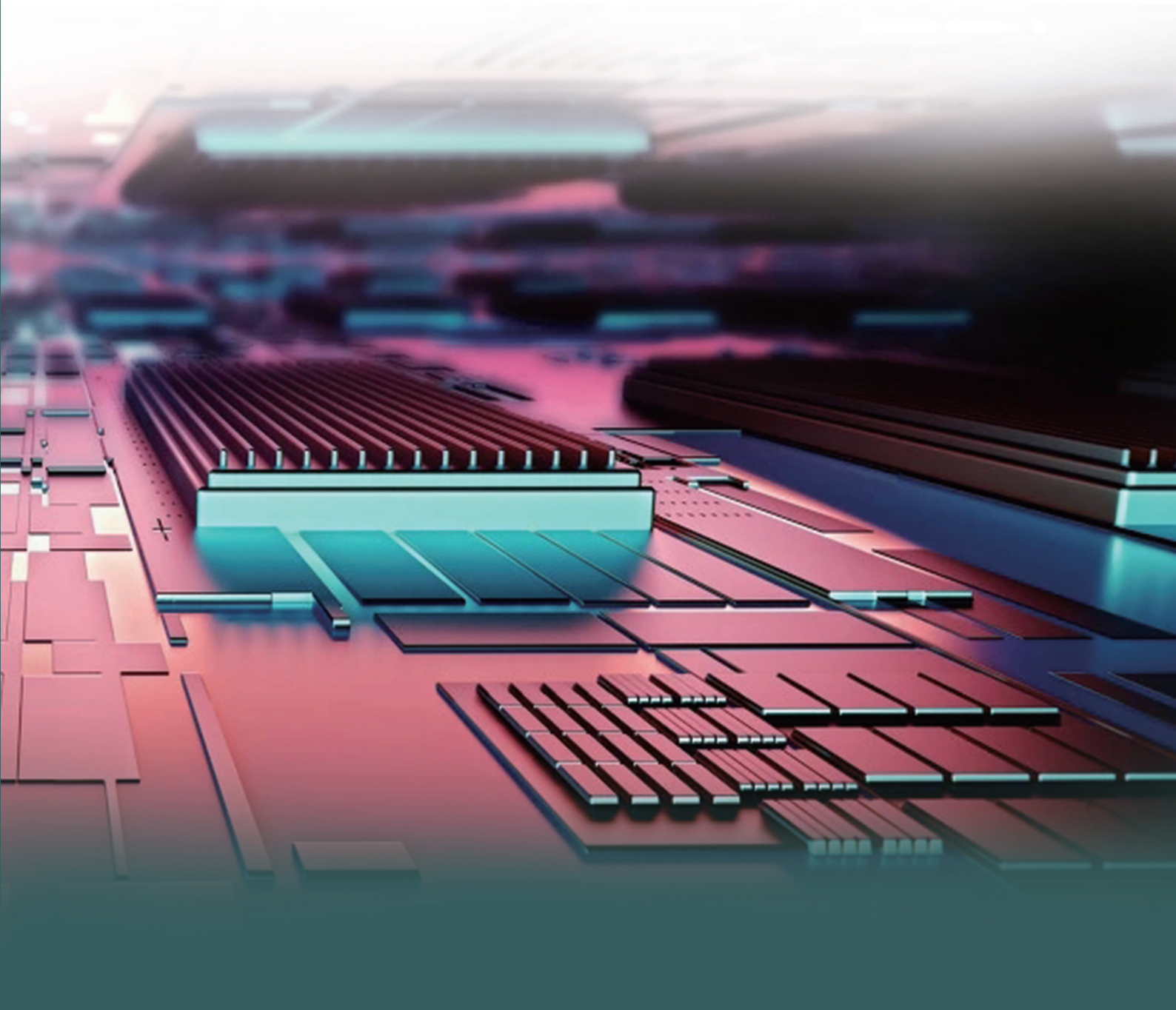
SYNERGY
PUBLISHING PTE. LTD.

Volume 6
Issue 01
April 2022

ISSN 2591-7110(Print) 2591-7129(Online)

Modern Electronic Technology



SYNERGY
PUBLISHING PTE. LTD.

Modern Electronic Technology

Aims and Scope

Modern Electronic Technology (MET) is an open access, peer-reviewed scholarly journal which aims to publish original research articles, reviews and short communications that covers all area of electronic engineering technology. MET emphasizes on publishing high quality papers, as well as aims to provide a source of information and discussion platform for engineers, researchers, and electronic professionals worldwide.

Subject areas suitable for publication include, but are not limited to the following fields:

- Microelectronics
- Nanoelectronics
- Electronic Materials Technology
- Structure and Nature of Semiconductor
- Digital Technology
- Automation System

Publishing Cycle

Quarterly

Journal Homepage

http://ojs.s-p.sg/index.php/met

Key Features

- Open Access
- High Academic Level Editorial Board
- Easy and Fast Submissions
- Double Blind Peer Review
- Rapid Online Publication of Articles upon Acceptance
- Outlet for Academic Institutions and Industry

Volume 6 Issue 1 · April 2022
ISSN 2591-7110 (Print) ISSN 2591-7129 (Online)

Synergy Publishing Pte.Ltd.

E-Mail: contact@s-p.sg
Official Website: www.s-p.sg
Address: 12 Eu Tong Sen Street,
#07-169, Singapore 059819

CONTENTS

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Preparation and Performance of CdZnTe Ray Detector

Jun Jiang*

Kunming Institute of Physics, Kunming, Yunnan, 650223, China

ARTICLE INFO

Article history
Received: 19 January 2022
Revised: 26 January 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Cadmium zinc telluride (CZT)
Room temperature detector
Crystal growth
Preparation process
Performance

ABSTRACT

γ-ray and x-ray detectors made by $Cd_{1-x}Zn_xTe$ alloy can gain high energy resolution and detect efficiency at room temperature due to its high atomic number, large energy gap and high density, which were well-developed recently. By well controlled of Cadmium partial pressure and compensatory doping technique, Φ90 mm $Cd_{1-x}Zn_xTe$ alloy obtained successfully ($\rho \geq 10^{11}$ Ω·cm) by an improved-Bridgman method. 3 mm × 3 mm × 3 mm CZT detector was made at Kunming Institute of Physics, which has energy resolution of 3.52% (FWHM) at room temperature when detect 59.54 KeV Am^{241} γ-ray source.

1. Introduction

As a technical branch of experimental nuclear physics, X-ray detector plays an important role in the development of nuclear physics. With the development of national economy, science and technology, compound semiconductor detector can work at room temperature γ Nuclear radiation detector with high detection efficiency and good energy resolution can be used in various detectors and spectrometers in astronomy, medicine, industry, military and other fields [1-3].

Cadmium zinc telluride (CD1 xznxte) crystal (hereinafter referred to as CZT when x value is not specifically referred to) is a new solid solution compound semiconductor developed from cadmium telluride (CdTe) crystal. CdTe as an X-ray and γ Ray detector materials has been widely studied since the 1970s. But its resistivity is too low to work at room temperature. Therefore, it is assumed that adding a certain amount of Zn into CdTe can continuously adjust the lattice constant between 0.6100 nm ~ 0.6428 nm, so as to adjust the lattice constant, increase the band gap and improve the resistivity. At the same time, the introduction of Zn increases the lattice strength and stacking staggered energy, reduces the dislocation density and the possibility of forming twins, so as to make up for the deficiency of CdTe crystal performance, CdZnTe crystal was developed.

Cadmium zinc telluride (CD1 xznxte) crystal is a new type of room temperature ternary compound semiconductor nuclear radiation detector material with excellent performance. It has high resistivity (about 1011 Ω/cm), large atomic number (48, 30, 52), high density (about 6 G/cm^3) and large band gap width. With the different content of Zn, the band gap width changes continuously from 1.4 eV (near infrared) to 2.26 eV (green light). X-ray at room temperature γ The ray energy resolution is good, and its energy range is 10K eV-6m eV. Compared with CdTe,

**Corresponding Author:*
Jun Jiang,
Kunming Institute of Physics, Kunming, Yunnan, 650223, China;
Email: jj.wjy @163.com

 DOI: https://doi.org/10.26549/met.v6i1.9507

CD1 xznxte crystal has wider band gap, so it has higher resistivity than CdTe, smaller leakage current at room temperature and higher energy resolution, so it is more popular [4].

2. CD1 Xznxte Crystal Growth

CdZnTe is a pseudo binary compound semiconductor material composed of CdTe and ZnTe. It has a sphalerite structure (as shown in Figures 1 and 2). The melting point varies from 1092 °C to 1295 °C due to the content of Zn. Due to the factors unfavorable to crystal growth such as high growth temperature, low thermal conductivity, strong ionic property and low stacking fault energy, it is difficult to grow CdZnTe crystal with good repeatability and high yield.

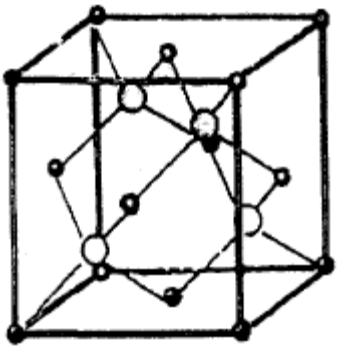


Figure 1. Structure of sphalerite

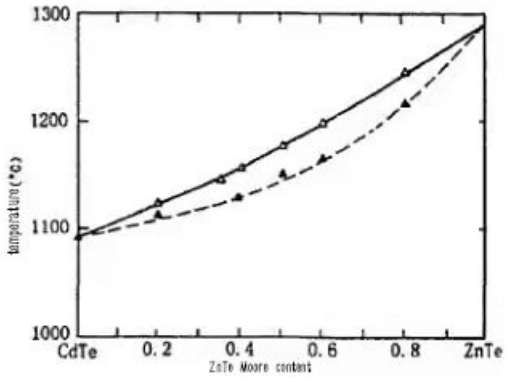


Figure 2. CZT pseudo binary phase diagram

At present, the commonly used methods include vertical gradient solidification method (VGFM), mobile heater method (THM), high pressure Bridgeman method (HPB), physical vapor transport (PVTM), horizontal Bridgeman method (HBM) and vertical Bridgeman method (VBM). The above methods have their own advantages and disadvantages [5]. At present, we mainly use the improved Bridgeman method to grow CZT crystals. It is a common melt growth method. Its growth process is actually a compound melt composed of Cd, Te and Zn, which decreases slowly in a temperature field with a certain temperature gradient and crystallizes continuously due to local supercooling nucleation.

For the ray detector, the material of the detector is required to have high resistivity (greater than 1010 Ω·cm) and small leakage current. The high Cd vacancy concentration in the crystal is the main reason for the reduction of infrared transmittance and resistivity. At the same time, the defect complex composed of vacancy, dislocation, Te deposition and impurities can play the role of donor or acceptor in the crystal, Affect its electrical properties and produce large leakage current [6,7]. In order to improve the photoelectric performance of materials, in order to improve the performance, we improve the purity of raw materials, strictly operate and reduce the pollution of foreign impurities in the process. The unique Cd pressure control technology, the optimization of preset seed crystal process and in-situ heat treatment technology are used to compensate doping (such as Cl doping). Appropriate wafer annealing can reduce the size and density of Te precipitation in the crystal. Specific measures are as follows:

(1) Additional Cd source growth to maintain the Cd vapor pressure balance of the growth system. The Cd vapor pressure near the growth temperature is obtained from the relationship between CD1 xznxte, alloy component partial pressure and component X, and the additional amount of Cd is obtained in combination with the ampoule volume.

(2) Slow growth and slow cooling reduce the possibility of Te inclusion or Te precipitation in the crystal. A slower growth rate and a lower interfacial temperature gradient are used to grow single crystals.

(3) Improve the purity of raw materials, strictly operate and reduce the pollution of foreign impurities in the process. The raw materials Cd, Zn and Te are further purified by secondary evaporation and zone dissolution process, and the purity can reach 7 N ~ 8 N. During synthesis, the temperature oscillates to eliminate the free gas in the melt. Finally, high purity CZT single crystal raw materials were obtained.

(4) Targeted compensation doping (such as C1 doping) can further improve the resistivity of crystal materials.

Now we have successfully grown the crystal diameter Φ 90 mm, resistivity ρ CD1 xznxte crystal ≥ 1011 Ω·cm. Infrared transmittance: > 60%, dislocation density: EPD < 1×10^4 cm^{-2}. Half peak width of X-ray diffraction: FWHM: 10 ~ 20 arcsec. Inclusions and sedimentary facies: particles less than 10 um and quantity less than 2×10^4 cm^{-2}. The X-ray morphology image is uniform and the lattice structure is relatively complete.

DOI: https://doi.org/10.26549/met.v6i1.9507

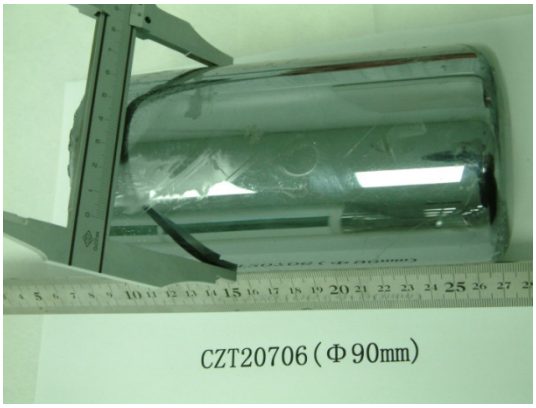


Figure 3. Growing Φ 90mm CD1 xznxte crystal

3. Detector Preparation Process

The preparation process of CdZnTe detector is: material selection - wafer processing - electrode fabrication - passivation - packaging.

In the preparation process of CdZnTe detector, the leakage current is the main source of noise for the detector. The surface leakage current formed by the surface conductive layer and the influence of electrode injection or blocking on the bulk leakage current are two important problems that seriously reduce the detection efficiency [8]. In addition, the internal electric field distribution and the composite centers in the body and near the surface are the factors affecting the current collection efficiency and leakage current. Therefore, the preparation process of the detector needs to be optimized. The specific manufacturing process of the detector is as follows:

(1) Select the selected CdZnTe material according to the designed size (3 mm × 3 mm × 3 mm) physical cutting;

(2) SiC (9) for μm, 3 μm) After mechanical grinding of the abrasive, use diamond suspension (1 μ m) Mechanical polishing;

(3) After polishing, the polished CZT wafer was chemically treated with 5% br + methanol (BM) and 2% br + 20% lactic acid + ethylene glycol (LB), and the corroded wafer was rinsed in methanol to remove the residual BR on the surface;

(4) Au electrodes were plated on both sides of the sample by magnetron sputtering;

(5) The non electrode sidewall of the sample was chemically passivated with koh-kcl and NHF/H_2O_2 aqueous solution.

(6) Assemble the detector on the prepared ceramic base for packaging.

The purpose of corrosion is to make the crystal surface clean, uniform and high finish. To achieve these, we must carefully select the type, concentration, corrosion time and temperature of the corrosion solution. Generally, br-ch30h solution is used for corrosion. After repeated experiments, the best corrosion condition of the wafer is determined as follows: use 5% bromomethyl alcohol solution as the corrosion solution and corrosion at 300 K for 30 s. When the wafer is corroded by bromomethanol for 30 seconds, the detector shows better I-V curve performance, and the resistivity of the detector is increased by 1-2 orders of magnitude compared with other corrosion times. This is because appropriate bromomethanol corrosion can remove the surface strain layer formed in the machining process and reduce the influence of surface defects and impurities. However, the corrosion solution is non-uniform corrosion, which will cause the deviation of stoichiometric ratio on the wafer surface (generally rich in TE). The interface layer can also form excessive leakage current in the device. For example, ch3chohco2h (lactic acid) + br-hoch2chzoh (ethylene glycol) corrosion solution can reduce this effect. After the wafer is corroded by bromine methanol, bromine removes the surface damage layer to make the surface smooth. If it is corroded again by lactic acid, the crystal surface will be more smooth. Further reduce the surface leakage current.

In the process of manufacturing semiconductor devices, it is inevitable to encounter the problem of metal semiconductor contact. Because semiconductor devices have to lead out electrodes or need to be connected with external circuits, even if the materials used have excellent properties, if there is no suitable contact, the performance of the device will be degraded. At least one of the two electrodes of the detector is designed to block contact, which can reduce leakage current and improve energy resolution. For high resistance materials, if the two electrodes are designed as ohmic contact, the response speed of the detector can be significantly improved, and the electrical performance test of high resistance materials is also inseparable from ohmic contact. Therefore, good ohmic contact is a necessary condition to ensure the normal operation of semiconductor devices.

We experimented with Au, in and C as contact materials, using the combination of c-czt-c, au-czt-au, au-czti-n and in CZT in, considering only from the work function, Φ In < Φ CZT, Φ Au > Φ CZT, Φ C > Φ CZT, c-czt-c and Au CZT Au can theoretically form single hole injection, Au CZT in is double injection or no injection, and in CZT in is single electron injection. From the I-V characteristics, single injection will form the space charge limiting current of i-v2 relationship. We first consider that as a p-type semiconductor, the metal with high work function is easy to form ohmic contact with it. Among Au, in and Al as electrode materials, AU

 DOI: https://doi.org/10.26549/met.v6i1.9507

has the largest work function. Therefore, we use the method of vacuum evaporation of Au and appropriate diffusion treatment to form ohmic contact.

After CdZnTe surface is treated with absolute ethanol and bromine containing corrosive solution, it is difficult to obtain a fresh surface without oxidation. Improper surface treatment will also aggravate the surface oxidation, which will affect the performance, yield and reliability of the device. For the detector, it is required to strictly control the surface of the material, so that the surface has uniform chemical ratio, no crystal defects, little surface oxidation and other surface contamination. Various charged particles adsorbed on the semiconductor surface and movable ions, fixed charges, trap charges, etc. in the semiconductor surface oxide layer can lead to the formation of surface space charge region, so as to increase the surface leakage current, change the electric field distribution in the surface layer, reduce the performance and stability of the detector, and obtain a detector with good and stable performance, A stable passive film shall also be deposited on the wafer surface. We passivate the wafer made above with koh-kcl and NHF/H_2O_2 aqueous solution to form a film with high resistivity on the wafer surface, which can effectively reduce the surface leakage current. At the same time, after forming a thin layer, it can effectively saturate the hanging bond on the wafer surface, so as to prevent surface pollution.

4. Performance Test

We tested the performance of the prepared CdZnTe detector in the following aspects:

4.1 Resistivity

According to the working principle of photoconductive device, two important factors affecting the performance of detector are material resistivity and trap concentration. The higher the resistivity of the detector material, the smaller the leakage current, and the lower the noise of the detector. Defects in the crystal will become carrier traps or recombination centers. Carrier traps will capture photogenerated carriers and form space charges, which will reduce the electric field intensity of the detector depletion layer and reduce the working stability of the detector. When the concentration of recombination center in the crystal is high, it will shorten the lifetime of photogenerated carriers, reduce the mobility Lifetime product of carriers, reduce the collection efficiency of photogenerated carriers, and reduce the energy resolution of the detector. And the performance of nuclear radiation detector becomes worse with the extension of time. Therefore, studying the electrical properties of CZT crystal and measuring its resistivity and trap concentration are important experimental means to evaluate the properties of materials.

We use Keithley 8002a high resistance tester to test the resistance of CdZnTe detector, and its resistance is 5.1 × 1010 Ω, the resistivity is 1.5 calculated according to the size of the detector × 1011 Ω·cm。

4.2 I-V Characteristics

A good current voltage characteristic (I-V characteristic) is a necessary condition for a high-quality detector. The leakage current of the detector is the main source of its noise. It directly affects the energy resolution and sensitivity in ray measurement. It is one of the important parameters of the detector.

The leakage current of detectors prepared under different process conditions varies greatly. Under the condition of adding working bias voltage and avoiding light, the current flowing in the detector when no particles are incident is called the leakage current of the detector. There are three main sources of leakage current: body leakage current, diffusion leakage current and surface leakage current. I-V characteristic is a simple method to judge the preparation process and the quality of finished products. In the experiment, we first judge whether the surface treatment is successful according to the leakage current at room temperature. When the leakage current is small. Then it is judged by the straight line fitting of I-V characteristic curve good or bad contact.

In the process of making semiconductor devices, it is inevitable to encounter the problem of metal semiconductor contact [9-11]. If both electrodes are designed as ohmic contact, the response speed of the detector can be significantly improved, and the electrical performance test of high resistance materials is also inseparable from ohmic contact. It can be seen from the I-V test curve in Figure 4 that the CZT electrode made by this method realizes ohmic contact.

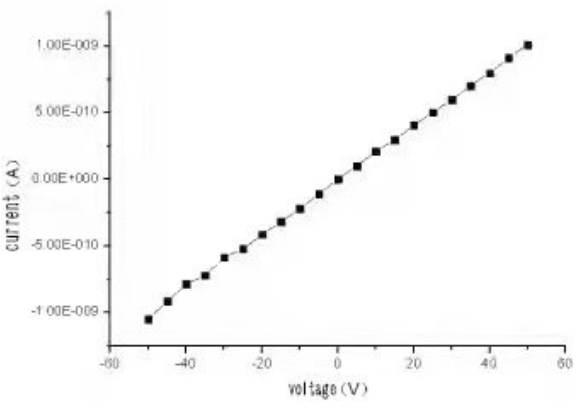


Figure 4. I-V characteristics under Au CZT Au contact

DOI: https://doi.org/10.26549/met.v6i1.9507

4.3 Energy Spectrum Response and Resolution

For an ideal detector, the pulse height generated by each X-ray with determined energy should be equal. For single energy rays, all pulse counts shall be on the same track number. In the actual detection process, the role of ray and crystal is more complex. Part of the incident photons do not exchange energy with the crystal completely, but directly pass through the detector; Some photons are scattered due to Compton effect; Some photons produce secondary radiation in the detector. These factors will increase the statistical fluctuation of the number of carriers produced by the ray, widen the spectral line of the energy spectrum and affect its resolution. In addition, the current fluctuation in the detector is also caused by the capture of carriers by traps and the recombination of electron hole pairs, which affects the resolution.

The resolution of the detector includes three aspects: spatial resolution, time resolution and energy resolution. Spatial resolution refers to the resolution of array multi-channel detector to spatial solid angle, which is usually expressed by linewidth that can distinguish a certain distance. In this study, only single channel detector is discussed, and the problem of spatial resolution is not involved. Time resolution refers to the time difference between two signals that the detector can distinguish. It requires the detector to have short response time and afterglow time.

Energy resolution is an important parameter to characterize the performance of detector, which refers to the ability to separate spectral lines with similar energy. It is often measured by the half height width or relative linewidth of the spectral line. One definition method is to express the energy resolution by the ratio of the half height width of the spectral peak to the energy of the spectral peak. Namely $ER = \frac{FWHM}{E_{peak}}$.

DC coupled preamplifier with working bias voltage of 300V is adopted for am241 γ. As shown in Figure 5, the main peak position of 59.54 keV is 250.42 channels, the half width of 59.54 keV peak is 8.83 channels, and FWHM = 2.10 kev (resolution is 3.52%).

4.4 Polarization Effect and Stability

II-VI compound semiconductor materials are considered to be prone to polarization effect when making detectors. This effect is due to the internal electric field generated by the charge accumulated in the detector, which weakens the external electric field. With the extension of bias time, the resolution and detection efficiency of the detector gradually decrease [12,13]. We kept the detector working for a long time and observed the energy spectrum response after 2 hours. It was found that there was no obvious polarization effect and the detector worked stably. At the same time, we retested the same detector after being stored in an ordinary laboratory environment for 6 months, and its performance did not change significantly.

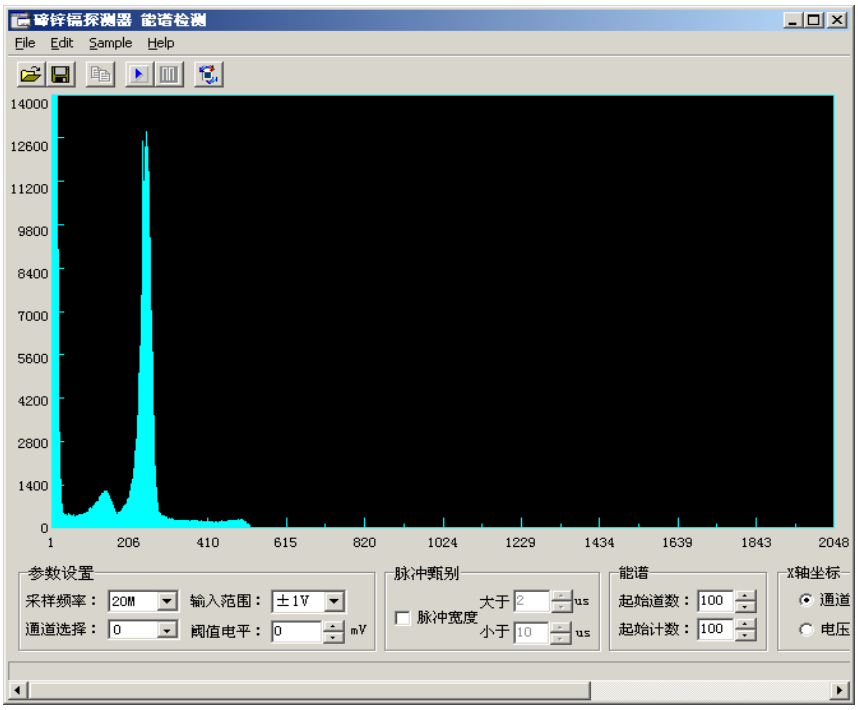


Figure 5. Energy spectrum response test of CZT detector to am241

 DOI: https://doi.org/10.26549/met.v6i1.9507

5. Conclusions

CZT crystals were grown by improved Bridgman method. Large size CZT crystals were grown by using Cd pressure control technology, optimization of preset seed crystal process and in-situ heat treatment technology (Φ 90 mm), high resistivity ($\rho \geq 1011\ \Omega \cdot cm$). A ray detector with no polarization effect, stable performance and good energy spectrum resolution is prepared by using a unique detector preparation process. The test results show that the detector has met the requirements of practical application.

References

[1] Bertolucci, E., Maiorino, M., Mettivier, G., et al., 2002. Nucl Instrument Methods. A487, 193.

[2] Czock, K.H., Arlt, R., 2001. Nucl Instrument Methods. A458, 175-182.

[3] Asahi, T., et al., 1996. Journal of Crystal Growth. 161, 20.

[4] Limousin, O., 2003. New trends in CdTe and CdZnTe detector for X-and gamma-ray applications. Nucl Instrument Methods. A504, 24.

[5] Li, W.W., Sang, W.B., 2004. Research progress of CdZnTe Nuclear radiation detector materials and devices. Shanghai Nonferrous Metals. 25(2), 88-89.

[6] Peter, K., 1990. The Pressure-Temperature-Composition Projection of the system Cadmium-tellurium. Crystal Res. Technol. 25, 1107.

[7] Hiroshi Kimura, M., Komiya, H., 1973. Melt Compositions of II-VI compounds during Crystal Growth in a Hight-Pressure Furnace. Journal of Crystal Growth. 20, 283-291.

[8] Rybk, A.V., Leonov, S.A., Prokhoretz, I.M., 2001. Influence of detector surface processing on detector performance. Nuclear Inst & Methods in Physics Research A. 458, 248-253.

[9] Morton, E.J., Hossain, M.A., Antonis, P.De., et al., 2001. Investigation of Au-CdZnTe contacts using photovoltaic measurement. Nuclear Instruments & Methods in Physics Research. A458, 558-562.

[10] Lachish, U., 1999. CdTe and CdZnTe semiconductor gamma detectors equipped with ohmic contacts. Nuclear Instruments and Methods. A436, 146-149.

[11] Ricq, S., Glasser, F., Garin, M., 2001. Study of CdTe ang CdZnTe dectors for X-ray computed tomography. Nuclear Instruments and Methods in Physics Research. 458, 534-543.

[12] Malm, H.L., Martini, M., 1974. Polarization Phenomena of CdTe nuclear Radiation Detectors. IEEE Transactions on Nuclear Science. 21, 322-330.

[13] Bell, R.O., Entine, G., Serreze, H.B., 1974. Time dependent Polarization of CdTe Gamma-ray Detectors. Nuclear Instruments & Methods. 117, 267-271.

DOI: https://doi.org/10.26549/met.v6i1.9507

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Automatic Sentiment Classification of News Using Machine Learning Methods

Yuhan Wang*

Chengdu Foreign Languages School, Chengdu, Sichuan, 643000, China

ARTICLE INFO

Article history
Received: 14 February 2022
Revised: 21 February 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Machine learning
Automatic classification of news sentiment
Specific measures

ABSTRACT

With the rapid development of social economy, the society has entered into a new stage of development, especially in new media under the background of rapid development, makes the importance of news and information to get the comprehensive promotion, and in order to further identify the positive and negative news, should be fully using machine learning methods, based on the emotion to realize the automatic classifying of news, in order to improve the efficiency of news classification. Therefore, the article first makes clear the basic outline of news sentiment classification. Secondly, the specific way of automatic classification of news emotion is deeply analyzed. On the basis of this, the paper puts forward the concrete measures of automatic classification of news emotion by using machine learning.

1. Introduction

In the context of the rapid development of modern technology, the number and scale of online resources are constantly improving, and most of these resources are in the form of text, which makes text classification become an important technical means to process this part of document data information. Therefore, the importance of computer technology is further highlighted, through the computer to carry out automatic classification of all kinds of text, is an important part of the field of natural language processing. Would expand the current automatic text classification to study, the main is adopted by the machine learning methods, and according to the current theme and content to all kinds of text analysis, in order to ensure the machine learning approach can effectively play in the emotional automatic classification of news out the actual effect, should be in-depth study of machine learning. Make sure it can categorize text based on emotion.

2. The Basic Overview of News Sentiment Classification

News and the comments on the content, the main can be classified two types, respectively, the subjective and objective, the subjective type mostly appeared in the news review. It not only belongs to a kind of objective description of the content of the event. It also introduces the author's own ideas and judgment, has a relatively strong emotional color. Thus, the main content of the news and the reported object are reviewed. The objective type is commonly seen in news, which is the objective descrip-

**Corresponding Author:*
Yuhan Wang,
Chengdu Foreign Languages School, Chengdu, Sichuan, 643000, China;
Email: 2126828166@qq.com

 DOI: https://doi.org/10.26549/met.v6i1.11132

tion of the news events. Pure objective news reports only belong to a rational form, and most news reports have a certain degree of subjectivity.

Li Liangrong, a professor at the school of journalism of Fudan University, believes that the application of journalistic professionalism in Contemporary Chinese journalistic practice includes the following three points: first, only the media that openly deliver information to the society can talk about journalistic professionalism, information communication of personal opinions, small-scale information. Opinion diffusion does not apply to journalistic professionalism. Second, journalistic professionalism only applies to professional news organizations. Third, journalistic professionalism is applicable to the transmission of serious news. Journalistic professionalism cannot be used to require gossip news.

In his article subversion or reconstruction: on "journalistic professionalism" in the new media environment, Zhong Danian, a professor of Communication University of China, clearly distinguished the respective characteristics of traditional news (organization, authority, objectivity and professionalism) and self media news (autonomy, fragmentation and emotion). With regard to journalistic professionalism in the new media environment, Zhong Danian believes that just as the traditional journalism has experienced the road of professional and standardized system construction from minority to mass, from politics and business to public, and from disorder to order, the new media journalism represented by we media is leading the transformation of media from communication to mutual broadcasting. They must also experience the growth process from news spontaneity to news consciousness. Therefore, the new media era is bound to construct a set of ideas, norms, professional ethics and skills to adapt to the new media news, which has become the journalistic professionalism in the new media era [1].

As a Chinese discourse, journalistic professionalism does not mean that it is completely separated from the western context. On the contrary, when discussing journalistic professionalism, western communication scholars will always consciously or unconsciously regard China as a reference, and there is no exception in the three years from 2015 to 2017. When Oppegaard published the article "the boundaries of Journalism: professionalism, practice and participation" on Journalism & Mass Communication quarter, he not only discussed the new practice of professionalism in theory, but also made a meaningful response to the voice in China. Graves and Lucas emphasize rediscovering and examining professionalism from a global perspective. In reinventing professionalism: Journalism and news in a global perspective, he believes that from the perspective of global information circulation, journalism and journalism are rewriting people's cognition of this industry and profession, which will also extend to the reshaping of professional standards. Under the new communication form dominated by new technology and characterized by social platform and public participation, every individual such as news practitioners and the public is constantly involved, occupied and penetrated into each other on the network nodes of news information production and transmission. When professional journalists and the public are transmitting information in the form of news, they are also defining what news is and realizing the understanding of news together [2].

Lu Ye, a professor at the school of journalism of Fudan University, and others put forward the concept of "liquid journalism" in the article "liquid Journalism: Rethinking new communication forms and journalistic professionalism". In liquid news, the identity and role of journalists are no longer stable, but constantly changing among professional journalists, citizen journalists and the public. Professional journalists and citizen journalists have their own news communities, but their different news production groups interact with each other to reconstruct their work objectives, time norms and the ideology of journalism. The article points out that if we regard journalistic professionalism not only as the expectation of good media publicity and the professional role of journalists, but also as an integral part of the social and cultural value system with free expression and public participation as the core, the concept of journalistic professionalism and its discourse practice will still be an important discourse resource to promote social progress, And it has new theoretical significance of universal care. According to Professor Lu Ye, journalistic professionalism and civic professionalism exist and influence each other in the new media era, and jointly promote the progress and development of the journalism industry [3].

Hu Yiqing of Nanjing University put forward the political economics research perspective of journalistic professionalism in his article "criticism of journalistic professionalism: A Perspective of communication political economics". From the perspective of communication political economy, journalistic professionalism is a concept and way of media enterprise management. However, in order to cover up its highly utilitarian practical function, it is usually advertised as the professional ethics and highest belief of the journalism industry. The historical causes are closely related to the penetration of American scientific management thought into all walks of life at the turn of the 19th and 20th centuries. Therefore, the challenge of contemporary citizen journalism to journalistic profes-

DOI: https://doi.org/10.26549/met.v6i1.11132

sionalism is a challenge to the traditional news management model in the 21st century. People always need professional news, but they don't necessarily need capitalist news enterprises.

In his article "the past and present life of journalistic professionalism", Wang Weijia pointed out that in China's journalistic practice, the louder the slogan of independence and professionalism, the lower the ability of journalists to grasp social problems as a whole. Moreover, under the market logic, excluding the awareness of political and social responsibility, there are a large number of negative phenomena such as vulgarity, rumor and sensationalism in the news industry. In this process, the negative results brought by the process of American news industrialization and its accompanying news thought appear more and more in the news practice around us. How to really return to our own life world to interpret and study the practical problems of journalism is the real direction of Journalism and the industry, rather than blindly emphasizing the journalistic professionalism under the essence of utilitarianism.

(1) Subjective type

In part of news comment, it is with the author's own content for the object and the report's views and attitudes, this part of the attitudes in most cases are embodied in the author's language, can be in the words of the author used to distinguish between specific emotional color and tone, which for sure evaluation report objects, and contains the praise the color is positive, And those that give negation and have negative feelings are negative.For example, there are several Sony-related headlines: SONY names a new CEO. With the appointment of foreign CEO, SONY entered a new stage of development. SONY's change of CEO is Japan's biggest funeral. In these headlines, the main content is about stringer's inauguration as the president of SONY, but in terms of words, it can clearly feel the author's own views and attitude. In the first title, it is a neutral point of view, without significant positive or negative meanings. The word "enter" is used in the second news title to express the author's affirmation of the news fact, which is a positive word. The third title uses the derogatory term funeral, which indicates that the author is not optimistic about Stringer's appointment, which is negative.

(2) Objective type

Even though the news content is mainly reported based on objective facts, from the perspective of the reported object, there is still a difference between positive and negative content of the specific report, which further reflects the attitude and opinion of the media and the author, and will also bring positive or negative impact on the reported object. Among them, the good influence can be called positive news, and the bad influence belongs to negative news. For example, there are several headlines: Procter & Gamble cosmetics was found to contain prohibited ingredients and Procter & Gamble donated 4 million yuan to Project Hope. Among the two headlines, the first one is negative news, while the second one is positive news [4].

3. The Specific Way of Automatic Classification of News Emotion

The text classification of news and comment content based on emotion belongs to a two-category classification problem in essence, in which the main target types are only positive and negative, which requires the adoption of machine learning with better performance to better achieve the emotional text classification of news.

(1) Naïve Bayes

Naïve Bayes Classifier is a probabilistic classifier, in which the distribution of specific features of the category and prior probability are taken as the basis, so as to accurately calculate the probability of which category the location document belongs to Naïve Bayes. The main feature is that it can assume the words appearing in the document and change them into a mutually independent state. Although this probabilistic method is relatively simple, it still belongs to a classification method with better application effect. In the process of classifying news based on emotion, text vector space model is also used to further show the document content, specifically as attribute value. This is going to be larger Naïve Bayes. The application efficiency of classifier, and the positive and negative categories in a document belong to the classification with the highest probability of words being observed in the document.

(2) Maximum entropy

This mode of maximum entropy is mainly to find a model with relatively uniform distribution on the premise of meeting the basic requirements of the system. Simply speaking, it is to recruit the maximum model of new entropy. Most of them will be all kinds of the facts of the known as the main restriction conditions, further, to calculate the maximum entropy probability and the probability distribution of concrete as the probability distribution of the main, can also be training focused on the data and information related to classification, described its effective as one of the main features of the whole series. These characteristics in the general case belong to binary function. For text classification based on emotion, feature words should be regarded as one of the specific features, and whether feature values belong to word frequency or binary values should be determined according to the ap-

 DOI: https://doi.org/10.26549/met.v6i1.11132

plication situation, so as to better adapt to the application requirements of text classification at the document level [5].

4. The Use of Machine Learning to Carry out Automatic Classification of News Sentiment Specific Measures

(1) Construct a red-black dictionary

Build standards. Whether at the document level or the word level, it is the words used in news content that determine emotional orientation. Therefore, in using machine learning approach to news emotion in the process of automatic classification, is one of the basic content of words to identify emotional tendency, in the process of development in recent years. However, both Chinese and English, have established a complete dictionary, which covers all the emotional tendency of words very hard, the main cause of this problem. It is because most words express different emotional tendencies in different contexts. Emotional words and subject words collected in thesaurus have great influence on the model in terms of emotional accuracy and vocabulary size, which further highlights the importance of thesaurus improvement. Therefore, we must construct a higher accuracy of thesaurus, enhance the emotion classification precision degree, and in the actual process of sentence level emotion classification, key content is the emotion classification of all kinds of subjective statement, as long as the objective statement and subjective statements effectively distinguish, can improve the emotion of classification model. In the recognition of subjective sentences, the importance of dictionaries is further highlighted. Subject words and emotional words in subjective sentences can be more accurately identified through a relatively complete thesaurus, so as to comprehensively improve the recognition accuracy and efficiency [6].

Construction measures. In the process of perfecting the dictionary, the most important is the emotional vocabulary and word glossary. This is because in this part of the word from word, have a significant negative or positive, it also makes this part is called "red and black dictionary", thesaurus is also contains the positive emotional words and negative emotional words thesaurus. Therefore, in the process of using machine learning methods, shall further of thesauri and emotion word table, to be perfect in the first place is given priority to captioned table, including some of the collected mainly news have a political colour words, such as the political or national government agencies more and bright color terms, including setting up polarity fields. In order to effectively show the emotional polarity of the theme words, the default +1 belongs to the weight of the positive theme words, and because the negative theme words have a large impact on the emotional tendency, the default -9 is the weight of the negative theme words. Followed by emotional vocabulary, collected in the emotional word table is some political tendency of adjectives, nouns or verbs and other parts of speech, but in the process of development in recent years, the word is also still is in the process of a kind of perfect, which despite the necessary fields, emotional words have polarity fields with the same settings, in order to express the emotional polarity of emotional words.

(2) Constructing a Chinese news corpus

In the process of emotional classification of news, machine learning should be used to select a good algorithm or a good model. In order to ensure its wide application in practical projects, it is necessary to build a corpus of higher quality. However, in the current social environment, there are relatively few corpora based on Chinese emotion classification, especially those based on Chinese news. Therefore, it is necessary to actively build a Chinese news corpus. And in the process of construction of emotional corpus, stand in the perspective of the frequent characteristics, sparse data has always been the main bottleneck of machine learning methods, and its as the main source of knowledge of sentiment analysis, more should establish a big emotional corpus, in the language specification, expected acquisition provides rules and expect to develop aspects of content.

Build a sentence-level corpus. In the actual process of construct sentence level corpus, shall specify the building of the basic standards, mainly is to collect all kinds of the political tendency of Chinese news information. This part will be collected by the data preprocessing, further eliminate the noise of the existence of text data, resulting in improved quality TXT file, because of the corpus of sentence level. Sentences should be taken as the main unit, and the standard of its internal corpus collection is relatively simple, as long as the Chinese news contains political tendencies, no matter it is positive or negative emotions.

Build a document-level corpus. In the Chinese news corpus of document level, the construction of the specific standards is to adopt the way of text reading, will the Chinese news text data collected information effectively is divided into positive and negative both types, and will be positive or negative category tag effectively reflected in the name of the file, it will also be able to better. At the same time, the level of document corpus and sentence level emotional corpus construction, the existing difference lies in the construction of more strict, it must be based on Chinese news area as the core, only such ability can effectively reflect the model with features, and in using artificial reading way to distinguish between Chinese news text in the process of emotional tendency. Political orien-

DOI: https://doi.org/10.26549/met.v6i1.11132

tation should also be accurately grasped so as to conduct a more comprehensive screening and ensure that machine learning can play a better role in automatic classification of news emotions [7].

5. Conclusions

In the current social environment, text classification based on emotion for news and comments can better help enterprises or individuals to take targeted measures to effectively reduce the emergence of all kinds of negative news in the media, so as to avoid negative impact on their reputation. Therefore, in order to further enhance the emotion classification efficiency, machine learning method, which will be taken to better the news comment on the surface of the positive and negative emotion classification, and in the process of using machine learning methods, also pay attention to build the red and black dictionary and Chinese news corpus, thus provide a more solid and strong support for the sentiment analysis.

References

[1] Oppegaard, B., 2016. Boundaries of Journalism: Professionalism, Practices and Participation. Journalism & Mass Communication Quarterly. 93, 3.

[2] Graves, L., 2016. Reinventing Professionalism: Journalism and News in a Global Perspective. New Media & Society. 18, 521-527.

[3] Michael, H., 2016. Journalism after All: Professionalism, Content and Performance—A Comparison between Alternative News Websites and Websites of traditional newspapers in German Local Media Markets. 16, 1062-1084.

[4] Jiang, Q.L., Chen, Z.H., Chen, X.J., 2021. Continuity and Change: An Analysis of the Current Situation of Emotion Research in Journalism in China. China Publishing. (10), 17-23.

[5] Lin, S.Q., Yu, Zh.T., Guo, J.J., 2020. Goldman Sachs Xiang. Journal of Kunming University of Science and Technology (Natural Science Edition). 45(06), 67-73.

[6] Li, T.C., Wang, H., Fang, B.F., 2019. Chinese news sentiment classification based on mic-cnn method [J]. Journal of Shanxi University (Natural Science Edition). 42(04), 746-754.

[7] Xu, Y., 2018. Analysis on the grasp of emotional scale of news under the new media environment. Guide to Journalism Research. 9(13), 84-85.

 DOI: https://doi.org/10.26549/met.v6i1.11132

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Research on Handwritten Chinese Character Recognition Based on BP Neural Network

Zihao Ning*

China University of Geosciences Automation College, Wuhan, Hubei, 430070, China

ARTICLE INFO

Article history
Received: 15 March 2022
Revised: 22 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Pattern recognition
Handwritten Chinese character recognition
BP neural network

ABSTRACT

The application of pattern recognition technology enables us to solve various human-computer interaction problems that were difficult to solve before. Handwritten Chinese character recognition, as a hot research object in image pattern recognition, has many applications in people's daily life, and more and more scholars are beginning to study off-line handwritten Chinese character recognition. This paper mainly studies the recognition of handwritten Chinese characters by BP (Back Propagation) neural network. Establish a handwritten Chinese character recognition model based on BP neural network, and then verify the accuracy and feasibility of the neural network through GUI (Graphical User Interface) model established by Matlab. This paper mainly includes the following aspects: Firstly, the preprocessing process of handwritten Chinese character recognition in this paper is analyzed. Among them, image preprocessing mainly includes six processes: graying, binarization, smoothing and denoising, character segmentation, histogram equalization and normalization. Secondly, through the comparative selection of feature extraction methods for handwritten Chinese characters, and through the comparative analysis of the results of three different feature extraction methods, the most suitable feature extraction method for this paper is found. Finally, it is the application of BP neural network in handwritten Chinese character recognition. The establishment, training process and parameter selection of BP neural network are described in detail. The simulation software platform chosen in this paper is Matlab, and the sample images are used to train BP neural network to verify the feasibility of Chinese character recognition. Design the GUI interface of human-computer interaction based on Matlab, show the process and results of handwritten Chinese character recognition, and analyze the experimental results.

1. Introduction

Today, with the increasing popularity of artificial intelligence, artificial intelligence has been gradually integrated into our lives, penetrating into various fields and leading the development of new technologies. As an important part of artificial intelligence, artificial neural network has made great progress in recent years. Artificial neural network has been widely concerned because of its strong function of searching for optimal solution by associative storage of things and strong self-learning ability. The multilayer feedforward BP (Back Propagation) neural network trained by error back propagation algorithm has good nonlinear mapping ability and good flexible network structure. At present, the research on BP neural network is relatively mature, so it is very meaningful to study BP

**Corresponding Author:*
Zihao Ning,
China University of Geosciences Automation College, Wuhan, Hubei, 430070, China;
Email: 3324536759@qq.com

DOI: https://doi.org/10.26549/met.v6i1.11359

neural network and apply it to an example of handwritten Chinese character recognition.

1.1 Research Background and Significance

As an important cultural wealth of the Chinese nation, Chinese characters bear a long history of Chinese culture, and at the same time are the most important language tools for our communication and expression. Therefore, the research on Chinese characters recognition is an important part of the development of social life, scientific culture, humanities communication and so on. After entering the 21st century, the development of the times requires that work and life be rapidly intelligent and information-based. In the past, manual recognition in Chinese character recognition, inputting characters into computer system, etc. increased the manual workload, reduced the work efficiency and productivity, and made people's work inconvenient. Therefore, the technical problems of handwriting Chinese characters for recognition need to be solved.

With the outbreak of the third industrial revolution, informationization, automation, intelligence and modernization have become the realistic needs of the rapid development of all walks of life. To a certain extent, the development of artificial intelligence is also a benchmark of national modernization. China has accordingly issued relevant policies to promote the development of artificial intelligence technology, such as "new infrastructure" and "internet plus", in an effort to lead the tide of artificial intelligence development. As a high-tech cutting-edge technology integrating many disciplines, artificial intelligence technology plays an important role in promoting the liberation of productive forces, improving work efficiency and helping industrial upgrading and transformation, and has a very far-reaching impact on the innovation of science and technology and the development of national economy. With the rapid development of computer technology, the field of deep learning and big data has also made great progress, and the field of artificial intelligence has also developed rapidly under its impetus. Deep learning has gradually become the focus of people's attention, and more and more scholars have begun to study the field of artificial intelligence. It makes deep learning popular in image recognition and other related fields.

Handwritten Chinese character recognition has attracted much attention in many fields, such as digital retrieval of documents, bank check processing, OCR (Optical Character Recognition) conversion, handwritten text input and other related fields. The introduction of computer vision technology enables users to quickly and conveniently input Chinese information, thus improving office efficiency and processing large quantities of Chinese information. Chinese Character Recognition is still an important problem in the field of image recognition. Although the recognition effect of the original "preprocessing+feature extraction+classifier" is acceptable, there are still many difficulties and shortcomings in the research of Chinese character recognition due to many confusing words, various types and complex structures. Handwritten Chinese character recognition, for standard Chinese characters, adds some problems such as writing standards, which makes handwritten Chinese character recognition still face many problems to be solved. After long-term research, deep learning put forward DBN (Deep Belief Network). DBN is very creative. It absorbs the training idea of unsupervised learning, enriches the neural network, rationalizes the model parameters, and can carry out hierarchical learning. Convolutional neural network, BP neural network and stack self-coding network, as well as network models that use different network structures to fuse for different tasks. The application of deep learning to character recognition has made great progress and breakthrough. Compared with traditional methods, it can solve the pain points in the process of Chinese character recognition, achieve better recognition effect and generalization ability, and has strong development potential.

Compared with printed Chinese characters, handwritten Chinese characters have the following characteristics:

(1) Basic stroke changes.

(2) Lian Bi phenomenon of strokes is very common in handwritten Chinese characters.

(3) The position of stroke writing and the relative position between the radicals are also different. At the same time, stroke length is different, and straight line bending is also very common.

Generally, the handwriting fonts we mainly write are running script, regular script and cursive script. By comparing the three Chinese characters, it can be seen that the glyphs and strokes of the same character are different, or even far apart. Among them, some people can't see what Chinese characters are written in cursive script. At this time, it is obviously difficult to ask the computer to recognize Chinese characters. Therefore, it is of great significance to study the computer automatic recognition of handwritten Chinese characters with different fonts.

1.2 Development and Present Situation of Domestic and Foreign Research

Handwritten Chinese character recognition is an important branch in the field of artificial intelligence and computer vision, and it is also one of the most challenging problems in OCR (Optical Character Recognition).

 DOI: https://doi.org/10.26549/met.v6i1.11359

Initial stage: OCR technology was formally put forward in 1928 and began to be studied, which experienced the research process from printing to handwriting [1]. In 1960, Casey and Nag of IBM (International Business Machines Corporation) successfully identified 1,000 printed Chinese characters by template matching, and achieved good results [2]. At the same time, the research of handwritten Chinese character recognition has also attracted extensive attention. Liu et al. of MIT put forward features based on strokes of Chinese characters, and the weight vector corresponding to each feature is determined by the learning process similar to perceptron [3].

Because there are many similarities between Chinese characters and Japanese to a certain extent, in 1977, Toshiba Company of Japan began to study the recognition of Chinese characters, and designed a system that can recognize more than 2,000 Chinese characters. Then Musashino Company designed a recognition system that can recognize hundreds of printed Chinese characters in one second with an error of 0.02% [5], and can also recognize more than 2,000 Chinese characters.

Stage: Around 1970, some people in China began to study the recognition of Chinese characters written by opponents. After all, Chinese characters are the most widely used in China. Moreover, for the writing methods, rules and norms of Chinese characters, the domestic understanding is more thorough, so after that, the research on handwritten Chinese character recognition is mostly concentrated in China. Nantong Institute of Electronics and Shenyang Institute of Automation in China have developed printed Chinese character recognition systems [6]. At this time, more and more ideas about printed Chinese character recognition have been put forward and introduced into the field of recognition: Nakano et al. began to study feature projection [7]. Yamamoto et al. developed a framework that should be based on multi-resolution matching [8]. At the same time, various variation methods are gradually applied to feature extraction of Chinese character recognition.

During this period, the development of various technologies also promoted the research of handwritten Chinese character recognition. For example, the Otus algorithm in image processing [9] is widely used in the process of converting gray image into binary image. Character adjustment [10] and character blur [11] were first proposed and studied as concepts, which also paved the way for the later Chinese character orientation feature extraction [12].

Development stage: Around 1980, the research on the recognition of printed Chinese characters began to have a qualitative leap. In 1985, Japan's Fuji Company released a reading product that can recognize various printed Japanese fonts [13]. At this time, China also launched the standard data set of commonly used Chinese characters, which contains 3755 most commonly used Chinese characters. Moreover, because Japanese characters are similar to Chinese characters to a certain extent, at that time, China also made great progress in the field of Chinese character recognition [14], and then slowly began the related research of off-line Chinese character recognition. At this time, algorithms such as neural network [15] and quadratic discriminant function have also made great progress, which also promoted the research of Chinese character recognition, and put forward several well-known models, such as hierarchical neural network [16] and implicit Markov model based on maximum mutual information standard [17]. Various algorithms and models make Chinese character recognition develop rapidly.

Re-development stage: Since 2000, Chinese recognition has made great progress in China, and in-depth research has been made in all aspects. The recognition of Chinese characters has gradually formed a three-step recognition processing mode: image preprocessing, feature extraction and pattern classification.

In the aspect of preprocessing, the elastic grid method [18] solves the problem of self-adaptive partitioning of Chinese characters with uneven overall distribution. The nonlinear method to regularize Chinese characters [19,20] also puts forward a solution to the problem of writing differences of handwritten Chinese characters.

In the aspect of feature extraction, the direction line element feature and gradient direction feature [21] also greatly improve the stability of direction feature.

In the aspect of pattern recognition, the design methods such as HMMs (Implicit Markov Model) [22], CNN (Convolutional Neural Network [23]) and LVQ (Learning Vector Quantization [24]) and the method of comprehensive use of multiple classifiers further improve the recognition rate.

Since 2006, many new handwritten Chinese character databases have been published gradually [25], and the research direction has begun to develop from single-character recognition to multi-character recognition, which makes the recognition more authentic [26]. At the same time, the combination of various neural networks and deep learning has begun to be applied to handwritten Chinese character recognition, which can adapt to various writing styles of handwritten Chinese characters with very good results.

As for the recognition application of BP neural network, in 2012, Ouyang Jun and others designed a new algorithm of license plate character recognition by combining the image processing method with BP neural network. Through a large number of BP neural network training

DOI: https://doi.org/10.26549/met.v6i1.11359

and simulation experiments, the recognition speed and accuracy of license plate were greatly improved [27]. In 2014, Liu Fang and others proposed an improved algorithm of character recognition with BP neural network, which improved the recognition rate and speed of characters in various noisy environments [28]. In 2016, Li Dan and others carried out multi-character handwritten character recognition with any number of character templates. The accuracy rate of handwritten characters is over 95% [29].

BP neural network has great advantages in handwritten Chinese character recognition because of its good generalization ability, fault tolerance ability, self-learning and adaptive ability and nonlinear mapping ability [30]. The model structure and optimization algorithm of BP neural network are constantly updated and improved, which also provides a good reference for off-line handwritten Chinese character recognition.

1.3 Research Content and Chapter Arrangement

On the basis of studying and analyzing the research background of BP neural network and handwritten Chinese character recognition at home and abroad, this paper illustrates the performance of BP neural network by training BP neural network from the aspects of Chinese character preprocessing, feature extraction method and the design of handwritten Chinese characters by BP neural network. The full text is divided into five chapters, which are as follows:

The first chapter is the introduction. Firstly, the research background and significance of this paper are summarized. Secondly, from the development process of handwritten Chinese character recognition and the development of BP neural network, the research status at home and abroad is expounded. Finally, the overall structure and chapter arrangement of the full text are introduced.

The second chapter is an overview of the preprocessing process of handwritten Chinese characters, which is introduced from the aspects of grayscale and binarization, smooth denoising, character cutting and normalization of handwritten Chinese characters, and gives the scheme steps of handwritten Chinese characters preprocessing.

In chapter 3, based on the preprocessed single character of handwritten Chinese characters obtained in chapter 2, the feature extraction methods of characters are analyzed in detail, and the most suitable image processing method is selected by comparing the results.

Chapter 4 illustrates the design of handwritten Chinese character recognition framework based on BP neural network. Firstly, the design and training process of BP neural network is analyzed. Secondly, the GUI interface is introduced. Then, the design of each module such as picture reading, preprocessing, network training and handwritten Chinese character recognition is introduced respectively. Finally, the results of handwritten Chinese character recognition are analyzed.

The fifth chapter is a summary of the work done in this paper, as well as the existing problems and future research ideas.

2. Preprocessing of Handwritten Chinese Character Images

The image used for recognition is usually a picture with multiple fonts, so it is necessary to preprocess the recognition samples to facilitate computer recognition. Moreover, for handwritten Chinese character recognition, the degree of optimization of image processing during preprocessing may directly affect the final recognition effect. Among them, the main processes of pre-processing include image gray processing, binarization, denoising, character cutting and so on.

2.1 Grayscale and Binarization of Chinese Character Images

2.1.1 Grayscale of Chinese Character Images

For an original picture, the content of the 256-color bitmap palette is very complicated. Therefore, we need to gray the image to reduce the data of the image. The process of converting a color image into a grayscale image through adjustment is called grayscale processing. The gray image of the image refers to a picture with the same R, G and B color components, and the values of the components are all between 0 and 255. There is no color difference in the whole gray scale image, only the difference in color brightness [31]. Where the gray value is large, the pixels will be brighter, and where the gray value is small, the pixels will be darker. Therefore, gray-scale processing is actually the process of quantifying the brightness value. Transformation matrix of RGB color system and YIQ color system is as follows:

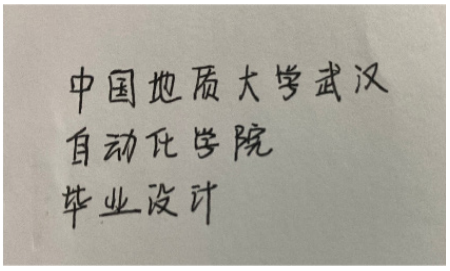
$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.523 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

Y represents the transparency of the picture, that is, the gray value of the color, I represents the hue, and Q represents the saturation. Y contains all the information of the gray scale image, so a gray scale image can be represented only by the Y component. According to formula (1), we can find the value of y component:

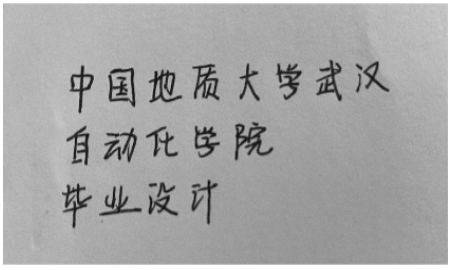
$$Y=0.299R+0.587G+0.114B \tag{2}$$

 DOI: https://doi.org/10.26549/met.v6i1.11359

Among them, the weights of R, G, and B are 0.3, 0.59 and 0.11, respectively. Figure 1 is a comparison chart of the effect of gray-scale processing of the original image.



(a) Grayscale the original image



(b) Grayscale result graph

Figure 1. Contrast chart of grayscale effect of color pictures

2.1.2 Binarization of Chinese Character Images

Before Chinese characters are segmented, it is necessary to binarize them to remove unnecessary information in the image. The binarized image can compress the image data and reduce the storage space. At the same time, it can also enhance the adaptability of the software. However, a lot of original information will be lost in the process of image binarization. How to retain the details of the original image to the greatest extent while binarizing is a big problem.

The function of transforming the gray scale image into binary image is:

$$g(x,y)=\begin{cases}0 & f(i,j)>T\\ 1 & f(i,j)\le T\end{cases} \quad (2.3)$$

where f(i,j) is the original image function, g(x,y) is the transformed image function, and t is the threshold value. Its function is to select the threshold, and then set the pixel to 255 or 0 according to the size relationship between the pixel in the image and the threshold.

There are many methods for image binarization, but their universality is usually not very good. Therefore, we should make a concrete analysis according to specific picture objects. The key to binarization of handwritten Chinese character recognition is to select a reasonable threshold. Now there are many methods for threshold selection, which can be divided into the following three categories:

(1) Local threshold method

The local threshold method is a method to determine the local threshold according to the pixel characteristics around the target pixel. When identifying characters with more disturbances, the local threshold method can handle these disturbances well. However, its disadvantages are: the implementation speed is relatively slow, and the connectivity of the processed whole image cannot be guaranteed.

(2) Whole threshold method

The whole threshold method can be divided into two categories: one is manually set, the other is determined by gray histogram.

1) manual setting

The artificial threshold is to get a reasonable threshold T according to the previous experience or our experimental results. If f(i,j)>T(i,j) is the background point, otherwise it is the pixel point on Chinese characters. This method is relatively simple, but it is not universal and has strong limitations. When the picture changes, the threshold cannot be automatically changed.

2) according to the gray histogram

This is to automatically determine the threshold through the gray image. Gray histogram can vividly describe the gray level of an image. The histogram should have two peaks, one corresponding to the background and the other corresponding to the Chinese characters. The threshold should be chosen at the valley between the two peaks. The deeper and steeper the valley is, the more obvious the binarization effect is.

(3) Dynamic threshold method

This method mainly selects the threshold according to the position of pixel coordinates and the surrounding gray values. Mainly used for processing images with low quality. However, for handwritten Chinese character images, this processing method takes too long, so it is rarely used.

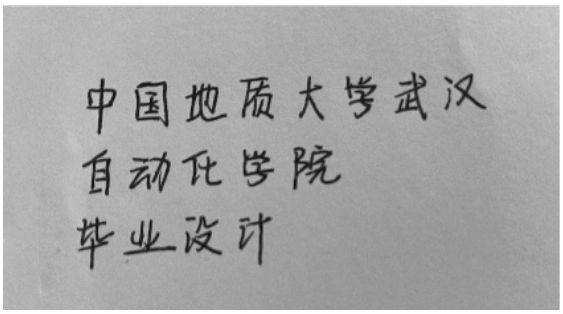
Generally speaking, the collected handwritten Chinese character images have high clarity and contrast, and the handwritten Chinese character part and background part can be clearly distinguished. Therefore, the overall threshold method can not only ensure the final image binarization effect, but also have relatively fast processing time. There will also be obvious differences in the effects after different threshold segmentation. If t is too large, interference will also be extracted; if t is too small, some information will be lost. Therefore, we choose the threshold according to the following rules:

If the gray level of the pixel is f (x, y) and the value of f (x, y) is 0~255, determine the threshold T:

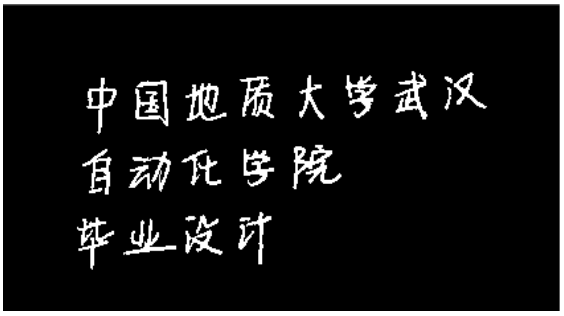
DOI: https://doi.org/10.26549/met.v6i1.11359

$$T = (f(x,y)_{\max} - f(x,y)_{\min}) / 3 \quad (4)$$

The binarized effect image is shown in Figure 2.



(a) Grayscale image before binarization



(b) The result graph after binarization

Figure 2. Binary image rendering

2.2 Smooth Denoising of Chinese Character Images

There may be various noises in the sample image. The process of eliminating these noise components in the sample image and smoothing the image is called image denoising. De-noising is a technology in image enhancement. By de-noising, useful information in an image can be highlighted, and the noise mixed in Chinese character input can also be eliminated [32]. However, there may be noise in the Chinese character pictures that need to be processed, but it is not obvious enough. If the noise is directly removed, the effect will not be very good. Therefore, if the noise is not obvious enough, it is necessary to supplement the noise first and then remove the noise, which will have better effect to some extent.

Image noise is mainly divided into internal noise and external noise. The image noise has three characteristics: superposition, correlation between noise and image, and irregular distribution.

Image denoising is to do: firstly, the lines and other important information of the image can not be damaged, resulting in information loss; The second is to make the picture to be identified clearer. There are two main ways to improve the image: First, aiming at the image noise, use a specific method to compensate the influence of noise to eliminate the noise, and make the compensated image as close as possible to the original image. This kind of method is also called image restoration. Second, it doesn't consider the cause of noise, only highlights the useful feature information of the image, and the improved image may not be completely consistent with the original image. It is mainly used to improve the identifiability of images.

In this paper, the median filter is used to remove the noise of the sample image. The median filter provides a good denoising ability for the random noise generated in. As shown in Figure 3, the 3×3 neighborhood, where p represents the pixel to be processed, assumes that the pixel value of this pixel is 150 at this time, and investigates the surrounding eight pixel values, such as 67, 97, 160, 270, 300, 250, 120, 180 respectively. Among them, the median value is taken as the new value of the pixel gray of the center point, so the pixel value of P should take the median value of 160 after the 9 numbers are sorted by size.

one	one	one
one	P	one
one	one	one

Figure 3. 3×3 neighborhood of pixel p

Considering that the handwritten Chinese characters used for recognition in this paper have little noise, the effect is not obvious or even worse after filtering. Therefore, we choose to use a smooth denoising step for the image. Partially smoothed denoising is shown in Figure 4.

(a) Binarization image before denoising

(b) Binarization image after denoising

Figure 4. Comparison before and after denoising

2.3 Chinese Character Segmentation

For Chinese character recognition, the original Chi-

 DOI: https://doi.org/10.26549/met.v6i1.11359

nese character pictures can't all be samples of Chinese characters one by one. It may be that a picture contains many Chinese characters, and handwritten Chinese characters vary from person to person. People don't always write according to the standard box, and there are great differences in position and writing size. Therefore, before handwritten Chinese character recognition, it is necessary to divide them into single Chinese characters of the same size to facilitate subsequent recognition.

At present, there are three main methods of character segmentation: the first one is based on structure, which is to analyze the structure between adjacent words to find the segmentation method; The second method is based on statistics, which is mainly used in the case of small differences in character widths. Find out the dividing line of characters through the overall distribution characteristics of characters. The third method is based on recognition, that is, before the image is segmented, identify all kinds of segmentation results of the image, and then identify the segmentation points according to the final segmentation results.

The character segmentation method adopted in this paper is vertical projection method:

At present, the vertical projection method is the most commonly used cutting method, which is simple to implement and faster than other methods. Firstly, the image area is projected in the row direction, and the statistical histogram of black pixels is obtained. The histogram part with Chinese characters will show a peak state, while the background part between Chinese characters will show a trough shape. After processing Chinese characters' rows, it can be divided into individual characters by the same projection in the column direction.

The specific implementation steps of single character segmentation are as follows:

First, read in the binarized picture, and extract the character height from the feature extraction link: d, and according to the mathematical formula of histogram:

$$D(k)=\sum_{k\geq 1}^{d} S(k) \tag{5}$$

Draw a horizontal projection histogram D, and then detect and analyze the peaks and valleys of the histogram;

Then, according to the pixel distribution of the horizontal projection histogram, a reasonable threshold P is set to separate the character area from the non-character area, that is, to distinguish the rising point and the valley bottom point in the histogram. Among them, the threshold selection criteria are:

First, average the histogram:

$$w=\frac{\sum_{k\geq 1}^{d} S(K)\cdot K}{\sum_{k\geq 1}^{d} S(k)} \tag{6}$$

The minimum value of histogram is:

$$e=MIN\sum_{k\geq 1}^{d} S(K) \tag{7}$$

Then average the sum of histogram average and minimum:

$$r=\frac{w+e}{2} \tag{8}$$

Get the optimal segmentation threshold r. After obtaining the best threshold, we scan from the left side of the Chinese character area to the rightmost side, compare the histogram of the k-th point with the threshold R. When D(k) is greater than the threshold, it is judged as the character area, that is, the k-th rising point, and create a sequence T to record the position of the rising point, which is denoted as T(k), and similarly, create a sequence Y to record the k-th valley bottom. When T(k) is less than or equal to the threshold value, it is judged as a non-character area, that is, the valley bottom point, and the number of statistical valley bottom points is recorded in the variable U, which is used to calculate the width of the k-th valley bottom, that is, D(k).

Figure 5 shows the projection histogram of handwritten Chinese characters.

Finally, each segmented single character is stored in the image and normalized, and the recognition result is displayed in the text dialog box.

The research object of this paper is relatively standard Chinese characters, so there are few cases of individual adhesion. Therefore, in order to ensure the processing speed and improve the system performance, the vertical projection method is used for segmentation. The segmentation effect of Chinese characters is shown in Figure 6.

2.4 Normalization

The segmented images are different in width and length of Chinese characters, so they need to be normalized to convert the images to be recognized into a unified standard form.

In handwritten Chinese character recognition, there are various changes such as size and position. Many normalization methods have also been proposed, which adjust handwritten Chinese character images to the same size by processing, so as to facilitate feature extraction and recognition.

In order to keep the original Chinese character features unchanged as much as possible, the normalization process

DOI: https://doi.org/10.26549/met.v6i1.11359

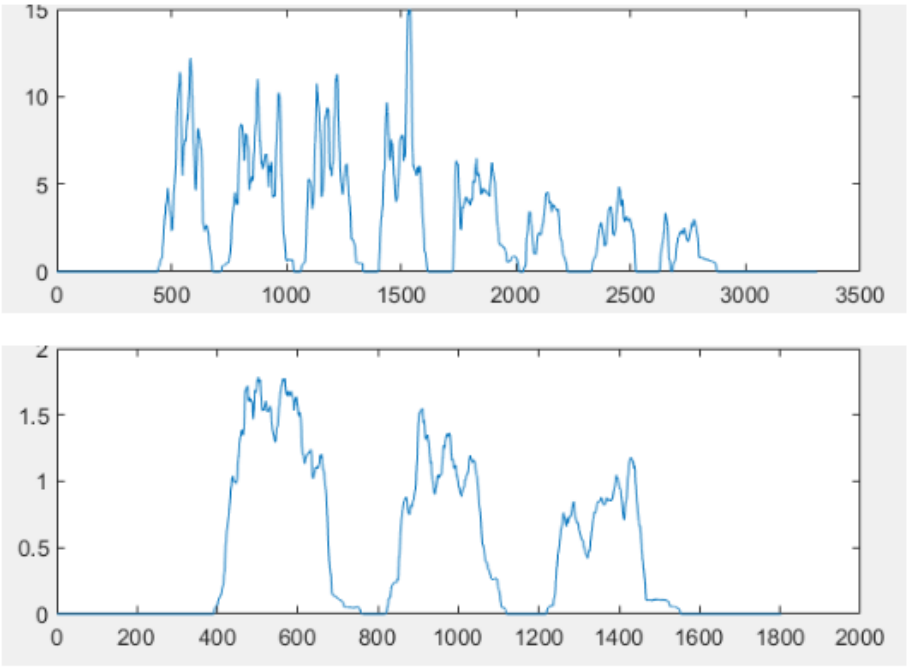


Figure 5. Projection histogram of handwritten Chinese characters

Figure 6. Character segmentation rendering

requires that the useful features contained in the image should not be changed. By normalizing the handwritten Chinese character images, the points belonging to the same category in the feature space can be closer. The processing of image normalization is mainly size normalization.

Character images of different sizes become Chinese characters of the same size after normalization. Methods The size of black pixels is normalized according to two directions, that is, the centroid GI, Gj of each Chinese character is calculated according to formula (9), then the divergence in I and J directions is calculated according to formula (10), and finally the Chinese character image is scaled up or down into a dot matrix of specified size.

$$\begin{cases} G_I = \sum_{i=D}^{U}\sum_{j=L}^{R} i.c(i,j) / \sum_{i=D}^{U}\sum_{j=L}^{R} c(i,j) \\ G_J = \sum_{i=D}^{U}\sum_{j=L}^{R} j.c(i,j) / \sum_{i=D}^{U}\sum_{j=L}^{R} c(i,j) \end{cases} \tag{9}$$

$$\begin{cases} \sigma_J^2 = \sum_{i=D}^{U}(\sum_{j=L}^{R} c(i,j)) * (i - G_I)^2 / \sum_{i=D}^{U}\sum_{j=L}^{R} c(i,j) \\ \sigma_I^2 = \sum_{i=D}^{U}(\sum_{j=L}^{R} c(i,j)) * (i - G_I)^2 / \sum_{i=D}^{U}\sum_{j=L}^{R} c(i,j) \end{cases} \tag{10}$$

In the formula, c(i,j) is a numerical lattice, and U, D, L and R are the four boundaries of Chinese characters, respectively. As shown in Figure 7, the images of Chinese characters are normalized by using 0-255, 0-1 and the normalization function provided by Matlab respectively.

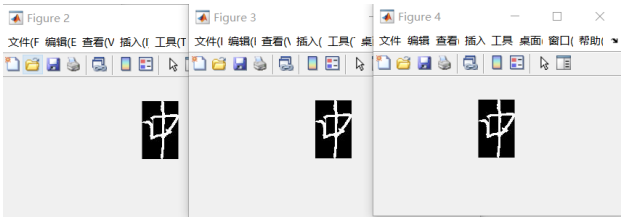


Figure 7. Three Normalization Processes

As can be seen from the above figure, the normaliza-

 DOI: https://doi.org/10.26549/met.v6i1.11359

tion effect for handwritten Chinese characters is basically the same, so this paper chooses the normalization function of Matlab to normalize.

2.5 Histogram Equalization of Chinese Character Images

2.5.1 Principle of Histogram Equalization

The histogram of an image is a statistical chart that reflects the distribution characteristics of pixels in an image. It represents the distribution curve of the image data that is dark on the left and bright on the right, instead of directly displaying the original data. The cost of image processing is relatively small, but sometimes good results can be obtained. Among them, the value of abscissa represents each pixel feature, and the value of ordinate represents the number of pixels of different color elements in the image.

The purpose of histogram equalization processing is to transform the gray values from the relatively concentrated gray areas through formula (11) according to the gray histogram of the original image, so that the gray values are evenly distributed in all gray areas. In this way, the range of gray values can be increased, thus achieving the effect of enhancing the overall contrast of the image.

$$h(v) = round(\frac{cdf(v) - cdf_{\min}}{(M \times N) - cdf_{\min}} \times (L-1)) \quad (11)$$

where m and n represent the number of pixels in the length and width of the image respectively; L is the gray scale; V is the pixel value of V in the original image; Cdf_{min} represents the minimum value in the cumulative distribution function.

Histogram equalization is roughly divided into three steps: first, count the number of each gray level in the histogram; Then, the histogram is shown; Finally, the new pixel value is calculated by the formula.

2.5.2 Histogram Equalization for Image Enhancement

For images with poor binarization effect, histogram equalization is carried out, which makes the gray values of pixels evenly distributed again, thus enhancing the image contrast and better selecting the threshold of image binarization.

Generally speaking, histogram can improve the quality of the image after image enhancement, so it is very useful for image processing. In this paper, histogram equalization is used to enhance individual images that are difficult to binarize to improve the recognition rate of images.

2.6 Summary of This Chapter

This chapter mainly introduces the preprocessing process of handwritten Chinese character images. Firstly, the image is processed by gray scale and binarization. For the image with poor effect, the histogram is enhanced before binarization. Secondly, the single Chinese character is segmented by line cutting and single character cutting, thus preparing for the next feature extraction.

3. Comparison and Selection of Handwritten Chinese Character Feature Extraction Methods

In the last chapter, after preprocessing the handwritten Chinese character image, a binary image of a single Chinese character is obtained, but it is obviously unreasonable to train and classify the binary image directly. First of all, too much image data and too much computation are unacceptable to the system; Secondly, if a lot of irrelevant information in the image is involved in the identification process, it will not only increase the workload, but also reduce the accuracy of identification, even make it impossible to classify. In any recognition process, the first step is to analyze the features and select the most representative features.

The most important thing for feature selection and extraction of handwritten Chinese characters is how to find the most representative features from all image features without losing the information of handwritten Chinese characters images.

3.1 Brief Introduction of Feature Extraction Method

In the recognition of images, it is unrealistic to classify the preprocessed images directly. The feature extraction process of Chinese characters is to find out the most representative feature vectors. There are many feature extraction methods for character recognition, which can be roughly divided into two categories. One is feature extraction method based on statistical features, and the other is feature extraction method based on structure. Statistics-based feature extraction can be extracted from binary images, while structure-based feature extraction can be extracted from the thinned skeleton after preprocessing, and the structural features of characters can better reflect local details.

Different features can show different aspects of the described things, and the extraction methods of statistical features and structural features have their own advantages and disadvantages. Statistics have strong adaptability to classifiers, but they are not sensitive to noise and have relatively good stability. However, structural features are sensitive to the changes of detail features, and can well distinguish small changes. However, structural features

DOI: https://doi.org/10.26549/met.v6i1.11359

are too sensitive to noise, and if noise is not handled well enough, there will be unpredictable recognition consequences.

Turn the extracted feature vectors of training samples into the input of the network, train the BP neural network, and then input the features of the samples to be recognized into the trained BP neural network, so that the characters can be recognized. For different recognition objects, feature extraction methods should be selected within the demand range. The following will summarize the commonly used feature extraction methods in Chinese character recognition.

Handwritten Chinese character feature extraction methods mainly include pixel-by-pixel feature extraction method, skeleton thinning method, elastic grid feature extraction method and many other methods, which can be selected according to our actual situation.

3.1.1 Feature Pixel by Feature Extraction Method

Feature-by-feature pixel extraction method is to select the sample image that has been binarized row by row and column by column, and take its feature value as 1 when the white pixel is found, or take its feature value as 0 when the black pixel is found. Then we can get a feature matrix composed of all sample feature values. This method is relatively simple, which is the most obvious for the training effect and can make BP neural network converge quickly, but its disadvantage is that it is not adaptable and insensitive to changes. Therefore, more training samples are needed to make it more adaptable.

3.1.2 Skeleton Thinning Feature Extraction Method

For handwritten Chinese characters, the same Chinese character written by different people will have a lot of recognition difficulties due to different writing methods and different line thicknesses. However, after thinning them, the line width of each Chinese character is unified to a fixed pixel width, so that the difference is not obvious. Then, the thinned Chinese character skeleton is used as a feature for training and recognition, which makes the system have certain adaptability. However, once the position deviation occurs in this method, it will be difficult to identify it.

The structural features of skeleton are mainly divided into trunk features and edge contour features:

(1) It refers to the main part of Chinese characters, mainly including the length and width of each Chinese character, the shape of strokes, etc. When extracting the features of handwritten Chinese characters, the first thing to determine is the changes of inflection points, endpoints and intersections of Chinese characters.

(2) Edge features: Edge features refer to the distance from the outermost periphery of each Chinese character to the picture boundary, etc.

3.1.3 Elastic Mesh Feature Extraction Method

If the grid is evenly distributed in the horizontal and vertical directions, we call it a uniform grid; However, if the Chinese characters are divided according to the strokes and concentration degree of Chinese characters, the grid which is non-uniform in both horizontal and vertical directions is non-uniform grid, which is also called elastic grid. Generally speaking, an elastic grid will be determined according to the histogram of Chinese characters in both horizontal and vertical directions. The feature extraction method of elastic grid is as follows:

(1) Firstly, determine the grid number n, m of the grid, and generally take n = m for convenience of calculation;

(2) Applying projection to obtain projection values Hom(i) and Vert(i) in the horizontal and vertical directions;

(3) According to the calculation formulas (12) and (13) of the net wires, determine the non-uniform net wires ii and ij in the horizontal and vertical directions;

$$\sum_{i=I_g}^{I_g} Horn(i) = \sum_{i=I_{g-1}}^{I_g} Horn(i), I_g = I_2, I_3, K, I_n \tag{12}$$

$$\sum_{i=I_g}^{I_g+1} \mathrm{Vert}(j) = \sum_{i=I_{g-1}}^{I_g} Vert(j), I_g = I_2, I_3, K, I_m \tag{13}$$

(4) Divide the Chinese character image into corresponding elastic grids according to the obtained values of ii and ij, and then calculate the proportion of black pixels in each grid.

(5) Finally, the feature vectors in each grid are combined to form a feature vector, which is used as the feature of Chinese characters. This vector is the elastic grid feature vector.

3.2 Selection of Handwritten Chinese Character Recognition Feature Extraction Method

3.2.1 Selection Principle of Handwritten Chinese Character Feature Extraction Method

Selecting appropriate feature vectors is the key step of character recognition, and the selection of feature vectors is also one of the core tasks of character recognition. The principle of feature selection is: firstly, we should adapt to our own samples, and secondly, we should choose a method with strong classification ability; Then the dimension of features should be reduced as much as possible to

 DOI: https://doi.org/10.26549/met.v6i1.11359

reduce the complexity and difficulty of system operation as much as possible, and improve the running speed and recognition accuracy of the whole system. In short, when all the requirements cannot be fully met, reasonable trade-offs and balances should be made, either by improving the recognition rate or by improving the running speed of the system, so as to meet their own requirements as much as possible.

For different recognition things and different data sets, different feature extraction methods can meet different demands. The best choice method is to select the most important details for recognition according to experience and knowledge. You can also compare different functions to find out the most classified information. It can also be verified by real simulation to determine the best method.

3.2.2 Selection of Feature Extraction Methods

By comparing the experimental results of feature extraction of samples with different methods (as shown in Figure 8), it is found that this paper is more suitable for feature extraction by pixel:

(a) Elastic mesh (b) Skeleton thinning (c) Pixel by pixel extraction

Figure 8. Schematic diagram of different feature extraction methods

Pixel-by-pixel feature extraction method has the best training effect for a single character, and its relative running speed is relatively fast, but its disadvantage is that it is insensitive to details changes.

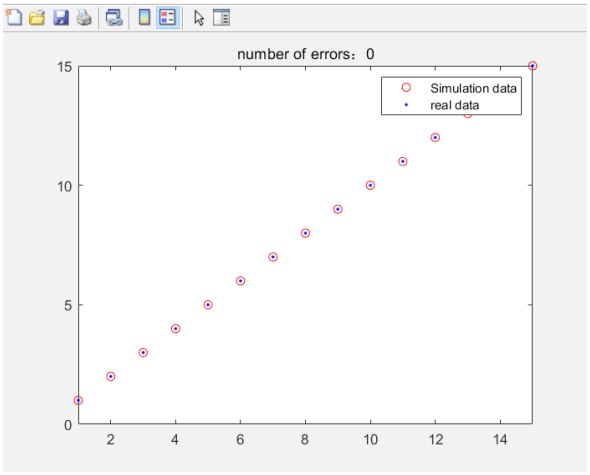


Figure 9. Error rate of pixel-by-pixel feature extraction and recognition

Skeleton thinning feature extraction can well reflect the details of Chinese characters, but the recognition rate is not very high because of too sensitive noise.

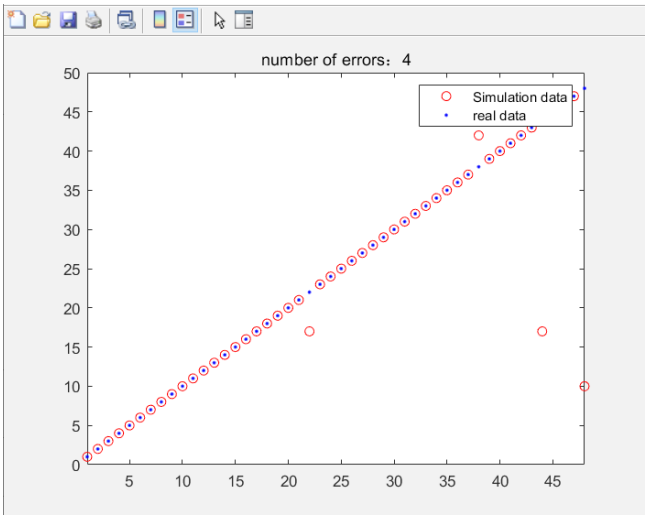


Figure 10. Error rate of skeleton refinement feature extraction and recognition

The feature extraction method of elastic mesh can well reflect the main part of Chinese characters, but it is also sensitive to noise, and it is easy to recognize errors in noisy images.

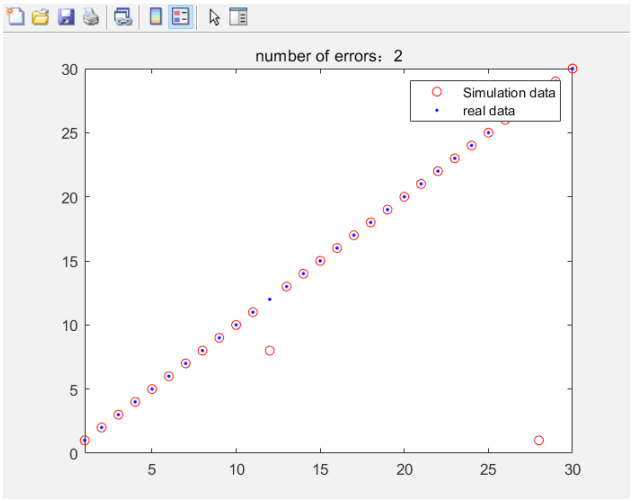


Figure 11. Error rate of elastic mesh feature extraction and recognition

Therefore, finally, the pixel-by-pixel extraction method with high recognition accuracy is selected, and the feature vector of each Chinese character is formed by scanning the binary image line by line and column by column, and the feature value of the Chinese character part is 1 and the feature value of the background part is 0. Finally, the obtained feature matrix is stored in the array matrix prepared for implementation. Provide data for the following BP

DOI: https://doi.org/10.26549/met.v6i1.11359

neural network training.

3.3 Summary of This Chapter

This chapter mainly introduces three feature extraction methods of handwritten Chinese characters: pixel-by-pixel extraction method, skeleton thinning feature extraction method and elastic grid feature extraction method based on the preprocessed images in Chapter 2, and compares the three methods for data sets, and finally selects the pixel-by-pixel extraction method with the highest recognition rate for this article as the feature extraction method of data sets. In order to prepare for the fourth chapter, the extracted feature vectors are sent to the neural network for training and recognition.

4. Handwritten Chinese Character Recognition Based on BP Neural Network

Handwritten Chinese character recognition is a kind of pattern recognition, which distinguishes different Chinese characters according to the features and details of each picture. Common methods of image recognition include support vector machine, template matching and neural network. The neural network has better self-learning ability, adaptive ability and very good fault tolerance. Multiple samples can be trained. Therefore, for the recognition of handwritten Chinese characters, this paper chooses BP neural network as a classifier to complete the recognition of handwritten Chinese characters.

4.1 BP Neural Network Model

BP neural network learning algorithm is a popular neural network learning algorithm at present. It is a multilayer feedforward network that corrects the error layer by layer through reverse transmission. Analyze the error between the results obtained from each training and the expected results, and then modify the weights and thresholds, step by step to get the model whose output is consistent with the expected results. The nonlinear I/O mapping relationship is reflected by the weight and structure of the network. At the same time, BP neural network can repeatedly adjust the weights and thresholds through training the known samples until the transmission relationship of the network reaches the output standard. The neurons in the same layer of BP neural network are independent without mutual connection, but the neurons in different layers are connected. In addition, the hidden layer can have one layer or many layers. Because the calculation method of the whole network goes forward layer by layer, it also belongs to the forward network [33].

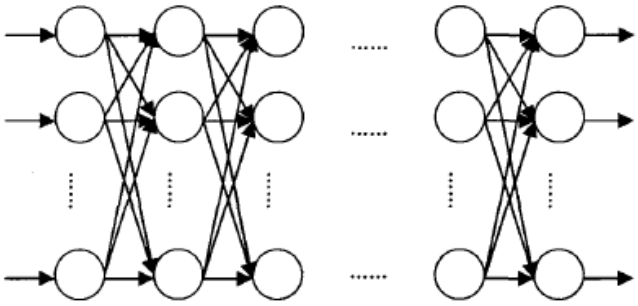


Figure 12. Topology of BP network

4.2 BP neural network learning method

BP neural network is divided into two stages: forward propagation and backward propagation, and it is a supervised learning algorithm. In the forward propagation stage, the input signal is input, then transferred to the hidden layer after the neuron weight function calculation and excitation function processing in the input layer, and the same is weighted by the hidden layer, then transferred to the output layer after the function calculation, and finally the output is obtained and compared with the error requirement. If it does not meet the requirement, it will be transferred from the output layer in the reverse direction layer by layer, and the weights and thresholds will be revised and adjusted repeatedly until the actual output of the network meets the error requirement standard.

The mathematical expression of BP neural network is as follows: assume that the input vector of the kth learning sample is Xk, that is, Xk=(xk1,xk2,…,xkm); The expected output vector of the k-th learning sample is Dk, while the actual output vector is Ok, which is represented as DK = (DK1, DK2, ..., DKM) and OK = (OK1, OK2, ..., OKM) respectively. Wji is the weight input from the ith neuron of the previous layer to the jth neuron of the next layer, and the threshold of the jth neuron is θj, and the excitation function of all neurons adopts S-type function:

$$f(x)=\frac{1}{1+e^{-x}} \tag{14}$$

Therefore there are:

When the neuron is the input layer unit, OK = xk;

(1) For the kth sample, the jth neuron can be described as:

$$Net_{Kj}=\sum_i W_{ji}O_{kj}+\theta_j \tag{15}$$

(2) The output of the J-th neuron is:

$$Net_{Kj}=\sum_i W_{ji}O_{kj}+\theta_j \tag{16}$$

(3) Error between actual output and expected output:

 DOI: https://doi.org/10.26549/met.v6i1.11359

$$E=\sum K_{\mathrm{k}}=(\sum(d_{kj}-O_{kj})^2)/2 \tag{17}$$

(4) The weight correction formula of BP network is:

$$\Delta W_{\mathrm{ji}}(\mathrm{t+1})=\eta\delta_{kj}O_{kj} \tag{18}$$

Type for the output layer using the formula:

$$\delta_{\mathrm{kj}}=(d_{kj}-O_{kj})(1-O_{kj})O_{kj} \tag{19}$$

For the hidden layer, the formula is used:

$$\delta_{\mathrm{kj}}=O_{ki}(1-O_{kj})\sum\delta_{km}W_{mj} \tag{20}$$

The learning rate η in the above equation is the step size of gradient search. The larger η value, the more intense the change of weight. In practical application, the maximum η value is taken under the condition of ensuring that it will not cause vibration. The parameter α is added to the weight formula in order to speed up the learning, and at the same time, it also controls the system not to be prone to oscillation. α indicates the influence degree of the past weight on the current weight:

$$\Delta W_{\mathrm{ji}}(t+1)=\eta\delta_{kj}O_{ki}+\alpha\Delta W_{ji}(t) \tag{21}$$

4.3 Design of Handwritten Chinese Character Recognition System Based on BP Neural Network

For the BP neural network, the actual output value of the neural network is related to its input value, the number of hidden layers, the thresholds and weights of each layer. Only the neural network with good weights and thresholds can make the actual output value and the expected output value of the network finally agree. The biggest feature of BP neural network is to adjust the weights and thresholds between networks through continuous training, so that the error between the output of the network and the expected output can reach the expected value.

This paper is a research on the background design of some handwritten Chinese characters based on BP neural network recognition, so the first task is to determine the structure of BP neural network for handwritten Chinese characters recognition.

4.3.1 Overall Design of BP Neural Network Training Framework

Learning and training are the key steps of BP neural network. Through training, the digital information in feature extraction can be stored in the network, that is, the network is used as a medium to establish a correspondence between input and output. The training process of BP neural network in this paper is shown in Figure 13:

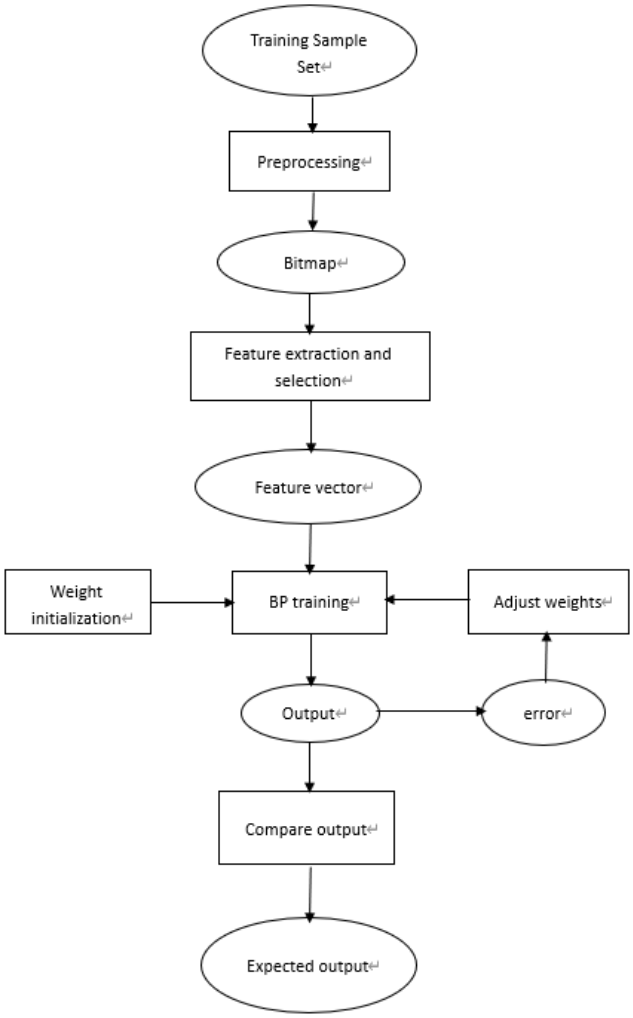


Figure 13. BP neural network training process

(1) Initialization, fix the weights and thresholds of BP neural network, because in Matlab, the initialization weights and thresholds of neural network are random. If these parameters are not fixed, the neural network will get different results every time, which is not uniform. Fix the random number by the following statement.

rng(152, 'twister'); The value in% fixed random number can be modified at will.

(2) According to the order of data set, the feature vector of the first picture is first read as the input of neural network, and the output of each node in each layer is calculated from the input layer.

(3) Calculating the error value between the current sample data and the corresponding output layer;

(4) Modify the weights and thresholds layer by layer from the output layer to the front;

(5) A round of training is conducted on the data set samples. Considering that one training may not reach the

DOI: https://doi.org/10.26549/met.v6i1.11359

final expected value, the input is processed once and the training is conducted for many times. The code is as follows.

Input =[Input Input Input Input]; % repeat training

Output=[Output Output Output Output];

(6) detecting whether the total network error meets the requirements, if so, ending the training, otherwise, returning to step (2) and continuing the training.

4.3.2 System Parameter Tuning of BP Neural Network

The most important thing to design a BP neural network is to choose various parameters. Determining the number of network layers, initial weights, learning algorithms and number of nodes is equivalent to determining the BP neural network. Although the selection of these parameters mainly depends on experience, there are certain guiding principles.

(1) Determination of hidden layers of BP neural network

The number of hidden layers of BP neural network is uncertain, but the more hidden layers, the smaller the system error. When the number of hidden layers is too large, the errors generated by the network may decrease, but with the increase of layers, the training time of the whole system will be longer, and the transmission errors between layers will also increase. BP neural network with only one hidden layer can be used to realize continuous functions in any interval, which has strong applicability. Therefore, this paper chooses to use a network structure with multiple neurons in a single hidden layer. That is, what we usually call the three-layer network structure.

(2) Transfer function of BP neural network

When training BP neural network with data set, it can be found that when using logarithmic transfer function as neuron training function in output layer and hidden layer, the error will be much smaller than that when using piecewise function or threshold function. Therefore, this design adopts logarithmic sigmoid function as the function of hidden layer and output layer.

(3) Selection of learning rate

Learning rate is also a very important variable for the training of neural network. Its value can affect the span process of each training of the network; If the learning rate is very small, it will take many iterations to reach convergence. If the learning rate is very high, the error requirement may fluctuate repeatedly, or even exceed the local minimum value, which may lead to the inability to converge. According to experience, the choice of learning rate will be between 0.01 and 0.8. This paper is set at about 0.05.

(4) Expected error

The setting of the expected error value should be selected according to the convergence of the network. Generally speaking, if the network is easy to converge, you can choose the smaller expected error. If the network is not easy to converge, the expected error value should be appropriately increased.

(5) Determination of the number of hidden layer nodes

For systems with many classifications, the requirements of hidden layer are relatively high. If there are too many nodes in the hidden layer, the training time may be too long. However, if the number of nodes is less selected, the identification accuracy of the network may be poor and the identification task cannot be completed. Generally speaking, according to the known number of input nodes and output nodes, we should train and compare the networks with different number of nodes to find out the one with the smallest error as the best number of nodes. In this paper, the growth method is used to train the neural network with fewer nodes, observe the change of learning error, and then slowly increase the number of nodes until a relatively satisfactory learning error is obtained.

Training the three-layer BP neural network has the following empirical formula to learn and select the number of nodes for reference:

$$h = \sqrt{i+o} + u \tag{22}$$

$$h = \log_2^i \tag{23}$$

$$h = \sqrt{io} \tag{24}$$

$$h = 2i + 1 \tag{25}$$

where I is the number of nodes in the input layer, H is the number of nodes in the hidden layer, O is the number of nodes in the output layer, and U is any integer between [1,10]. The training results are shown in the following table.

Table 1. Relationship between Number of Hidden Layer Nodes and Error

The number of neurons in the hidden layer	training error	test error
5	1.25441	1.1563
7	0.89514	0.9215
9	0.61452	0.6927
11	0.580315	0.6896
13	0.552773	0.6835
15	0.455114	0.6575
17	0.387725	0.6483
19	0.269714	0.4527
21	0.173878	0.6538
23	0.168583	0.47
25	0.168479	0.6571

According to the above table, the following conclusions can be drawn:

1) The number of nodes in the hidden layer can reduce the training error, but after 19 nodes, the error starts to

 DOI: https://doi.org/10.26549/met.v6i1.11359

rise, indicating that its generalization ability has changed, which is comparatively available. The number of nodes between 19 and 21 can be selected as the number of nodes in the hidden layer.

2) The error is always very large, which can be improved by adjusting the initial weight and learning rate. After experiments, it is decided that the learning rate is 0.05, which can achieve the expected goal.

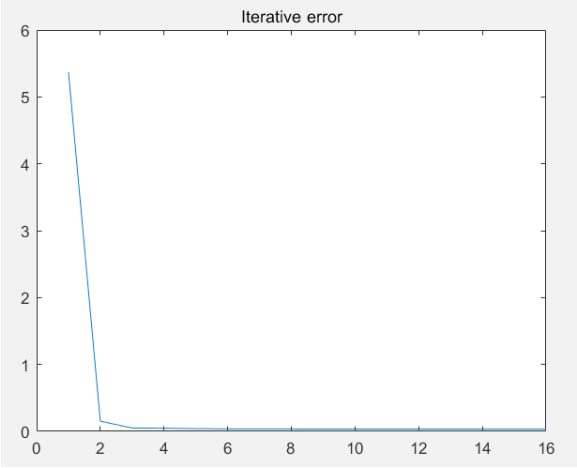


Figure 14. Iterative error of neural network when the number of hidden layer nodes is 21

4.3 Simulation and Implementation of Handwritten Chinese Character Recognition System Based on Matlab

For different Chinese character recognition systems, except for the feature extraction method and recognition method, the other steps are a relatively similar process. The recognition process of this paper is: firstly, by reading Chinese character pictures, pre-processing the pictures to be recognized, training the BP neural network with the pictures in the data set, and then, after post-processing, comparing the final recognition results and outputting them as strings.

Based on the pretreatment of the first two chapters and the principle of BP neural network, the feasibility of handwritten Chinese character recognition by BP neural network is verified, and the results of handwritten Chinese character recognition are shown. Recognition includes the steps of graying, binarization, interference removal, segmentation, network training, handwritten Chinese character recognition and so on.

4.3.1 Design of GUI Interface

GUI interface is a graphical user interface for us to execute applications in the computer. We can get the steps we want to achieve in the display window by selecting different buttons and menu options to realize interpersonal interaction. Making GUI interface is mainly through building the corresponding event-driven system, that is, clicking the corresponding button will execute the corresponding program segment, and return the corresponding information in the window, and then wait for the next operation of the user. As shown in Figure 15.

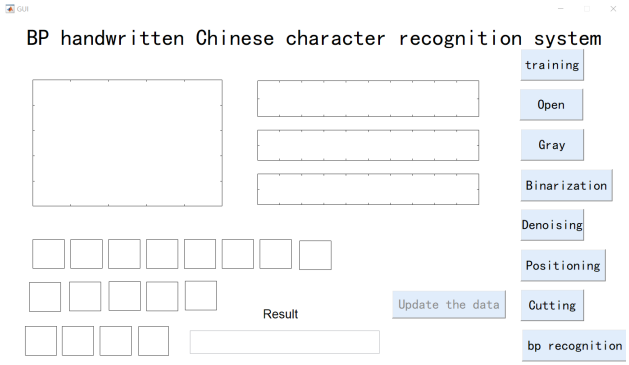


Figure 15. GUI interface

4.3.2 Handwritten Chinese Character Data Set Acquisition Module

The establishment of data set is mainly to select representative Chinese characters to train BP neural network. The training of data set is a necessary condition for Chinese character recognition, so before starting to recognize Chinese characters, the data set should be established first.

At present, there are about 4,000 Chinese characters commonly used in China, and handwritten Chinese characters that vary from person to person, even those written by the same person, may be different, so the collected characters should reflect universality.

In this paper, 17 Chinese characters, such as Chinese, Chinese, local, qualitative, big and academic, are collected to establish the data set. The establishment process is as follows:

(1) Save the handwriting in jpg format to the computer;

(2) preprocessing the sample to obtain a binary image;

(3) Cut into individual Chinese characters and store them in the data set;

(4) Select a part of the segmented binary handwritten Chinese character image as a data set to train the neural network, and the collected samples are shown in Figure 16;

Handwritten Chinese character recognition system based on BP neural network can be roughly divided into three parts: input image, preprocessing of handwritten Chinese characters and BP neural network recognition of Chinese characters after preprocessing. Among them, the input is data set and test set image. Firstly, the BP neural network is trained with the preprocessed feature vectors

DOI: https://doi.org/10.26549/met.v6i1.11359

of the data set, and then the trained BP neural network is used to identify the Chinese characters in the test set. In this paper, the segmented single-character binary image to be identified is transformed from left to right and from top to bottom in turn, and transformed into the feature vector of each character. If the pixel value is 255, the characteristic value is 1, and if the pixel value is 0, the characteristic value is 0. Finally, the feature vectors are input into the trained neural network for recognition. The system has buttons in the GUI interface for each step, which can clearly reflect the identification process of each step.

4.3.3 Chinese Character Preprocessing Module

Preprocessing: It mainly includes the processes of image gray processing, binarization, denoising, image segmentation and normalization. Its main contents have been described in detail in Chapter 2.

(1) First, the Chinese character pictures are read into the system, then grayed and binarized, and the blank background of the font is set to 0, and the part with black font is set to 1. Therefore, the image can be expressed as a binary data matrix composed of 0 and 1.

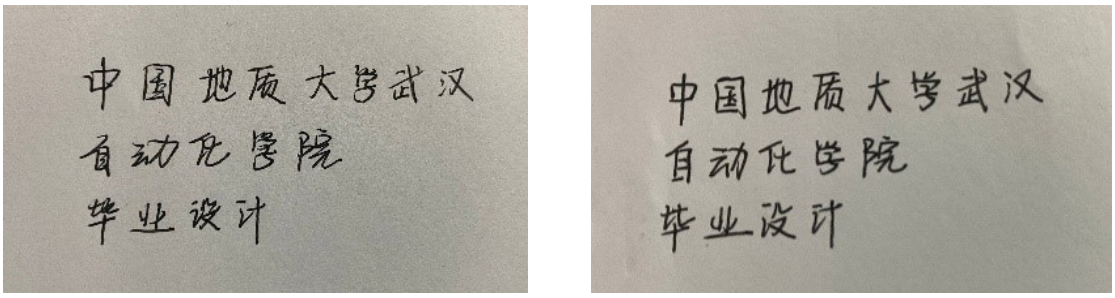


Figure 16. Partial Sample Data Map

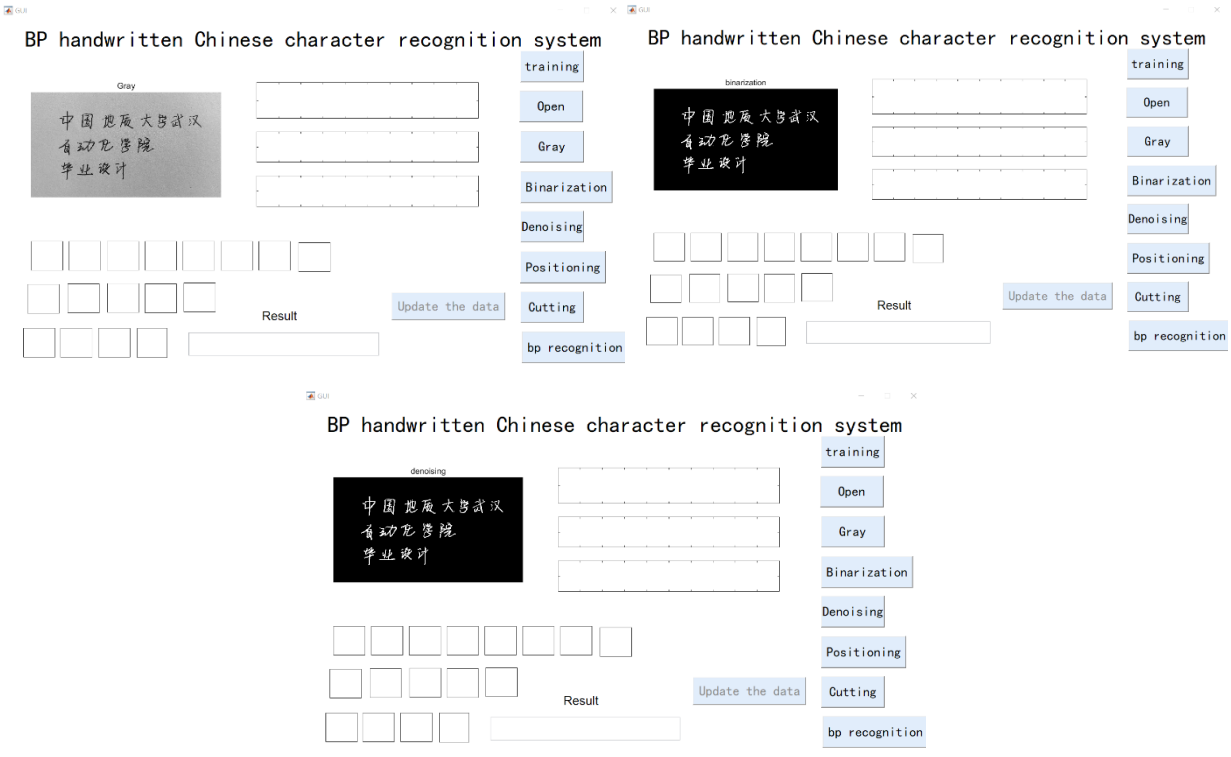


Figure 17. Pretreatment process

Figure 18. Character cutting process

(2) Then cut the binary image. First, get the characters of each line, then cut each line separately, number each single character and display it in the interface of the system.

4.3.4 Network Training and Identification Module

In the part of neural network training, the training process of neural network is also a key step in recognition. Through the training process, the feature information of each Chinese character can be connected with the network, that is, the feature input of each Chinese character can establish a corresponding relationship with the output of neural network, and the corresponding feature output can be obtained by inputting a feature value, thus achieving the purpose of recognition. The training process of BP neural network has been explained in Chapter 3.

According to Figure 19, the accuracy of network identification after training can basically meet the requirements, and the identification function can be completed. The recognition result shown in Figure 20 can be obtained.

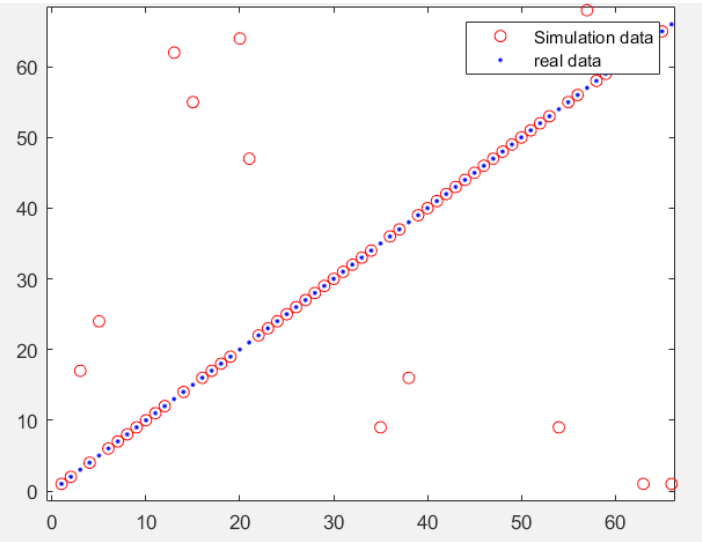


Figure 19. Comparison between simulation data and real data

DOI: https://doi.org/10.26549/met.v6i1.11359

Figure 20. Identification results

4.4 Analysis of Experimental Results

4.4.1 Test Results

In order to verify the performance of the handwritten Chinese character recognition system, this paper selects 10 test pictures for recognition verification. When the samples in the data set and the samples in the test set are used for recognition, two results will be obtained:

(1) When the sample of the data set is used as the recognition image, the recognition accuracy can reach 100%, and the simulation value is basically consistent with the output value.

(2) When the test set picture is used as the recognition image, there will be the possibility of recognition errors. In this example, there are three recognition errors in 17 Chinese characters, as shown in Figure 21.

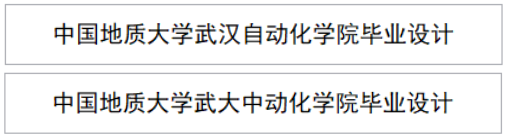


Figure 21. Partial recognition results

4.4.2 Analysis of System Identification Performance

Table 2. Sample recognition rate

sample source	number of correct	correct rate(%)	Error rate (%)
data set	17	100	0
test set	15	94	6

Through multiple groups of sample tests, it can be seen from Table 2 that for the data set, the correct rate of samples can reach 100%, while there are errors in the training set, among which, the error rate of "Chinese" and "self" is the highest, and "self" is often identified as "medium".

By comparing the experimental results, we can see that the reasons for the decline of recognition accuracy are as follows:

(1) Data set samples are still insufficient. In the case of less training, BP neural network can't be well trained, and it is very effective to build a representative data set, which can quickly improve the network performance.

(2) The characteristics of handwritten Chinese characters change relatively greatly. For two similar Chinese characters, there may be recognition errors due to different writing styles. The method of feature extraction can be optimized to stabilize the system performance.

(3) The network structure is relatively simple. For a single BP neural network to recognize handwritten Chinese characters, it is still very difficult to adapt to various styles of Chinese characters. The combination of BP neural network and other networks can be used to improve the recognition rate.

According to the above experimental results, we can know that the number of data sets and the diversity of data in them will be an important factor affecting the final recognition result of BP neural network. When the data set sample is large enough, it will have better recognition effect after learning by BP neural network. At the same time, the training parameters of BP neural network are also an important influencing factor, and the rationality of the parameters can also determine the final recognition result.

 DOI: https://doi.org/10.26549/met.v6i1.11359

4.5 Summary of This Chapter

This chapter mainly introduces the simulation and implementation of handwritten Chinese character recognition in Matlab software. Firstly, the composition of each part of handwritten Chinese character recognition system is introduced. Then it introduces the design of GUI interpersonal interface, and the recognition process of handwritten Chinese characters through GUI interface. Finally, based on the test results and the training data of BP neural network, the influencing factors of handwritten Chinese character recognition are analyzed. The simulation results show that the handwritten Chinese character recognition system based on BP neural network can complete the recognition function and has high recognition accuracy.

5. Summary and Prospect

5.1 Summary

Handwritten Chinese character recognition has always been a very important direction in the field of pattern recognition. The research on handwritten Chinese character recognition is not only helpful to our study and life, but also promotes the development of pattern recognition. For a long time, as Chinese characters are the common words of the Chinese nation, foreign scholars have not studied them very deeply, and the uncertainty of handwritten handwriting also makes the recognition of handwritten Chinese characters even more difficult. Therefore, it is of great significance and value to study off-line handwritten Chinese character recognition and improve the accuracy of handwritten Chinese character recognition.

This paper studies the construction and learning process of BP neural network, the design and implementation of handwritten Chinese character recognition based on BP neural network. Through the pretreatment of Chinese characters, learning with BP neural network, and finally verifying the feasibility of recognizing handwritten Chinese characters with BP neural network through Matlab recognition simulation. The work is divided into the following four parts:

(1) The research background and significance of handwritten Chinese character recognition, the historical development of handwritten Chinese character recognition and some research processes are systematically described, and some examples of Chinese character recognition by different methods are listed.

(2) The pretreatment process of handwritten Chinese characters is introduced in detail. Comprises five steps of gray scale processing of handwritten Chinese character pictures and histogram enhancement, binary processing, smooth denoising, Chinese character cutting and normalization of some handwritten Chinese character pictures, and then stores the pictures.

(3) Analyze the feature extraction methods of handwritten Chinese characters, and select the most suitable method for this paper by comparing the advantages and disadvantages of different methods.

(4) In this chapter, the framework of written Chinese character recognition based on BP neural network is described in detail. The model and parameter characteristics of BP neural network are summarized, and then the design scheme of BP neural network is put forward. The modules of GUI man-machine interaction interface in Matlab are explained. Finally, the simulation of handwritten Chinese character recognition process based on BP neural network is studied. And the simulation results are analyzed. Firstly, it is explained that the platform used in the simulation test is Matlab. Then, image preprocessing and network training are carried out on the sample pictures. Finally, the feasibility of recognizing handwritten Chinese characters by BP neural network is verified through recognition, which has good practical significance.

5.2 Outlook

Because of the time and energy constraints, this paper still has many shortcomings in recognizing handwritten Chinese characters by BP neural network. The following are the deficiencies and further planning:

(1) There are too few samples used to train BP neural network in this paper, which leads to the inability to accurately identify other Chinese characters after the same network training, and has certain limitations. After that, it can be improved from the following aspects: First, increase the training sample set for training BP neural network; Second, some limitations in improving procedures.

(2) The preprocessing process of Chinese characters may need to be further refined, or different preprocessing methods may be tried to improve the feature extraction method to improve the recognition efficiency and accuracy.

(3) At present, only Chinese characters in simple background can be recognized, and the recognition effect of block letters is very good. After that, we should try to see if we can recognize Chinese characters in complex background and Chinese characters with different fonts.

(4) In the next step, BP neural network can be combined with other classifiers to find a more efficient combination of classifiers to supplement the shortcomings of BP neural network.

DOI: https://doi.org/10.26549/met.v6i1.11359

References

[1] Zhao, Zh.Ch., Luo, Z., Wang, P.Y., et al., 2020. Summary of research on network image classification algorithm based on depth residual. Application of Computer System. 29(1), 18-25.

[2] Li, X.L., 2018. Printed Chinese character recognition based on BP neural network. Think Tank Era. 161(45), 193+195.

[3] Wan, L.P., Lan, X.G., Zhang, H.B., et al., 2019. Summary of deep reinforcement learning theory and its application. Pattern recognition and artificial intelligence. 32(1), 67-81.

[4] Liu, F., 2014. Research on character recognition algorithm based on improved BP neural network. Value Engineering. 33(10), 206-207.

[5] Li, D., 2016. Multi-sample handwritten character recognition software based on BP neural network. 37(7), 103-108.

[6] Yamamoto, S., Nakajima, A., Nakata, K., 1973. Chinese character recognition by hierarchical pattern matching. Pattern Recognition, International Conference. 187-196.

[7] Sun, Q., 2007. Offline Handwritten Chinese Character Recognition System. Nanjing: Nanjing University of Aeronautics and Astronautics.

[8] Zhang, Y., Li, E.L., 1999. Research on preprocessing algorithm in off-line handwritten Chinese character recognition. Journal of Shenyang University of Technology. 21(6), 534-535.

[9] Wang, Q., Zhao, R.Ch., 2001. Standardization and Evaluation of Handwritten Chinese Characters. Data Collection and Processing. 16(2), 227-232.

[10] Gao, Y.Y., 2004. Summary of unconstrained handwritten Chinese character segmentation methods. Computer Engineering. 30(5), 144-145.

[11] Ding, S., Su, C., Yu, J., 2011. An optimizing BP neural network algorithm based on genetic algorithm. Artificial Intelligence Review. 36(2), 153-162.

[12] Song, M.L., 2002. Median filtering algorithm for removing image noise. Journal of Luoyang Normal University. (5), 67-69.

[13] Lin, K.K., Song, G.X., 2004. Comparison and improvement of image denoising methods. Journal of Xidian University (Natural Science Edition). 31(4), 627-628.

[14] Zhang, X., Bengio, Y., Liu, C., 2017. Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark. Pattern Recognition. 61, 348-360.

[15] Xue, B.R., Yang, J.Y., Lou, Zh., et al., 1999. Eliminating burrs and burrs of handwritten Chinese characters. Journal of Nanjing University of Science and Technology. 23(2), 49-52.

[16] Zhuang, Y., Liu, Q., Qiu, C., et al., 2021. A handwritten Chinese character recognition based on convolutional neural network and median filtering. Journal of Physics: Conference Series. 1820(1), 012162.

[17] Lian, W., Xu, B.Zh., 2007. A new feature extraction method in handwritten Chinese character recognition-elastic grid directional decomposition feature. Journal of Circuit and System. 2(3), 8-13.

[18] Ma, Sh.P., Xia, Y., Zhu, X.Y., 2006. Research on handwritten Chinese character recognition based on directional line element method. Journal of Tsinghua University. 24-27.

[19] Wan, L.Y., Chen, J.X., Wang, W.P., et al., 2006. Research on image recognition based on BP neural network. Journal of Wuhan University of Science and Technology. 29(3), 277-279.

[20] Ning, B., Chen, J., Tan, J., 2019. The handwritten Chinese character recognition uses convolutional neural networks with the google net. International Journal of Pattern Recognition and Artificial Intelligence.

[21] Xu, Y.Sh., Gu, J.H., Tao, Zh., et al., 2011. Handwritten character recognition based on improved BP neural network. Communication Technology. (5), 106-109.

[22] Huang, Sh.Q., Zhao, Zh.Y., Sun, L.B., 2017. Improvement of BP neural network algorithm. Science and Technology Innovation Herald. 14(20), 146-147.

[23] Ren, H.Y., Yuan, B.Sh., Pastoral, 2010. Uyghur Online Handwritten Character Recognition Based on BP Neural Network. Microelectronics and Computer. 27(8), 238-241.

[24] Yu, H., Cao, L., Li, Q.Y., 2009.. Improvement of BP neural network algorithm and its application in handwritten Chinese character recognition. Journal of Jiangxi Normal University (Natural Science Edition). 33(05), 598-603.

[25] Murru, N., Rossini, R., 2016. A Bayesian approach for initialization of weights in backpropagation neural net with application to character recognition. Neurocomputing. 193, 92-105.

[26] Srivastava, N., Hinton, G., Krizhevsky, A., et al., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 15(1), 1929-1958.

[27] Redmon, J., Divvala, S., Girshick, R., et al., 2016. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Com-

 DOI: https://doi.org/10.26549/met.v6i1.11359

puter Vision and Pattern Recognition. 779-788.
[28] Chen, K., 1977. Iterated path integral. Bulletin of The American Mathematical Society. 83(5), 831-879.
[29] Liu, B., 2015. Application of improved image matching method in Chinese character recognition. Guangzhou: Jinan University.
[30] Lian, W., Qin, J.Zh., 2004. Research on Gabor feature extraction method of elastic grid for handwritten Chinese character recognition. Computer Application Research. (12), 163-165.
[31] Wang, F., 2021. Identification method of station state information based on image processing and neural network. Railway Communication Signal. 57(3), 60-63.
[32] Chen, J.H., Yang, L.L., Zhao, Y.J., et al., 2021. Application of deep neural network model in logging electrical imaging image processing. Electronic Measurement Technology. 44(4), 138-143.

DOI: https://doi.org/10.26549/met.v6i1.11359

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Study of Wireless Sensor Network Based on Optical Communication: Research Challenges and Current Results

Xinrui Li Dandan Li*

Jiangsu Linyang Energy Co.,Ltd, Nantong, Jiangsu, 226299, China

ARTICLE INFO

Article history
Received: 19 March 2022
Revised: 26 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Underwater Wireless Sensor Network (UWSN)
Visible Light Communication (VLC)
ACO – OFDM
DCO – OFDM
Line of Sight (LoS)

ABSTRACT

With the rapid developments of commercial demands, a majority of advanced researches have been investigated for the applications of underwater wireless sensor (WSN) networks. Recently optical communication has been considered for underwater wireless sensor network. An experimental set-up for testing optical communication underwater has been provided and designed in present papers to maximize the energy coupled from these displacements to the transduction mechanism that converts the mechanical energy into electrical. The true case has been considered by measuring diffuse attenuation coefficients in different seas. One stand out potential optical communication method, Visible Light Communication (VLC) has been talked and several communication methods are compared from many points of view, for example attenuation in salt water. The evaluation of modulation techniques for underwater wireless optical communications has been displayed, and further how the data collection and storage with an underwater WSN is introduced. In this paper current researches for an (UWSN) based on optical communication are studied, in particular the potential VLC method and comparisons of VLC with other optical communication approaches. Underwater challenges would be analyzed by comparing a sort of communication methods, applied in underwater. Future work will be developed at last.

1. Introduction

High efficiency Underwater Wireless Sensor Networks (UWSN) is hard to be designed for underwater wireless communication applications. What is known to us is that the medium water has a significant attenuation for Radio Frequency (RF) radio waves. In order to degrade this problem, sonar communication, infrared wireless communication, optical wireless communications, based on laser and Visible Light Spectrum have been investigated in [1,3,9]. Due to the limited resource of spectrum and its big advantages of VLC system, an application of VLC on UWSN stands out, which can be used to detect the environmental condition in deep sea, lift condition of marine organisms etc. This paper analyses the challenges by applying the optical communication techniques and describes the details of current researches and derived results on VLC in UWSN. Following that, the future work for enhancing VLC based on LEDs will be explored.

This paper is organized as follows: Section II depicts the (UWSN) challenges by comparing existing telecommunication techniques and a short introduction of structure of UWSN. Section III describes current result achieved and potential research challenges on UWSN system. Section IV concludes this paper and declares future work.

**Corresponding Author:*
Dandan Li,
Jiangsu Linyang Energy Co.,Ltd, Nantong, Jiangsu, 226299, China;
Email: 1251910060@qq.com

 DOI: https://doi.org/10.26549/met.v6i1.11372

2. Underwater Challenges

Wireless sensor network (WSN) (Figure 1a) is the spatially distributed autonomous sensor to monitor physical or environmental conditions, (e.g. temperature) and to cooperatively pass their data through the network to a main location. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance. To produce low-cost and tiny sensor nodes is a major challenge in WSN and a majority of the nodes are still in the research and development stage, particularly their software and also inherent to sensor network adoption is the use of very low power methods for radio communication and data acquisition. However the most interesting part is about Underwater WSNs (UWSNs) (Figure 1b).

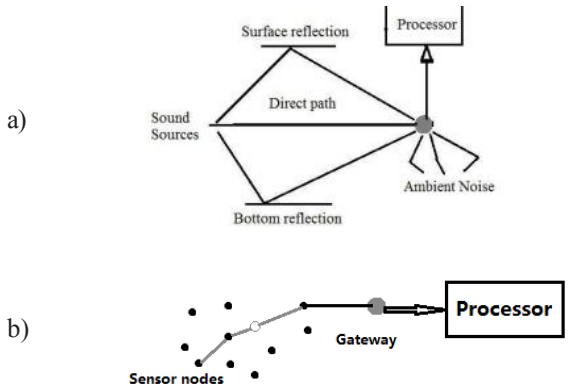


Figure 1. Wireless sensor network architecture, a) typical multi-hop WSN, b) typical Underwater WSN.

Recently, there are more and more projects and papers become interested in UWSN applications, such as acoustic communication [1], infrared wireless communication [2], and Visible Light Communication [3], etc. They have the great advantages on monitoring marine biology, deep-sea archaeology, and pollution detection and explore the life conditions of marine organisms in unknown sea-area, etc. The high frequency radio waves are strongly attenuated in water, especially in electrically more conductive salt water and the available radio modules such as Bluetooth or WLAN (802.11) operate around 2.4 GHz. In order to compensate shortages, the visible light underwater communication is a state-of-art technique, based on on-off shelf LEDs or laser LEDs. High speed transmission, high data rates and high transmission reliability are the noticeable tags of VLC system. In [5], a novel platform consists of static and mobile underwater sensor nodes and a nature inspired UWSN named as Smart Plankton is introduced in [1], where it figures out the design concept of UWSN. In order to develop an UWSN it is necessary to build an artifact that can survive, communicate, sense and cooperate in the underwater harsh and demanding environment where high pressure, corrosion, fouling and bio-erosion from colonizing organisms are threats to structure integrity and functionality. The comparison of the four communication systems is displayed as Table 1.

Table 1. Comparison between available transmission systems for underwater wireless communication

Property of medium	Radio	Acoustic	Infrared	Visible Light
Available bandwidth	Limited	Strictly limited	middle	Quite large
Transmission rate	Middle	Low	Middle	High
Range	Limited	Small	Middle	Large
Path loss	High	High	High	High
Dominant Noise	Other users	Marine organisms	Sunlight	Sunlight
Cost	High	High	Low	Low
Complexity	High	High	Low	Low
Security	High	Low	High	High

3. Current Result Achieved and Potential Research Challenges on UWSN System

In [5], a novel platform for underwater sensor networks to be used for a long-term monitoring of coral reefs and fisheries is illustrated. The nodes including both static and mobile communication point-to-point using a novel high-speed optical communication system. Moreover, it studies and shows the processes data collection; storage and retrieval work in an UWSN. A transmission protocol based on optical UWSN has been shown in [16], where the optical PHY layer structure is produced as Figure 2. Finally an experimental set-up has been created to test the optical PHY layer.

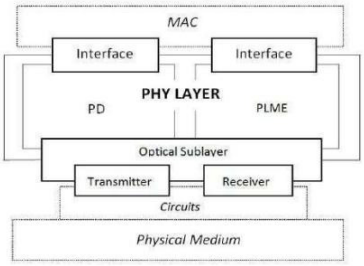


Figure 2. Optical PHY Layer structure

An experimental set-up for testing the optical communication underwater is investigated. Variability of water optical parameters and basic design considerations are further studied in [2,8]. The channel modeling shown as Figure 3 and performance evaluation using Vector Radiative Transfer Theory based on Underwater Wireless Opti-

DOI: https://doi.org/10.26549/met.v6i1.11372

cal Communication is developed and simulated with a lot of waveforms both transmitted and received added. The complete research based on visible spectrum optical communication for underwater applications has been referred in [3], where both the experimental result in air and under water are presented and compared.

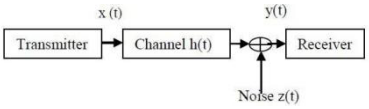


Figure 3. Additive noise channel model

1) Acoustic Communication on UWSN

Acoustic communication underwater application is also investigated in [5]. As we know, the acoustic communication is a broadcast communication system, where an XR2206 integrated circuit generates the transmitted signal. The transmission protocols are both amplitude modulation (ASK) and pulse position modulation (PPM), which can compensate the reflections. The results show that the acoustic communication is usable up to a distance of 15 m at a data rate of 41 bit/s and using PPM can achieve a higher data rate since the pulse period can be reduced. Finally the amount of data can be collected with a static sensor node over an extended period of time, depending on the maximum data storage capacity of a sensor node and the battery lifetime.

2) Optical Communication on UWSN

The following several paragraphs display us the current researches and results obtained based on optical communication using UWSN. As stated before, a mobile underwater sensor network is built in [5], where the mobile node can locate and hover above the static nodes for data mulling and can perform network maintenance functions, namely deployment, relocation and recovery. In addition, the hardware and software architecture of this UWSN is described.

Both optical communication and acoustic systems are used in [5]. The former is preferred with the potential high speed and high theoretical bandwidth, although is restricted to short-range line of sight applications. In current sensor node design [5], it has been proved each node can accumulate up to 512 Kbyte of data with a reduction time period and can realize the full duplex communication. The hardware of optical communication system used in [5] consists of a transmitter, Luxeon 5 LXHL-PM02, a 532 nm (green) LED with 10%-15% power efficiency and a receiver, PDB-C156 high speed PIN photodiode with 8 mm^2 surface areas. A green dichroic filter is imposed on photodiode to attenuate frequencies other than green. The transmission protocol is based on 4-PPM, such as NRZ, SIR, FIR and VFIR and the system is adapted by using a start pulse for each type, which can be seen from Table 2.

Table 2. Time intervals between high value pulses in our modified VFIR optical communication format.

State	Time Interval (μs)
Data bits 00	4.0
Data bits 01	4.5
Data bits 10	5.0
Data bits 11	5.5
End of Byte	6.0

The built UWSN is adapted with the optical system to determine the maximum communication range and coverage area. The results manifest that the maximum range in the clear water is longer with 8 times than that in the turbid water with the coverage area 30◦. In addition, a longer range has been obtained by applying a 60 mm F1.06 to the optical transmitter, displaying in Table 3.

Table 3. Results obtained with the UWSN system adapted with optical communication

Distance (m)	Received packets	Missed packets	Packet Success Rate
2.1	199	0	100%
4.3	199	0	100%
5.3	166	2	99%
6.4	199	0	100%
7.0	199	8	96%

The available bandwidth of underwater acoustic channels is quite limited because of absorption and dramatically depends on both transmission range and frequency. Referring to the physical layer, a large amount of studies for adapting and extending the existing link layer, routing and transport layer protocols are declared to perform a good acoustical communication.

The vector radiative transfer theory (VRTT) explains the behavior of wave propagation and scattering in a discrete random medium. It provides also characteristics of water and particles in an underwater environment, dividing them into two main parts according to the effects caused by the different materials, absorption and scattering individually.

3) Research Challenges

In this section, challenges of realizing underwater wireless sensor network based on optical communication will be presented. First, a much more efficient algorithm or protocols between PHY layer and MAC layer for underwater optical communication is not easy to figure out.

 DOI: https://doi.org/10.26549/met.v6i1.11372

Since the limited bandwidth of light sources (LEDs) and the channel dispersion, the data rate is limited. Referring to the cost of communication systems, typical underwater wireless sensor networks (UWSN) are expensive, however the use of such kinds of underwater manned or unmanned vehicles is quite popular for current typically communication systems. Moving to the carriers, the optical attenuation coefficients directly impacts on the performance of these communication systems, such as transmission range, error rate, and power efficiency. Besides, the water clarity changes in time, which is a permanent natural challenge, limiting the whole system performance. To build a low cost efficient UWSN using visible light communication approach is a big challenge, however, it is feasible due to the pre-distortion or pre-equalizers techniques, which can perfect compensate the non-linearity and bandwidth limited problems. More efficient modulation and demodulation techniques should be considered to the applications on UWSN to increase the data rate and mitigate the channel dispersion, such as asymptotically-clipped optical OFDM (ACO-OFDM), DC-biased OFDM (DCO OFDM) and flip-OFDM. Meanwhile the cost increases due to the high complexity of Optical OFDM. Compared with DCO-OFDM and Flip-OFDM, ACO-OFDM requires less optical power to achieve the same bit error rate but decrease the spectrum efficiency since only odd sub-carriers carry information. Due to the potential harmful effects on organism eyes, the flickers, namely inter-frame flicker and intra- frame flicker, should be mitigated by applying some possible methods, such as idle/visibility patterns, coding methods, increasing the optical clock rate, depending on what standard and modulation techniques are used. Although the advanced Optical OFDM techniques could bring the significant benefits for communication systems, the property of high PAPR may aggravate the channel dispersion and dynamics.

4. Conclusions and Future work

This paper is based on the information collection from several specific scientific papers with respect to the current researches and results of UWSN using Visible Light Communication. It consists of five sub-sections. The first section towards the general description and motivations of this paper. The basic structure of this paper has been explained at the end of first sub-section. The second one focuses on the basic introduction of VLC indoor application technique, followed by the underwater communication challenges in section III, where there is a brief introduction of WSN and the comparison of present transmission protocols underwater applications in [1,2,4-6,8]. Section IV presents the current researches, current results, and challenges of applying VLC on UWSN.

According to the current researches on UWSN, the optical OFDM with multiple transmitter systems can be used on UWSN in the future instead of using conventional methods, e.g. OOK and PPM, to compensate the dispersion. Besides, Multiple-In Multiple-Out (MIMO) technology can also be investigated on UWSN to increase the data rate. Last but not least, the dimming support should also be reminded to save the optical power due to the daytime and night. What are well known to us from VLC applications in mobile network is that light brightness control can be achieved using the continuous current reduction (CCR) and the pulse-width modulation (PWM) dimming techniques.

References

[1] Anguita, D., Brizzolara, D., Ghio, A., et al., 2008. Smart plankton: a nature inspired underwater wireless sensor network. Fourth International Conference on Natural Computation. IEEE.

[2] Smart, J.H., 2005. Underwater optical communications systems part 1: variability of water optical parameters. MILCOM 2005-2005 IEEE Military Communications Conference. IEEE.

[3] Schill, F., Zimmer, U.R., Trumpf, J., 2004. Visible spectrum optical communication and distance sensing for underwater applications. Proceedings of ACRA.

[4] Sui, M.H., Yu, X.Sh., Zhang, F.L., 2009. The evaluation of modulation techniques for underwater wireless optical communications. Communication Software and Networks. ICCSN'09. International Conference on. IEEE.

[5] Vasilescu, I., Kotay, K., Rus, D., et al., 2005. Data collection, storage, and retrieval with an underwater sensor network. Proceedings of the 3rd international conference on Embedded networked sensor systems. ACM.

[6] Jaruwatanadilok, S., 2008. Underwater wireless optical communication channel modeling and performance evaluation using vector radiative transfer theory. IEEE Journal on Selected Areas in Communications 26(9), 1620-1627.

[7] Akyildiz, I.F., Vuran, M.C., 2010. Wireless sensor networks (Vol. 4). Hoboken: Wiley. CrossRef MATH.

[8] Giles, J.W., Bankman, I.N., 2005. Underwater optical communications systems. Part 2: basic design considerations. MILCOM 2005-2005 IEEE Military Communications Conference. IEEE.

[9] Akyildiz, I.F., Pompili, D., Melodia, T., 2005. Underwater acoustic sensor networks: research challenges.

DOI: https://doi.org/10.26549/met.v6i1.11372

Ad hoc networks 3.3. 257-279.

[10] Rajagopal, S., Roberts, R.D., Lim, S.K., 2012. IEEE 802.15. 7 visible light communication: modulation schemes and dimming support.Communications Magazine, IEEE, 50(3), 72-82.

[11] Hranilovic, S., Kschischang, F.R., 2004. Short-range wireless optical communication using pixilated transmitters and imaging receivers. Communications, 2004 IEEE International Conference on. IEEE.

[12] Cui, K., 2012. Physical layer characteristics and techniques for visible light communications. University of California, Riverside.

[13] Stefan, I., Elgala, H., Haas, H., 2012. Study of dimming and LED nonlinearity for ACO-OFDM based VLC systems. 2012 IEEE Wireless Communications and Networking Conference (WCNC). IEEE.

[14] Armstrong, J., Schmidt, B.J.C., 2008. Comparison of asymmetrically clipped optical OFDM and DC-biased optical OFDM in AWGN. IEEE Communication Letters 12.5. 343-345.

[15] Fernando, N., Hong, Y., Viterbo, E., 2011. Flip-OFDM for optical wireless communications. Information Theory Workshop (ITW), IEEE.

[16] Anguita, D., Brizzolara, D., Parodi, G., 2009. Building an underwater wireless sensor network based on optical: communication: research challenges and current results. Sensor Technologies and Applications, 2009. SENSORCOMM'09. Third International Conference on. IEEE.

 DOI: https://doi.org/10.26549/met.v6i1.11372

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Research of Paraphrasing for Chinese Complex Sentences Based on Templates

Zhongjian Wang* **Ling Wang**

Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China

ARTICLE INFO

Article history
Received: 19 March 2022
Revised: 26 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Complex sentence
Associated word
Paraphrasing template

ABSTRACT

Based on the paraphrasing of Chinese simple sentences, the complex sentence paraphrasing by using templates are studied. Through the classification of complex sentences, syntactic analysis and structural analysis, the proposed methods construct complex sentence paraphrasing templates that the associated words are as the core. The part of speech tagging is used in the calculation of the similarity between the paraphrasing sentences and the paraphrasing template. The joint complex sentence can be divided into parallel relationship, sequence relationship, selection relationship, progressive relationship, and interpretive relationship's complex sentences. The subordinate complex sentence can be divided into transition relationship, conditional relationship, hypothesis relationship, causal relationship and objective relationship's complex sentences. Joint complex sentence and subordinate complex sentence are divided to associated words. By using pretreated sentences, the preliminary experiment is carried out to decide the threshold between the paraphrasing sentence and the template. A small scale paraphrase experiment shows the method is availability, acquire the coverage rate of paraphrasing template 40.20% and the paraphrase correct rate 62.61%.

1. Introduction

Natural language has been widely concerned by domestic and foreign scholars. Many languages, whether written or verbal language have different expressions, Chinese is no exception. With the rapid development of computer and Internet, the massive sentence needs to be processed, including a large number of complex sentences, which requires us to paraphrase the sentence of the imminent.

According to a simple classification of the complexity of paraphrasing sentences, we can paraphrase the simple sentences and sentence rewriting. The study of simple sentence paraphrasing is relatively common, and the complex sentence paraphrasing relates to a lot of lexical and syntactic parsing, it is difficult to implement because of the need for a higher level of language processing techniques.

Careful review of a large number of documents, we found that Chinese sentence research is basically at the grammar level, the operation, a formal model of the building, representation of mathematical form and algorithm procedure and practical research are less. Especially the paraphrasing of Chinese sentence, few results can be operated in the field of natural language processing.

2. Analysis of Complex Sentence Theory

2.1 The Classification of Complex Sentence

In this paper, the classification of complex sentences is basically based on the sentence grammar literature, but

**Corresponding Author:*
Zhongjian Wang,
Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China;
Email: zhongj_w@126.com

DOI: https://doi.org/10.26549/met.v6i1.11421

not limited to grammar rules. According to the research needs, we can take to increase, delete, and summarize the grammar of the sentence structure, in order to facilitate the implementation of the paraphrase.

Generally speaking, simple sentence contains a subject and a predicate part, complex sentence is made up of two or more than two sentences, clauses can be subject-predicate sentence, also can is a non subject-predicate sentence.

The division of the grammar studies of sentence category, there are many differences. These differences make sentence category without a clear unified standard [1]. In this paper, the classification of complex sentences is based on Jiaoyan Jia [2], which puts the sentences into the joint complex sentence, subordinate complex sentence and multiple complex sentence in three categories. The joint sentence and compound sentence contains five kinds of small class.

The joint complex sentence can be divided into parallel relationship, sequence relationship, selection relationship, progressive relationship, and interpretive relationship's complex sentences. The subordinate complex sentence can be divided into transition relationship, conditional relationship, hypothesis relationship, causal relationship and objective relationship's complex sentences. Joint complex sentence and subordinate complex sentence are divided to associated words.

Multiple complex sentences are sentences that contain two or more relations which is one of the most difficult to be rewritten and is very low in terms of overwrite coverage.

2.2 Complex Sentence Semantic Analysis

Complex sentence semantic analysis results containing the word segmentation, part-of-speech tagging, and the grammar of the sentence structure analysis. Word segmentation and part-of-speech tagging is the first step of rewriting, for of complex sentence word segmentation and part-of-speech tagging, this paper adopts the ICTCLAS [3] segmentation software. Part-of-speech tagging is judging each word of the sentence in given grammatical category, determining its part-of-speech and labeling process [4].

The research target of this paper is mainly tag complex sentence that compared with tag complex sentence, no-marked complex sentence's paraphrasing is difficult, so we only extract the main part of the sentence to paraphrase such as object, predicate and subject.

The first category is the joint complex sentence of complex sentences. The joint complex sentence includes parallel, sequence, selection, progressive and interpretive complex sentences. Parallel complex sentence is composed of several clauses, each clause shows one thing, a kind of situation, a phenomenon or a particular aspect of a thing.

3. Complex Sentence Paraphrasing Strategy

On the basis of simple sentence paraphrasing, we try to paraphrase the complex sentences that use template method. Through construct corpus as the resources necessary to paraphrase complex sentences and through the simple sentences template combined to achieve complex sentences paraphrasing, and then expand the corpus size to further paraphrase complex sentences to lay the foundation for further study.

There are a lot of theoretical research on complex sentences, as mentioned in the literature [5] proposed three methods for long sentences into short sentences which are dispersion method, iterative method and segmentation method.

The basic principle of paraphrasing is the same for the simple sentence and complex sentence which is to paraphrase the sentence structure without changing the meaning of the sentence, we will use the following 4 kinds of sentence paraphrasing strategy:

1) Extract the sentence trunk, extraction of the main components for no-marked complex sentence and complex sentence having many clauses.

2) The sentence in a complex sentence merged into an attributive clause, other clauses remain unchanged. The sentence is a set of clauses in the same or similar structures, the scattered sentence is a set of sentence structure irregular.

3) On the basis of simple sentences, we add the two clause positional inverted which exchange the front and rear position between the two clauses. Simple sentence paraphrasing strategy includes the replacement, deletion, addition, repetition and locomotion of words.

4) For a sentence with metaphor, human and other rhetorical methods, we change the non obvious, ambiguous words to the obvious modification.

3.1 Template Extraction

In the process of rewriting template extraction, we use the above methods, or combination of several methods. The following template "[]" has two kinds of the contents, one is part of speech, the other is the associated word and its part of speech, there is a comma in the "< >", there is a replaceable associated word in the "{}". This thesis selects P and Q as the variable, the variable P and Q are characterized as follows:

1) P and Q are just symbols, representing different sentence elements.

 DOI: https://doi.org/10.26549/met.v6i1.11421

2) The contents of P and Q can be a sentence, phrase, word, punctuation or the combination of the above.

3) In the same sentence template for the original sentence and the paraphrasing sentence, P in the original sentence template and correspondingly in the paraphrasing template is the same sentence components. Similarly, Q in the original sentence template and correspondingly in the paraphrasing template is the same sentence components.

Paraphrase the complex sentence template extraction method:

A. Word segmentation and part of speech tagging on sentence segmentation system using ICTCLAS.

B. Each word and its part of speech in the complex sentence respectively compares with each word and its part of speech in a template of in the template library, if there is the same word and the same part of speech components between sentence and template. We use position label to replace the extracted words, at the same time, the position of the label and the extracted words are in the same position in the original sentence. If there is not the same word and the same part of speech components between sentence and template, then the loop terminates.

C. In a complex sentence, the non extracted parts are bundled into a whole between the two positions. That is bundled into a block. The word is a block that had been extracted and it is the same key word with the template.

Case 1 Original sentence: 如果讳疾忌医 , 就可能小病拖成大病。

Word segmentation, part of speech tagging:

如果 /c 讳疾忌医 /i , /w 就 /d 可能 /v 小 /a 病 /n 拖 /v 成 /v 大 /a 病 /n 。/w

The template matching with the original sentence is:

[如果 /c]+[/i]+{, /w}+[就 /d]+[/v]+[/a]+[Q]

The ingredients contained in Q are: { /n, /v, /v, /a, /n }

As shown in Figure 1 and Figure 2, complex sentence is divided into 7 blocks, ingredients 1 to 7. Figure 1 and Figure 2 in the contents of the corresponding relationship, Figure 2 is a diagram of sentence components. Among them, the composition of one to six, and the template in the same key words. Elements 7 is the uncertain variables in the template, the component 7 contains the part of speech bundled into a whole as a variable Q, the contents of ingredients 1 to 7 are arranged in the order of the original sentence.

Componen1	Componen2	Componen3	Componen4	Componen5	Componen6	Componen7

Figure 1. Sentence composition block diagram

如果/c	/i	，/w	就/d	/v	/a	Q

Figure 2. Sentence component diagram

The template has the following four categories:

1) It doesn't contain variable template, template does not contain P or Q.

2) It has a variable template, the template is only one Q or a P.

3) It has two variable templates has P and Q, or P1 and P2, or Q1 and Q2.

4) It has three variable templates contains P1, P2 and Q, or P and Q1, Q2.

The template with several uncertain variables is more complex, template extraction in the process must be refined template, this reduces the template coverage rate.

The following is a template for the paraphrasing of different associated words:

Example 1 contains the word " 如 果 " complex sentence paraphrase.

Original sentence: *如果我听妈妈的话，我就不会拉肚子了。*

Original sentence template:

[如果 /c]+[/r]+[/v]+[P]+<, /w>+[/r]+{[就 /d], [就要 /d], [就是 /d]}+[Q]

Paraphrasing sentence template:

[/r]+[/v]+[P]+<, /w>+[/r]+{[就 /d], [就要 /d], [就是 /d]}+[Q]

[假如 /c]+[/r]+[/v]+[P]+<, /w>+[/r]+{[就 /d], [就要 /d], [就是 /d]}+[Q]

[/r]+[Q]+<, /w>+[如果 /c]+[/r]+[/v]+[P]

Paraphrase the template corresponding to paraphrase the sentences as follows:

我听妈妈的话 , 我就不会拉肚子了。

假如我听妈妈的话 , 我就不会拉肚子了。

Example 2 contains the word " 只有……才 " complex sentence paraphrase

Original sentence: 只有国家强盛了 , 才不会受欺负。

Original sentence template:

[只有 /c]+[/n]+[P]+<, /w>+[才 /d]+[/d]+[/v]+[Q]

Paraphrasing sentence template:

[唯有 /c]+[/n]+[P]+<, /w>+[才 /d]+[/d]+[/v]+[Q]

[只有 /c]+[在 /c]+[/n]+[P]+[的 /u]+[条件 /n]+[下 /f]+<, /w>+[才 /d]+[/d]+[/v]+[Q]

Paraphrase the template corresponding to paraphrase the sentences as follows:

唯有国家强盛了 , 才不会受欺负。

只有在国家强盛了的条件下 , 才不会受欺负。

3.2 Paraphrasong Process

In order to improve the success paraphrasing rate, input of complex sentences need to match template in the templates library, by sentence similarity calculation to find the appropriate paraphrase template. We need to set a similar

DOI: https://doi.org/10.26549/met.v6i1.11421

level, the similarity threshold which determine by preliminary test.

We put forward an improved algorithm based on similarity calculation and paraphrase the flow chart as shown below:

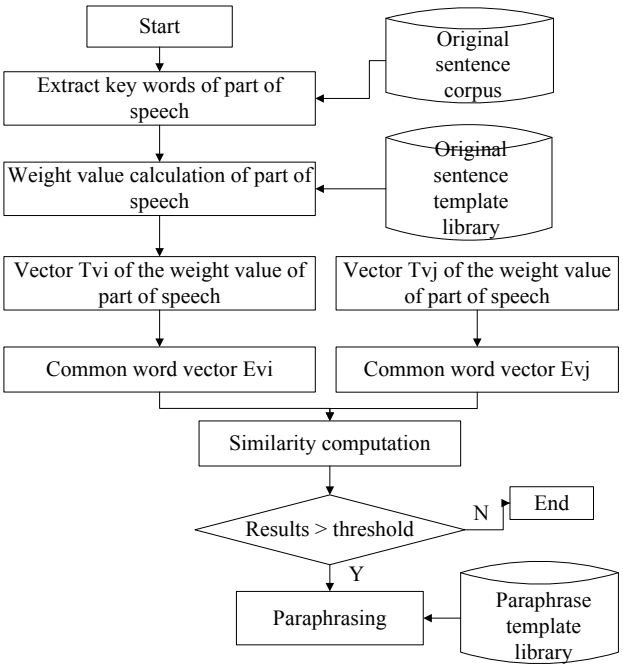


Figure 3. Paraphrase the flow chart

As shown in Figure 3, we calculate the similarity:

First of all, the sentence and the template, we extracted keywords and take vector representation. A given complex sentence T_i vector representation of T_i={m_1, m_2, m_3,..., m_n}, the number of Ti words called vector of length Ti, m_1 to m_n is T_i keyword words.

Secondly, we will introduce the calculation method of the keyword weight value. The initial weight value of each word is 1/n, weights constitute the vector called the weight value vector. Keywords vector's length is Len(-T_i), key words in this method are the sentence elements contained in the template which also include punctuation marks.

Next, we will introduce the method of calculating the common word vector. Given two sentences T_i and T_j, k and n are the length of the vector, respectively, in the T_i and T_j, among them k<=n. Every word of the mi for T_i={m_1, m_2, m_3,..., m_k}, If mi is also present in vector T_j={m_1, m_2, m_3,..., m_K}, the vector of the same words in T_i and T_j is called public word vector. This public word vector and keyword vector are the same, they are expressed as $E_{i,j}$={e_1,e_2,...,e_p}.

Finally, the similarity between the sentence and the paraphrasing template is calculated, similarity degree formula is shown below:

$$Sim(T_i, T_j) = \frac{\sum_{k=1}^{p} v_k + \sum_{l=1}^{p} v_l}{\sum_{i=1}^{n_1} v_i + \sum_{j=1}^{n_2} v_j} \times \frac{2Len(T_i)}{Len(T_i) + Len(T_j)} \quad (1)$$

In (1), v_k represents the value of item K in the common word vector Evi.

In this formula, the calculation method of the weight value is as follows:

If any one of the key words w_i in T_i or the synonym of the keyword appears in T_j, and in T_j and T_i, w_i and w_{i-1} are equal or are synonymous with each other, and the corresponding weights of Tbi value bi to increase the α times, in the same way, in T_j and T_i, w_i and w_{i+1} are equal or are synonymous with each other, the corresponding weights of Tbi value bi also increase α times, after several tests to determine the alpha =1.3. If the w_i not in T_j, the Tbi corresponding weight value remains the same.

After a lot of preliminary experiments we got a paraphrasing threshold of 0.7598, the similarity of input complex sentence template and template library of up to 75.98%, we can paraphrase the sentence according to the template.

4. Paraphrasing Experiment and Results Analysis

4.1 Experiment Procedure

We randomly selected 1500 sentences with associated words from the joint and compound complex sentence, corresponding template sentence is 603. We use the word segmentation software to carry out word segmentation and part of speech tagging, the original sentence corpus is a sentence that has been marked by word segmentation and part of speech tagging.

The experimental process is divided into two steps, one is needed to create a database, two is to write programs.

4.2 Experimental Results Analysis

In the process of manual checking paraphrasing results, we found that the small errors in the template have a great impact on the paraphrasing results. The absence of spaces will not only make a serious error in paraphrasing the results, but also the lack of spaces of different locations in the same template can lead to a lot of different errors in the result. The absence of a comma and period has a negligible effect on the correct rate of paraphrasing. Error types are the following, respectively, give examples:

(1) The original sentence missing comma in the template, such errors account for 77% of the total errors,such as Figure 4.

 DOI: https://doi.org/10.26549/met.v6i1.11421

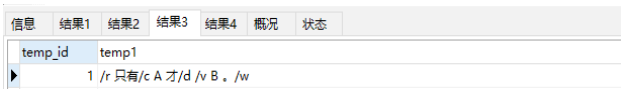


Figure 4. Original sentence template

Figure 5. Error type 1

Paraphrasing results with two comma is because the original sentence and paraphrasing template each have a comma, there is no period in the original sentence template, adds a full stop to the variable A in the process of program processing, a comma in the paraphrasing template is also added to the paraphrasing result, so there are two comma in the result, as Figure 5.

(2) Phrase collocation error, this error accounted for 18% of the total error, as Figure 6.

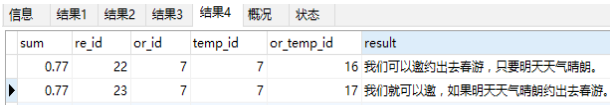


Figure 6. Error type 2

Without considering the clause phrase collocation, the sentence did not exchange the position and previous clauses together, programming is not reasonable.

(3) Long sentence similarity is low, this error is 5% of the total error, as Figure 7.

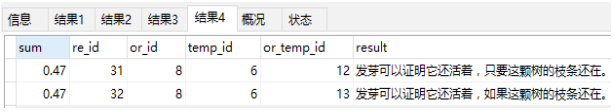


Figure 7. Error type 3

The experimental data included rewriting correct rate, the template coverage, and the rate of not being rewritten, the following specifically introduces the calculation method of all kinds of data.

We define the total number of sentences as Psum, in the result, the total number of sentences to be rewritten is Pasum, paraphrase the correct number of sentences is Rres, one of the original sentences only corresponds to a correct paraphrasing sentence. The total number of templates is Tsum.

The proportion of the sentence that has not been paraphrased is shown in the (2):

$$NPrate = \frac{Psum - pasum}{Psum} \times 100\% \quad (2)$$

Paraphrase correct rate calculation as shown in the (3):

$$PRrate = \frac{Rres}{Psum - pasum} \times 100\% \quad (3)$$

The formula for calculating the template coverage is shown in (4):

$$Trate = \frac{Tsum}{Psum - Pasum} \times 100\% \quad (4)$$

According to paraphrase the correct sentence and the total sentence compared, the proportion of sentences that have not been rewritten by the (2) is 7%. The correct rate of rewriting is 62.61%, which is obtained by the (3). The template coverage rate was 40.2%, which was obtained by the (4).

5. Conclusions

This paper presents the method of paraphrasing Chinese sentence based on template, by building to associated words as the core of the corpus, provides the basis for sentence paraphrasing. The experimental results show the effectiveness of the method and its deficiency.

The template coverage rate and correct rate is the key to paraphrase the sentences based on template. In the process of rewriting the sentences, we need a further deeper level of syntax and semantic analysis of sentences, get a more efficient paraphrasing template, raise paraphrasing accuracy and template coverage.

References

[1] Rinaldi, F., Dowdall, J., Moll, D., et al., 2003. Exploiting Paraphrases in a Question Answering System. Proceedings of Workshop in Paraphrasing at ACL2003, Sapporo, Japan.

[2] Li, W.G., Liu, T., Zhang, Y., et al., 2005. Automated Generalization of Phrasal Paraphrases from the Web. The 3rd International Workshop on Paraphrasing. Jeju Island, South Korea. pp. 49-57.

[3] ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System): http://www.ictclas.org/index.html.

[4] Zhao, Sh.Q., Liu, T., Yuan, X.Ch., et al., 2007. Automatic Acquisition of Context-Specific Lexical Paraphrases. Proceedings of IJCAI, Hyderabad, India. pp. 1789-1794.

[5] Wang, Z., Wang, L., 2010. Paraphrase of Chinese Sentences Based on Associated Word. ASIA-ICIM 2010, Wuhan, China.

DOI: https://doi.org/10.26549/met.v6i1.11421

SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Japanese-Chinese Machine Translation of Japanese Determiners Based on Templates

Ling Wang Zhongjian Wang*

Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China

ARTICLE INFO

Article history
Received: 19 March 2022
Revised: 26 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Japanese determiners
Similarity calculation
Machine translation
Translation templates

ABSTRACT

The machine translation of Japanese sentences with determiners, like "shi-ka...nai", "tyoutto...dakedeha", "tada...dake" and so on, are more special and regular on sentences structure. The research collects and classifies the Japanese sentences which contain the determiners. The classification is carried out by according to the characteristics of Japanese sentences and translation habit of Chinese sentences. Through further abstraction and simplification, translation templates are extracted by gathering grammar rules information, studying syntax and analysis the collocation mode of sentences. Those determiners express confirmed meaning, and the corresponding translation Chinese sentences have the same characteristic. By analyzing the sentence characteristics with determiners and formalizing the sentences structure, the translation templates are abstracted. By investigating the structure characteristic of original sentences with translation templates, the similarity algorithm was defined. The threshold value of the similarity calculation was obtained by preliminary experiments, and the experiments of Japanese-Chinese translation are carried out by a small corpus. The experimental results for several kinds of Japanese sentences with determiners show the translation accuracy rate is 68.6%, template coverage rate reach 83.3%. At last, through the analysis for the translation errors, following conclusion is drawn: the results of morphological analysis are erroneous, because the error of word segmentation the part of speech tagging also are erroneous, result in the grammar structure cannot match with templates; the original sentences are long and especially complex sentences; the templates are too complicated; the similarity calculation method needs to discuss further, and so on.

1. Introduction

The research of the machine translation has a long history [1], although the great progress has been achieved, because the complexity of natural language there are still a lot of difficult problems. There are various machine translation techniques, such as rules based, examples based, statistical machine translation. Usually, the different machine translation techniques have different advantages; they are used according to the different problem domain [2]. Some machine translation methods are as follows.

Example-based machine translation methods are trained from bilingual parallel corpora, which contain sentence pairs. Sentence pairs contain sentences in one language with their translations into another language. Statistical machine translation methods are applied to generate results using bilingual corpora. There are three different models of statistical machine translation, statistical word based, statistical phrase based and statistical syntax based models.

**Corresponding Author:*
Zhongjian Wang,
Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China;
Email: 3591256684@qq.com

 DOI: https://doi.org/10.26549/met.v6i1.11422

The other method is rule based machine translation. This method needs a large amount linguistic rules, which are building considering both the languages source language and target language. The rule based machine translation method is based on linking the structure of the given input sentence with the structure of the target output sentence, preserving their unique meaning. The method uses large collection of manually developed rules used for mapping source language into target language text. These can be edited to improve translations [3].

Paper [6] proposes a method for translating English sentences to Malayalam by rule based method. The main process is mediated by bilingual dictionaries and rules for converting source language structures into target language structures. The rules used in this approach are prepared based on the parts of speech tag and dependency information obtained from the parser. In their method, the transfer link rules are used for generating target structure. Morphological rules are used for assigning morphological features.

Any kind of translation method has advantages and shortages, the rule based machine translation are also no exception. There are also a lot of researches about the machine translation method based on templates.

Such as paper [5] proposed a method based on templates, the method is used to improve the performance of patent machine translation. They created more than 600 templates manually and integrated it to a rule-based MT system. Their evaluation experiment shows that the translation quality of 40% test text is improved.

Template-based machine translation is widely used in the machine translation of limited field or specific issue, such as automatic patent translation. Because in those translation problem, there are enormous amount of jargons, those jargons have uneven word frequency distribution and data sparseness, there are serious problems when applying statistical method to automatic template acquisition.

In this paper, we will discuss the translation of Japanese sentences, which contains determiners. Japanese sentences which contain the determiners are a kind of frequently used sentence. Those sentences have special determiners, rigorous grammar, and strong structural characteristics. So it is very suitable for template-based translation method which having particularly effective in the phrase-level alignment of parallel corpora.

2. The Translation Process of Japanese Sentences with Determiners

Template-based machine translation is a kind of intuitive representation method. It does not require massive knowledge of linguistics, and labor cost is low and easy to get translation templates by using corpus.

First of all, it needs to be explained, the determiners in this paper are "…sika…nai", "…nikagitte…", "tada…… dake", "…tekosohajimete…", "…kagiri…", "…nomida", "…nitodomaru", etc. In this paper, we deal with the sentences which express that that's all there is. Those sentences use the especially auxiliary word to express restriction of scope, estimation of quantity etc. For those sentences, we collect data, classify the sentences, and use the tool of Japanese morphological to carry out morphological analysis, and use the morphological analysis results to do grammar analysis and sentences structure analysis, then extract translation templates.

2.1 Classification of Japanese Sentences with Determiners

We collect and classify the Japanese sentences which contain the determiners. The classification is carried out by according to the characteristics of Japanese sentences and translation habit of Chinese sentences. The parts of sentences with determiners are listed as following in Table 1.

The Table 1 indicated nine kinds of sentences and each kind has its corresponding Chinese translation keyword.

The " しか…ない " is used to indicate that there is nothing else. For example " これしかない。(There's nothing but this.)".

The phrase "… ただ … だけ …" express various degree of amounts. For example, sentences like " 私はひとつだけ食べました。(I only ate one.)", " この乗車券は発売当日のみ有効です。(This boarding ticket is only valid on the date on which it was purchased.)". A particle that is essentially identical both grammatically and in meaning to " だけ " is " のみ ". " だけ " is used in regular conversations and " のみ " is usually only used in a written context.

2.2 Morphological Analysis of Japanese Sentences with Determiner

The Japanese text is same as Chinese text; there are no spaces between words. The Japanese text is comprised of three main written characters: Hiragana, Katakana, and Kanji. The Japanese morphological analysis includes following works, such as segmenting text into words, part-of-speech tagging, get dictionary forms for inflected verbs and adjectives and extracting readings for kanji. In this paper, WinCha [7] is used to analyze Japanese text. In fact, the morphological analysis results by WinCh include more information, but we only use the results of word segmentation and part of speech tagging.

DOI: https://doi.org/10.26549/met.v6i1.11422

Table 1. The sentences with determiners

No.	The sentences	The determiners	The Chinese translation
1	これは僕しか知らない話だ。	…しか…ない…(…しかない)	只有…；仅仅…
2	その日に限って、帰りが早かった。	… に限って …	唯独…；只有…；仅…；只限于……
3	彼よりほか知っている者はいない。	… よりほかはない …	只有…；只好…
4	ちょっと読んでだけでは、分かりません。	… ちょっと … だけでは …	只是…；光…
5	彼女はただ笑うだけで、答えません。	… ただ … だけ …	只是…；仅仅…
6	自分でやってこそはじめて分かる。	… てこそはじめて …	只有（唯有）…才能…
7	生命のつづくかぎり祖国のために尽くす。	… かぎり …	只要…就…；除非…就…
8	あとは返事を待つのみだ。	… のみだ	只有…；唯有…
9	単に希望を述べたにとどまる。	… にとどまる	只是…（而已）

We sum up nine kinds of the Japanese sentences as following.

- The particle “ だけ ” is used to express that there is all there. “ ただ ” is used with “ だけ ”, to emphasize expression meaning. “ ただ ” is used with “ だけ ”, just apple (and nothing else) to express more stronger meaning than only “ だけ ”.
- “ しか…ない ” is used to indicate that there is nothing else. The sentence express always negative. For example, “ これしかない。 There is nothing but this.” “ しか ” has an embedded negative meaning while “ だけ ” doesn’t have any particular nuance.
- “ に限って ” are consist of “ 限っ ” and “ て ”, to express specifically things and specifically scope.
- “ よりほかはない ” are consist of “ より ”, “ ほか ”, “ は ” and “ ない ”, to express the meaning that affirm this and negate other else.
- “ちょっと…だけでは” are consist of “ ちょっと ”, “ だけ ”, “ で ” and “ は ”, to express the meaning that some things happen in a particular situation. sometimes only “ ちょっと…だけ ” is used.
- “ てこそはじめて ” express the meaning when only a certain kind of situation occurs, the other kind of situation may appear.
- “ かぎり ” express the limit and range of things, structure adverbial phrase. Represents the maximum extent of competence, degree, knowledge and so on. Example, 力のかぎり戦ったのだから、負けても悔しく思いません。（ Has gone all out to fight, so even if it is lost, I will not regret it. ）
- “ のみだ ” is consist of “ のみ ” and “ だ ”. A particle that is essentially identical both grammatically and in meaning to “ だけ ” is “ のみ ”. However, unlike “ だけ ”, which is used in regular conversations, “ のみ ” is usually only used in a written context.
- “ にとどまる ” express the limit, the scope or the degree are limit in the scope that describe by the words before “ に ”. such as an example, “ 単に希望を述べたにとどまる。/I just expressed my hope.”

2.3 The Grammar Analysis and the Template Abstract

To get abstract form of Japanese-Chinese translation pair, morphological analysis of Japanese sentences and the Chinese target sentences are carried out meanwhile. By comparing the part of speech tagging of bilingual sentences, we extract grammar structure expressed by POS, key words and part of variables. We summarize the sentences’ grammar structure as following in Table 2. Each kind of the sentence is illustrated by part of speech tagging of corresponding words.

Through further abstraction and simplification, translation templates are extracted by gathering grammar rules information, studying syntax and analysis the collocation mode of sentences. The Japanese sentence patterns are a lot and its form of expressions are abundant. Especially, the sentences of including determiners are variety, various styles. In this paper emphasizes describes on translation of determiners, especially the determiners for list in Table 1.

According to the results of sentence morphological analysis, we summarize and formalize the grammar structure of sentences, thus extract the translation templates. Some examples are shown in Figure 1.

Example 1:

これはあの店でしか売っていない。
これ/名詞-代名詞-一般/は/助詞-係助詞/あの/連体詞/店/名詞-一般/で/助詞-格助詞-一般/しか/助詞-係助詞/
売っ/動詞-自立/て/助詞-接続助詞/い/動詞-非自立/ない/助動詞/。/記号-句点

Original sentence template: N１＋は＋N２＋で＋しか＋V＋ない
Extract translation template: N１+只有+N2+V】

Figure 1. A translation example

 DOI: https://doi.org/10.26549/met.v6i1.11422

We must consider much more factors when extract the translation templates because the words sequence of sentences is too long and consist of various words with different part of speech tagging. To get a translation template that can represent feature of sentence exactly, we take words as a basic unit to abstract the translation templates. The part of translation templates are shown in Table 2.

Table 2. The extraction of translation templates

Japanese Determiners	Chinese translation	Abstracted translation templates
しか…ない、しかない	只有...，仅有....	N+は+NP+しか…N+が+ない/N 除了 NP 之外没有 N N が A1+A2+しか+V+ない/因为+N＋A１,只能+V+A2
に限って、に限り、に限る、に限らず	唯有...；只有...；仅...；只限于...	N1+に+限って,N2+が+A/只限于+N1+N2+A N1+に+限って,N2+は+V/只限于+N1+V+N2
よりほかはない、よりほかには…ない	只有...;只好...	N１＋よりほか+VP＋N２＋は＋ない/除+N１＋之外没有 N２＋VP N１＋V１＋には＋N２＋を＋V２＋ほかはない/V１＋N1＋只有+N２
のみだ、	只有...唯有...	N１＋のみならず、N２＋も＋N/不仅+N1,而且 N2 N１＋は＋N２＋A+のみならず、N3 も＋A/N1+N2+不仅+A，而且＋N3+A
にとどまる、にとどまらず	只是...(而已)	N１＋は＋N２＋にとどまる/只限于，N１＋只限于+N２ N１＋を＋V＋にとどまる/只限于+V＋N１

3. System Implementation and Translation Experiments

We developed the system and carried out the evaluation experiments by collection sentences from Japanese textbook.

3.1 Translation Process

The translation process is shown in Figure 2.

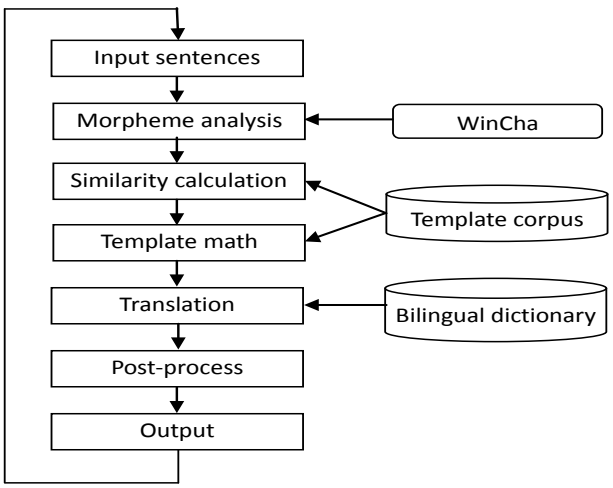


Figure 2. The translation Process

The translation process is described as follows, for the input Japanese sentences, morphological analysis is carried out. Here the WinCha is used. We only take the part of speech tagging as basic unit to formalize the grammar structure of sentence, remain the determiners and particles, got a formalizing sentence grammar structure. Then we compare it with translation templates; calculate similarity of the formalizing sentence grammar structure with translation template. When the similarity calculation value is greater or equal to the threshold value that is determined by preliminary experiments, translation template are selected. The translation is carried out by using the bilingual dictionary. Here the first entry in bilingual dictionary is used. Figure 3 is an example of translation.

3.2 Experiments and Evaluation

To evaluate the proposed method, it is necessary to carry out experiments. Through we collect 480 Japanese sentences with determiners from elementary Japanese textbook, use they as experimental data. We use 100 sentences to do preliminary experiment, to get the threshold. The similarity of source sentence with template is calculated by formula (1) as following.

$$TemSim[\%]=\begin{cases}0 & DW(T,S)=0\\ \dfrac{\alpha\cdot KW(T,S)+\beta\cdot RW(T,S)}{KW(T,S)+RW(T,S)}\times DW(T,S)\times 100 & DW(T,S)\neq 0\end{cases} \quad (1)$$

where ***KW*** is the number of part of speech tagging of the being original sentence to compare with the template; ***RW*** is the number of particle of the being original sentence to compare with the template; ***DW*** is a fixed value. When the original sentence and template contain same determiners ***DW*** is equal to 1, else 0. ***TemSim*** is the similarity of the being original sentence to match with the template. Here

DOI: https://doi.org/10.26549/met.v6i1.11422

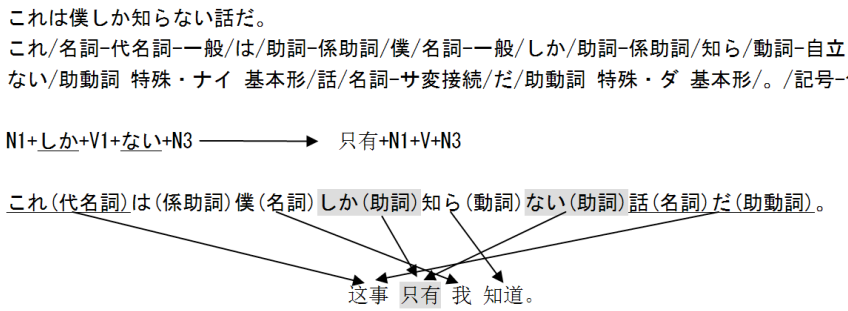


Figure 3. An example of translation

α and β are weighted parameter. The optimum coefficients are decided by the preliminary experiments using Greedy method, α=2 and β=5.

About the evaluation of experimental results, we judge the readability of translation sentence according to word order and expression meaning correctness. For each translation sentence, a score is given between 1 and 0. Here 1 is correct and 0 is error. The translation correct rate is calculated by formula (2).

$$TCR[\%] = \sum_{i=1}^{sum} \frac{S_i}{sum} \times 100 \quad (2)$$

Here TCR is the translation correct rate, Si is the score of the *i*th translation sentence, and sum is the total number of using in the evaluation experiment.

The coverage rate of translation templates (CRT) is calculated by formula (3), equal to the quotient of the template number which is used in translation experiment and the total number of translation template.

$$CRT[\%] = \frac{used\ template\ number}{the\ total\ template\ umber} \times 100 \quad (3)$$

By judging the score of each translation sentence, using formula (2) to get the TCR is 68.6%, and by formula (3) CRT is 83%. The reason that TCR is lower will be illustrated in the next section.

3.3 Analysis of Erroneous Results

We got a lower translation correct rate. Through analysis the result of translation, the summing up is as following.

1) The results of morphological analysis are erroneous. Such as Figure 6 and Figure 7. Because the error of word segmentation the part of speech tagging also are erroneous, result in the grammar structure cannot match with templates.

2) The original sentences are long and especially complex sentences. The result of morphological analysis contains various part of speech tagging, clause structure.

3) The templates are too complicated.

4) The similarity calculation method needs to discuss further.

Word boundary ambiguity cannot be ignored when dealing with Japanese sentences. Because word boundaries are not clear in Japanese, some mistakes occur when morphological analysis is carried out. For example, Figure 4 and Figure 5 show.

しばらく	シバラク	しばらく	副詞-助詞類接続		
し	シ	する	動詞-自立	サ変・スル	連用形
か	カ	か	助詞-副助詞／並立助詞／終助詞		
滞在	タイザイ	滞在	名詞-サ変接続		
でき	デキ	できる	動詞-自立	一段	未然形
ない	ナイ	ない	助動詞	特殊・ナイ	基本形
。	。	。	記号-句点		

Figure 4. A mistake of "sika" morphological analysis

表通り	オモテドオリ	表通り	名詞-一般
の	ノ	の	助詞-連体化
みか	ミカ	みか	名詞-固有名詞-人名-名
路地	ロジ	路地	名詞-一般
の	ノ	の	助詞-連体化
隅々	スミズミ	隅々	名詞-一般
まで	マデ	まで	助詞-副助詞

Figure 5. A mistake of "nomi" morphological analysis

According to error analysis of experiment result, the translation error rate that occurs owing to the mistakes of morphological analysis is 15%. The reasons of translation error are shown as Table 3.

Table 3. The ratio of the various translation error

Error reason	Word segment & POS error	Template error	Computing error of similarity	Others
Proportion [%]	15	10.5	5.8	0.7

The template matching error includes wrong template extraction and the inaccuracy of similarity calculation is 10.5% and 5.8% respectively. This shows the need for some longer sentences the proposed approach should be further research, and the similarity calculation also needs further improvement. The others error 0.7% is the transla-

 DOI: https://doi.org/10.26549/met.v6i1.11422

tion errors except for the first three items listed in Table 3.

4. Conclusions

The advantage of template-based approach is that it can carry out research in the lack of resources, but the shortcoming is expenditure of manual work in the extraction of translation templates. So the method is suitable to use the translation of the Japanese sentences those have some special vocabularies and special language phenomenon. Error analysis pointed out the existing problems of the proposed approach and disposing the details, in the future study we plan for further search on extraction of template and the improvement of similarity calculation method.

References

[1] Chen, J.R., 2013. New Approach to Translation Technologies (In Chinese). Journal of Southwest Jiaotong Universit (Social Science Edition). 6(14), 109-113.

[2] Chen, Y., Zhang, P.H., Ren, L.H., 2013. A Review on Machine Translation (In Chinese). Value Engineering. (1), 174-176.

[3] Feng, Zh.W., 2010. Machine Translation: from rule based technology to statistic based technology. CTPF, China Translation Profession Forum.

[4] Rajan, R., Sivan, R., Ravindran, R., et al., 2009. Rule Based Machine Translation from English to Malayalam. Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09. International Conference on Date of Conference.

[5] Zhang, D.M., Liu, X.D., Ji, Y.H., 2013. Chinese-English patent machine translation based on templates(In Chinese). Application Research of Computers. 7(30), 2044-2046.

[6] Wang, Zh.J., 2012. Machine Translation of Japanese-Chinese for Conjunctive particles. Applied Mechanics and Materials.

[7] ChaSen's Wiki. http://chasen.naist.jp/hiki/ChaSen/.

DOI: https://doi.org/10.26549/met.v6i1.11422

SYNERGY
PUBLISHING PTE. LTD.
Tel: +65 65881289
E-mail: contact@s-p.sg
Website: ojs.s-p.sg

2591-7110
01
9 772591 711221