

# Research Summary

September 2012 – September 2014

**Shudong Hao**

Artificial Intelligence Laboratory  
Beijing Forestry University  
shudongh@acm.org  
<http://shudonghao.com/>

---

## Supervisors

|                             |                             |
|-----------------------------|-----------------------------|
| Prof. Dengfeng Ke           | Prof. Yanyan Xu             |
| Chinese Academy of Sciences | Beijing Forestry University |
| dengfeng.ke@ia.ac.cn        | xuyanyan@bjfu.edu.cn        |

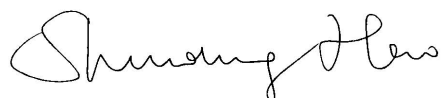
## Contents

---

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>2</b>  |
| <b>2</b> | <b>Methodology</b>  | <b>4</b>  |
| 2.1      | Erroneous Character Detection and Correction<br>(Character Perspective) . . . . . | 6         |
| 2.2      | Content Perspective . . . . .   | 7         |
| 2.3      | Topic Perspective . . . . .   | 9         |
| 2.4      | Language Fluency Perspective . . . . .  | 10        |
| 2.5      | Automated Scoring . . . . .   | 12        |
| <b>3</b> | <b>Outlook</b>  | <b>13</b> |
| 3.1      | Limitations of the current ACES . . . . .   | 13        |
| 3.2      | Deep Learning . . . . .   | 14        |
|          | <b>Acknowledgements</b>   | <b>16</b> |

## **Abstract**

Since September 2012, I co-founded the first Artificial Intelligence Laboratory at Beijing Forestry University, as a research assistant and a team leader. Fortunately, I was supervised by Prof. Dengfeng Ke from Institute of Automaton, Chinese Academy of Sciences and Prof. Yanyan Xu from Beijing Forestry University. The research interests of the group I lead are Machine Learning (ML) and Natural Language Processing (NLP), and the project is called “Automated Chinese Essay Scoring” (ACES). Specifically, the task is to find or invent some new ML and NLP methods to automatically score Chinese written essays in the language tests. The project has harvested the fruits in the past almost three years, and it is very promising.

A handwritten signature in black ink, appearing to read 'Shuang He', with a stylized, cursive script.

# 1

---

## Introduction

---

Automated Essay Scoring is a representative cross-disciplinary task, combining various fields, such as psychology, artificial intelligence, linguistics, and so forth [1]. The first automated essay scoring system can be traced back to 1960s when Ellis Batten Page developed the PEG system. With the trend of artificial intelligence and its applications in industry, scoring system attracted much interests of researchers and educators. At present, automated English scoring system is mature. For example, E-rater is used for The Test of English as a Foreign Language (TOEFL) and the Graduation Record Examination (GRE), and its human-machine agreement has been improved significantly [2].

As Chinese becomes increasingly popular, the scale of the tests of Chinese language, like The Minorities-oriented Chinese Level Test (MHK), skyrocketed as well. According to the statistics of MHK, the number of test-takers in Xinjiang Province was 190,000 in 2009. Writing is the most important part of Chinese tests, because it reflects the language proficiency of the test-taker. Therefore, automated Chinese essay scoring system is urgently needed [3].

Unfortunately, the difficulties in processing Chinese language using intelligent system hamper the development, leading to that the research

南京市长江大桥。

Segmentation 1: 南京／市长／江大桥。

(Daqiao Jiang the major of Nanjing.)

Segmentation 2: 南京市／长江大桥。

(The Yangtze River Bridge in Nanjing.)

**Figure 1.1:** One sentence can have two ambiguous meanings according to different segmentations.

is very rare and the progress is very slow. First, in NLP, *word* is the atomic element for almost any task of natural language understanding. But Chinese does not provide explicit delimiters, like spaces in English, raising the problem called “segmentation”: the same sentence with different segmentations can be understood in different ways. We use Figure 1.1 to show this problem. Second, the problem is how to understand the sentence based on the result of the segmentation, and this is where the NLP algorithms come in. In the field of automated Chinese essay scoring, it is a more complicated problem because it involves processing and understanding the essays from different levels or perspectives. We will explain this in the next section. Finally, after the machine has understood the language, the real problem (which is also the magical part) is how to give these essays their scores, and this is where the ML algorithms come in.

In the next sections, we will explain the methodologies of the philosophy of our Automated Chinese Essay Scoring project (ACES) in detail, and look ahead to the future of ACES. We will take MHK as the example in this research summary, since all the datasets used in our project are the essays from this test.

# 2

---

## Methodology

---

Briefly, automated Chinese essay scoring follows the criteria of human scoring, consisting of four levels: *character*, *word*, *sentence* and *paragraph* [4].

- **Character** : Use correct Chinese characters;
- **Word** : Use appropriate and exact words.
- **Sentence** : Use grammatically correct sentences; and
- **Paragraph** : Organize sentences, paragraphs into an essay logically.

The criteria above, actually, are interrelated and are often combined to yield new scoring perspectives. For example, the word level and paragraph level can be connected to score essays from content perspective. From sentence level, we may detect the language fluency of the sentences and whether these sentences are well-fit into the context. Moreover, paragraph level can be divided into different perspectives as well. For example, from this level, one can score essays from the perspective of topic or that of logical structure.

### **Philosophy of ACES**

In ACES, our philosophy is to simulate the human raters as much as possible. When scoring an essay, a human rater will consider all the perspectives – from character and word to sentence and paragraph – to give a comprehensive evaluation. Therefore, we expect to develop or discover algorithms to deal with them separately and to achieve relatively high precision. Then these scores from different aspects will be combined in a certain way to give an overall score. For example, Regularized Latent Semantic Indexing (RLSI) from topic modeling is an algorithm to grade essays from topic perspective; Contextualized Latent Semantic Indexing (CLSI) can grade essays in terms of language fluency. They score essays separately, but a more ideal model is to combine them in a system.

SCESS is our first system, which grade essays from content perspectives. We are also working on algorithms to combine scores from different perspectives in SCESS in the future.

Inspired by these analysis above, we study the concrete methods from different perspectives.

## 2.1 Erroneous Character Detection and Correction (Character Perspective)

Scoring essays only from character perspective is far from being enough, although using correct Chinese characters is the most basic requirement. Practically, it takes only a small proportion in human scoring. But it is not-trivial in automated Chinese essay scoring, because all the data analysis and methods are based on correct characters and the most possible segmentations.

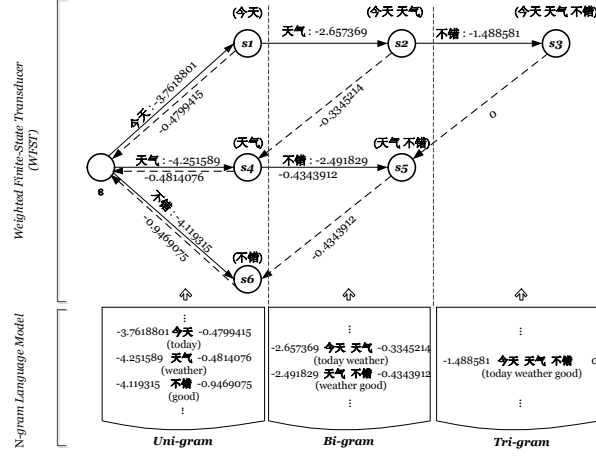
Before this, there were some studies on this topic [5]. The most obvious shortage is that the current algorithms use "static" methods, meaning that in each step in the algorithm they consider all the possible solutions without filtering. This will lead to the serious problem of efficiency and effectiveness.

Considering the deficiencies of these algorithms, we combine  $N$ -gram Language Model, which is a very common tool in NLP, and Weighted Finite-State Transducer (WFST).  $N$ -gram Language Model records the probabilities of the words appeared in a very large-scale corpus, based on statistical learning algorithms and Markov assumptions in NLP. Thus, we could choose the best segmentation solution and the correct character appeared in a sentence. The intuition behinds this is that the probability of the correct sentence is higher than that of the wrong sentence. WFST is also used for Speech Recognition successfully, proposed by Mehryar Mohri [6]. When we put the terms from  $N$ -gram Language Model on WFST, the problem of segmentation, erroneous detection and correction becomes a kind of dynamic decoding algorithm. Figure 2.1 illustrates how to convert an  $N$ -gram Language Model to WFST.

The segmentation in WFST becomes a problem of node expansion, and what we need is the best path from starting state  $\epsilon$  to a certain ending state. Therefore, we input a sentence without segmentation, or with some erroneous characters, and the algorithm will take the sentence from the starting state, transducing the states through arcs, filtering unpromising candidates at each step immediately, to a certain ending state. Finally we get the segmented or correct sentence.

This work was published first on the International Conference on





**Figure 2.1:** We use a simple example to show how to convert  $N$ -gram Language Model to WFST

Document Analysis and Recognition 2013, Washington DC, USA, and it is the first publication of this project. Based on this method, Contextualized Latent Semantic Indexing (CLSI) is also studied, which will be described in the section “Language Fluency Perspective”.

## Publication

Automated Error Detection and Correction of Chinese Characters in Written Essays Based on Weighted Finite-State Transducer.

Shudong Hao, Zongtian Gao, Mingqing Zhang, Yanyan Xu, Hengli Peng, Kaile Su, and Dengfeng Ke.

In *Proceedings of International Conference on Document Analysis and Recognition 2013* (ICDAR 2013), pp.763–767, Washington DC, USA.

## 2.2 Content Perspective

After several experiments, we see that WFST is powerful to segment Chinese sentences effectively, leading to the convenience of language processing. We will focus on concrete language understanding, which is the core of automated Chinese essay scoring.

The first method we notice is Latent Semantic Analysis (LSA), which is very classical in the field of Information Retrieval (IR) and NLP [7]. There has been a large amounts of studies applying LSA to the English essay scoring, but the related research on Chinese essay is still absent. Therefore, we experiment on LSA.

First we segment essays into words and organized them into a matrix  $\mathbf{D} \in \mathbb{R}^{M \times N}$ . Singular Value Decomposition (SVD), a common matrix factorisation method with low-rank approximation, is the underlying algorithm of LSA, which can construct a lower-dimensional and less-noisy semantic space:

$$\begin{aligned} \mathbf{D} &\approx \mathbf{U}\mathbf{\Sigma}\mathbf{V} \\ &\approx \begin{bmatrix} \begin{bmatrix} u_{11} \\ \vdots \\ u_{M1} \end{bmatrix} & \cdots & \begin{bmatrix} u_{1M} \\ \vdots \\ u_{MK} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_K \end{bmatrix} \begin{bmatrix} \begin{bmatrix} v_{11} & \cdots & v_{1N} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} v_{K1} & \cdots & v_{KN} \end{bmatrix} \end{bmatrix} \end{aligned} \quad (2.1)$$

where  $\mathbf{U} \in \mathbb{R}^{M \times K}$  and  $\mathbf{V} \in \mathbb{R}^{K \times N}$  are orthogonal matrices, and matrix  $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$  is a diagonal matrix with extra zero where the elements are in descending order.

Although the result showed that LSA performed well, a serious problem emerged: the computation efficiency. As previously stated, the scale of MHK is rapidly increasing. If we still use conventional LSA, the computing time will be very long, and requires large memory storage. Seeking a new faster algorithm is the goal of this stage.

Incremental Singular Value Decomposition (ISVD) was firstly used for image processing successfully [8]. We adopted this algorithm to NLP field and published the Incremental Latent Semantic Analysis (ILSA) and its application in automated Chinese essay scoring. The experimental results show that ILSA effectively addresses the problem of computation: it requires much lesser memory usage and shorter computing time. Moreover, it improves the final scoring accuracy (an index used for evaluating a scoring system).

The first publication is *Automated Essay Scoring Using Incremental Latent Semantic Analysis*, which is completed by my teammate Mingqing Zhang. Considering that WFST and ILSA are enough to

build a scoring system to score essays from content perspective, we improve WFST with other functions and build a system called SCESS. That is the very initial version of the final construction of a mature scoring system. In the future, we will continuously develop and refine SCESS.

### Publications

Automated Essay Scoring Using Incremental Latent Semantic Analysis.  
Mingqing Zhang, Shudong Hao, Yanyan Xu, Dengfeng Ke, and Hengli Peng.  
*Journal of Software*, 9(2), pp.429–436, 2014.

SCESS: A WFSA-based Automated Simplified Chinese Essay Scoring System with Incremental Latent Semantic Analysis.  
Shudong Hao, Yanyan Xu, Dengfeng Ke, Kaile Su, and Hengli Peng.  
*Natural Language Engineering*, Cambridge University Press, UK.  
To appear.

## 2.3 Topic Perspective

Scoring essays from only one perspective is not comprehensive, so we need to search for other methods from ML, IR or NLP to score essays from other perspectives. At this point, the field of Topic Modeling came to our view [9]. Topic modeling methods discover the latent topic structures in large article collections, and thus, it is appropriate to adopt them to our task. Intrinsically, essay scoring system needs to learn the latent topics in an essay collection, and recognize the possible topics in the collection. Latent Dirichlet Allocation (LDA) is a representative algorithm [10], but according to the experimental results, this method does not perform well. A new method published recently became our first choice, called Regularized Latent Semantic Indexing (RLSI) [11].

RLSI starts from the objective function:

$$\min_{\mathbf{U}, \mathbf{V}} \left( \|\mathbf{D} - \mathbf{UV}\|_F^2 + \lambda_1 \|\mathbf{U}\|_1 + \lambda_2 \|\mathbf{V}\|_F^2 \right) \quad (2.2)$$

where  $\mathbf{D} \in \mathbb{R}^{M \times N}$  is the matrix of our essay collection as defined before, and  $\mathbf{U} \in \mathbb{R}^{M \times K}$  reflects the relations between  $N$  documents and  $K$  latent topics. Inherently, it is a new method of matrix factorisation with low-rank approximation, like SVD in LSA.

The experimental results show that this method outperforms LDA, and therefore it is an appropriate method in automated Chinese essay scoring from topic perspective. This work is published on International Conference on Pattern Recognition 2014, as a pilot work in this field.

### Publication

Automated Chinese Essay Scoring From Topic Perspective Using Regularized Latent Semantic Indexing.

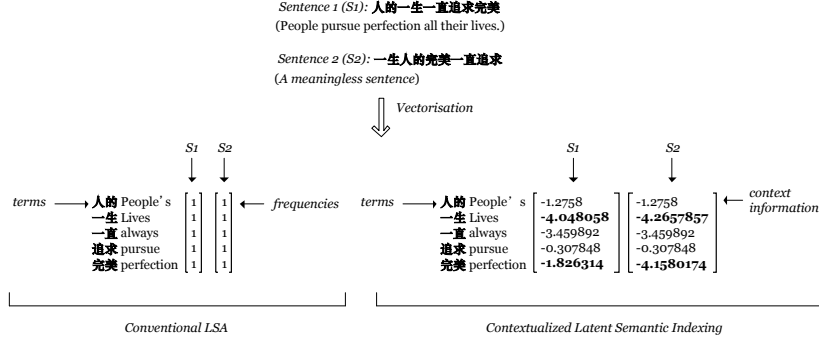
Shudong Hao, Yanyan Xu, Hengli Peng, Kaile Su, and Dengfeng Ke.

In *Proceedings of International Conference on Pattern Recognition 2014* (ICPR 2014), pp.3092–3097, Stockholm, Sweden.

## 2.4 Language Fluency Perspective

The perspective of language fluency is a relatively new but less concerned perspective. During our study, we found that conventional LSA reveals some problems. For example, although it can construct semantic space and analyse the contents by the word usage appeared in an essay or an essay collection (namely, *what are these words?*), it is unable to reveal the context background (namely, *how are these words used?*). Finally, the system based on conventional LSA may give the same score to two totally different essays: one uses words logically and in correct order, while the other does not. In this stage, we need to figure out some algorithms to detect the language fluency. Figure 2.2 shows the intuition.

From this point, we propose a new method called Contextualized Latent Semantic Indexing (CLSI), an improved LSA method.  $N$ -gram Language Model and WFST are also used in this method, because they provide probabilities of words in a large corpus. Previously in conventional LSA, one has to use some techniques, like WFST, to segment the Chinese essays into words and arranged them to a matrix  $\mathbf{D}$ , recording the occurrence of each word appeared in each essay. In our method, instead of occurrence, we fill the matrix with context information extracted from  $N$ -gram Language Model, during the period of segmenting in WFST. Thus, although two sentences use the same words, their representations of vectorisation differ.



**Figure 2.2:** A simple comparison between conventional LSA and CLSI.

We conducted various experiments. The results show that the performance of our method is better and more stable than conventional method. As a tentative experiment, we used WFST to segment sentences and extract context information. This work, called Genuine Contextualized Latent Semantic Indexing (Genuine CLSI), has been submitted to the *Conference of the North American Chapter of the Association for Computational Linguistics*. After that, we modified the Genuine CLSI by using WFST to correct erroneous characters, segment sentences and extract context information, called Modified Contextualized Latent Semantic Indexing (Modified CLSI) and conducted more experiments. This work has been described in great details and has been submitted to the *ACM Transactions on Asian Language Information Processing* as a full research paper.

## Publications

Contextualized Latent Semantic Indexing: A New Approach Applied in Automated Chinese Essay Scoring.

Shudong Hao, Bin Huang, Yanyan Xu, Dengfeng Ke, and Kaile Su.

*ACM Transactions on Asian Language Information Processing*, Association for Computing Machinery (ACM), USA.

Under review.

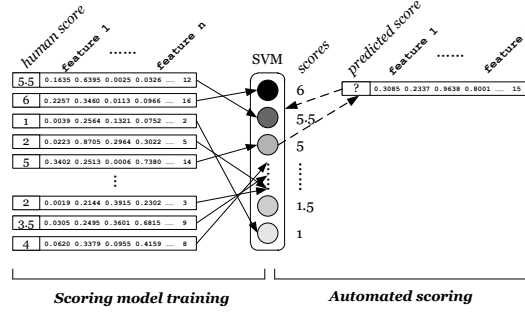
Contextualized Latent Semantic Indexing for Automated Chinese Essay Scoring.

Shudong Hao, Bin Huang, Yanyan Xu, Dengfeng Ke, and Kaile Su.

*Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HTL 2015).*

Under review.

## 2.5 Automated Scoring



**Figure 2.3:** The SVM-based scoring model.

After analysing, each essay will obtain a representation of vectorisation, *essay vector*, and each dimension of the essay vector is regarded as a feature. For the essays in the training set, their human scores will be the classification labels. From this point of view, we can see the automated scoring as a kind of multi-class classification in ML. In all of our works, we use Support Vector Machine (SVM) as the tool to score essays automatically, as shown in Figure 2.3. Specifically, we use LIBSVM<sup>1</sup>, a popular toolkit, to complete this task [12].

<sup>1</sup>available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

# 3

---

## Outlook

---

### 3.1 Limitations of the current ACES

With the upcoming the last year of my college, the project “Automated Chinese Essay Scoring” (ACES) will also step into the last stage under my leadership. There are several problems left. First, we see that the scoring method from the sentence level is still absent. Although CLSI can detect the language fluency of sentences, it must be used in an essay collection. What is more, it cannot detect the grammar errors. Using dependency parsing will be a promising method. In addition to NLP related methods, ML is still very useful to perform the automatic scoring in real. Bayesian classifier, perceptron, and other novel matrix factorisation algorithms like Multiresolution Matrix Factorization (MMF), are worth trying and studying. Additionally, adapting SVM to deal with unbalanced dataset is also worth studying, because the distribution of human scoring in our dataset is like a normal distribution, instead of a balanced distribution.

### 3.2 Deep Learning

Back to the philosophy we introduced in Chapter 2, we see that a huge challenge the conventional semantic extraction methods present is how to combine them. That is to say, we may use various methods to score essays from various perspectives (word usage, content, topic, *etc.*), but how to combine them so that for each essay we will get a score from comprehensive scoring perspectives?

Deep learning is a very active field recently in NLP. In 2003, Bengio *et al.* published the paper about deep learning in language modeling [13]. In 2011, Collobert *et al.* explored a brand new field of NLP, and developed many basic NLP tasks using deep learning [14]. Since then, there are increasingly more publications about deep learning for various NLP tasks in top conferences like ACL and EMNLP.

In the near future, we will be trying to develop ACES systems using Convolutional Neural Networks (CNN). There are two reasons we try CNN. First, in [15], CNN with dynamic pooling has been successfully used to model sentences. This model can handle with various sentence length and dynamically determine the pooling size. For ACES, essays are not of the same length, so CNN could deal with this problem. Second, CNN could involve many convolutional layers with pooling (subsampling) layers. If we consider this architecture as a simulator of human brain, this model will be very appropriate. For human beings, to understand an essay, we need to process it from “lower” level to “upper” level – characters, words, sentences, paragraphs and topics. Therefore, as the feed forward algorithm in CNN goes deeper, it will simulate this processing. For example, after the first convolutional layer and pooling layer, CNN will capture the organization of word usage in an essay; then using it as an input, the second convolutional and pooling layer will also capture the structure of sentences, and this processing will go on. Finally, in the last layer, CNN will capture the structure of the essay in a comprehensive manner, and do our task – score essays. This architecture is shown in Figure 3.1.

Though this architecture is still being studied and experimented, this framework is very promising because that is how we human beings understand natural languages. Specifically in ACES, it may solve our



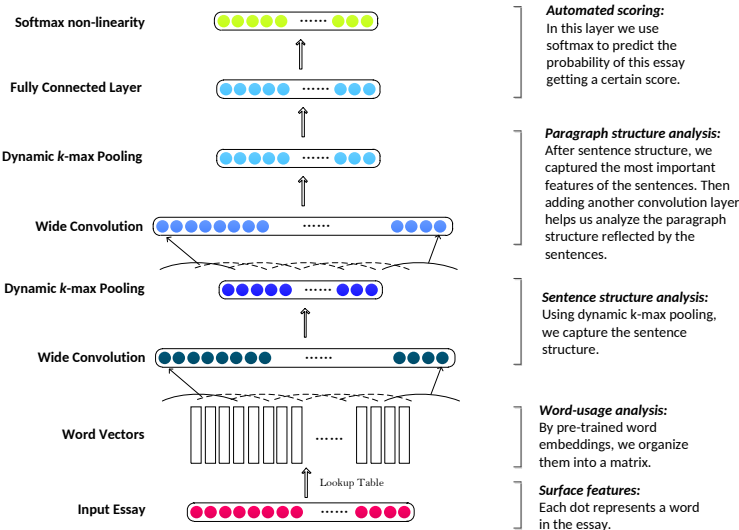


Figure 3.1: Proposed CNN architecture for ACES.

problem: we do not need to design algorithms from each perspective to score essays and combine them. Using CNN, we may just use a one-set manner and get the final score.

## Acknowledgements

---

I would like to thank Prof. Yanyan Xu who introduced me to Prof. Dengfeng Ke. For the past two years, I grew up from a student who only knows about coding and programming and reading, to a beginner researcher. I think I have learned more with the help of them than in classes. I learned how to read a paper, how to make a presentation, how to spark an idea from common sense. I still remember the first time I attended the meeting of our project without understanding what they were talking about. I still remember the first time I wrote the script of my paper which drove Prof. Xu and Prof. Ke crazy. I still remember the first time I felt the great joy the acceptance of my paper brought. I still remember the first time I attended an international conference with my teammate and also best friend Mingqing Zhang in USA.

But the most important thing for me, I think, is that they help me step into the field of Machine Learning. This project is more industry- or application-oriented. It is also a very good starting point to take this chance to know about Machine Learning, by the algorithms we adopted and designed. I would like to thank Prof. Yanyan Xu and Prof. Dengfeng Ke once more, for their patient instructions on these problems, which led me to see a spectacular view of Computer Science.

Finally I would like to thank my mother, who is supporting me all the time. It is her patience, wisdom and spirit that encourage me to go on this long, long road of my academic career.

## Bibliography

---

- [1] Mark D. Shermis and Jill C. Burstein. Automated essay scoring: A cross-disciplinary perspective. Psychology Press. 2003.
- [2] Yigal Attali and Jill Burstein. Automated Essay Scoring With E-rater v.2.0. *Journal of Technology, Learning and Assessment*, 2006 (4).
- [3] Yanan Li. Automated essay scoring for testing Chinese as a second language. PhD thesis, Beijing Language and Culture University. 2006.
- [4] Hengli Peng. The Minorities-oriented Chinese Level Test. *China Examinations*, pp.57–59, 2005 (10).
- [5] Kun Wang, Chengqing Zong, and Keh-Yih Su. Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. *ACM Transactions on Asian Language Information Processing*, Article 7, 11(2), 2012.
- [6] Mehryar Mohri. Weighted Finite-State Transducer Algorithms: An overview. *Formal Languages and Applications*, 2004(3).
- [7] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic

- Analysis. *Journal of the American Society for Information Science*, pp.391–407, 1990(41).
- [8] Matthew Brand. Incremental Singular Value Decomposition of Uncertain Data with Missing Values. In *Proceddings of European Conference on Computer Vision 2002*, pp.707-720, 2002.
- [9] David M. Blei and John Lafferty. *Topic Models*. Text Mining: Classification, Clustering, and Applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.
- [10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent Dirichlet Allocation*. *Journal of Machine Learning Research*, pp.993–1022, 2003(3).
- [11] Quan Wang, Jun Xu, Hang Li, and Nick Craswell. Regularized Latent Semantic Indexing: A New Approach to Large-scale Topic Modeling. *ACM Transactions on Information Systems*, Article 5, 2013 (31).
- [12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, Article 2, 2011.
- [13] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. *A Neural Probabilistic Language Model*. *Journal of Machine Learning Research*, pp.1137–1155, 2003(3).
- [14] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. *Natural Language Processing (Almost) from Scratch*. *Journal of Machine Learning Research*, pp.2493–2537, 2011(12).
- [15] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.