Advances in Social Science, Education and Humanities Research, volume 403
International Academic Conference on Frontiers in Social Sciences and Management Innovation (IAFSM 2019)

Hidden Space and Segmented Labelling in Text Visualization

Xiaoguang ZHU[1,a,*], Xin CAI[2,b] and Peiyao NIE[1,c]

[1]School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, P. R. China;

[2]Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, College of Information Science and Technology, Donghua University, Shanghai, P. R. China;

[a]xizhu@mail.sdufe.edu.cn, [b]xcai@dhu.edu.cn, [c]pynie@sdufe.edu.cn

[*]Corresponding author

Keywords: hidden space, segmented labelling, feature reduction, text visualization, model interpretation

Abstract. Text visualization can interpret large size documents with various linguistic units and glyphs. Each unit owns its advantages of intuition and precision, which could be visualize under different space efficiencies. For example, histogram of word frequency is an intuitive glyph but not precise, and word embedding could be optimized globally but not intuitive. Previous studies have applied many linguistic units and glyphs with implicit combinations, but lack an approach to align those units explicitly to sustain interpretability and predicability. In another side, ever growing methods of feature reduction and selection require a framework to compare and interpret hidden spaces with regard to large volume documents. To align and visualize linguistic units intensively, we proposed a visualization method to interpret document with its distribution on continuous space. Also, the accuracy of the segmented labelling is compared with the min-max and entropy methods. The result shows that: 1) our visualization is flexibility and efficiency to exhibit large volume documents; 2) of feature selection accuracy, the segmented labelling has comparable advantage on various parameters of hidden space.

1. Introduction

Interpretation of document and statistical model is an important task in knowledge discovery applications. From traditional topic model to recent deep neural network, studies have proposed many approaches to improve the interpretation with text visualization methods[1,2,3,4].

Most interpretation approaches have been proposed toward two components: one is labelling with abstract words or categories, and another is visualization with discernable glyphs. For large volume documents, efficient interpretation relies on three components: 1) feature reduction to discover novel hidden space with sustained predictability; 2) feature selection or labelling to generate prominent and readable features; 3) visualize them with various and discernable glyphs. In statistics, feature reduction will be more applicable if original space was sparse, especially, if applied supervised model and specific cost function.

When labelling hidden variables, words are ranked along each dimension with different criteria of similarity and coherence. A dimension may denote a topic, a centroid or a hidden feature. For classification models, feature selections are usually conditioned on response variable to sustain accuracy and performance of the classification [4,1]. In previous visualizations, many approaches have been proposed to visualize linguistic units with different glyphs and morphemes, which are usually sensitive to document volume heterogeneously if data hasn't been aggregated [5,6,7]. For example, histogram of word frequency is not sensitive to the document volume, which means, even the volume is changed from 10^2 to 10^6, we still could keep figure size of the histogram and the original level of discernment. In contrary, matrix plot of hidden space is sensitive to the document

volume. When plotting a matrix, one document added will cause a new row appended to the figure.

This size sensitivity promoted approaches of sampling, feature reduction and projection from original space to language resources and hidden spaces. As linguistic theory indicated, two linguistic units may connote the same concept or entity in language resources, e.g., explicit feature selection or semantic analysis can connect a linguistic unit to a term in encyclopedia [8]. In perspective of cognitive association, word usually contains prototype in a semantic field, hence, many continuous spaces have been pre-trained to enhance discriminant models, e.g., *GloVe*. For different tasks, the hidden space may be trained separately, and be initialized with a pre-trained continuous space, e.g., neural networks may exhibit stochastic issue which has been named convergent learning [9]. This means, selected features should be interpreted along specific statistical model to distil knowledge of the original feature space.

In nutshell, powerful feature reduction and labelling requires a swift approach to interpret them with readability and predicability. This interpretation will confront issues to align different linguistic units and interpret the predicability of labelling methods.

Given a scatter plotting, we can segment its axis with regard to linearity of the dimensions. In previous studies, labelling methods assumed the linearity of different hidden spaces, e.g., the linearity has been validated in several topic models [10,2]. One essential source of the linearity is the non-negative of bag-of-word model. In contrary, most visualizations of neural network depend on weight analysis and labelling of exemplars, which present attentions or prominent features with accented areas [9,3].

In summary, large volume documents and convergent learning requires a text visualization with space efficiency and aligned linguistic units to interpret the documents and feature reductions efficiently. Therefore, we propose a text visualization method with segmented labelling to interpret documents with global distribution and aligned labelling along each dimension in hidden space. Through several experiments, we validated the space efficiency and comparable accuracies of our method.

The outline of the paper is as follows. Section [overview] design symbols and frameworks to support our method, and explain space efficiency and interpretability of various glyphs and figures. Section [labelling] is our method and rationales. Section [experiments] validate our method from examples and comparisons in classification tasks. The last two sections are discussion and conclusion.

2. Overview

2.1. Symbols

The symbols are derived from the general process of language modelling and statistical learning, which are listed in table 1.

Table 1: Primary symbols

Symbol	Denotion
$D = \{d_1, \dots, d_N\}$	A set of documents.
$V = \{v_1, \dots, v_M\}$	Vocabulary of the D from word or other linguistic units.
$X \in \mathbb{R}^{N \times M}$	Aligned design matrix from the D, V, e.g., word bag model.
$H^i \in \mathbb{R}^{N \times K^i}$	A hidden matrix, from feature reduction or neural network
H^{ei}, H^{li}	Denote explicit space and embedding, usually in first layer of network.
$H^i(d) \in \mathbb{R}^{N \times K^i}$	A hidden vector from input document d which could be a single word.
$\beta^i \in \mathbb{R}^{K^{i-1} \times K^i}$	Weights related to H^i, e.g. eigenvectors in PCA.
$L^i \in \mathbb{R}^{K^i \times K\prime^i}$	Labels of the matrix H^i.
$L^i \in \mathbb{R}^{K^i \times K\prime^i \times K\prime\prime^i}$	Segmented labels of the matrix H^i.

For a specific layer, we will use H to denote its hidden matrix, and use β to denote its weight matrix which connects the H to its previous layer. In text processing, *alignment* implies *design*. This means, we select linguistic unit at first, like word or phrase, and then, align document with the selected unit to generate design matrix.

Based on the table, we can use subscript to index matrix along priority of N, M, K. For example, $H_i(d)$ will denote ith row of a hidden space of document d, and $H_{:,j}(d)$ will denote its jth column.

2.2. The common framework

Text visualization has four components: pre-processing, feature reduction, labelling and plotting. The first and the last components are stabilized under current maturity of information processing. Hence, we will explain the middle two components with examples in a unified framework. Feature reduction and labelling will generate H, L from a design matrix X. As a dense matrix, H could be inspected by the real numbers itself. Within a geometry space, many distance functions could be applied to transform the H to a graph, and read out global information from layouts of the graph.

In detail, the four components can be ordered linearly as follows.

1. Pre-processing. Extract design matrix from a document set. The matrix could be enriched from pre-trained models or language resources H^e, H^l.
2. Feature Reduction. Transform the matrix to a hidden space, e.g., matrix decomposition or optimized neural network model.
3. Labelling. Select prominent words or other linguistic units to represent the hidden space with interpretability.
4. Plotting. Exhibit the variables in table 1 with glyphs.

The unified framework and independence between the components can increase compatibility of our method to resolve neural network, topic model, or even clusters. Our segmented labelling is an extension of the labelling approaches with two criteria are being inherited.

2.3. Problem statement and general approach

Novel feature reductions and neural models are being developed to tackle intricate problems in language processing. For example, language resources and recurrent neural model have been widely applied to enrich the original feature space. From the large vocabulary and the stochastic convergence of complicated models, the optimization may converged to lot of local optimums with indiscernible accuracies. The in-discernibility means that many optimized feature spaces may achieve same level of accuracy and interpretability. Therefore, many labelling methods were proposed to generate coherent labels which can discover more stable hidden topics.

Generally, the generated labels are set of words, and labelling methods are classified in studies [11]. For neural network model, prominent features or attentions could be calculated from values of optimized weight and hidden matrix. In another word, it is decidable that a specific node is effected by features which affects the node with weights and activation functions. Hence, many applications have visualized network networks with weight matrix directly. For complex activations, like attention and tensor production, the weight can be reflected by accents on an exemplar.

In summary, topic model and neural model have been labelled and visualized with may approaches, but the labels haven't been attached to the figure with unified interpretation. Therefore, we align the labels along dimensions in hidden space, and interpret the sustained accuracy with tasks of classification and language modelling.

3. Segmented Labelling

Of definition, labelling is a kind of description, such as, assign a short description to each of the topical clusters to facilitate interpretations of the topics [12], and assign description to dimensions

of the hidden space [13]. When predicting response variable, the topics or hidden space could be represented by the assigned labels.

Previous studies have proposed labelling methods along different criteria of relevance and discriminative [11]. With confirm of the criteria, we will extend the labelling methods through segmentation of range of each hidden dimension. Given a specific hidden space H, the segmentation will divide its dimension to areas, and label each area through the following methods which are implemented from the criteria.

In detail, our design rationale behind the segmented labelling is as follows. Firstly, in hidden space, first several dimensions can load primary variance and predicability, like slope of eigenvalues and mean decrease in impurity indicated. Secondly, additive mechanism of neural network could support composition of moderate features to predict target value. In applications about language expression, words are semantically composed and comprehended with dependence on different functions like content word and function word. Hence, heterogeneous features could achieve higher accuracy in classification tasks. Thirdly, there is a dual phenomenon of the additive mechanism that was presented as feature homogeneous in traditional feature selection methods [2,14]. Whence association or language resources are considered, features may overlap themselves under a semantic field. For example, it's unnecessary to select all words of beauty, pretty, gorgeous, if we have thesaurus or literacy to associate them together.

3.1. Relevance

3.1.1. Methods

Relevance criteria will select prominent words with maximal values. In a hidden matrix $H(V) \in \mathbb{R}^{M\times K}$, the first label of the j[th] hidden dimension will be: $L_j = argmax_i H(V)_{i,j}$. In previous studies, it has been implemented with: zero-order relevance [11], integrated gradients to attribute predictability [15], and viewed with $tanh$(memory vector)[3].

The traditional calculations are equations 1 and 2, which could select labels with hidden matrix or conditional probabilities to response variable.

$$L_j = \underset{v\subset V}{argmax} \sum_{v_i\in v} |H_j(v_i)| \tag{1}$$

$$L_j = \underset{v\subset V}{argmax} \sum_{v_i\in v} p\,(y=1|H_j(v_i)) \tag{2}$$

where, $v \in \mathbb{R}^{K'}$ denotes selected labels for each dimension. Then, each H will be labelled by $L \in \mathbb{R}^{K\times K'}$ along its K dimensions.

Based on the equations, our extensions are equation 3 or 4, which divide each dimension to K'' areas, and label each area with K' words. Let a_k denotes the k[th] area of a dimension, then, its labels will be generated as follows.

$$L_{j,k} = \underset{v\subset V}{argmax} \sum_{v_i\in v} |H_j(v_i)| \times I(H_j(v_i) \in a_k) \tag{3}$$

$$L_{j,k} = \underset{v\subset V}{argmax} \sum_{v_i\in v} p\,(y=1|H_j(v_i)) \times I(H_j(v_i) \in a_k) \tag{4}$$

where, $L \in \mathbb{R}^{K\times K'\times K''}$, $v \in \mathbb{R}^{K'}$, and the operator $a \times b$ means element-wise product between two vectors.

3.1.2. Characteristics

The segmentation with relevance could sustain accuracy with following reasons.

1. Magnitude. In hidden matrix, a larger element will contribute more to the accuracy of prediction. For a hidden layer H^i in an optimized model, its succeed prediction $P(y|H_j^i)$ will be different to responses, otherwise, the model will lost its predictability.
2. Composition. General features could be combined to activate a specific semantic, e.g., combination of lot of different colours could activate semantic of colourful. General words have large range of connotation and efficiency of composition when discern objects. Hence, they usually have higher frequency than specific features at terminal areas. As an essential process to discern objects, we usually compose general features to activate different specific semantics. Length of a document will imply its dependence to the general features. For example, if we assign each semantic a specific feature, then, document should have length of one word, which could discern a small subset of responses. With increase of general features, the document length will increase, and more composed features will appear, e.g., deviation from an established pattern versus eccentricity.

If we introduce structured language resources into the labelling, then, text comprehension could be transformed to a logical inference. Hence, the segmentations will be transformed to logical terms in the new extended conceptual or taxonomic space. The most trivial approach is introducing ontological commitment, which maps the general features to specified terms in ontology system.

3.2 Discriminative

3.2.1. Methods

In topic model, labels could be selected through distance between word vector and topic vector, that is: $L_j = am_v sim(H(v), H(D))$. In previous studies, it has been implemented with: relevance score [11], cosine [16], select nearest neighbours in embedding [13], χ^2 and generic title generation [17].

Similar to labelling of clusters, the segmentation implied a partition of documents, hence, we can label an area through the document subset in that area. Here, a document subset will function as a centroid in clustering. With the equation 5, the selected words could discriminate its D^k from the other documents.

$$L_{j,k}=\underset{v\subset V}{argmax}\sum_{v_i\in v} sim\,(H(v_i), H(D^k)) \tag{5}$$

$$D^k=\{d_i|H_j(d_i)\in a_k\} \tag{6}$$

Where, $sim(v_i, D^k)$ denotes similarity function between word and document subset, which may implemented as K-L divergence, mutual information or $\chi^2(word, centroid)$[17]. The $H(v), H(D)$ could be substituted by other spaces, such as kernel and explicit semantic space [14,8,17]. For their substitutional spaces, our method could sustain the same calculation and interpretation.

3.3.2. Characteristics

The similarity function $sim()$ could be extended to enhance coherence and discernment of the selected labels. To analyse utility of the discriminative method, we divide the accuracy along the segmented areas, where, each area includes a document set. Rationale of the division is as follows. If each area has severe unbalanced distribution along response variable, and the selected labels can predict that which area a document is belongs to, then, the labels will achieve a high accuracy for the predication task.

In figure 1, the plotting could be segmented to areas, e.g., $a_{1,6}$ denotes a rectangle area. If we have exhibited response values as colours, like figure 1 did, we could count classes in each area, e.g., $a_{1,6} = [5,1]$ means that this area includes five positive documents and one negative document. Let A denote all areas, then, we could conclude that accuracy of selected labels could be approximated by $\frac{1}{N}\sum[m(a)|a \in A]$. This approximation is applicable to multiple classes predication as well.

Given this approximation through divided areas, the segmented discriminative labelling could sustain the original accuracy through predicability of the segmented areas.

4. Experiment and Evaluation

In this section, we will validate compatibility and accuracy of our method through examples and comparison of accuracies of classification tasks. With three datasets, we will evaluate our methods as follows: 1) validate the unification and compatibility by analysis of examples and its adaption to feature reduction methods; 2) compare labelling performances between our methods and feature selections in previous studies.

4.1 Data

Common tasks in language processing are selected to validate our method, which are listed as follows. All datasets have relatively large volume of document and vocabulary, hence, require feature selection to interpret their learning models.

1. Sentiment of Movie review data(MRD). Collections of movie review documents labelled with respect to their overall sentiment polarity (positive or negative).
2. Sentiment of Amazon product review(AMZ). This dataset contains product reviews and metadata from Amazon. We applied overall star as response variable which contains five classes.
3. Tagging of Stack-exchange platform. This dataset originates from the Stack Exchange data dump, and applied in Kaggle competition of *Transfer Learning on Stack Exchange Tags*.

The prediction and sequence tagging have been widely applied in platforms through user-generation contents or marketing analysis. That means, the platform documents are necessary to be interpreted with global distributions.

4.2 Evaluated Methods

This section analyses the parameters and cost functions which reflected their discrepancies at implementation level. For accuracy analysis, we construct interfaces of the compared labelling methods in table 3, which are capable to adapt various implementation of the criteria in section [labelling], and sustain the same interpretation of labels.

4.2.1. Feature Reduction

A design matrix could be transformed to a new feature space with feature reductions, e.g., $V' \in V$ with higher frequencies, or $s' \in \{sim(d_i, d_j)\}$ like Gaussian process. In those reductions, the generated hidden spaces are usually continuous, i.e., with a real number matrix $h = x\beta$. Our method could accept variety methods of reduction and transformation, if the features have been aligned and projected to response values. In detail, table 2 presented several methods that could be interpreted identically.

Table 2: Feature Reduction and Cost Function

Name	Parameter	Objective
PCA	$u \in \mathbb{R}^{m \times k}$	$m\|uu^T x - x\|_2$
SNE	$s \in \mathbb{R}^{m \times k}$	$m \sum \sum KL\,(f_{x,i,j} \parallel f_{y,i,j}); f_{x,i,j} \propto e(-\|x_i - x_j\|^2)$.
PLS	$\alpha \in \mathbb{R}^{m \times k}$	$mCorr(y, X\alpha)var(X\alpha)$
NPLM	$h_w \in \mathbb{R}^k, \beta \in \mathbb{R}^{k \times n}$	$m \sum D\,(h_w \beta, p(w\|w_c))$
DNN	$f(), \beta$	$m\|f(X, \beta) - y\|_2 + \|\beta\|_p$

4.2.2. Feature Selection

Based on the analysis in [labelling], we compare our methods to previous methods of global selection and dimensional labelling, which are summarized in table 3. The global selections are ranking features with explicit correlation or information theory. The correlation could be extended with implicit similarity to adapt co-linearity and kernel function, e.g., the nearest neighbour and iterative selection [18]. In topic model and our segmentation, labels are generated along hidden dimensions. Compare to the global labelling, the dimensional labelling has the same information theoretical foundation, but considered coherences that inner or inter dimensions. In addition, semantic field of a dimension could be interpreted explicitly like categorization in dictionary learnings.

Notice that we ignored methods about the best subset selection. In their methods, select a feature requires a process of optimization and estimation, hence, it is not suitable for huge feature space.

Table 3: Feature Selection and Cost Function

ID	Name	Parameter	Cost
ENT	Info. Theory	K'	$L = \underset{\tau \subset [0,\ldots,\|V\|]}{am} \sum I\,(X_{:,i}, Y)$
MMX	Min-Max	K'	$L_i = \underset{\tau \subset [0,\ldots,\|V\|]}{am} \sum \|H_{i,:}\|_1$
SE+*	Segmented	K', K''	our methods, see section [labelling]

In the table, K' is limiting the vector τ, i.e., $\tau \in \mathbb{R}^{K'}$.

The information theoretic mutual information can be expressed by entropy:
$I(X, Y) = H(X) + H(Y) - H(X, Y)$.

where, the first two terms denote entropy, and the third one denote divergence.

Given those methods, we will reduce feature space with neural network model. Given an optimized hidden space, we will select labels with the methods and identical sizes to compare their accuracies.

4.3 Compatibility

In this section, we present two document sets with our method *SE+P* and *SE+D*. From the examples, we will explain their unified interpretation and space efficiency. And, how could those labels could be comprehended to support concept learning and inference.

4.3.1. Example 1: MRD

In this example, we study hidden space of *MRD* sentiment analysis. By choosing a low dimensional layer in an optimized neural network, we can generate figure 1, which contains movie reviews and labels from method *SE+P*. The figure implied a hidden space of words which are calculated from iterative dot product. The iteration is similar to cascading activation of neuron system, and the lefter layer are affect its righter layer through a weight matrix.

Compare to previous plottings, the figure could exhibit global distribution with alignment from documents to their labels. In the figure 1, each area implies a semantic field which indicated by

labels on corresponding axes. From the middle areas to the terminals, the labels are becoming emotional to predict the response values. For example, the extreme negative words include emotions of `pathetic, meandering`, probably ironic of `anything` and transition of `nevertheless`. In concept learning, those words could be integrated to language resources to enhance accuracy of sentiment analysis. In applications like market analysis, the figure are capable to exhibit their primary emotions with frequencies of words and documents.

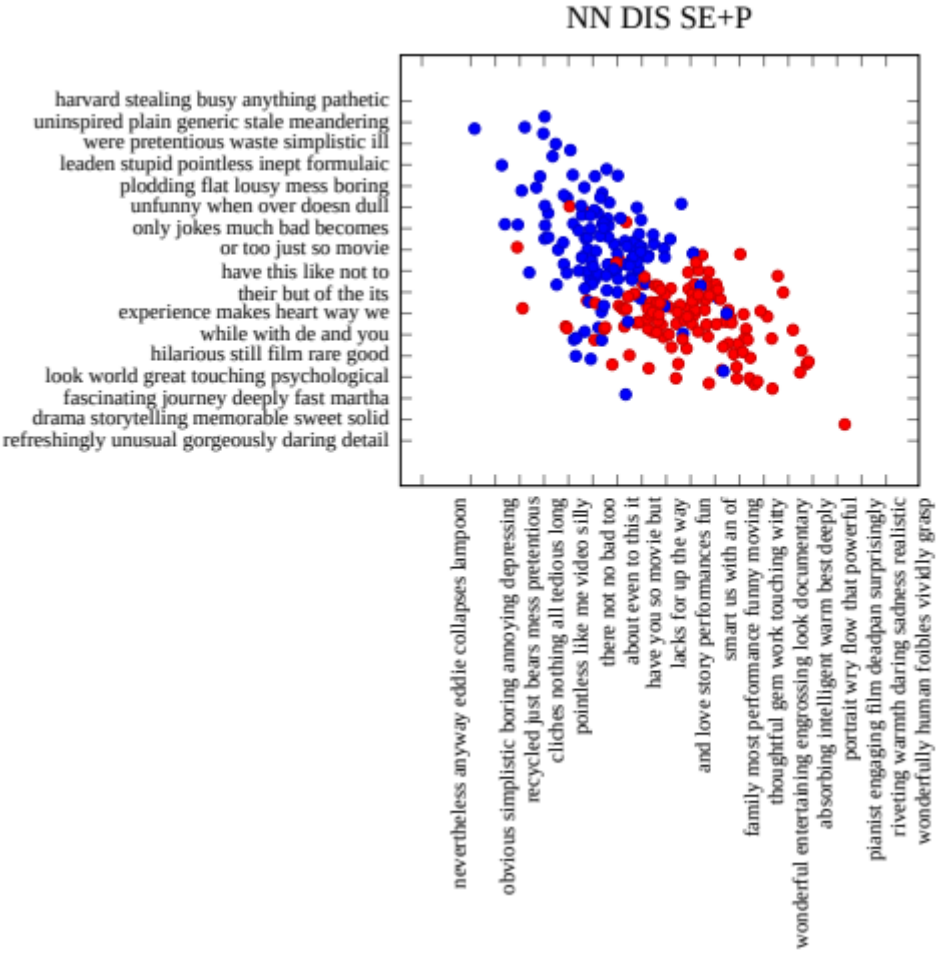

Figure 1: Example of MRD with Segmented Labelling

From the activation function and mislabelled documents, we can inspect patterns of the optimization as well. For a label and binary classification, its extremity on axis could reflect activation levels of response variable, e.g., we could directly assign emotional strength quantitatively for their selected labels. Given the colours of response values, even without prediction value, we still could recognize several mislabelled documents confidently. Their documents could be read and compared with tools of hover and attached summaries, like bokeh supplied.

4.3.2. Example 2: Tagging

This example exhibited the capacity to visualize latent variables of recurrent neural networks when recognize tags inner sequence. For resurrect neural networks, the labelling has ignored supplementary variables like memory in recurrent neural network, but focused on the embeddings and the variable of *is_tag*. In this case, we only predict the response variable of *is_tag*, hence, the selected labels are expected to have more functional words which could indicate the tags with different magnitude. Specifically, figure 2 presented the *stack* dataset with tagging predications, where, the colour are denotes territories of *biology* and *crypto*.

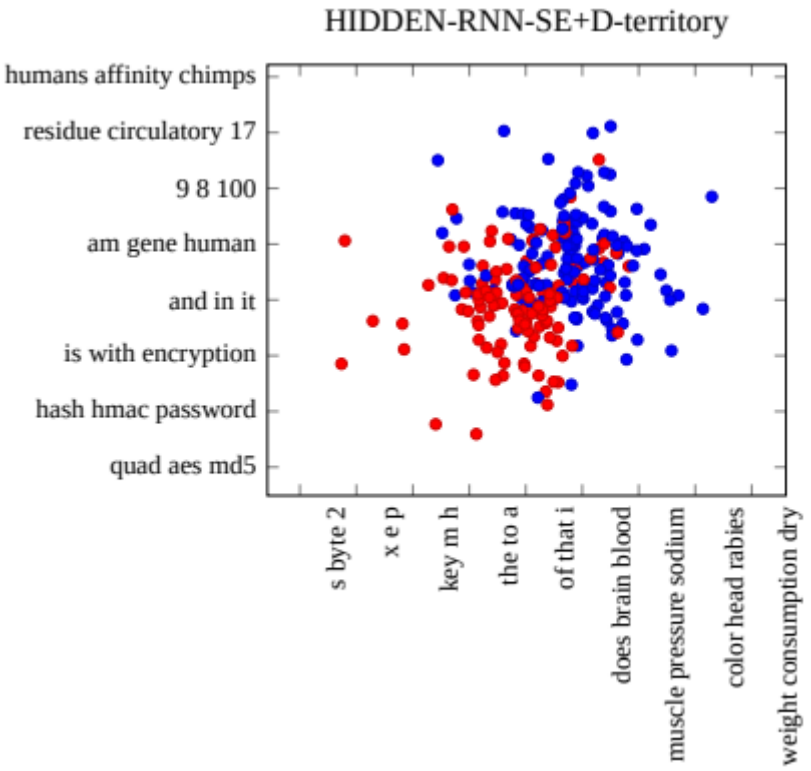

Figure 2: Example of Stack Exchange Tagging

As the figure reflected, the middle area are preferred functional words, such as *and, in, of*, also, there are articles and numbers. Of territories, terminals have common terminology, such as *md5, rabies*, which are indicate tags with linguistic attributes of collocation.

4.4 Comparative Evaluation

Given the feature selection methods in table 3, in this section, we compare their accuracies to validate utility of our methods. The comparisons are implemented as follows. For the feature reduction and selection, the *ENT* is independent to hidden space, so, we directly selected them with parameter K'. The *MMX* and *SE* require optimized hidden space, so, we trained a neural network with following settings: middle layers of {16,2}, *tanh* as activation function, *Adam* algorithm as gradient descent, and *softmax* as cost function. Next, labels could be selected with parameters H, K', K'' and the optimized models. In our method, the final label number is decided by both K', K'', and to align the dimensions, we set segmentation of *SE* as $K''^{i} = 10$, and increase K'^{i} to acquire more labels.

For the results, figure 3 presented *f1* accuracy of *MRD* with regard to train set and valid set. Naturally, the curves are increasing with more labels being selected, but our methods have more vibrations. For accuracies, we see that *SE+D* performs better than others, and *SE+P* is inferior. In addition, when labels are small, e.g., less than 50, our methods are express lesser accuracy. The reason is, when labels are not capable to be composed, the pure prominent feature, like strong emotions, are better to predict the response. Reason of low accuracy of *SE+P* is that local prominent features are not coherent with response variables. In another word, their features are selected with independence to the response values. Hence, without compositionality, the *SE+P* is vibrated more than other methods.

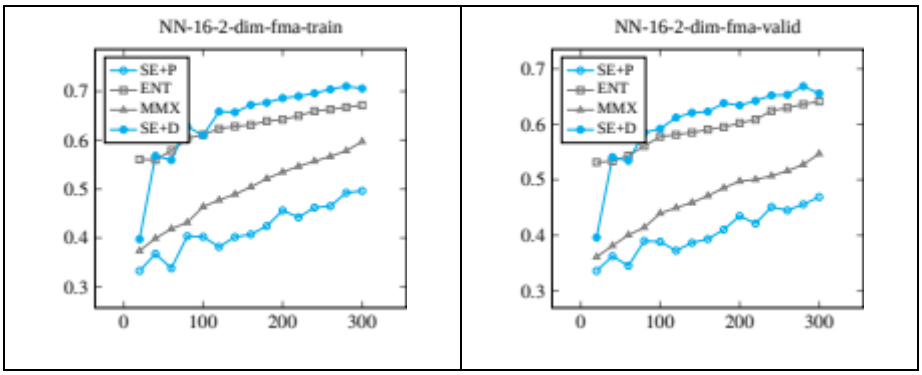

Figure 3: Accuracy of MRD

Figure 4 presented accuracies of *AMZ* dataset. Here, we can get similar conclusion of the comparable and superior of our method. And, with more number of classes, all the curves are vibrated more than *MRD* results. Even they are vibrate more, within this specific model, our *SP+D* still presented significant accurate than other methods. This means, we can confirm the utility of the foundations of accuracy division of the *SP+D*, which has been analysed in characteristics of section [discriminative].

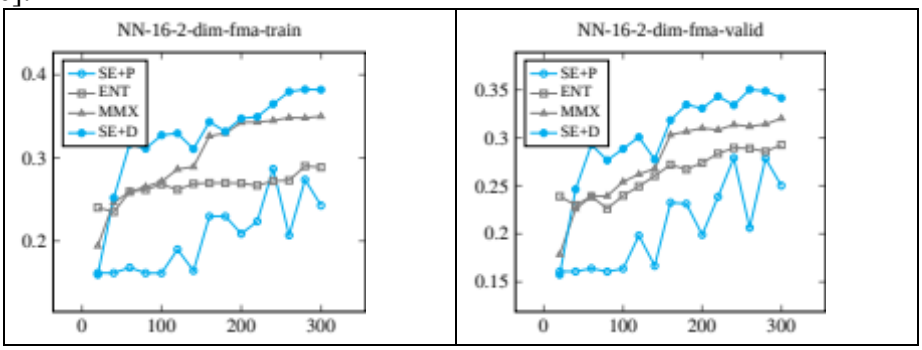

Figure 4: Accuracy of AMZ

Based on the experimental results in the figures, we summarize our findings as follows.

1. Compare to the previous methods, *SE+D* presented stability and comparable accuracy in the tasks of sentiment analysis and star prediction. *SE+P* has less accuracy that composed features with independence of response variables.
2. All the methods presented vibration along dimensions. Of our methods, *SE+D* have slightly more vibration, but stable when features are relative large. *SE+P* is vibrate a lot with its independence.

4.5 Stability Analysis

To validate the characteristics and rationales that proposed in section [labelling], we will perform feature selections with different parameters. Essentially, the stochastic convergence will be effected by the structure of hidden layers. In addition, learning rate and regularization will affect the accuracy in a minor level. Hence, we optimize neural models with different network structures, and analyse their labelling accuracies of the two datasets.

Figures 5 and 6 presented results of the parameter stability, where, title identified hidden layer nodes, x-axis is label number, y-axis is accuracy, and each line reflects a labelling method.

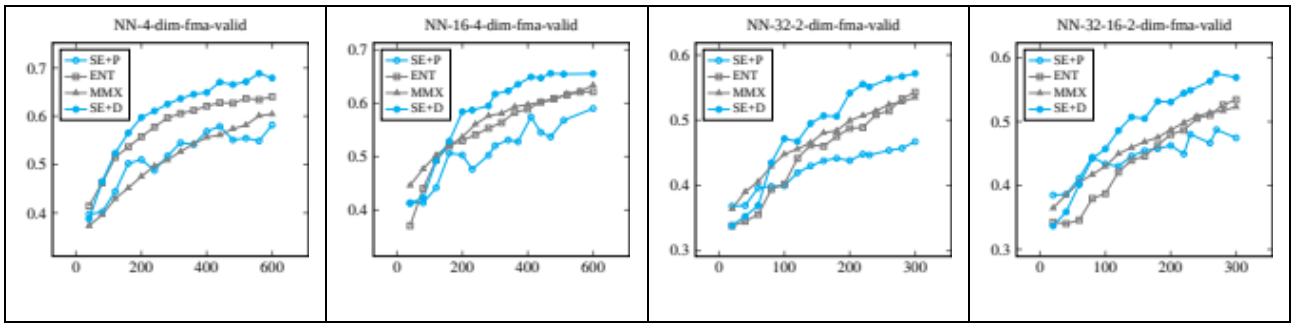

Figure 5: Accuracies of MRD with Different Network Structures

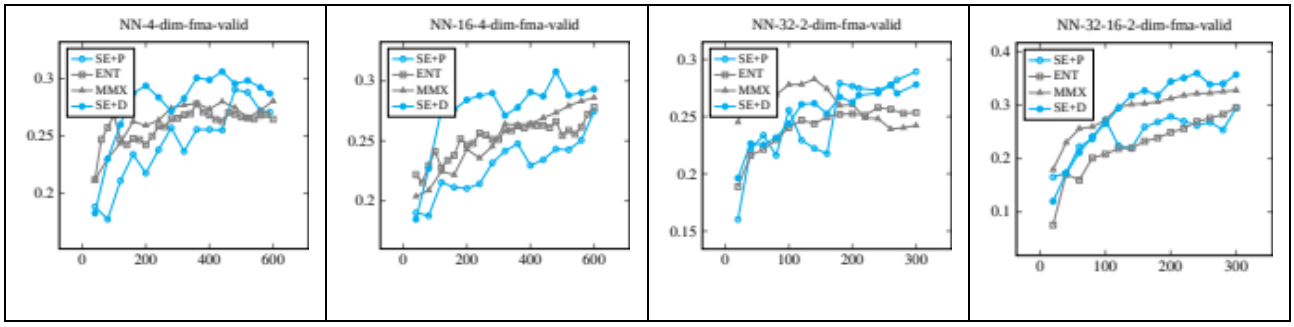

Figure 6: Accuracies of AMZ with Different Network Structures

From the results, we can conclude as follows. Firstly, the methods sustained stable rank in most models, i.e., rank of *SE+D, MMX, ENT, SE+P*. In different network structures, the accuracies has similar pattern with compare to the above structure of *16_2* hidden nodes. And, our method *SP+D* presented best accuracy in most cases. Under the general pattern, the detail discrepancy of initial parameters confirmed that neural network has characteristic of stochastic convergence. Hence, it's valuable that our *SP+D* sustained accuracy on those diverged optimums.

Secondly, we confirmed the rationales of our method. One rationale is that several dimensions can hold majority of information. Compare the structures between **_2* and **_4*, we could say that no domination exists between their two structures. Therefore, our visualization is capable to present majority information along response variables. Another is compositionality of moderate features. 'SP+D' quick increase on the left side means that composition is working with number of label of area is increased to a specific amount.

Thirdly, we can discover a pattern of classification, where, many classes or origination may disturb optimization. It's obvious that two classes *MRD* has smoother curves than the *AMZ* which has more classes and primate reviews.

5. Conclusion

We have proposed a method of text visualization with segmented labelling that built on hidden spaces and feature selection criteria of relevance and discriminative. In the visualization, documents could be interpreted with global distribution in hidden space and segmented labels which could be projected to original features or response variables directly. Through experiments, we validated that: 1) our visualization is flexible and space efficient for many reduction methods; 2) for classification tasks, segmented labelling could achieve better accuracy in most cases, with advantages of stability and alignment.

In applications, our method presented efficiency of sharing and compatibility for different interfaces. For neural network models, our method is helpful to resolve its patterns and predicability. In addition, the selected labels could be codified to thesaurus with its aligned documents and response values. As the sustained accuracies validated, the thesaurus could support concept learning

and inference in many applications.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (no. 2232017D-13) and the National Natural Science Foundation of China (no. 61603089).

References

[1] D. Ramage, C. D. Manning, S. Dumais, Partially labeled topic models for interpretable text mining, in: SIGKDD - KDD '11, ACM Press, San Diego, CA, 2011, p. 457.
[2] J. H. Lau, D. Newman, S. Karimi, T. Baldwin, Best Topic Word Selection for Topic Labelling, in: Coling, ACL, Beijing, China, 2010, pp. 605–613.[3] A. Karpathy, J. Johnson, L. Fei-Fei, Visualizing and Understanding Recurrent Networks, in: ICLR, San Diego, CA, USA, 2015.
[4] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, N. A. Smith, Sparse Overcomplete Word Vector Representations, in: ACL-IJCNLP, ACL, Stroudsburg, PA, USA, 2015, pp. 1491–1500.
[5] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, X. Tong, H. Qu, Textflow: Towards better understanding of evolving topics in text, IEEE Trans. Vis. Comput. Graph. 17 (12) (2011) 2412–2421.
[6] M. Hu, K. Wongsuphasawat, J. Stasko, Visualizing Social Media Content with SentenTree, IEEE Trans. Vis. Comput. Graph. 23 (1) (2017) 621–630.
[7] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, N. Elmqvist, ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding, IEEE Trans. Vis. Comput. Graph. 24 (1) (2018) 361–370.
[8] E. Gabrilovich, S. Markovitch,Computing semantic relatedness using wikipedia-based explicit semantic analysis, in: IJCAI, Vol. 20, IJCAI Organization, Hyderabad, India, 2007.
[9] Y. Li, J. Yosinski, J. Clune, H. Lipson, J. Hopcroft, Convergent Learning: Do different neural networks learn the same representations?, in: ICLR, San Diego, CA, USA, 2015.
[10] M. Omar, B.-W. On, I. Lee, G. S. Choi, LDA topics: Representation and evaluation, J. Inf. Sci. 41 (5) (2015) 662–675.
[11] Q. Mei, X. Shen, C. Zhai, Automatic labeling of multinomial topic models, in: SIGKDD - KDD '07, no. March, ACM Press, New York, New York, USA, 2007, p. 490.
[12] R. Takanobu, M. Huang, Z. Zhao, F. Li, H. Chen, X. Zhu, L. Nie, A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning, in: IJCAI, IJCAI Organization, California, USA, 2018, pp. 4403–4410.
[13] S. Bhatia, J. H. Lau, T. Baldwin, Automatic Labelling of Topics with Neural Embeddings, in: Coling, ACL, Osaka, Japan, 2016, pp. 953–963.
[14] I. Hulpus, C. Hayes, M. Karnstedt, D. Greene, Unsupervised graph-based topic labelling using dbpedia, in: WSDM '13, ACM Press, New York, New York, USA, 2013, p. 465.
[15] S. Ueda, S.-P. Quek, T. Itioka, K. Murase, T. Itino, Phylogeography of the Coccus scale insects inhabiting myrmecophytic Macaranga plants in Southeast Asia, Popul. Ecol. 52 (1) (2010) 137– 146.
[16] W. Kou, F. Li, T. Baldwin, Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors, in: Inf. Retr. Technol., no. Chapter 20, 2015, pp. 253–264.
[17] Y. H. Tseng, Generic title labeling for clustered documents, Expert Syst. Appl. 37 (3) (2010) 2247– 2254.
[18] S. Ingram, T. Munzner, Dimensionality reduction for documents with nearest neighbor queries, Neurocomputing 150 (2015) 557–569.