（式(2)）を求め，式(1)のように，その中の最大値$\mathrm{BoWsim}(S_n, C)$を素性とする．

$$\mathrm{BoWsim}(S_n, C) = \max_{1\le i\le n-1} \cos(S_n, c_i) \qquad (1)$$

$$\cos(S_n, c_i) = \frac{S_n \cdot c_i}{\|S_n\|\|c_i\|} \qquad (2)$$

- **KL-divergence (KL) :** KL-divergence は確率分布同士の分布の差を表す距離的尺度である．n 番目の文S_nとそれ以前の文の集合Cとの KL-divergence を求めるため，単語wが文S_n，ならびにそれ以前の文の集合Cに現れる出現確率を$p_s(w), p_c(w)$ で表現する．$r(w)$は単語wが出現する回数を表す．ただし，S_n とCのどちらかに，ある単語wが出現しない場合，ゼロ除算問題が起こる．そのため，S_n に出現していない単語wがCに出現している場合には，S_n の単語wの出現確率に重みαをスムージングのため加算する．逆の場合も同様に行う．また，S_n に出現した単語でCに出現しなかった単語には重みβを加算することで，よりCとの差を強調させる．式(3)のようにp_s とp_c の KL-divergence を求め，素性とする．SW，CWはそれぞれS_n，Cに出現した単語の集合である．実験では，$\alpha = 0.1, \beta = 10$ とした．

$$\mathrm{KL}(p_s, p_c) = \sum_w p_s(w) \cdot \log \frac{p_s(w)}{p_c(w)} \qquad (3)$$

$$p_s(w) = \begin{cases} \dfrac{r(w)+\alpha}{\sum_w p_s(w)}, & w \notin SW \text{ and } w \in CW \\ \dfrac{r(w)+\beta}{\sum_w p_s(w)}, & w \in SW \text{ and } w \notin CW \\ \dfrac{r(w)}{\sum_w p_s(w)}, & \text{otherwise} \end{cases}$$

$$p_c(w) = \begin{cases} \dfrac{r(w)+\alpha}{\sum_w p_c(w)}, & w \in SW \text{ and } w \notin CW \\ \dfrac{r(w)}{\sum_w p_c(w)}, & \text{otherwise} \end{cases}$$

- **Word-Vectors similarity (WVsim) :** 単語分散表現 (distributed word representation) は，単語を多次元空間のベクトルとして表現することができる． ここでは学習済みの GloVe [6](50 次元)を用いて，文を構成する単語の単語分散表現を抽出し，その平均ベクトルで文を表現する．n 番目の文Sに対する GloVe の平均ベクトルをS_n，同様に表現したそれ以前の文の集合を$C\{c_1,\cdots,c_{n-1}\}$とする．このとき式(4)のようにS_nとC の要素全てとのコサイン類似度を求め，その中の最大値$\mathrm{WVsim}(S_n, C)$を素性とする．

$$\mathrm{WVsim}(S_n, C) = \max_{1\le i\le n-1} \cos(S_n, c_i) \qquad (4)$$

- **Hausdorff Distance similarity (HDsim) :** ハウスドルフ距離 (Hausdorff distance) は，2 つの集合間の距離である．文を単語の集合と考えると，文と文の距離は

† 豊橋技術科学大学 情報・知能工学専攻
‡ 豊橋技術科学大学 情報・知能工学系

それぞれの単語分散表現の集合間のハウスドルフ距離で表すことができる．2 つの集合 X と Y のハウスドルフ距離は次式で求められる．ここでは $d(x,y)$ にユークリッド距離を用いた．また，単語分散表現には学習済みの GloVe(50 次元)を用いた．

$$dH(X,Y) = \max\{\max_{x\in X}\{\min_{y\in Y} d(x,y)\}, \max_{y\in Y}\{\min_{x\in X} d(x,y)\}\}$$

n 番目の文Sに対する単語分散表現の集合をS_n，同様に表現したそれ以前の文の集合を$C\{c_1,\cdots,c_{n-1}\}$とする．このとき式(5)のように単語分散表現の集合 S_n とC の要素全てとのハウスドルフ距離を求め，その中の最大値HDsim(S_n, C)を素性とする．

$$\text{HDsim}(S_n, C) = \max_{1\le i\le n-1} dH(S_n, c_i) \qquad (5)$$

- **Sentence Vector similarity (SVsim)：** 文や文章レベルの分散表現を獲得する研究[7]に着目した素性である．ここでは gensim[8]に実装されている doc2vec を用いて 2009 年から 2015 年の Associated Press English news articles で学習されたモデル(300 次元)[9]を使用した．n 番目の文Sに対する文の分散表現をS_n，同様に表現したそれ以前の文の集合を$C\{c_1,\cdots,c_{n-1}\}$とする．このとき式(6)のようにS_n とC の要素全てとのコサイン類似度を求め，その中の最大値SVsim(S_n, C)を素性とする．

$$\text{SVsim}(S_n, C) = \max_{1\le i\le n-1} \cos(S_n, c_i) \qquad (6)$$

- **Recursive Feature Concatenation (RFC)：** Lee[4]の手法では，文S_n の素性にストリーム中で 1 つ前の文 S_{n-1} の素性を連結させる方法が用いられていた．ここでは 1 つ前に限らず，3 つ前までの文の素性を連結した素性とする．

4. 評価実験

提案手法の有効性を検証するために評価実験を行った．以下にデータセット，評価方法，実験結果について述べる．

データセットには TREC 2003, 2004 Novelty Track Data を用いた．このデータは各 50 個の話題のニュースストリームで構成されている。訓練データに 2003 年のデータ 15,557 文、テストデータに 2004 年のデータ 8,343 文を用いる．回帰のための学習器には Kernel Ridge Regression (KRR) を使用する．カーネルは RBF カーネルを使用する．実装には scikit-learn[10]を使用し，グリッドサーチによってパラメータ調整を行った。評価尺度には F 値 を用いた。

実験結果を表 1 に示す．ベースラインは，Unique Words (UW)のみ，Bag-of-Words similarity (BoWsim)のみ，および KL-divergence (KL)のみの 3 種類の素性それぞれで閾値処理を行い，新規性判定を行う方法とする．

比較手法は，UW，BoWsim，KL，Word-Vectors similarity (WVsim)，Recursive Feature Concatenation (RFC)素性を用いた学習器による回帰として実験を行った．また，手法の横の値はその手法に用いた素性の次元数である．表 1 から，提案素性である Hausdorff Distance similarity (HDsim)と Sentence Vector similarity (SVsim)を導入した場合に，優れた結果を得られたことがわかる．

表 1：実験結果

素性	F 値	手法
Unique Words (UW) 《1》	0.621	ベー
Bag-of-Words similarity (BoWsim) 《1》	0.614	スラ
KL-divergence (KL) 《1》	0.617	イン
UW + BoWsim + KL + WVsim 《4》	0.624	比較
UW + BoWsim+ KL + WVsim + RFC 《16》	0.625	手法
UW + BoWsim + KL + HDsim 《4》	0.625	
UW + BoWsim + KL + HDsim + RFC 《16》	**0.628**	提案
UW + BoWsim + KL + SVsim《4》	0.624	手法
UW + BoWsim + KL + SVsim + RFC 《16》	0.627	

5. おわりに

本研究では，ニュースストリームを入力とした，文単位での新規性判定を行った．2 つの文の距離を表す方法として，単語分散表現を用いたハウスドルフ距離素性（HDsim）と文の分散表現を用いた素性（SVsim)を提案した．これらの素性を用いて，回帰モデルでニュースストリームからの新規性判定を行った．評価実験では，単語分散表現の平均ベクトルで文を表現して類似度を測る素性よりも，文を単語分散表現の集合と考えてハウスドルフ距離を求める素性や，文の分散表現を用いて類似度を測る素性のほうが優れた結果を得られることがわかった．

参考文献

[1] Stephen Blott, Oisin Boydell, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Gareth J. F. Jones, Noel Murphy, Noel O'Connor, Alan F. Smeaton, Barry Smyth, and Peter Wilkins, "Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC-2004", (2004).

[2] Micheal Gamon, "Graph-based text representation for novelty detection", In Proceedings of the North American Chapter of the Association for Computational Linguistics, NAACL'06, pp.17-24, (2006)

[3] Margarita Karkali, Francois Rousseau, Alexandors Ntoulas, and Michalis Vazirgiannis, "Efficient Online Novelty Detection in News Streams", In Proceedings of the Web Information Systems Engineering 2013, WISE'13, Part I, LNCS 8180, pp. 57-71, (2013).

[4] Sungjin Lee, "Online Sentece Novelty Scoring for Topical Document Streams", In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP'15, pp. 567-572, (2015).

[5] Natural Language Toolkit (NLTK), http://www.nltk.org/

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP'14, pp.1532-1543, (2014).

[7] Quoc V. Le, and Tomas Mikolov, "Distributed Representations of Sentences and Documents", In Proceedings of International Conference on Machine Learning 2014, pp.1-9, (2014)

[8] Genism, topic models for humans, https://radimrehurek.com/gensim/

[9] Jey Han Lau and Timothy Baldwin, "An Empirical Evaluation of doc2vec with Pratical Insights into Document Embedding Generation", In Proceedings of the 1st Workshop on Representation Learning for NLP, pp.78-86, (2016).

[10] Scikit-learn, http://scikit-learn.org/stable/