# Analysis on Chinese quantitative stylistic features based on text mining

**Renkui Hou and Minghu Jiang**
Laboratory of Computational Linguistics, School of Humanities, Tsinghua University, Beijing, China

## Abstract

In this article, data mining was selected to examine whether some linguistic features, taking parts of speech (POS) for instance, can be used as Chinese quantitative stylistic feature. It can be also said that the purpose of this article is to explore the method to determine the Chinese quantitative stylistic features. Texts of different styles, which are news, science, official, art, TV conversation, and daily conversation styles, were selected to establish the corpus for our study. Text vectors characterized by POS were analyzed by principal component analysis and clustered by agglomerative hierarchical clustering method. The results of them indicate that POS can be used as a distinctive feature of texts. Then, support vector machine was adopted to establish classification model on training data and precision and recall rates to validate the results of text classification. Random forest was selected to compute the importance of POS, i.e. the contribution to classification, and text vectors characterized by important POS were clustered and classified consequently. The results of the experiments show that POS can be taken as Chinese quantitative stylistic feature, and the results of clustering and classification are preferably taking the 60 most important POS as the character of texts.

**Correspondence:**
Renkui Hou, Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China.
**E-mail:**
hourk0917@163.com

## 1. Introduction

Style refers to the principles and rules generated in communication and followed by speakers and listeners, the essence of which is to adjust the psychological distance between the communicators (ShengLi 2010). Style and linguistic performance are interdependent and cannot be tenable without each other. To some extent, the research of stylistic features and rules amounts to that of the governing rules behind linguistic performance, which may contribute to the guidance of actual linguistic performance. To generate natural and fluent utterance, it is necessary to follow stylistic rules.

With the help of some stylistic linguistic means, people involved in a verbal communication can construct a type of stylistic discourse by means of certain expressions mixed with large amount of other neutral linguistic materials according to a certain context (Dechun 2000). The use of a kind of linguistic means has its inevitability and contingency, which can be described with statistical probability in mathematics. Therefore, style can be measured through a quantitative analysis approach which is more objective and scientific in explaining linguistic features of style. The inevitable linguistic means can be seen as the quantitative feature of the correspondingly style. As the confirming of the

Digital Scholarship in the Humanities, Vol. 31, No. 2, 2016. 
doi:10.1093/llc/fqu067 Advance Access published on 8 December 2014

stylistic features, the corresponding style can be defined as the set of quantitative features (Stamatatos 2001.).

### 1.1 Research question and conclusion

This article aims to examine whether Part of Speech can be taken as a Chinese stylistic feature selecting principal component analysis (PCA), text clustering, and classification as our approaches. The results of PCA and text clustering preliminarily indicate that POS can be used as the distinctive feature of different style texts. The support vector machine (SVM) was adopted to establish classification model on training texts, and the classification result on test texts was validated. The experimental results can help us determine whether the POS can be taken as a stylistic feature. Then, we adopted random forest to compute the importance of parts of speech (POS), i.e. the contribution to the classification. Then, the same clustering and classification process were performed. The POS, which make the clustering and classification perform better, can be selected as stylistic features and further influence the style class of texts by means of the comparison of experiments results.

### 1.2 The structure of this article

The related research about quantitative stylistics will be introduced in the second part, and after that we talk about the establishment of the corpus in this study. In the fourth part, we focus on the experimental process including PCA, text clustering, text classification, and the importance of POS by random forest. Finally, we discuss experimental results and further research.

## 2. Related Research

Quantitative analysis applied to linguistic research was first adopted by English mathematics De Morgan in 1851. With word length taken as statistic features, he conducted a quantitative research on the style of text. It is generally believed that Biber is the first linguist who used the quantitative method to study the style (Biber 1986, 1988, 1992, 1995). Markov and Zipf made great contributions to the quantitative study of language (Liu 2009), and Yule marked a starting point of using modern statistics in quantitative stylistic research (Juhan 2005). Many studies gave evidence to show that the distributed differences of linguistic structures in different styles are objective (Biber 1988; Swales 2001; Stamatatos et al., 2001; Mario and Constantinos, 2006; Takafumi et al., 2012). Biber (1995) pointed out the use of statistical methods in style processing has proved to be a reliable approach. Mannion and Dixon (2004) pointed out that in the case of Goldsmith, then, sentence-length may be considered a reliable stylistic marker. Hoover (2002) pointed out that analyses based on frequent word sequences constitute improved tools for authorship and stylistic studies. Compared with the research abroad, the quantitative feature studies of styles of Chinese texts involve the study on Harry Potter series (Chen 2007), the study on 'reading news' and 'saying news' from the perspective of quantitative linguistics (Hou 2010), the study of Chinese quantitative stylistic features based on corpus (Huang 2007), the comparative study on reading the news and saying the news (Liu et al., 2011), and so on.

Former research on quantitative linguistic features and data mining mainly focuses on selecting different linguistic elements as vector features for text clustering and classification, and examining the results to improve feature selection and algorithm. For instance, Liang et al. (2006) proposed a language model adaptation based on the classification of a trigram's language style feature. Meng and Hou (2009) select discourse markers as vector features to classify texts. Gao and Feng (2011) construct text vectors for clustering based on word class dependency relations. Feldman et al. (2009) select POS histogram as vector feature and take PCA to reduce dimension for genre classification. Biber (1993b) provides statistics showing differences in the frequency of usage of different POS or syntactic structures for fiction versus exposition, focusing on ambiguous cases. Some research (Iyer and Ostendorf, 1999; Schwarm et al., 2004) has shown part-of-speech differences associated with different types of conversational speech, news text, and e-mail. Zhang (2012) pointed out that POS can be used as the distinctive feature of Chinese written

style adopting quantitative methods and correspondence analysis. We adopted text mining as an approach to examine whether POS can be taken a Chinese quantitative stylistic feature, to compute the importance of POS and classified texts characterized by POS.

## 3. Corpus Establishment

The class of style needs to be considered in order to explore whether the POS can be used as a Chinese quantitative stylistic feature. In this light, researchers of different periods have conducted a wealth of research. According to Dechun (2000), style is divided into conversational and written style, with the latter being subdivided into the artistic and the practical style which includes scientific, official, report, and political styles. All the aforementioned styles are commonly used in practical linguistic performance. Yuan and Li 2005 classify style into seven classes—conversational, official, scientific, news, literary and art, lecture, as well as advertisement styles. Shengli (2010) thinks that style is a polarized opposite continuum with formal written style and daily colloquial style in two poles, and other styles lying in between. The discrete class of style is adopted for the convenience of study although we accept the theory that style is a polarized opposite continuum.

Distinction and opposition make style. Only by comparison can we discover the features of language used in different styles and the quantitative features of style. In combination with the style class studies of predecessors, texts of news, science, official, art and daily conversation, and TV conversation styles were selected to establish corpora.

Texts of *News Co-broadcasting* were selected as news style texts. Research papers in the disciplines of computer, agriculture, and economics were selected as scientific style texts. Surveys, white papers, decisions, and notices were selected as official style texts. Both of them were downloaded from the Internet. Prose written by *Zhu Ziqing*, *Yu Qiuyu*, *Lin Yutang*, and *Lu Xun* was selected as the corpus of artistic style. *Behind the Headlines with Wentao* was selected as the corpus of tv conversational style with every day's program as a text. Colloquial style was composed of some dialogs from the sitcom 'I Love My Family'; downloaded from the Internet, each one of the 120 episodes was treated as a text.

Biber (1990, 1993a) pointed out that a text length of 1000 words is adequate for representing the distributions of many core linguistic features of a stylistic category, it is possible to represent the distribution of many core linguistic features of a stylistic category based on relatively few texts from each category.

A Chinese lexical analysis system created by the Institute of Computing Technology of Chinese Academy of Science (hereafter abbreviated as ICTCLAS) was adopted for text segmentation and part-of-speech tagging. ICTCLAS has been acknowledged with a high accuracy of 97.58%, a recall rate of over 90% for the recognition of unknown words based on role tagging, and a recall rate of approximate 98% for the recognition of Chinese names.[1] The POS were subdivided by ICTCLAS according to the need of machine learning and the classification of the POS of contemporary Chinese language, for example, where 把 and 被 are viewed as a subclass of preposition.

The occurrence frequencies of Part of Speeches were calculated to establish text vectors.

## 4. Experiments

### 4.1 Principal component analysis

Firstly, PCA was adopted to initially examine the relationship of the text vectors characterized by POS. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (i.e. accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e. uncorrelated with) the preceding components. The relationship of texts, see the Figure 1, was

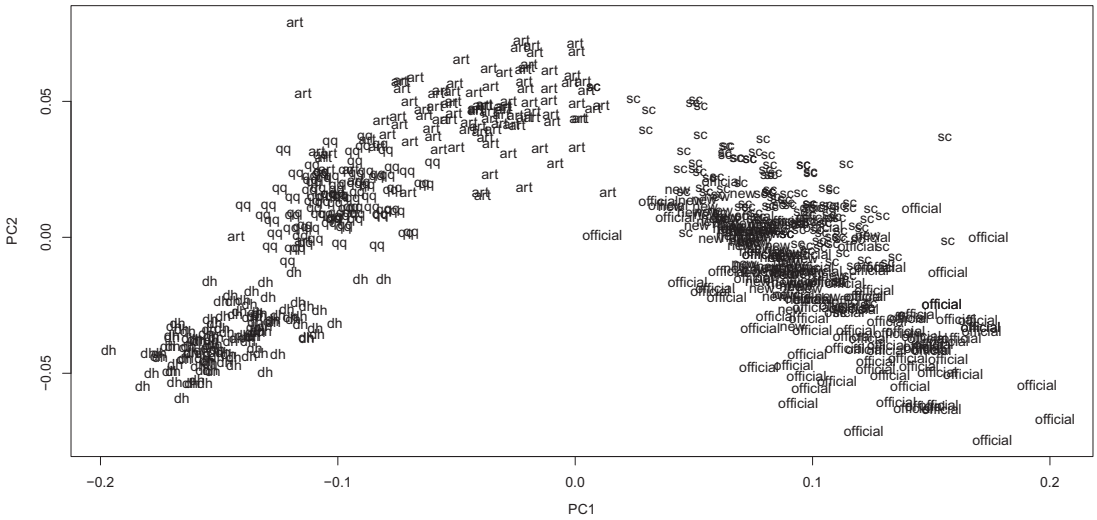


**Fig. 1** The results of PCA, ''dh'' stands for daily conversation text, ''qq'' for tv conversation text, ''art'' for artistic text, ''sc'' for scientific text, ''new'' for news text, ''official'' for official text

**Table 1** The importance of principal components

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| Standard deviation | 0.1048 | 0.03097 | 0.02122 | 0.01888 | 0.01778 | 0.01498 |
| Proportion of variance | 0.7658 | 0.06690 | 0.03142 | 0.02485 | 0.02205 | 0.01566 |
| Cumulative proportion | 0.7658 | 0.83273 | 0.86415 | 0.88900 | 0.91105 | 0.92671 |

obtained by the PCA, the importance of principal components is shown in Table 1.

In the Table 1, only the six principal components with the largest possible variance were listed. From that, we can see the cumulative proportion of variance of first two principal components surpasses 80%, and that of the first three components achieves 86%. So, the relationship of texts, obtained by PCA and seen from Figure 1, is reliable. From Figure 1, we can see that the daily conversation, as well as TV conversation and art, texts have the compact internal cohesion and are far from the other texts. There is more overlap between the texts of other three styles, especially news and science styles.

3D was selected to demonstrate the relationship between texts, i.e. the text was represented by first three principal components with the highest variance. From Figure 2, we can see that the texts of science and news overlapped in figure 1 are separated, and the texts of science, news, and official are far from the other styles texts.

Combined with Figures 1 and 2 and Table 1, the relationship between the six style texts can be preliminary observed. There is a good internal similarity between texts of each style and separation between texts of different styles. It can be thought that POS can distinguish the texts of different styles.

## 4.2 Clustering analysis

Clustering is a common data-mining method. Clustering, which does not know the number and relation of clusters on data in advance, is an unsupervised learning algorithm and partitions data into groups (clusters) such that the data points in a cluster are more similar to each other than the points in different clusters according to the data

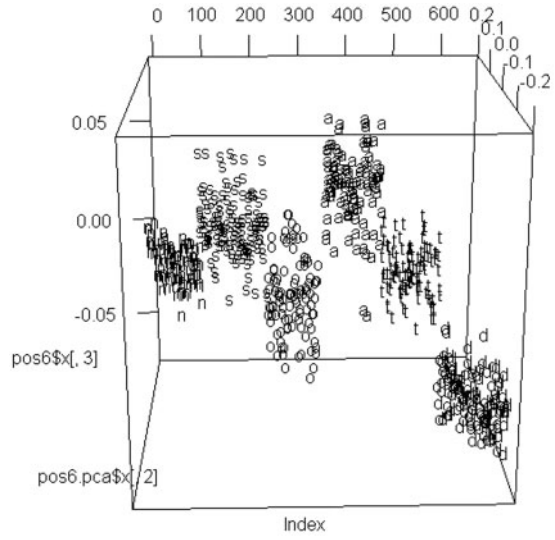


**Fig. 2** The results of PCA, ''n'' stands for news text, ''s'' for science text, ''o'' for official text, ''a'' for artistic text, ''t'' for tv conversation text, ''d'' for daily conversation text

themselves and their relation. Generally speaking, a better clustering result can be defined as a high similarity between the data points of the same cluster and a greater variance between the data points of different clusters (Manning and Schutze, 1999). Text clustering was herein not the goal but an approach to determine whether POS can serve as a stylistic feature. If the clustering result was satisfactory, the POS can be considered as a stylistic feature.

Agglomerative hierarchical clustering method was adopted to cluster the texts of different styles in this section. From each text as a cluster, this clustering method proceeds successively by merging the most similar smaller clusters into larger ones until one cluster is remained. The Euclidean distance was selected to measure the distance between text vectors. The sum of deviation square, was used to calculate the similarity between clusters. The number of cluster need not to be appointed in advance before texts are clustered. From the result of hierarchical clustering, as well as PCA, we can roughly observe the relationship between the texts. The result of the algorithm is a tree of clusters, called dendrogram, which shows the relationship between clusters and how the clusters are related. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained (Halkidi et al., 2001). From this cutting, the table of clustering result can be obtained. Since clustering algorithms discover clusters, which are not known a priori, irrespective of the clustering methods, the final partition of data requires some sort of evaluation in most applications (Rezaee et al., 1998). The supervised and unsupervised methods, i.e. entropy and cophenetic correlation coefficient (CPCC) were selected to validate the text clustering result.

The dendrogram, seen in Figure 3, was obtained after the text vectors were clustered. From Figure 3, we can see that the texts are clustered into approximate six clusters. The Table 2 was obtained by cutting the dendrogram to see the composition of every cluster. Each cluster is mainly composed of one style texts, for instance, cluster 1 is news style texts and cluster 5 is daily conversation texts. It can be seen in Table 2 that each cluster has low entropy. The weighted entropy of all clusters is 0.306, and the CPCC between cophenetic matrix of the dendrogram and distance matrix of the texts is 0.81. This means the clustering result is satisfactory. From the hierarchical clustering result, it can be seen that POS can be taken as a distinctive feature of these six styles. If the dendrogram was cut five clusters, science and news style texts were merged together. It means that the usage of POS between science and news styles is similar. This also explained that these two style texts are overlapped in PCA.

## 4.3 Text classification

Different from clustering, classification algorithm, which knows the number and relation of class, is a supervised learning method. It builds classification model on training data to judge the class of unknown data point. SVM is one of the widely used classification algorithm and can be expressed as a convex optimization problem, and therefore we can take advantage of the existing efficient algorithm to find the global minimum of the objective function. SVM was selected to classify the texts vectors characterized by POS. The precision and recall rates were selected to validate the text classification result.

According to 3:2, the texts of each style were divided into training and test data. The classification model was built on the training data by SVM. Then,

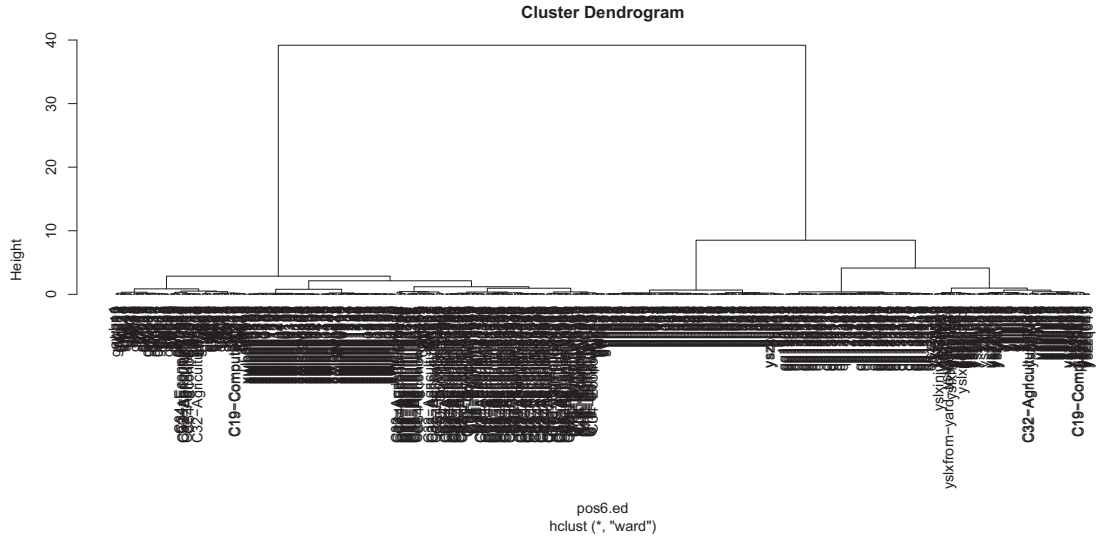


**Fig. 3** The dendrogram of hierarchical clustering

**Table 2** The result of hierarchical clustering

| Labels | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Daily conversation | 0 | 0 | 0 | 0 | 120 | 0 |
| Offical | 3 | 21 | 0 | 81 | 0 | 0 |
| TV conversation | 0 | 0 | 0 | 0 | 0 | 101 |
| Science | 0 | 121 | 4 | 9 | 0 | 0 |
| News | 99 | 1 | 0 | 0 | 0 | 0 |
| Art | 0 | 0 | 103 | 0 | 2 | 1 |
| Entropy | 0.191 | 0.660 | 0.230 | 0.469 | 0.121 | 0.079 |

the test texts were classified by the classification model, and the classification result is shown in Table 3.

It can be seen in the Table 3 that the classification result is satisfactory, in which the precision rates of all style texts surpass 93%, especially that of conversation and news style texts (100%), and the recall rates surpass 94%, especially that of conversation and art style texts (100%). This result verified the above conclusion: POS can be taken as a quantitative stylistic feature. It also indicated that the POS can be used a feature of text classification.

## 4.4 Importance of POS

In the above experiments, all of the POS were selected to represent the text of different styles. We did not take into account whether every POS was helpful to text clustering and classification. Some POS may influence classification performance if they were selected as classification features. Occam's razor tells us that it is unnecessary to choose all POS as features if fewer POS can reach better classification result. Furthermore, the POS with little contribution to classification may be mistaken as the quantitative stylistic features for distinguish different style texts and lead to an over-fitting problem.

Random forest is an ensemble classifier making use of decision tree as base classifier (Breiman 2001). This algorithm selects a subset of input features to form each of the training set. The subset can

**Table 3** The classification result of SVM

| | Daily conversation | Official | TV conversation | Science | News | Art | Recall |
|---|---|---|---|---|---|---|---|
| Daily conversation | 48 | 0 | 0 | 0 | 0 | 0 | 100% |
| Official | 0 | 40 | 0 | 1 | 0 | 1 | 95.24% |
| TV conversation | 0 | 0 | 41 | 0 | 0 | 0 | 100% |
| Science | 0 | 2 | 0 | 51 | 0 | 1 | 94.44% |
| News | 0 | 1 | 0 | 0 | 39 | 0 | 97.5% |
| Art | 0 | 0 | 0 | 0 | 0 | 43 | 100% |
| Precision | 100% | 93.02% | 100% | 98.07% | 100% | 95.55% | |

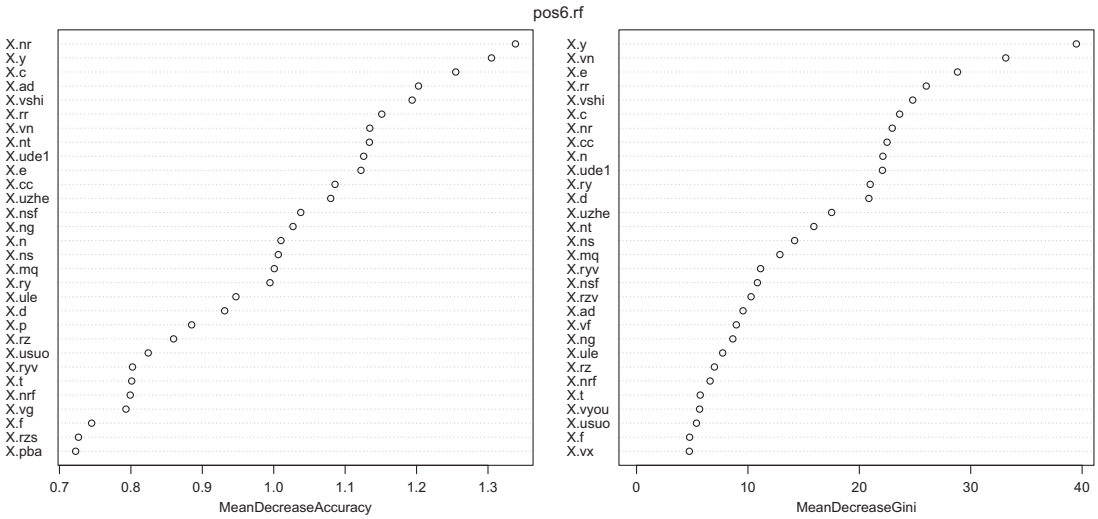


**Fig. 4** Importance of POS (expressed in MeanDecreaseAccuracy and MeanDecreaseGini)

be selected randomly or according to the recommendations of field experts. Some studies show that the performance of this method is very good for the data set containing a lot of redundant features. Random forest can be used for classification, regression, and to compute the importance of properties, i.e. the contribution to classification. We select random forest to compute the contribution of POS to style classification. After training, you can see which POS are more important.

Random forest was adopted to build classification model on training data. The importance of POS is shown in Figure 4.

In Figure 4, MeanDecreaseAccuracy measures the reduced degree of accuracy predicted by random forest when the value of a variable can be changed into a random number. The higher MeanDecreaseAccuracy of one variable indicates the greater importance of it. MeanDecreaseGini computes the impurity-level influence of variables to compare the importance of them. Similarly, the higher MeanDecreaseGini of one variable indicates the greater importance of it. There is a slight difference between them, but the difference is not obvious.

It can be seen from the Figure 4 that the importance of each POS is different. This is consistent with our intuitions. MeanDecreaseAccuracy was selected as the importance value of POS in next section.

## 4.5 Clustering and classification taking different important POS

In this section, we selected POS with different importance to represent texts, and then clustered

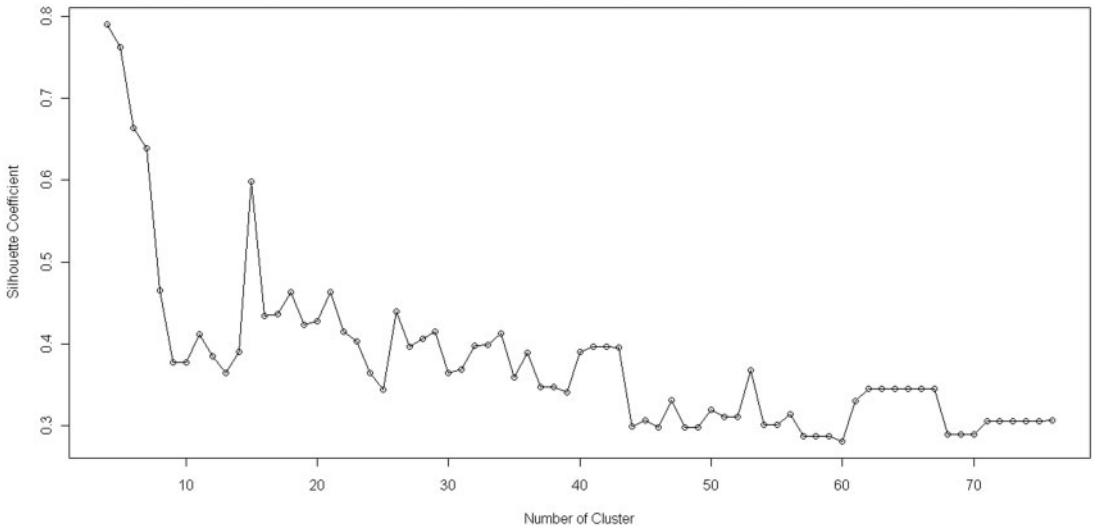


**Fig. 5** The weighted entropy of hierarchical clustering of texts represented by different POS

and classified them to observe their performance respectively.

Firstly, all POS were ordered by the importance and then were selected to represent texts. Secondly, the texts were hierarchical clustered, and the weighted entropy of all six clusters was obtained to examine the clustering result. Thirdly, the POS with minimum importance value was removed from the text vectors. Repeating the processes 2 and 3, we can observe that how many POS should be selected in order to the weighted entropy is least, i.e. the hierarchical clustering result is best. The corresponding relation between the number of POS and weighted entropy of all six clusters is shown in Figure 5.

In Figure 5, the weighted entropy of hierarchical clustering is minimal when the most 60 important POS, with importance value greater than 0.4, were selected to represent text. The composition of each cluster in this clustering is shown in Table 4. The weighted entropy was 0.281 and less than 0.306 obtained from hierarchical clustering when the texts were represented by all POS.

The 60 most important POS were selected to establish text vectors. The texts were divided into training and test data according to 3:2. Classification model was built on training data by SVM and was used to classify test data, and the result of classification is shown in Table 5.

Combined with Tables 4 and 5, it can be observed that the results of hierarchical clustering and classification, when the 60 most important POS were selected to establish text vectors, are better than that of selecting all POS. According to the principle of Occam's razor, we think that these 60 POS should be selected when the texts were classified according to style.

## 5. Conclusions and Further Study

In terms of text clustering and classification, previous research mainly concentrates on algorithm itself which deals with how to improve algorithm and how to select text vector features for improving clustering and classification effect, etc. This research focuses on whether some certain language means, taking POS for instance, can be taken as stylistic features to distinguish texts of different styles with the research tools of text mining, including PCA, text clustering, and classification. Bringing text clustering and classification into quantitative stylistics offers new thoughts for studying the quantization research of text genre and work style and deepens

**Table 4** The hierarchical clustering result taking 60 most important POS

| Labels | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Daily conversation | 0 | 0 | 0 | 0 | 120 | 0 |
| Official | 3 | 21 | 0 | 81 | 0 | 0 |
| TV conversation | 0 | 0 | 0 | 0 | 0 | 101 |
| Science | 0 | 125 | 2 | 7 | 0 | 0 |
| News | 99 | 1 | 0 | 0 | 0 | 0 |
| Art | 0 | 0 | 103 | 0 | 2 | 1 |
| Entropy | 0.191 | 0.649 | 0.136 | 0.401 | 0.121 | 0.079 |

**Table 5** The classification result on test data of SVM

| | Daily conversation | Official | TV conversation | Science | News | Art | Recall |
|---|---|---|---|---|---|---|---|
| Daily conversation | 48 | 0 | 0 | 0 | 0 | 0 | 100% |
| Official | 0 | 40 | 0 | 1 | 1 | 0 | 95.24% |
| TV conversation | 0 | 0 | 41 | 0 | 0 | 0 | 100% |
| Science | 0 | 1 | 0 | 53 | 0 | 0 | 98.15% |
| News | 0 | 0 | 0 | 0 | 40 | 0 | 100% |
| Art | 0 | 0 | 0 | 0 | 0 | 43 | 100% |
| Precision | 100% | 97.56% | 100% | 98.15% | 97.56% | 100% | |

the study of quantitative linguistics to some extent. Maybe several articles have already proved that POS can be used to quantify Chinese stylistic features (Zhang 2012), our study pointed out that not all POS can be used to do it and contribute to stylistic classification of texts, and can obtain better classification result using more important POS.

POS were selected and their occurrence frequencies were counted for establishing text vectors. The PCA, text clustering, and classification results tell us that POS can be taken as the quantitative stylistic feature of different styles selected in this paper.

The importance of POS, computed by random forest, tells us that not all of POS are helpful to classification. The results of experiments show that clustering and classification results on text vectors with the most 60 important POS are better. Occam's razor principle tells us that given two models with the same generalization error, a relatively simple one is more preferable than a complex one. We think that more important POS are preferable in text classification.

Choosing more texts of more different styles to establish corpus, with the above analysis methods, to examine whether POS can be taken as a distinctive feature and obtain better clustering and classification results, is our further research.

Meanwhile, choosing sentence length, word length, the sentence-initial word POS, sentence category, syntactic characteristics, and other language elements to establish text vectors, the above analysis method is used to examine whether they are taken as distinctive features of different styles. Selecting more language sample of more different styles to establish corpus, PCA and clustering analysis were adopted to examine whether these language features can be combined to better represent discourse of certain style. Selecting random forest to examine importance of these features, we can choose different feature combinations to achieve better classification. In addition, based on the determination of the quantitative stylistic features, the existing style class system should be considered to optimization because they are not enough in natural language processing. With in-depth study of natural language processing, text genre classification will demonstrate an increasingly important role. For example, speech recognition models and syntactic parsing models

trained on *News Co-broadcasting* corpus cannot achieve the desired accuracy on conversation, such as *Behind the Headlines with Wentao*, because of the different styles they belong to. This is the reason for us to study stylistic feature extraction and stylistic classification of texts.

## Acknowledgement

We would like to thank the anonymous LLC reviewers for their insightful and valuable comments. Their suggestions have greatly improved the earlier manuscript of this paper.

## Funding

This work was supported by the National Natural Science Fund (61171114 and Key Fund (61433015), and National Social Science Major Fund of China.

## References

**Biber, D.** (1986). Spoken and written textual dimensions in English: resolving the contradictory findings. *Language*, **62**(2): 384–413.

**Biber, D.** (1988). *Variation across Speech and Writing*. Cambridge, England: Cambridge University Press.

**Biber, D.** (1990). Methodological issues regarding corpus-based analysis of linguistic variations. *Literary and Linguistic Computing*, **5**: 257–69.

**Biber, D.** (1992). The multidimensional approach to linguistic analyses of genre variation: an overview of methodology and finding. *Computers in the Humanities*, **26**(5/6): 331–47.

**Biber, D.** (1993a). Representativeness in corpus design. *Literary and Linguistic Computing*, **8**(4): 243–57.

**Biber, D.** (1993b). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2): 219–41.

**Biber, D.** (1995). *Dimensions of Register variation: A Cross-Linguistic Comparison*. Cambridge, England: Cambridge University Press.

**Breiman, L.** (2001). Random forests. *Machine Learning*, **45**(1): 5–32.

**Chen, L.** (2007). *The Stylometric Analysis of the Harry Potter Series (I-VI)[D]*. Dalian, China: Dalian Maritime University.

**DeChun, W. and RuiDuan, C.** (2000). *YUTIXUE[M]*. . GuangXi Nanning: Guangxi Education Press.

**Feldman, S., Marin, M. A., Ostendorf, M., and Gupta, M. R.** (2009). Part-of-speech histograms for genre classification of text In. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Washington, DC, pp. 4781–4.

**Halkidi, M., Batistakis, Y., and Vazirgiannis, M.** (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, **17**: 107–45.

**Hoover, D. L.** (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*, **17**(2): 157–80.

**Gao, S. and Feng, Z.** (2011). Research on text clustering based on dependency treebank. *Journal of Chinese Information Processing*, **25**(3): 59–63.

**Hou, R.** (2010). *A Computational Stylistic Analysis of "MaBin Reading" and "News Broadcast"*. Beijing, China: Communication University of China.

**Huang, W.** (2007). *Quantitative Study of Chinese Stylistic Features Based on Corpus*. Beijing, China: Communication University of China.

**Iyer, R. and Ostendorf, M.** (1999). Relevance weighting for combining multi-domain data for n-gram language modeling. *Computer Speech & Language*, **13**(3): 267–82.

**Juhan, T.** (2005). Stylistics, author identification. In Köhler, R., Altmann, G. and Piotrowski, R. (eds), *Quantitative Linguistics: An International Handbook*. Berlin: Walter de Gruyter, pp. 368–86.

**Kohler, R., Altmann, G., and Piotrowski, R. G.** (2005). *Quantitative Linguistics: An International Handbook*. New York: Walter de Gruyter, pp. 368–86.

**Liang, Qi., Zheng, F., Mingxing, X., and Wenhu, W.** (2006). Language model adaptation based on the classification of a trigram's language style feature. *Journal of Chinese Information Processing*, **20**(4): 68–74.

**Liu, H.** (2009). *Dependency Grammar from Theory to Practice*. Beijing: Science Press.

**Liu, Y. and Hu, F.** (2011). A comparative study of stylistics between "Reading News" and "Talking News". *Language Teaching and Linguistic Studies*, **1**: 97–104.

**Manning, C. and Schutze, H.** (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

**Mannion, D. and Dixon, P.** (2004). Sentence-length and authorship attribution: the case of Oliver Goldsmith. *Literary and Linguistic Computing*, **19**(4): 497–508.

**Mario, C-B. and Constantinos, C.** (2006). A stylometric analysis of newspapers, periodicals and news scripts. *Journal of Quantitative Linguistics*, **13**(2/3): 285–312.

**Meng, X. and Hou, M.** (2009). Research on stylistic feature of the discourse markers and its application. *Journal of Chinese Information Processing*, **23**(4): 34–9.

**Rezaee, R., Lelieveldt, B. P. F. and Reiber, J. H. C.** (1998). A new cluster validity index for the fuzzy c-Mean. *Pattern Recognition Letters*, **19**: 237–46.

**Shengli, F.** (2010). On mechanisms of register system and its grammatical property. *Studies of the Chinese Language*, **5**: 400–12.

**Schwarm, S., Bulyko, I., and Ostendorf, M.** (2004). Adaptive language modeling with varied sources to cover new vocabulary items. *IEEE Transactions on Speech and Audio*, **12**(3): 334–42.

**Stamatatos, E., Fakotakis, N. and Kokkinakis, G.** (2001). Automatic text categorization in terms of genre and author. *Computational Linguistics*, **26**(4): 471–95.

**Swales, J. M.** (2001). *Genre Analysis, English in Academic and Research Setting*. Shanghai: Shanghai Foreign Language Education Press.

**Takafumi, S., Shuntaro, K., and Akiko, A.** (2012). Stylistic analysis of text submissions to Japanese Q & A communities. *Journal of Quantitative Linguistics*, **19**(4): 262–80.

**Yuan, H. and Li, X.** (2005). *Outline of Chinese Style*. Beijing: The Commercial Press, China, P3.

**Zhang, Z.** (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, **8**(1): 209–40.

## Note

1 http://http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/200706/t20070628_2121143.html, 2013-1-31.