Intra-Topic Variability Normalization based on Linear Projection for Topic Classification

Quan Liu[†], Wu Guo[†], Zhen-Hua Ling[†], Hui Jiang[‡], Yu Hu[†§]
[†] National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, China
[‡] Department of Electrical Engineering and Computer Science
York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada
[§] iFLYTEK Research, Hefei, China
emails: quanliu@mail.ustc.edu.cn, guowu@ustc.edu.cn, zhling@ustc.edu.cn
hj@cse.yorku.ca, yuhu@iflytek.com

Abstract

This paper proposes a variability normalization algorithm to reduce the variability between intra-topic documents for topic classification. Firstly, an optimization problem is constructed based on linear variability removable assumption. Secondly, a new feature space for document representation is found by solving the optimization problem with kernel principle component analysis (KPCA). Finally, effective feature transformation is taken through linear projection. As for experiments, state-of-the-art SVM and KNN algorithm are adopted for topic classification respectively. Experimental results on a free-style conversational corpus show that the proposed variability normalization algorithm for topic classification achieves 3.8% absolute improvement for micro-F_1 measure.

1 Introduction

Topic classification is now faced with the problem of enormous variability between documents due to the exponential growth of free-style unstructured texts in recent years. This paper treats variability as differences between text documents and aims at reducing the intra-topic document variability for better topic classification. There are various factors to cause the intra-topic variability problem, such as the different language usages of different persons (Chambers, 1995; Fillmore et al., 2014). In free-style conversations experimented in this paper, different people would use very different words to express their opinions. Therefore, documents in a same topic could be quite different because of the intra-topic variability problem.

In this work, we are interested in finding a robust document representation strategy to address the intra-topic variability problem. Traditional method represents document by a high-dimensional TF-IDF vector based on the bag-of-word approach (Salton and McGill, 1986; Salton and Buckley, 1988). However, the TF-IDF feature reveals little semantic similarity information between terms, which would increase the differences between intra-topic documents when different words are used. Beyond the TF-IDF strategy, there are two class of techniques, i.e., unsupervised technique and supervised technique for document representations. The unsupervised technique includes some latent semantic analysis methods. The typical method is Latent Semantic Indexing (LSI) while the features estimated by LSI are linear combinations of the original features (Deerwester et al., 1990; Wang et al., 2013). Meanwhile, the popular Latent Dirichlet Allocation (Blei et al., 2003; Morchid et al., 2014) algorithm was proposed to represent document by a generative probabilistic model (Blei et al., 2003). Moreover, in recent years, many neural network based methods have been investigated for document representations (Hinton and Salakhutdinov, 2006; Srivastava et al., 2013; Le and Mikolov, 2014). For example, in (Le and Mikolov, 2014), a model called *paragraph vector* was designed to represent each document by a dense vector while the vector is trained by predicting all words in the corresponding document. On the other hand, supervised technique for document representation includes some discrim-

Proceedings of NAACL-HLT 2016, pages 441–446,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

inative approaches, e.g., Linear Discriminant Analysis (Berry et al., 1995; Chakrabarti et al., 2003; Torkkola, 2004) and supervised latent semantic indexing (Sun et al., 2004; Chakraborti et al., 2007; Bai et al., 2009). Meanwhile, some improved linear analysis methods were proposed for encoding documents with a reliable similarity information (Yih et al., 2011; Chang et al., 2013). However, all those works for document representation paid little attention to the variability of intra-topic documents. Therefore, they could hardly solve the intra-topic variability problem in a direct way.

This paper makes a preliminary investigation to deal with the intra-topic variability problem. The main purpose of this work is to find a new feature space with minimized intra-topic variability. An objective criterion is constructed for optimization. Mathematically, we make use of the topic label information of the training set to create a weighting matrix, and then sum over all the differences of intra-topic documents. Then a robust feature space with minimized intra-topic variability is generated by solving the optimization problem with effective KPCA based algorithm. Finally, we accomplish the variability normalization operations for the baseline features. We also employ the linear discriminant analysis as a supplementary algorithm. As for experiments, state-of-the-art SVM and KNN algorithms are employed for topic classification. System performances are evaluated on a challenging free-style conversational database.

The rest of this paper is organized as follows. In section 2, we introduce the proposed variability normalization algorithm for topic classification in detail. After it, section 3 presents experimental setup and results. Finally, conclusions and future work would be given in section 4.

2 Variability Normalization Algorithm

2.1 Motivation for variability normalization

This work aims to find a robust document representation strategy for topic classification. The proposed algorithm is motivated by the Nuisance Attribute Projection (NAP) algorithm in speaker verification field (Solomonoff et al., 2005; Solomonoff et al., 2007). We firstly make a **linear variability removable assumption** for document representation. Mathematically, given a document, it could be denoted by a column vector $\mathbf{x}$ with dimensionality of d as follows

$$\mathbf{x} = \mathbf{x}_t + \mathbf{x}_v \tag{1}$$

where $\mathbf{x}_t$ denotes the useful signal information in current document, $\mathbf{x}_v$ stands for the remaining noise. It is very difficult to model the noise signal in a document since it could come from various sources. Therefore, in this paper, we focus on the noise created by the variability among intra-topic documents. Our goal is to find a new document representation through linear projection:

$$\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x} \tag{2}$$

where $\mathbf{P}$ is the projection matrix. Since the goal of this paper is not dimensionality reduction, the dimensionality of the new document representation is the same as the source document representation. Therefore, the size of $\mathbf{P}$ is $d \times d$. This paper proposes to learn $\mathbf{P}$ by minimizing the following intra-topic variability

$$\mathcal{Q} = \sum_{i,j} w_{ij} \|\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \tag{3}$$

where w_{ij} is the *i*-th row and *j*-th column element of a weighting matrix $\mathbf{W}$ created in this work. The matrix is determined by the topic label information of training set as follows

$$w_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to a same topic} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

2.2 Variability normalization algorithm

For deriving the variability normalization algorithm, we follow the work of (Solomonoff et al., 2007) and re-write the projection matrix $\mathbf{P}$ by the variability space (denoted as a unit vector $\mathbf{v}$ here) as follows

$$\mathbf{P} = \mathbf{I} - \mathbf{v}\mathbf{v}^{\mathrm{T}} \tag{5}$$

where $\mathbf{I}$ is a ($d \times d$) dimensional identity matrix. Combining (3) and (5), we get

$$\mathcal{Q} = \sum_{i,j} w_{ij} (\|\mathbf{x}_i - \mathbf{x}_j\|^2 - (\mathbf{v}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j))^2). \tag{6}$$

Since the first part of $\mathcal{Q}$ in (6) is independent on $\mathbf{v}$, we discard it and create the final criterion

$$\mathcal{Q} = -\sum_{i,j} w_{ij}(\mathbf{v}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j))^2). \quad (7)$$

Unfolding (7) by linear operation, we get

$$\mathcal{Q} = 2\mathbf{v}^{\mathrm{T}}\mathbf{X} \cdot (\mathbf{W} - \mathrm{diag}(\mathbf{W} \cdot \mathbf{1})) \cdot \mathbf{X}^{\mathrm{T}}\mathbf{v}. \quad (8)$$

where $\mathbf{X}$ denotes the training set matrix, each row of $\mathbf{X}$ represents one document vector, $\mathbf{1}$ is a vector with all elements equal to 1. Minimizing (8) is equivalent to solving the flowing eigenvalue decomposition problem

$$\mathbf{X} \cdot (\mathrm{diag}(\mathbf{W} \cdot \mathbf{1}) - \mathbf{W})\mathbf{X}^{\mathrm{T}}\mathbf{v} = \lambda\mathbf{v}. \quad (9)$$

Here we apply the idea of KPCA (Solomonoff et al., 2007; Schölkopf et al., 1997) to solve (9). Denoting $\mathbf{v}$ by a new vector $\mathbf{Xu}$, finding $\mathbf{u}$ turns to solving a generalized eigenvalue problem in kernel space as

$$\begin{cases} \mathbf{KZKu} = \lambda\mathbf{Ku} \\ \mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \\ \mathbf{Z} = \mathrm{diag}(\mathbf{W} \cdot \mathbf{1}) - \mathbf{W}. \end{cases} \quad (10)$$

The variability space is then constructed by selecting a set of eigenvectors corresponding to the d_1 largest eigenvalues.

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_{d_1}] \quad (11)$$

Finally, a $(d \times d)$ projection matrix is obtained by combining (5), (11) and $\mathbf{v} = \mathbf{Xu}$.

Based on this variability normalization algorithm, the baseline document vectors could be transformed to a new feature space with minimized intra-topic variability. The main procedure to implement intra-topic variation normalization could be divided into the following steps:

- Generate sample matrix $\mathbf{X}$ using the whole n documents of training set.
- Construct weighting matrix $\mathbf{W}$ according to (4) with the use of topic label information.
- Estimate a projection matrix $\mathbf{P}$ by solving the aforementioned eigenvalue problem.
- Transform all documents to new feature space through linear projection according to (2).

It should be noticed that after making feature transformation by the proposed variability normalization algorithm, the dimensionality of document representation has not been changed. This is different with all the existing dimensionality reduction methods since our goal is to re-define the feature representation space for topic document representation. To prove the effectiveness of the proposed algorithm, this paper presents experimental results on a challenging conversational dataset.

3 Experiments

In this section, we evaluate the proposed variability normalization method in a typical topic classification problem. We will firstly introduce the experimental setup, including dataset, evaluation criteria and system description. After it, all the experimental results would be reported in detail.

3.1 Experimental setup

3.1.1 Dataset

The data set used in this paper is the text transcripts of free-style conversational speech database, Fisher English corpus released by LDC, which contains 11699 recorded conversations (Cieri et al., 2004). This corpus is collected from 40 different topics, and each document includes relatively a distinct topic (e.g. "Comedy", "Smoking", "Terrorism", etc.) as well as topics covering similar subject areas (e.g. "Airport Security", "Bioterrorism", "Issues in the Middle East"). This paper randomly chooses 60 documents and 50 documents per topic for the training set and testing set respectively. Another 50 documents for each topic are randomly selected to for the development set.

3.1.2 Evaluation criteria

We use two types of criteria to make a comprehensive evaluations for this work. The first evaluation creterion is F_1 measure corresponding to the recall and precision rates for a typical classification system. In detail, we would report micro-average F_1 and macro-average F_1 results. In consideration of topic classification is similar to topic verification, we choose equal error rate (EER) to be the second criterion, which is the equal value of miss probability and false probability.

3.1.3 System description

Module	Methods
Text processing	stop-word removal, stemming
Representation	TF-IDF feature
Classification	KNN, SVM algorithm

Table 1: Baseline system modules for topic classification.

This paper constructs several systems for comparison. The configurations of our baseline system are shown in Table 1. Porter algorithm (Porter, 1980) is adopted for word stemming after stop-words removal. Then a vocabulary with 19534 unique words is determined according to the occurrence frequency information of training set. Documents in the baseline system are represented by using the popular TF-IDF term weighting strategy (Salton and Buckley, 1988). Two popular algorithms SVM and KNN are used for classification separately. The SVM classification is implemented using the LIBSVM toolkit (Chang and Lin, 2011).

Based on the baseline system, descriptions of other systems are given as below.

(1) *LSI*: documents are represented in latent semantic space estimated by the LSI algorithm (Deerwester et al., 1990) based on the baseline features.

(2) *LDA*: document features are transformed by linear discriminant analysis. We select 50 eigenvectors for the low dimensional feature space.

(3) *VarNorm*: document features are transformed from the baseline TF-IDF vectors by the approach proposed in this paper. We select 60 eigenvectors for generating the project matrix.

(4) *VarNorm-LDA*: system combined *VarNorm* with *LDA*, which employs feature transformation operations twice on the original TF-IDF document features. The number of eigenvectors for *VarNorm* and *LDA* are set to 60 and 50 respectively.

All the parameters suggested in this paper are tuned on the development set. However, the eigenvector number is not restricted to 50 or 60. It is recommended to set the eigenvector num from 45 to 75 since we have 40 topics for experiments.

3.2 Experimental Results

3.2.1 Variability normalization performance

According to (3), we compare the intra-topic variability for the baseline and the *VarNorm* system. The difference for variability calculation is whether to use the projection matrix $\mathbf{P}$ or not. Figure 1 shows the intra-topic variability on 40 topics of the training set. The vertical axis represents the variability for each topic, while the horizontal axis stands for 40 topics in the conversation corpus. As we can see clearly, the variability of baseline system is high. After conducting variability normalization, it could be reduced effectively.

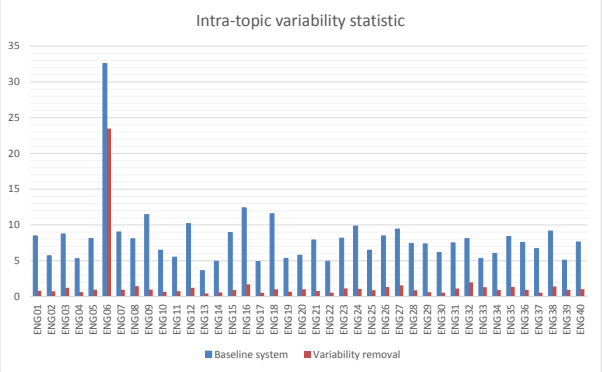

Figure 1: Variability normalization performance.

After making detailed analysis, we find for the topic ENG06, the theme is "*Hypothetical Situations: Perjury – Do either of you think that you would commit perjury for a close friend or family member?*", the variability among documents from this topic is largest in the whole corpus. However, for the topic ENG13, "*Movies: Do each of you enjoy going to the movies in a theater, or would you rather rent a movie and stay home? What was the last movie that you saw? Was it good or bad and why?*", the variability is the lowest. This is the difference between common topics and infrequent topics. Since people would use various words to express their ideas, it is reasonable to find the variability problem is more serious for infrequent topics than common topics.

3.2.2 Classification Results using KNN

Experimental results using KNN classification algorithm are given in Table 2. The results show that, compared to the baseline system, the variability normalization system *VarNorm* achieves 2% absolute F_1 improvement, and 29% relative improvement for EER. When taking the variability removing as a preliminary process, and employing LDA as the secondary transformation, the system *VarNorm-LDA* achieves the best performance. The EER is im-

Table 2: Classification results using KNN algorithm

System	EER	macro-F_1	micro-F_1
Baseline	6.10	84.72	83.15
LSI	4.25	86.24	85.45
LDA	4.49	88.94	88.15
VarNorm	4.30	86.46	85.60
VarNorm-LDA	**2.51**	**90.29**	**90.00**

Table 3: Classification results using SVM algorithm

System	EER	macro-F_1	micro-F_1
Baseline	3.40	88.86	88.40
LSI	3.35	89.59	89.25
LDA	3.05	90.81	90.55
VarNorm	**2.90**	**91.04**	**90.80**
VarNorm-LDA	**2.50**	**92.28**	**92.15**

proved by 65% relatively, and the micro-F_1 measure is improved by 6.85% absolutely. The reason for this performance is straightforward. Since the proposed algorithm effectively reduce the differences among intra-topic documents, the LDA algorithm would be more easier and effective to maximize the ratio of between-class-variance to within-class-variance.

3.2.3 Classification Results using SVM

Similarly, the experimental results using SVM classification algorithm are shown in Table 3. The baseline performance is better than system using KNN algorithm. The improvements achieved by *LSI* in KNN sytem almost vanish here, while the *VarNorm* system keeps its improvement. The *VarNorm* system even works better than the *LDA* system, with nearly 15% relative improvement on EER, and 3.4% absolute improvement on micro-F_1 measure. The best results are obtained by the *VarNorm-LDA* system. There are 36% relative improvement for EER, and 3.75% absolute improvement for micro-F_1 measure.

4 Conclusions and Future Work

In this paper, we investigated the intra-topic variability problem for topic classification. The major contribution of this work is that we proposed a effective variability normalization approach for robust document representation. An optimization problem was constructed after making a linear variability removable assumption. In order to take a deep insight into the performance of the proposed variability normalization algorithm, we conducted experiments on a challenge free-style conversation corpus. Experimental results based on the SVM and KNN classification algorithm all confirmed the robustness of the proposed approach. As a conclusion, the variability normalization algorithm could be used as a front-end feature transformation strategy, and we also suggest to combine it with linear discriminant analysis algorithm or some other algorithms to further improve system performances.

Further study will investigate the adaptive methods for constructing robust feature spaces. We would also combine this work with more document representations methods as well. Moreover, it would be very interesting to extend and combine our work to some novel unsupervised machine learning techniques, like the work of (Zhang and Jiang, 2015) while they proposed a model for high-dimensional data by combineing a linear orthogonal projection and a finite mixture model under a unified generative modeling framework.

Acknowledgments

This work was supported in part by the Science and Technology Development of Anhui Province, China (Grants No. 2014z02006) and the Fundamental Research Funds for the Central Universities (Grant No. WK2350000001). At the same time, we want to give special thanks to the anonymous reviewers for their insightful comments as well as suggestions.

References

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. 2009. Supervised semantic indexing. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 187–196. ACM.

Michael W Berry, Susan T Dumais, and Gavin W O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Soumen Chakrabarti, Shourya Roy, and Mahesh V Soundalgekar. 2003. Fast and accurate text classifica-

tion via multiple linear discriminant projections. *The VLDB Journal*, 12(2):170–185.

Sutanu Chakraborti, Rahman Mukras, Robert Lothian, Nirmalie Wiratunga, Stuart NK Watt, and David J Harper. 2007. Supervised latent semantic indexing using adaptive sprinkling. In *IJCAI*, pages 1582–1587.

Jack K Chambers. 1995. *Sociolinguistic theory*. Blackwell.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. 2013. Multi-relational latent semantic analysis. In *EMNLP*, pages 1602–1612.

C. Cieri, D. Miller, and K. Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *LREC*, pages 69–71.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Charles J Fillmore, Daniel Kempler, and William SY Wang. 2014. *Individual differences in language ability and language behavior*. Academic Press.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Mohamed Morchid, Richard Dufour, and Georges Linares. 2014. A lda-based topic classification approach from highly imperfect automatic transcriptions. *LREC14*.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

B. Schölkopf, A. Smola, and KR. Müller. 1997. Kernel principal component analysis. In *Artificial Neural Networks-ICANN'97*, pages 583–588. Springer.

Alex Solomonoff, William M Campbell, and Ian Boardman. 2005. Advances in channel compensation for svm speaker recognition. In *ICASSP*, pages 629–632.

A. Solomonoff, W. M. Campbell, and C. Quillen. 2007. Nuisance attribute projection. *Speech Communication*.

Nitish Srivastava, Ruslan R Salakhutdinov, and Geoffrey E Hinton. 2013. Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865*.

Jian-Tao Sun, Zheng Chen, Hua-Jun Zeng, Yu-Chang Lu, Chun-Yi Shi, and Wei-Ying Ma. 2004. Supervised latent semantic indexing for document categorization. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 535–538. IEEE.

K. Torkkola. 2004. Discriminative features for text document classification. *Formal Pattern Analysis & Applications*, 6(4):301–308.

Quan Wang, Jun Xu, Hang Li, and Nick Craswell. 2013. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Transactions on Information Systems (TOIS)*, 31(1):5.

Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *CoNLL*, pages 247–256. Association for Computational Linguistics.

Shiliang Zhang and Hui Jiang. 2015. Hybrid orthogonal projection and estimation (hope): A new framework to probe and learn neural networks. *arXiv preprint arXiv:1502.00702*.