From Image to Translation: Processing the Endangered Nyushu Script

TONGTAO ZHANG, ARITRA CHOWDHURY, and
NIMIT DHULEKAR, Rensselaer Polytechnic Institute
JINJING XIA, Tsinghua University
KEVIN KNIGHT, Information Sciences Institute, University of Southern California
HENG JI, and BÜLENT YENER, Rensselaer Polytechnic Institute
LIMING ZHAO, Tsinghua University

The lack of computational support has significantly slowed down automatic understanding of endangered languages. In this paper, we take Nyushu (simplified Chinese: 女书; literally: “women’s writing”) as a case study to present the first computational approach that combines Computer Vision and Natural Language Processing techniques to deeply understand an endangered language. We developed an end-to-end system to read a scanned hand-written Nyushu article, segment it into characters, link them to standard characters, and then translate the article into Mandarin Chinese. We propose several novel methods to address the new challenges introduced by noisy input and low resources, including Nyushu-specific feature selection for character segmentation and linking, and character linking lattice based Machine Translation. The end-to-end system performance indicates that the system is a promising approach and can serve as a standard benchmark.

Categories and Subject Descriptors: I.2.7 [Artificial Intelligence]: Natural Language Processing—Language Resources
General Terms: Documentation, Languages
Additional Key Words and Phrases: Endangered Languages, Nyushu, Recognition

ACM Reference Format:
Tongtao Zhang, Aritra Chowdhury, Nimit Dhulekar, Jinjing Xia, Kevin Knight, Heng Ji, Bülent Yener and Liming Zhao. 2015. From image to translation: processing the endangered Nyushu script. ACM V, N, Article A (January YYYY), 17 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

During the past 20 years, civilizations all over the world have witnessed cultural assimilation, which has led to language endangerment. It was predicted that 90 percent of mankind’s languages would meet their doom by the end of this century [Krauss 1992]. China, one of the countries that can boast of an ancient civilization, has over 5,000 years of long and mysterious history. Currently 292 living languages are spoken by 56 recognized ethnic groups. There are more than 85 endangered languages spoken by non-Han minorities in southwestern China [Bradley 2005; Zhao and Song 2011], including Nyushu [Huang 1993], [Zhao and Zhang 2014], Muya [Huang 1985], Namuyi [Zhao and Zhang 2014], Zhuang [Li 2005], Erusu [Sun 1983], Pimu [Sun et al. 2007], Naxi [He and Jiang 1985] and Shuishu [Zhang 1980]. As living “fossils”, these languages are precious and valuable to multiple research disciplines such as studying the origin and evolution of writing systems, preserving unique cultures (e.g., the affliction literature and the patio salon culture reflected in the gender-specific Nyushu, and the matrilineality system reflected in Mosuo), religions (e.g., Dongba is mainly used...
to write scriptures), music (e.g., folk songs in Zhuang), ethnicity and civilization. As Mandarin Chinese (*Putonghua*) is widely taught in the whole country as a national standard, and used by young people belonging to minority ethnicities to work in urban Han areas or to communicate with tourists, the population that masters and uses these languages regularly keeps decreasing and aging. Most masters are over 60 years old, and even their own expertise in these languages is declining [Zhao and Zhang 2014]. Apart from the decline in the usage of these languages, the other major crisis is with regards to the veracity of documents written in these languages. Some local businessmen have been forging characters and articles (e.g., randomly take components and characters to create fake documents) in these languages for the sake of commercial profit. These languages are thus suffering from the lack of standards from documentary orthography and the inadequate electronic support. As a consequence it is getting more convoluted to understand the works of the masters and to properly convey the past efforts of linguistic experts on these documents. Fortunately, in the past decade, extensive linguistic research efforts have been made toward manually documenting and translating traditional articles, and constructing dictionaries for some of these languages. However, such efforts are very expensive without computational support. Moreover, in a long-term endangered language-related project, consolidated standards and rules are required and objective evidence should be provided to ensure the quality and accuracy of recognition, translation and standardization.

One such language under threat of extinction is Nyushu (Simplified Chinese: 女书; literally "women’s writing") [Zhao 1995; 2004a; 2004b; 2005; 2008]. Nyushu was specially created for women in the ancient Hunan province of China. From ancient time to the mid 20th century, Chinese women were not allowed to receive education [Zhao 2004b], and a woman who was publicly literate would bring shame to her family. Thus, many knowledgable women, though behind the scenes, created and used this coded language to communicate with each other. As the only gender-specific language, Nyushu was kept alive by being passed from one generation of women to the next generation of women. They wrote this script on books, paper, paper fans and handkerchiefs by embroidering, engraving, stamping and calligraphy using bamboo or brush pens. The secrecy of this language lay in the fact that it appeared as normal decoration to men who did not understand or know of its existence. Figure 1 presents Nyushu script on a paper fan and a handkerchief.

Compared to their Chinese counterparts, Nyushu characters are thin, elegant, and generally presented in italicized diamond shapes. Chinese characters are converted into uniform dots, vertical bars, italicized bars and arcs in Nyushu. For example, the character “heart” is written as “心” in Chinese and “\(\text{卍}\)” in Nyushu. Each Nyushu character represents a phonogram in the Jiangyong dialect, and can be mapped to a cluster of Chinese characters which share similar pronunciations but different meanings (which is called *Rebus* in Chinese Linguistics). Therefore, there are many fewer Nyushu characters (about 800) compared to Chinese which has approximately 3,500 common characters. In fact, half of the Nyushu language experts can only chant the language, but do not have
the same level of expertise in writing the language. Moreover, neither was Nyushu promulgated
publicly, nor was it taught formally in schools. The masters of Nyushu lacked the opportunity to
receive higher education, so there are usually many variants and misspellings in Nyushu handwriting.
Another new challenge comes from the low quality of the scanned images, as we can see from
the example documents in Figure 1. The old articles were vulnerable to the bad conditions (pres-
sure, temperature, etc.) and thus not well preserved. Most documents were captured with old-style
roll film cameras and then the photos were scanned to computers several years later. Therefore,
recognition and translation methods must be robust enough to handle noisy data.

In comparison to other dominant languages such as Chinese with large labeled corpora available
for training, Nyushu’s resources are extremely insufficient for state-of-the-art Optical Character
Recognition (OCR) and Statistical Machine Translation (SMT). [Bai and Huo 2005] utilizes more
than 1 million annotated characters for OCR training in order to achieve 98.24% accuracy for Chi-
inese handwritten characters, but there is no annotation with equivalent size for Nyushu characters.
Moreover, stains on the articles, low resolution of the photos and distortion from scanners all increase
the difficulty for OCR. More importantly, stroke recognition algorithms [Liu et al. 2004; Liu et al.
2013] that have been employed for character recognition cannot be utilized in this context because
such recognition systems require users to literally “write” a character (e.g., writing capitalized “Λ”
by adding a “’” after writing “Λ”). This allows the algorithm to track the strokes including lengths,
directions, positions and order However, all Nyushu documents are scanned images and, thus, very
little knowledge about the order of the strokes in each character can be extracted. In place of the
traditional stroke-recognition algorithms, we propose a more refined character-segmentation-based
recognition algorithm specific to Nyushu. Given a scanned image of the document, our algorithm
segments out individual characters from the image, and links these characters to the standard Nyushu
characters. The primary advantage of this segmentation algorithm lies in its capacity to negate the
necessity for an extensive understanding of the specifics of the language and the variability across
hand-written documents, so that both known and novel features can be fully exploited. In addition,
we propose a novel lattice representation to compute the visual similarity between each pair of
hand-written characters and standard characters. The we apply language modeling to jointly select
the best linking and translation hypotheses. The data, resources and algorithms we created will be
useful to serve as a benchmark for researchers in Computer Vision, Image Processing and NLP
communities to work on this endangered language. Moreover, methods from our paper can also
be applied to other endangered languages that share similar characteristics with Nyushu. Also, for
more languages with scanned handwritten documents only, our methods can help bridging the gap
between raw images and clean digitalized texts.

In this research, using Nyushu as a case study, we present the first attempt to develop an end-to-
end system to fully understand a scanned hand-written document in an endangered language. We
propose a pipelined system that (1) automatically recognizes characters and segments a document
into characters using contextual information to augment image processing techniques; (2) links
the identified characters to the standard Nyushu characters; and (3) translates them into Mandarin
Chinese based on visual, semantic and contextual evidence.

2. RELATED WORK

Most of the previous work on endangered languages focused on “documenting” (e.g., [Bird 2009;
Bird et al. 2014; Uliński et al. 2014]) or constructing resources (e.g., [Riza 2008; Lam et al. 2014;
Benjamin and Radetzky 2014]), instead of “describing” these languages in terms of lexical, syntactic
and semantic analysis [Beale 2014]. A few exceptions include the automatic extraction of grammat-
ical information from interlinear text [Bender 2008; Bender et al. 2013; Bender et al. 2014] and
Machine Translation [Bird and Chiang 2012; Ombui et al. 2014]. However, none of these studies
covered methods for automatic typing from scanned documents to digitalized text files. Moreover,
none of the aforementioned methods have been applied to non-Mandarin-Chinese languages, even
though fairly involved linguistic research has been conducted on these languages since the 1960s.
Compared to traditional Machine Translation work, end-to-end translation of the endangered language takes noisy scanned documents (images) instead of clean digitalized texts as input. One of the primary difficulties of working with non-Mandarin-Chinese languages is that most corpora consist of scanned images of the original documents. Thus, a conversion mechanism is required to transform texts from the scanned images into machine-readable formats. Research efforts have been dedicated to developing image segmentation algorithms that can segment individual characters from the scanned documents [Zhao et al. 2003; Han et al. 2005; Fu et al. 2006]. The next step after segmentation is to map them to their standard counterparts. This step is performed via the computation of features on the characters; these features help distinguishing between characters and being robust to subtle variations in different versions of the same character presented in the database [Trier et al. 1996]. We will show the importance of embracing Image Processing, Computer Vision and linguistic knowledge and resources into the translation procedure. As the final module of the pipeline, our lattice representation based machine translation system determines candidates across Nyushu and Chinese characters whereas the representation [Wang et al. 2012] provides the ranking of probabilities of candidates of Chinese characters only.

3. DATA AND RESOURCE OVERVIEW

We start by converting some of the following data and linguistic resources created by Zhao et al. [Zhao 1995; 2004a; 2004b; 2005; 2008] into machine readable form.

3.1. Nyushu Corpus

In our work, the corpus (Table I) we use includes 651 documents (222,200 characters) from (Zhao, 2005), which is an authorized collection of 90% of existing Nyushu articles in multiple genres such as poems and biographies. This collection is drawn exclusively from original genuine Nyushu documents, ranging from the classical manuscripts inherited from ancient times to the articles by the last generation of Nyushu masters up until the end of the 20th century. The variability and vastness of the data is an indicator of its thoroughness and authenticity. Each document is handwritten by Nyushu masters (or language experts), without any punctuation and titles and has manual Mandarin Chinese translation which will be used to evaluate our Machine Translation output. All Nyushu documents are raw scanned images and have not been manually aligned with their Chinese translation.

<table>
<thead>
<tr>
<th>Author</th>
<th>Anon.</th>
<th>Yinxian Gao</th>
<th>Nianhua Yi</th>
<th>Huanyi Yang</th>
<th>Yanxin He</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of documents</td>
<td>64</td>
<td>180</td>
<td>80</td>
<td>130</td>
<td>197</td>
<td>651</td>
</tr>
<tr>
<td># of characters</td>
<td>34.8K</td>
<td>62.1K</td>
<td>49.7K</td>
<td>36K</td>
<td>39.6K</td>
<td>222.2K</td>
</tr>
</tbody>
</table>

In our work, the corpus (Table I) we use includes 651 documents (222,200 characters) from (Zhao, 2005), which is an authorized collection of 90% of existing Nyushu articles in multiple genres such as poems and biographies. This collection is drawn exclusively from original genuine Nyushu documents, ranging from the classical manuscripts inherited from ancient times to the articles by the last generation of Nyushu masters up until the end of the 20th century. The variability and vastness of the data is an indicator of its thoroughness and authenticity. Each document is handwritten by Nyushu masters (or language experts), without any punctuation and titles and has manual Mandarin Chinese translation which will be used to evaluate our Machine Translation output. All Nyushu documents are raw scanned images and have not been manually aligned with their Chinese translation.

3.2. Standard Nyushu Characters

There are many variants and misspellings in the Nyushu corpus. For example, all of the following four hand-written characters represent “安 (peace)”: 方, 方, 方, and 方. [Zhao 2004a] manually normalized all characters into clusters based on the pronunciation and meaning, and identified the most common character in each cluster as the canonical “standard” character, and the rest were considered as “variants” of the standard character. For the above example, the most frequent character 方 is extracted as the standard one. 398 standard Nyushu characters were finalized and a Unicode
representation was proposed for each standard character [Zhao 2004a]. In this study we exploit this valuable resource as the target set to link hand-written Nyusu characters and also use it to verify the visual features for character comparison.

3.3. Nyusu to Mandarin Chinese Dictionary

We utilize a Nyusu to Mandarin dictionary manually created in [Zhao 2004b; 2004a], which includes the alignment between 398 standard Nyusu characters and 2641 Chinese characters with frequency information, derived from the Nyusu corpus. This dictionary also records the frequencies of Nyusu characters representing Chinese characters, for example, ‘’ 1 ’’ appears as ‘’+’’ 802 times and ‘’丿’’ 309 times. Because Nyusu is a phonogram language, the mapping between Nyusu characters and Mandarin Chinese characters is many-to-many. Each Nyusu character is designed to represent a cluster of 1-15 Chinese characters sharing similar pronunciations. For example, both Nyusu characters, ㄦ and ㄢ, share the same pronunciation [pa13] in Jiangyong dialect, and so they can both be translated into either the Chinese character “被” (quilt) or the character “婢” (maidservant).

4. APPROACH

4.1. Overview

Figure 2 depicts the flow diagram of our system to understand a Nyusu document.

4.2. Nyusu Character Segmentation and Identification

Given an image of a scanned hand-written Nyusu document, our first step is to automatically segment it into individual characters. This is a hard problem because of the following reasons:

1. Noisy background in the scanned documents.
2. Inherent variability in the handwriting across various authors.
3. Issues related to the text such as overlap between characters, and disjoint components belonging to the same character.

We designed a multi-stage segmentation algorithm as illustrated in Figure 3.

Thresholding. We start by applying Otsu’s thresholding method [Otsu 1979] to separate the foreground handwritten characters from the background paper (or parchment). This is an intensity-based approach, where we dynamically identify an intensity threshold that can accurately segment out the characters and their components from the noisy background. With this method we are able to retrieve most of the components of each character as shown in Figure 3a.

Column Identification. The thresholded images are then passed to the next step for identifying columns. Since the documents are composed of characters organized in columns read from top to bottom, and right to left, the first step in the segmentation algorithm entails extracting individual columns of text. A vertical “scanning window” slides along the x-coordinate (horizontal direction) and the value of the pixels within the window are summed.
For a column with characters this sum will be non-zero, whereas for the background, which after thresholding only consists of black pixels, the sum will be approximately 0. The intuition for this “scanning window” comes from the fact that in the horizontal direction the centroids of characters within the same column vary minimally. We compare the \( x \)-coordinate of the local maximum and minimum on the curve and if the difference \( \delta > \tau \), where \( \tau \) is a threshold parameter which is empirically set to 20 pixels, we can make the maximum as a centroid (center point or line of a patch) and the minimum as a boundary of columns. Figure 3a shows this procedure. The columns are stored from right to left with decreasing \( x \)-coordinate values.

**Character Segmentation.** The segmentation of an individual character is carried out by associating the character’s components together so as to form a single complete character. This is accomplished by measuring the distance between the core component of the character and its auxiliary components and if the distance (\( L_2 \) norm) between the centroids of these components is less than \( \eta \), which is empirically set to 10 pixels, the components are thus associated to form a single character.

However, since the documents are handwritten, it’s highly possible that unequal spacing exists between two consecutive characters. Therefore, multiple characters could be identified as a single character according to the \( \eta \) threshold. In order to alleviate this problem, we perform the following two steps:
(1) After identifying that the character obtained from the previous threshold consists of multiple characters, we try to associate components by statistically determining a threshold distance $\epsilon$ as an upper bound for coupling or decoupling characters. The threshold was set to $35$.

(2) The second stage consists of splitting multiple character components by computing the overlap percentage between the characters. We observed that components belonging to the same character overlap in the $y$-coordinate. Based on this observation, when the overlap between components in the $y$-coordinate is below a threshold, we consider the components to be part of different characters. The threshold for overlapping was set as $5$ pixels.

**Character Normalization.** Once the characters have been segmented from a document, we normalize these characters (Figure 3c) to make every extracted character conform to a predetermined size and position, thereby reducing the variation in shape.

The images are mapped onto a standard image plane in order to ensure that features extracted from all images have the same dimensionality. Let the input image be denoted as $f(x, y)$, where $x$ and $y$ are the horizontal and vertical coordinates. The normalized image is denoted as $g(x', y')$.

The coordinates in the normalized image plane can be obtained from coordinates in the original coordinate system by the following mapping:

$$
x' = x'(x, y), \quad y' = y'(x, y).
$$

To simplify this further, we only consider a 1-D normalization; therefore our mapping reduces to

$$
x' = x'(x), \quad y' = y'(y).
$$

We map the coordinates from the input image to the normalized image using coordinate discretization and pixel interpolation.

By discretizing the binary image the mapped coordinates $(x', y')$ are approximated with the closest integral coordinates $(\lceil x' \rceil, \lceil y' \rceil)$. The pixel value $f(x, y)$ is mapped to all pixels in the normalized image ranging from $(\lceil x'(x) \rceil, \lceil y'(y) \rceil)$ to $(\lceil x'(x+1) \rceil, \lceil y'(y+1) \rceil)$.

The normalized image is constructed as a square of side $L$, where $L$ is set to $64$. However, normalized images may not be completely filled because the character extracted could be elongated in either the vertical or horizontal dimension. To alleviate this problem, we partially preserve the aspect ratio of the input image using the technique of Aspect Ratio Adaptive Normalization (ARAN) [Liu et al. 2000] and align the boundaries (the end of stroke projections) of the input image to the boundaries of the normalized image. The aspect ratio of the normalized image, $g$, is considered as a continuous function of the aspect ratio of the input image, $f$, and we have $g = \sqrt{\sin(\pi f/2)}$, where $f = W_1/H_1$ if $W_1 < H_1$ and $f = H_1/W_1$ otherwise, $H_1$ and $W_1$ are the height and width of the input image. If the input image is vertically elongated, the vertical dimension in the normalized image plane is filled. The horizontal dimension is centered and scaled according to the aspect ratio $(L \cdot g)$, and vice versa if the horizontal dimension is elongated. The coordinate mapping functions are as follows:

$$
x' = \frac{L}{W_1} x, \quad y' = \frac{L}{H_1} y.
$$

The linear normalization function aligns the boundaries (the end of stroke projections) of the input image to the boundaries of the normalized image.

### 4.3. Handwritten to Standard Nyushu Character Linking

Following segmentation, the next stage is linking each character to a standard Nyushu character. We extract the features from the segmented characters and utilize a classification method for linking.

**Input and Output.** The linking system is a classification problem: given a training instance matrix $X = [x_1, x_2, \ldots, x_n]^\top$ and labels $L(x_i) \in \{1, 2, \ldots, C\}$, where $i \in \{1, 2, \ldots, n\}$, and $C$ is the total number of characters in the database, we create a label matrix $Y \in \mathbb{R}^{n \times C}$ where each entry
$Y_{ij} = 1$ if $L(x_i) = j$ and $Y_{ij} = 0$ otherwise, where $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, C\}$. We also train a model $F(\cdot)$ such that $Y = F(X)$.

Then we obtain a confidence matrix with $Y_t = F(X_t)$, where $X_t = [x_{i+1}, x_{i+2}, \ldots, x_{i+l}]^\top$ is a test instance matrix and $Y_t \in \mathbb{R}^{l \times C}$. We can assign labels with $L(x_i) = \arg \max_j y_{ij}$, where $i \in \{n+1, n+2, \ldots, n+l\}$, $j \in \{1, 2, \ldots, C\}$ and $y_{ij} = [y_{i1}, y_{i2}, \ldots, y_{iC}]$ denotes a row in the confidence matrix $Y_t$. Our task is to predict and assign the testing instances $X_t = [x_{i+1}, x_{i+2}, \ldots, x_{i+l}]^\top$ with $L_j = \arg \max_j y_{ij}$ with $Y_t = [y_{1j}, y_{2j}, \ldots, y_{lj}]$. A mapping function $F(\cdot)$ is trained from $X$ and $Y$ where $F(X) \rightarrow Y$ and the prediction procedure can be conducted with $F(X_t) \rightarrow Y_t$.

Specifically, in our task, the vectors in matrix $X$ and $X_t$ are the feature vectors of the segmented characters, and each label represents one standard Nuysu character. A segmented character is considered “linked” to its standard counterpart which is represented by the label assigned to the segmented character. Moreover, We also refer to $Y_t$ as lattice to store multiple linking hypotheses and record it for later use. As an intermediate system in the whole pipeline, the linking system may receive upstream errors and also generate its own errors especially when we merely use the assigned label (with the maximum probability or confidence) for the downstream Nuysu-to-Chinese Machine Translation. A lattice contains multiple linking hypotheses. For example, in our experiments, the 30th character in image $\text{X}$ is linked to $\text{Y}$ with the highest probability of 12.7%, unfortunately this linking is incorrect. However, the character with the second highest probability of being linked to the segmented character is $\text{Y}$, and this is an accurate match. Thus, the lattice helps in increasing the linking accuracy by allowing the linking to be done with a range of potential matches rather than the single best match (we will verify this claim later in the experiment section 5).

**Feature Extraction.** We designed image-based statistical features as described in Table II. This results in a feature vector $x_i$ of 715 dimensions.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional Gradients</td>
<td>640</td>
<td>8 gradients with Kirsch operator [Kirsch 1971] plus overall gradient magnitude and direction from 64 8-pixel×8-pixel patches</td>
</tr>
<tr>
<td>Density Vector</td>
<td>64</td>
<td>Density of the pixels in each patch</td>
</tr>
<tr>
<td>Entropy</td>
<td>8</td>
<td>Entropy is assigned as sum of its value and its $k$ neighbors</td>
</tr>
<tr>
<td>Fractal Dimension</td>
<td>1</td>
<td>Complexity of the shape of the largest component</td>
</tr>
<tr>
<td>No. of Objects</td>
<td>1</td>
<td>Number of objects and connected components</td>
</tr>
<tr>
<td>Orientation</td>
<td>1</td>
<td>Major axis of the ellipse that encompasses the entire character</td>
</tr>
</tbody>
</table>

We utilize a feature selection technique based on spectral graph theory called Spectral Feature Selection (SPECS) [Zhao and Liu 2007] to identify the most important features. This method uses spectral information of the normalized Laplacian of the feature matrix. Given the feature matrix $X = [x_1, x_2, \ldots, x_n]^\top$, which consists of $n$ feature vectors, where $n$ is the number of character images, and $x_i \in \mathbb{R}^m$. We use the Gaussian radial basis function (RBF) kernel [Tsuda and Schölkopf 2004] as a similarity measure to construct the similarity matrix $S$, which calculates the correspondence of each feature vector to every other feature vector. A graph $G$ is constructed where each vertex corresponds to a feature vector and the weights $w_{ij}$ on each of the edges correspond to the entries in $S$. Given $G$, its adjacency matrix $W$ is defined as $W(i, j) = w_{ij}$.

The graph $G$ can be represented by its normalized Laplacian, $L$, calculated as $\hat{L} = D - W$; $L = D^{-\frac{1}{2}} \hat{L} D^{-\frac{1}{2}}$, where $D$ is the degree matrix constructed using the diagonal of the adjacency matrix, $W$. According to graph theory, the structural information of a graph can be obtained from its spectrum. This method employs the spectrum of the graph to measure feature relevance and
thus realizes spectral feature selection. Two functions are obtained from the normalized cut of the graph, and these functions in conjunction with the spectrum of the graph form three feature-ranking functions. The SPEC feature selection algorithm yields a subset of 203 features for each character image.

We further reduce the feature set by removing correlated features using Pearson’s correlation coefficient [Pearson 1895]. After the removal of redundant features, the feature vector is reduced to 175 features.

**Linking Algorithm.** Since this is a multi-class classification problem where the number of training instances from each class is unbalanced, methods such as Support Vector Machines (SVMs) [Cortes and Vapnik 1995] and its variants suffer from overfitting due to the rigid requirement for balancedness among instances, and are also time-consuming. Therefore, we choose Random Forest [Breiman 2001] as the classification algorithm to avoid overfitting. Random forest creates a bunch of randomized decision trees and utilizes a majority vote to determine the prediction. The input vector of a test instance will enter the root of each tree and finally reach the leaves through non-leaf nodes where trained rules and conditions are applied. The leaves store the labels which connect to a standized character. We count the output of the trees and produce the highest vote from all trees as the linked character for a test instance.

With the random forest method, we design two different learning approaches: one approach is the blind test, and the other approach is the leave-one-out methodology. For the blind test approach, we simply train with a single random forest consisting of 1,000 decision trees, and test all the instances with it; while for the leave-one-out cross-validation, we use one instance as a test instance and train a random forest model consisting of 1,000 decision trees on the rest of the instances (training instances plus other test instances) during each iteration.

### 4.4. Nyushu to Chinese Translation

The final step is to translate the Nyushu character sequence $n$ into a Mandarin Chinese sequence $c$. The translation procedure is a character-for-character substitution, which we attack with the noisy channel model: \[ \arg \max_c P(c|n) = \arg \max_n P(n|c). \]

We train a Mandarin Chinese character-bigram model $P(c)$ on manual translations of the Nyushu corpus (excluding the testing set), interpolating bigram (0.8 weight) with unigram (0.2 weight) probabilities. We derive $P(n|c)$ from the Nyushu/Mandarin dictionary. We cast these models as finite-state machines [Knight and Al-Onaizan 1998; Casacuberta 2001; Kumar and Byrne 2003], and use the Carmel toolkit [Graehl 1997] to decode the 1-best Nyushu sequence into its most likely Chinese translation. Finite-state machinery lets us alternatively decode the whole Nyushu lattice. In that case, we first exponentiate the lattice probabilities (factor of 8), as Nyushu lattice production results in fairly flat distributions. For example, Figure 4 shows the probabilities of the lattice’s top-10
Nyushu standard-character guesses for a particular image fragment (the fourth Nyushu character in Yi001 lattice). Notice that the 1-best guess only outscores the 2-best by a factor of 1.5 (0.092/0.062), and the rest of the probabilities are also within range. Therefore, the disambiguation decision is effectively turned over to the Mandarin Chinese language model, which has strong opinions about the final sequence. Raising the lattice probabilities to the 8th power increases its discriminatory power within the noisy channel - the 1-best guess now outscores the 2-best by a factor of 23. If we view the noisy channel as a log-linear model, this amounts to assigning a weight of 8 to the lattice feature, and a weight of 1 to the language model feature.

5. EXPERIMENTAL RESULTS

5.1. Data

As the annotation process is time consuming and human experts are scarce, we focus on Nianhua Yi’s documents for our experiments. We take the first five documents (Yi001-Yi005) as the test set and the rest as the training set. We collect 4281 correctly segmented and annotated characters to train the linker. Note that there are only 160 unique characters in Yi001-Yi005 instead of 398 for the whole set, for the sake of simplicity, all instances in the training set cover these 160 classes. We will use Yi003 (Figure 5a) as a walk-through example to show the input and output of each step and analyze the errors.

5.2. Segmentation

Each test document is a poem consisting of 12 sentences of 7 characters. The segmentation module successfully segments 380 correct characters from 420 characters, with a 90.48% accuracy. Figure 5b shows the segmentation result of Yi003, with errors marked with gray frames.

Comparing the counterparts between Figure 5a and 5b, we can conclude that there are two major challenges: the first challenge is to associate disconnected components together to build up the complete character, such challenge will trigger omission error and we mark them with solid frames.
in Figure 5b; the second challenge is to detect the negligible spacing (boundary) between two characters, it leads to association error and we mark the error with dotted frame in Figure 5b.

5.3. Linking

Figure 6 shows the ground truth and linking errors with the leave-one-out approach for the document Yi003. Figure 7 presents the linking accuracy of the top 150 hypotheses. The linking (identification) accuracy for the 1-best achieves approach 84% (353/420) and can be pushed to 90% with a deeper detection depth, which provides promising room for Machine Translation to select the best hypotheses. By comparing the linking results with the segmentation results in Figure 5b, we can see all segmentation errors are propagated into the linking step, except the fourth character in the
leftmost column: – the segmentation module misses two small dots in the character, but our visual features still successfully link it to the correct standard character.

Moreover, among the whole 420 test instances, we categorize the main linking errors as follows:

1. Two characters with similar structures are mistakenly grouped into the same cluster: Most of the incorrect clusters consist of this particular type of error. Humans with knowledge of Chinese Hanzi and Nyushu are able to distinguish these characters, however, the computer is unable to make this distinction purely on the basis of structures identified in the image (such as strokes). There may be different strokes between characters, but other similar strokes give the impression of a very similar framework for both characters. Figure 8a shows some examples illustrating this problem.

2. Partially similar structure can also cause some errors. This type of misclassification error is similar to the previous error with the exception that instead of a shared structure between two characters, a character or a component of one character has a similar framework with a component of the other character. This type of misclassification is difficult to identify visually (by humans) due to the difficulty in finding the similar component(s). This problem is illustrated through Figure 8b.

3. Occasionally, characters are linked because one of their components is the same. Therefore, failing to identify the different parts of two characters may lead to a misclassification. The errors are illustrated in Figure 8c.

4. In a few cases, particularly in clusters that have multiple characters, all 3 types of errors are possible, as illustrated in Figure 8d.

5.4. Translation

We evaluate our Nyushu to Mandarin Chinese Machine Translation using the following metrics:

1. **Character Accuracy**: The percentage of characters that are translated correctly.
2. **Word Recall**: The percentage of words in gold standard that appear in the system output.

<table>
<thead>
<tr>
<th></th>
<th>Child Acc.</th>
<th>Word Acc.</th>
<th>BLEU</th>
<th>Child Acc.</th>
<th>Word Acc.</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Best w/o LM</td>
<td>35.0</td>
<td>33.5</td>
<td>16.8</td>
<td>61.2</td>
<td>53.9</td>
<td>33.6</td>
</tr>
<tr>
<td>1-Best + LM</td>
<td>49.5</td>
<td>38.6</td>
<td>24.9</td>
<td>72.8</td>
<td>70.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Lattice + LM</td>
<td>50.7</td>
<td>45.7</td>
<td>28.9</td>
<td>76.0</td>
<td>72.0</td>
<td>56.4</td>
</tr>
</tbody>
</table>

ACM Journal Name, Vol. V, No. N, Article A, Publication date: January YYYY.
Table III summarizes the overall performance of the test set.
Table IV presents the output of various methods with leave-one-out approach for Yi003.

Table IV: Translation results from leave-one-out linking for Yi003, errors are bolded.

<table>
<thead>
<tr>
<th>Old Standard</th>
<th>Baseline LM</th>
<th>Baseline+LM</th>
<th>Learning LM</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>上帝原下之主</td>
<td>上帝原下之主</td>
<td>上帝原下之主</td>
<td>上帝原下之主</td>
<td>(1) had no brothers or sisters.</td>
</tr>
<tr>
<td>我们原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>My grandma (twice) had the empty house and was not happy.</td>
</tr>
<tr>
<td>王下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>My mom was young and stayed on the empty house.</td>
</tr>
<tr>
<td>你们下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>and asked my little uncle to take her (and me) home.</td>
</tr>
<tr>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>Two uncles took very good care of her (and me).</td>
</tr>
<tr>
<td>我们下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>Two aunts understood my grandma very well.</td>
</tr>
<tr>
<td>你们原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>My uncles' wives were born in educated families.</td>
</tr>
<tr>
<td>我们下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>and they know politeness and rules very well.</td>
</tr>
<tr>
<td>他们原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>原下之主</td>
<td>They often gathered together and comforted my grandma.</td>
</tr>
<tr>
<td>他们下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>下之主</td>
<td>and made her feel like she was still staying with my grandma as before.</td>
</tr>
</tbody>
</table>

Table V: Example of translation errors corrected by the language model.

<table>
<thead>
<tr>
<th>Words</th>
<th>Without Language Model</th>
<th>With Language Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngram</td>
<td>pronunciation</td>
<td>ngram</td>
</tr>
<tr>
<td>交竿</td>
<td>cross grandfather</td>
<td>Jiao Ye</td>
</tr>
<tr>
<td>搓绳</td>
<td>move mother</td>
<td>Yi Nang</td>
</tr>
<tr>
<td>蜈蚣</td>
<td>baby-sitter combines</td>
<td>Mo Bing</td>
</tr>
<tr>
<td>下心</td>
<td>below heart</td>
<td>Xia Xin</td>
</tr>
<tr>
<td>子下</td>
<td>grandson below</td>
<td>Sun Xia</td>
</tr>
<tr>
<td>四次</td>
<td>four who</td>
<td>Si Shui</td>
</tr>
<tr>
<td>出城</td>
<td>out success</td>
<td>Chu Cheng</td>
</tr>
<tr>
<td>空父</td>
<td>empty father</td>
<td>Kong Fu</td>
</tr>
<tr>
<td>伐宾</td>
<td>do together</td>
<td>Zuo Qi</td>
</tr>
<tr>
<td>成研</td>
<td>become talk</td>
<td>Cheng Jiang</td>
</tr>
</tbody>
</table>

We can clearly see that using the language model significantly improves the performance of all three measures. In particular, it successfully fixes many errors derived from similar pronunciations. These poems were written in a very terse style, so a character-bigram model is sufficient to capture some common phrases. Table V presents some examples corrected by the language model.

The lattice representation efficiently propagates multiple hypotheses into the language model for re-ranking, and achieves further improvement especially on word recall. For example, the fourth character in the second sentence of Yi003 is linked to the standard character (father or wife) with the highest probability of 9.1% and linked to the second hypothesis (father) with a probability of 8.9%. The language model successfully promoted the second hypothesis to the top because the phrase "空房 (empty house)" appears much more often than "空父 (empty father)" or "空妇 (empty wife)" in the training corpus.

Many of the remaining errors are out-of-vocabulary person names such as “义顺朝 (Yi Shunchao)”, “义Sing君 (Yi Xijun)”, “何光慈 (He Guangci)” and “唐丽英 (Tang Liying)”. In fact many of them were written in characters similar to their original mandarin Chinese forms. Some errors occur on the characters which share similar pronunciations and semantic categories with the correct translations. For example, “谷 (grandfather, pronounced as Ye)” was mistakenly translated into “多 (father, pronounced as Die)”; “急言 (crazy words, pronounced as Wang Yan)” was mistakenly translated as “狂言 (rumor, also pronounced as Wang Yan)”. There is another very interesting error. “小舅 (little uncle, pronounced as Xiao Jiu)” was mistakenly translated as “细旧 (thin and old, pronounced as Xi Jiu)”. Although the Mandarin pronunciations are slightly different, but their Jiang-yong dialect pronunciations are identical. This phenomena indicate that dialectic features should be
taken into consideration when processing and translating phonogram languages such as Nyushu, although current work does not include these features. Nevertheless, such errors are still tolerable because after chanting out the translations the audience can understand the correct meanings based on the contexts.

6. CONCLUSIONS AND FUTURE WORK

Using Nyushu as a case study, we presented an end-to-end system to tightly couple computational methods from both NLP, Image Processing and Computer Vision, and linguistic knowledge derived from field studies for deep understanding of endangered languages. Linguists such as Zhao et al. [Zhao 1995; 2004a; 2004b; 2005; 2008] have devoted more than thirty years to field studies in extreme locations without adequate funding support. The data sets and resources they collected have provided a valuable and authentic jump start for our system development. On the other hand, they have done extensive manual work which would have been dramatically expedited by the inclusion of “machine” in the loop. In the future we will improve and extend our methods to understand other endangered languages.

Acknowledgement

We sincerely appreciate Yang Gao and Tsinghua SRL Team of Rescuing Nyushu for generously providing and exhaustively cleaning up and aligning Nyushu documents which are utilized in the paper. We also appreciate Srinivas Nambirajan for helpful ideas and discussions regarding character segmentation and linking.

REFERENCES


Stephen Beale. 2014. Time to change the “D” in “DEL”. In Proc. of ACL2014, Workshop on ComputEL.


From Image to Translation: Processing the Endangered Nyushu Script

Response to Comments

Thank you very much for your encouraging and insightful comments. Based on the valuable recommendations and suggestions, we revised the manuscript and extended our paper to a full length (15-page) version. For each of your questions or comments, we provide point-by-point responses that are as follows:

Regarding Reviewer 1’s comments

Q: They state that the documents are noisy but they use a basic Otsu’s threshold for segmenting it.  
A: We mention the term “noisy” by indicating that not only image quality, but also handwriting styles, individual interpretation and authors’ mistakes reside in the dataset; therefore, Otsu’s method is simple but effective enough to handle the noise in image quality. Other modules, such as translation module mentioned in Section 5 handles noise caused by uncertainty among individual interpretation.

Q: In section 4.2, the caption “character identification” could be misunderstood. I think the authors refers to character segmentation. Later, The character identifier is well explained based on gradient features and a random forest classifier.  
A: Thank you for the suggestion on renaming the section. It makes it easier to explain our procedures clearly and succinctly.

Q: In section 4.3 the linking algorithm is difficult to follow as it is difficult for a reader to re-do that. Perhaps could be improved by giving more details.  
A: We re-organized and added more details about the linking algorithm such as clarifying the blind-test and leave-one-out approaches.

Regarding Reviewer 2’s comments

Q: The authors have a good overview of related advances in document image analysis, but seems overlook the major advances in character string recognition integrating character segmentation and classification. Particulay, the proposed lattice representation of character hypothesis is very similar to that poplarly used in character string recognition. For a reference, see Q.-F. Wang, F. Yin, C.-L. Liu, Handwritten Chinese text recognition by integrating multiple contexts, IEEE Trans. Pattern Analysis and Machine Intelligence, 34(8): 1469-1481, 2012.  
A: Thank you for the recommendation about the citation. We have investigated it and added it to the related work section. In this work, authors use the lattice representation to provide a ranked list of candidates of Chinese characters to recognize the handwritten Chinese characters. In our work, the major difference is that our lattice representation is a reference for the machine translation to determine Chinese characters from a list of possible Nyushu characters.

Q: The classification (linking) algorithm random forest has something to clarify. Does the forest of 1,000 trees deal with binary or multi-class classification?  
A: The 90.48% (380/420) accuracy is for segmentation modules, which signifies that the system can segment 380 correct Nyushu handwritten characters from the scanned 420-character documents. The graph in Fig. 6(a) (Now Fig. 7) shows a 84% (353/420) accuracy of identification, which signifies that the system can correctly identify 353 standard Nyushu characters from documents containing 420 characters.

Regarding Reviewer 3’s comments

Q: It is interesting that a detail description about ‘Nyushu’ has been provided in the introduction part of the paper to give a clear idea about ‘Nyushu’ to the readers. In the 1st paragraph of introduction part, authors have written that “Some local groups have been forging characters and articles in these...
languages”. What kind of local group do the authors mean to say here. Please explain how do the local groups forge the characters and articles in those languages?

A: For the forging issue, “local groups” refer to local businessmen (e.g., souvenir manufacturers and shopkeepers). They are forging a lot of characters and documents by randomly picking up some components and characters to make up some articles, and most of their customers are tourists who are not able to identify or distinguish them. Such action is misleading and harms the veracity. Also we have updated the phrases to clarify the “local group”.

Q: In the 2nd paragraph of introduction part, it is written that “At that time, women were not allowed to receive education”. What time frame has been talked about here when women were not allowed to receive education?

A: Generally speaking, most women in China were not allowed to be educated until 1950s, and Nyushu was invented and used about 500 years ago. Thus the timeframe here can be between 16th century and the middle 20th century. We have clarified the time frame in the latest submission.

Q: In the last paragraph of introduction part, it is written that “In place of the traditional stroke-recognition algorithms, we propose a more refined character-segmentation-based recognition algorithm . . . .” I would request them to provide some references of traditional stroke-recognition algorithms here. Please explain in what respect their algorithm is different than traditional stroke-recognition algorithms.

A: Stroke recognition systems (especially those in smartphones) require users to literally “write” a character, e.g. writing the capitalized “A” by adding a “.” after writing “A”. The systems will track the strokes including the lengths, directions, positions and orders of strokes. However, in Nyushu’s situation, what we have is merely a static image for a character and we are not able to reconstruct stroke information from it. We have updated the Introduction section to describe this in more detail.

Q: It is written in the last paragraph of introduction part that ‘. . . links these characters to the standard Nyushu character’. Is standard character mean complete character here? Please mention.

A: Standard characters are mentioned in Section 3.2. Linguists have been manually clustering the most used characters and filtering out some mistakes and outliers. The character clusters are considered as standard ones. Linguists have also created standard font for the characters which are pending for Unicode committee’s approval. For instance in Section 3.2 the ᕰ has a standardized font of ᕰ

Q: The proposed approach (section 4) is motivating and clearly written. In character identification of subsection 4.2, a threshold distance 10 has been used. Is this threshold distance fixed through the entire process for all samples?

A: For the distance threshold issue, the threshold is empirically fixed for a certain author’s handwriting. In this paper, we take Yi’s handwriting as our training and test sample set, and 10 is the best choice. We have updated section 4.2 to explain this in a better way.

Q: The identification accuracy obtained in an experiment is encouraging. From 420 characters the segmentation module successfully identified 380 characters. They may explain briefly the possible reasons of misidentification in a separate sub section and also they may put their opinion how to reduce this misidentification rate.

We extend subsection 5.3 with error examples and reasons for linking system. We provide detailed analysis on errors by comparing mistaken characters with ground truth ones.