Three-stage Short Text Language Identification Algorithm

Maimaitiyiming Hasimu¹,²,³*, Wushour Silamu¹,²
¹ School of Information Science and Engineering, Xinjiang University
  Urumqi, Xinjiang 830046, China
² Key Multi-lingual Laboratory of Xinjiang, Urumqi, Xinjiang 830046, China
³ Hotan Teachers College, Hotan, Xinjiang 848000, China
*Correspondence: mamtimin116@163.com, wushour@xju.edu.cn

ABSTRACT: Text on the internet is written in different languages and scripts, and a language identification system is used to analyze and identify them. To improve the performance of text language identification, this paper proposes a three-stage short text language identification algorithm. The script of a given text is identified in the first stage of the algorithm. The language group to which it belongs, consisting of languages written in the same script, is identified in the second stage. In the third stage, the specific language of the given text is recognized from within the language group. Experimental results showed that our proposed method improves the accuracy of text language identification systems stage by stage, reduces the time and the size of the feature set needed to make a prediction, and achieves optimal accuracy.

Subject Categories and Descriptors
[H.5.2 User Interfaces]: Natural language; [I.2.7 Natural Language Processing]

General Terms:
Language Processing, Language Identification

Keywords: Language Identification, Character n-gram, Script Identification, Language Group Identification

Received: 24 September 2017, Revised 28 October 2017, Accepted 7 November 2017

1. Introduction

Text language identification (TLI) is the process that attempts to classify a text in a language to one in a predefined set of known languages [1]. TLI is often the first step in many text processing systems. It is widely used in text mining, information retrieval, speech processing, and machine translation [2-4]. Although TLI is often portrayed as a solved problem, there is considerable room for improvement in terms of selecting, reducing, and weighting features, reducing training and testing times, and increasing the accuracy of the identification of similar languages.

Languages are written in different scripts. Each script in Unicode has a defined code range. This information helps us identify different parts of a script within a document [5]. Scripts are easy to identify based on form because their characters have different code points in Unicode. Hence, there is no need to analyze all languages in TLI when identifying a language in text: the script of the text can be identified in advance, making it easier to identify the language from among a group of languages written in the same script.

Languages belong to different families or groups. Languages in a group are related to one another as they have a common ancestral or parental language, and errors in TLI often occur in the case of similar languages in the same language group. Languages in such a family are similar in their vocabulary and structure [6, 32]. We can use this fact to discriminate different language groups written in the same script to narrow the range of identification of TLI.
Social media has become ubiquitous in recent years, and its content often consists of short multilingual text. Improving the accuracy of TLI can help the acquisition, search, and analysis of social media data. The main desiderata of a TLI system are high-speed real-time processing, efficiency, minimal storage, and robustness against textual errors [7]. In light of these preferences, we propose a three-stage short TLI algorithm in this paper. The script of a given text is identified in the first stage, and the language group to which the text belongs—consisting of languages written in the same script—is detected in the second. In the third stage, the specific language of the text within the language group is recognized. This procedure helps shorten the range of identification, and can help reduce processing time, remove noisy features, reduce the size of the feature set, and hence improve identification accuracy.

2. Related Work

Language identification (LI) is generally viewed as a form of text categorization. The authors of [8,9] provided a survey of research on the identification of written languages. In TLI, statistical models can be generated using the number of words [4] or letters in the given text [2], n-gram statistics [3,10], or a combination of the two [2]. Statistical methods require no prior linguistic knowledge and are highly accurate. The dominant statistical approach used in the literature is the character-based n-gram model, which is superior to the word-based model for small text fragments and performs equally well on large fragments. It is also tolerant to errors in text, and is easy to create and compute for any given text. Hence, most TLI systems use character n-grams [2,3,10]. We have therefore restricted the feature sets we use to character trigrams.

Many TLI studies have been carried out on a large number of languages with different scripts. The authors of [11] used the character n-gram language profile based on the most frequent character n-grams in each language. They introduced an ad-hoc “out-of-place” ranking distance to classify specific texts as belonging to a language. This system recorded an accuracy of 99.8% with an n-gram profile of 400 and text consisting of more than 300 words in 14 Indo–European languages, but accuracy decreased when identifying short texts. In [12], the authors compared eight distance measures for LI using combinations of different types of character n-grams. They analyzed 38 languages with different scripts and belonging to different language groups, and compared corpora of different sizes. They verified that decreasing text size can reduce the accuracy of TLI. The authors of [13] compared three distance measure-based methods, naive Bayes, and a support vector machine with three datasets containing different numbers of languages and sizes of documents. Experimental results showed that the TLI task becomes considerably more complex for larger numbers of languages, shorter documents, higher class skew, and multilingual documents. The author of [10] used the character n-gram model to identify 923 languages, with impressive results. Experimental results verified that confusion errors often occur among languages in the same family. To improve the accuracy of TLI online, the authors of [14] proposed a byte-sequence-based HTML parser and an HTML character entity converter for webpages prior to LI. They tested their method on webpages featuring 182 languages, where TLI accuracy increased from 86.99% to 94.04%. The authors of [4] proposed a word length algorithm for the identification of under-resourced languages. They tested it on 15 languages, and found that confusion errors occurred among closely related languages. In [2], the authors proposed five high-frequency approaches and conducted comparative tests using 11 similarity measures combined with several types of character n-grams. Their datasets consisted of 32 languages, and their experimental results verified that LI is more accurate on small datasets containing fewer languages. The authors of [15] used relative entropy to discriminate language similarity; they analyzed 27 languages written in the Roman script, and the results showed that increasing the size of training data can increase the accuracy for bigram and unigram models; moreover, when few languages are involved, the accuracy was higher. The authors of [3] analyzed factors influencing the accuracy of text-based LI. Their datasets consisted of 11 official languages of South Africa belonging to two language families. The experimental results verified that a vast majority of errors result from confusions within the same language family. The authors of [16] used five algorithms for LI experiments on 30, 60, and 90 languages, and verified that increasing the number of languages reduces the accuracy of TLI systems, and increases the time needed for prediction and training. In [17], the authors define two quantitative distances to measure how far apart two languages are. They compare the distance between two languages in forty-four languages. They found that in many cases languages within the same family or sub-family have low distances as expected. Although the above-mentioned studies analyzed languages belonging to different scripts or families, they did not classify text into different scripts or families during training and testing.

Some researchers have studied how to cluster natural languages but their methods do not apply to TLI. The authors of [18] explored the correlation between the frequency of bigrams and trigrams for each of nine European languages. The results do tend to adhere to Indo-European family tree which have been proposed by historical linguists. In [19], the authors selected character five-grams and measured similarities among documents in 31 languages to reveal a similarity-based clustering of languages. They concluded that accurate family groups can be formed by grouping together languages with similar scores. However, they verified only their system’s impressive performance at language discrimination, and did not provide any data pertaining to LI. The authors of [20] compared 108 languages and clustered them using character trigrams and the most frequently used words, and reported that the former yielded better results. The clustering results indicated that language comparisons
based on simple orthographic profiles can yield genealogical relations among languages, and can be used to detect similar languages written in the same script.

Some researchers have recently studied the identification of languages belonging to the same group. In [21, 33], the authors proposed a two-stage identification technique for similar languages. Their proposed language group prediction algorithm predicts the group to which a given language belongs, and then discriminates among languages within it. This language group classifier used character four-grams as features, and exhibited excellent performance. However, its prediction time was longer in tests than in training. The authors of [22] proposed a similar technique that uses a simple token-based maximum-entropy classifier to predict language groups. They did not provide any testing data to evaluate the efficiency of their method. On the Internet, similar languages are often presented as mixed with other languages belonging to same group. Hence, the above-mentioned approaches cannot be directly applied to TLI tasks.

Every script in Unicode has its own code range. The authors of [5] used this to traverse every letter in a document to find the starting and ending code points in a given script, detect different parts of a script in the document, and distinguish the language used in each. In LI, however, they analyzed all languages in the system rather than those using the same script.

Some studies have been conducted on the prediction, clustering, and identification of languages belonging to the same group, but there is a lack of comparative research to assess the efficiency of LI following script identification or language group prediction in TLI. Therefore, we aim to address this gap using the proposed method.

3. Preliminaries and Methods

TLI is a classification process, and our proposed method performs this classification in three stages: script identification, language group identification (LGI), and LI.

3.1 Script Identification

Each script in Unicode has a defined code range, and this allows us to detect different parts of a script in text or a sentence. Unlike the algorithm proposed in [5], we use the regular expression matching method to identify different scripts. Based on the code range of scripts in Unicode, we created a regular expression for every script. The proposed script identification stage consists of the following steps:

**Step 1:** Remove punctuation and nonalphanumeric items from the text.

**Step 2:** Ensure that the remaining text matches the regular expression of the script.

**Step 3:** Calculate the length of each matching result and, if the length is nonzero, save it in a list. Sort the list in order of descending length.

**Step 4:** Select the top item in the list as the main script of the text and return it for use in the LGI and the LI stages.

3.2 Classification Method and Evaluation

A range of classification algorithms have been used for TLI. Of these, support vector machines (SVMs), multinomial naive Bayes (NB), and logistic regression (Maxent) have yielded the best results [3,21,22]. Therefore, we chose these three classifiers for LGI and LI. The SVM is among the most effective classification algorithms and linear SVM is one of the most successful text classification algorithms [23]. The naive Bayes classifier is a simple probabilistic classifier based on Bayes' theorem. Multinomial NB is a widely accepted model for text classification [24]. In experiments, we used scikit-learn’s multinomial NB, LinearSVC, and the logistic regression toolbox with default parameter settings, excluding $C = 180$ for LinearSVC.

To test the proposed method, we used the F1 score widely used for text classification. The F1 score is defined as the harmonic mean of precision and recall. Precision is defined as the ratio of correct categorization of text to the total number of attempted classifications. Recall is the ratio of correct classifications of text to the total numbers of labeled data items in the test set. A good classifier is assumed to have a high F1 score, which indicates that the classifier performs well with respect to both precision and recall:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

3.3 Character-level n-gram Feature

An n-gram is a sequence of n consecutive letters. The n-gram-based approach for TLI divides text into character strings of equal size [11]. It is assumed that some languages use certain n-grams more frequently than others. This idea is based on Zipf’s law, which states that the size of the r-th largest occurrence of an event is inversely proportional to its rank r [25]. An example of the decomposition of the sentence “good boy” into character n-grams is shown in Table 1. The symbol “-” represents space and is used to capture the beginning and end of words.

<table>
<thead>
<tr>
<th>N-gram type</th>
<th>N-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>g, o, o, d, h, b, o, y</td>
</tr>
<tr>
<td>Bigram</td>
<td>-g, go, oo, od, -b, bo, oy, y-</td>
</tr>
<tr>
<td>Trigram</td>
<td>-go, goo, ood, -d-, -bo, -boo, boo, ook, ok-</td>
</tr>
<tr>
<td>Quadgram</td>
<td>-goo, good, oood, -od-, -d-bo, -boy, boy-</td>
</tr>
</tbody>
</table>

Table 1. Example of the decomposition of a sentence into character n-grams
4. Experiments and Results

All experiments were conducted using software written in Python 3.5. The execution environment was Microsoft Windows 10 on a computer with 32 GB of RAM and a 3.4 GHz CPU.

4.1 TLI Corpora

We selected the Leipzig Corpora Collection [26,27]. The corpora were identical in format, and similar in size and content. They contained randomly selected sentences in the language of the corpus. The sources were newspapers or text randomly collected from the Web and split into sentences.

In this study, we analyzed three widely used scripts: Latin, Cyrillic, and Arabic. According to their use on Wikipedia [28-30], we selected 51 widely used languages shown in Table 2, with 10,000 sentences for each. Of these, 49 languages were selected from the Leipzig Corpora Collection and the other two were based on Kazakh and Kirghiz, which have Arabic script (bolded items in Table 2). Sentences in these two languages were collected from relevant websites. We also used language family-related information on Wikipedia [6] and divided languages using the same script into different language groups. Tables 2–5 provide the description of the datasets. In Table 2, we list the names of the languages, their relevant ISO language code as described in [31], and the average sentence length in the corpora for each language; the minimum length for most languages was approximately 20 characters.

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>AVG</th>
<th>Min</th>
<th>Language</th>
<th>Code</th>
<th>AVG</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>afr</td>
<td>101</td>
<td>19</td>
<td>Uzbek</td>
<td>uzb</td>
<td>114</td>
<td>20</td>
</tr>
<tr>
<td>Azerbaijani</td>
<td>aze</td>
<td>109</td>
<td>18</td>
<td>Vietnamese</td>
<td>vie</td>
<td>111</td>
<td>20</td>
</tr>
<tr>
<td>Catalan</td>
<td>cat</td>
<td>133</td>
<td>19</td>
<td>Bashkir</td>
<td>bak</td>
<td>95</td>
<td>20</td>
</tr>
<tr>
<td>Czech</td>
<td>ces</td>
<td>107</td>
<td>20</td>
<td>Belarusian</td>
<td>bel</td>
<td>101</td>
<td>18</td>
</tr>
<tr>
<td>Danish</td>
<td>dan</td>
<td>114</td>
<td>19</td>
<td>Bulgarian</td>
<td>bul</td>
<td>110</td>
<td>20</td>
</tr>
<tr>
<td>German</td>
<td>deu</td>
<td>119</td>
<td>20</td>
<td>Chuvash</td>
<td>chv</td>
<td>82</td>
<td>20</td>
</tr>
<tr>
<td>English</td>
<td>eng</td>
<td>125</td>
<td>20</td>
<td>Kazakh</td>
<td>kaz</td>
<td>119</td>
<td>20</td>
</tr>
<tr>
<td>Faroese</td>
<td>fao</td>
<td>101</td>
<td>20</td>
<td>Kirghiz</td>
<td>kir</td>
<td>104</td>
<td>19</td>
</tr>
<tr>
<td>French</td>
<td>fra</td>
<td>127</td>
<td>22</td>
<td>Macedonian</td>
<td>mkd</td>
<td>126</td>
<td>19</td>
</tr>
<tr>
<td>Western Frisian</td>
<td>fry</td>
<td>90</td>
<td>20</td>
<td>Mongolian</td>
<td>mon</td>
<td>105</td>
<td>18</td>
</tr>
<tr>
<td>Indonesian</td>
<td>ind</td>
<td>117</td>
<td>19</td>
<td>Ossetian</td>
<td>oss</td>
<td>96</td>
<td>20</td>
</tr>
<tr>
<td>Icelandic</td>
<td>isl</td>
<td>99</td>
<td>35</td>
<td>Russian</td>
<td>rus</td>
<td>115</td>
<td>20</td>
</tr>
<tr>
<td>Italian</td>
<td>ita</td>
<td>126</td>
<td>19</td>
<td>Yakut</td>
<td>sah</td>
<td>87</td>
<td>20</td>
</tr>
<tr>
<td>Malay</td>
<td>msa</td>
<td>143</td>
<td>19</td>
<td>Serbian</td>
<td>srp</td>
<td>105</td>
<td>19</td>
</tr>
<tr>
<td>Dutch</td>
<td>nld</td>
<td>102</td>
<td>20</td>
<td>Tatar</td>
<td>tat</td>
<td>113</td>
<td>20</td>
</tr>
<tr>
<td>Norwegian Nynorsk</td>
<td>nno</td>
<td>104</td>
<td>20</td>
<td>Tajik</td>
<td>tgk</td>
<td>121</td>
<td>18</td>
</tr>
<tr>
<td>Norwegian Bokmål</td>
<td>nob</td>
<td>114</td>
<td>20</td>
<td>Ukrainian</td>
<td>ukr</td>
<td>113</td>
<td>20</td>
</tr>
<tr>
<td>Polish</td>
<td>pol</td>
<td>106</td>
<td>20</td>
<td>Arabic</td>
<td>ara</td>
<td>120</td>
<td>19</td>
</tr>
<tr>
<td>Portuguese</td>
<td>por</td>
<td>107</td>
<td>20</td>
<td>Kazakh (ara)</td>
<td>kaz</td>
<td>132</td>
<td>11</td>
</tr>
<tr>
<td>Romanian</td>
<td>ron</td>
<td>124</td>
<td>18</td>
<td>Kirghiz (ara)</td>
<td>kir</td>
<td>55</td>
<td>11</td>
</tr>
<tr>
<td>Slovak</td>
<td>slk</td>
<td>111</td>
<td>20</td>
<td>Persian</td>
<td>fas</td>
<td>112</td>
<td>19</td>
</tr>
<tr>
<td>Slovenian</td>
<td>slv</td>
<td>121</td>
<td>20</td>
<td>Kurdish</td>
<td>kur</td>
<td>175</td>
<td>21</td>
</tr>
<tr>
<td>Spanish</td>
<td>spa</td>
<td>133</td>
<td>20</td>
<td>Pushto</td>
<td>pus</td>
<td>109</td>
<td>19</td>
</tr>
<tr>
<td>Swedish</td>
<td>swe</td>
<td>109</td>
<td>17</td>
<td>Uighur</td>
<td>uig</td>
<td>120</td>
<td>20</td>
</tr>
<tr>
<td>Turkmen</td>
<td>tuk</td>
<td>104</td>
<td>20</td>
<td>Urdu</td>
<td>urd</td>
<td>115</td>
<td>19</td>
</tr>
<tr>
<td>Turkish</td>
<td>tur</td>
<td>108</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Average lengths of sentences of every language in the corpora (unit is character)
4.2 Script Identification

To evaluate our script identification method, we estimated the main script of each sentence and calculated the script identification results for the corpus of each script. Since our data was extracted from the Leipzig Corpora Collection, newspapers, and websites, some sentences might have contained the contents of two or more scripts. Tables 6–8 show the results of the identification of the main script of the sentences. We find that sentences in Arabic and Cyrillic contained some contents from Latin script, and those in Latin contained contents from Cyrillic script. Hence, we needed to identify the main script of a sentence before identifying its language and remove content from other scripts from the sentence. Our script identification algorithm yielded ideal efficiency in terms of time, taking only a few microseconds to identify the scripts of 510,000 sentences. This did not influence the training and testing times of the LI system.

Table 3. Latin script dataset

<table>
<thead>
<tr>
<th>Language Group</th>
<th>ISO Code List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo–European/Germanic</td>
<td>dan, fao, isl, nno, nob, swe, afr, deu, eng, fry, nld</td>
</tr>
<tr>
<td>Altaic/Turkic</td>
<td>tur, tuk, uzb, aze,</td>
</tr>
<tr>
<td>Indo–European/Italic</td>
<td>cat, fra, por, spa, ita, ron,</td>
</tr>
<tr>
<td>Indo–European/Balto–Slavic</td>
<td>slv, pol, cze, slk</td>
</tr>
<tr>
<td>Austroasiatic/Vietic</td>
<td>msa, ind</td>
</tr>
<tr>
<td>Austroasiatic/Vietic</td>
<td>Vie</td>
</tr>
</tbody>
</table>

Table 4. Arabic script dataset

<table>
<thead>
<tr>
<th>Language Group</th>
<th>ISO Code List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo–European/Indo–Iranian</td>
<td>fas, kur, pus, urd</td>
</tr>
<tr>
<td>Altaic/Turkic</td>
<td>uig, kaz*, kir*</td>
</tr>
<tr>
<td>Afro–Asiatic/Semitic</td>
<td>ara</td>
</tr>
</tbody>
</table>

Table 5. Cyrillic script dataset

<table>
<thead>
<tr>
<th>Language Group</th>
<th>ISO Code List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo–European/Balto–Slavic</td>
<td>bel, rus, ukr, bul, mkd, srp</td>
</tr>
<tr>
<td>Altaic/Turkic</td>
<td>chv, sah, bak, kaz, kir, tat</td>
</tr>
<tr>
<td>Indo–European/Indo–Iranian</td>
<td>tgl, oss</td>
</tr>
<tr>
<td>Altaic/Mongolic</td>
<td>mon</td>
</tr>
</tbody>
</table>

Table 6. Script identification results for the Latin dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Latin</th>
<th>Cyrillic</th>
<th>ID</th>
<th>Latin</th>
<th>Cyrillic</th>
<th>ID</th>
<th>Latin</th>
<th>Cyrillic</th>
<th>ID</th>
<th>Latin</th>
<th>Cyrillic</th>
</tr>
</thead>
<tbody>
<tr>
<td>tur</td>
<td>1.0</td>
<td>0</td>
<td>nno</td>
<td>1.0</td>
<td>0</td>
<td>nld</td>
<td>1.0</td>
<td>0</td>
<td>ind</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>tuk</td>
<td>1.0</td>
<td>0</td>
<td>nob</td>
<td>1.0</td>
<td>0</td>
<td>cat</td>
<td>1.0</td>
<td>0</td>
<td>slv</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>uzb</td>
<td>0.7525</td>
<td>0.2475</td>
<td>swe</td>
<td>1.0</td>
<td>0</td>
<td>fra</td>
<td>1.0</td>
<td>0</td>
<td>pol</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>aze</td>
<td>1.0</td>
<td>0</td>
<td>afr</td>
<td>1.0</td>
<td>0</td>
<td>por</td>
<td>1.0</td>
<td>0</td>
<td>ces</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>dan</td>
<td>1.0</td>
<td>0</td>
<td>deu</td>
<td>1.0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>0</td>
<td>slk</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>fao</td>
<td>1.0</td>
<td>0</td>
<td>eng</td>
<td>1.0</td>
<td>0</td>
<td>ita</td>
<td>1.0</td>
<td>0</td>
<td>vie</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>isl</td>
<td>1.0</td>
<td>0</td>
<td>fry</td>
<td>1.0</td>
<td>0</td>
<td>msa</td>
<td>1.0</td>
<td>0</td>
<td>ron</td>
<td>0.9999</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
### 4.3 Language Group Identification (LGI)
LGI consisted of preprocessing, feature selection, feature weighting, and training and testing.

#### 4.3.1 Preprocessing
To evaluate the proposed LGI, we used the 10-fold classification method. We used sklearn’s Kfold toolbox to split the corpus of each language into 10 parts, and conducted training and testing 10 times. Nine blocks were used for training and other for testing each time. The average test score was used as the final test score for classification. Following this, we cleaned the corpus of each script, and removed contents from other scripts as well as nonalphanumeric items from the text. We then extracted trigrams or bigrams to create an n-gram profile for each language group. Tables 9–10 show the number of trigrams for different sizes of corpora for each script. We can conclude from this that increasing the number of sentences in a language corpus increases the number of n-grams. The number of n-grams of every script was smaller than that of the entire TLI system. The number of n-grams of every language group was smaller than that of the relevant script as well. Narrowing the classification range reduced the number of combinations of n-grams, and helped remove noise and reduce feature size.

#### 4.3.2 Feature Selection and Weighting
In this work, we selected the frequency distribution of n-grams as the feature selection method for LGI and LI. We calculated the frequency of each n-gram in the corpus of

---

<table>
<thead>
<tr>
<th>ID</th>
<th>Cyrillic</th>
<th>Latin</th>
<th>Arabic</th>
<th>Cyrillic</th>
<th>Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td>chv</td>
<td>0.9918</td>
<td>0.0082</td>
<td>0.9991</td>
<td>0.9009</td>
<td>0.0051</td>
</tr>
<tr>
<td>sah</td>
<td>0.9964</td>
<td>0.0036</td>
<td>0.9999</td>
<td>0.0001</td>
<td>0.9977</td>
</tr>
<tr>
<td>bak</td>
<td>0.8665</td>
<td>0.1335</td>
<td>0.9373</td>
<td>0.0627</td>
<td>0.9956</td>
</tr>
<tr>
<td>kaz</td>
<td>1.0</td>
<td>0</td>
<td>0.9984</td>
<td>0.0016</td>
<td>0.9993</td>
</tr>
</tbody>
</table>

Table 7. Script identification results for the Arabic dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Arabic</th>
<th>Latin</th>
<th>Cyrillic</th>
<th>Arabic</th>
<th>Latin</th>
<th>Cyrillic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ara</td>
<td>0.9964</td>
<td>0.0036</td>
<td>0</td>
<td>pus</td>
<td>0.9946</td>
<td>0.0054</td>
</tr>
<tr>
<td>fas</td>
<td>0.9952</td>
<td>0.0048</td>
<td>0</td>
<td>urd</td>
<td>0.9978</td>
<td>0.0022</td>
</tr>
<tr>
<td>kur</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>uig</td>
<td>0.9987</td>
<td>0.0011</td>
</tr>
</tbody>
</table>

Table 8. Script identification results for the Arabic dataset

---

<table>
<thead>
<tr>
<th>Language corpora size (sentence number)</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic script (includes 8 languages)</td>
<td>28252</td>
<td>34892</td>
<td>42105</td>
<td>46685</td>
<td>50673</td>
</tr>
<tr>
<td>Cyrillic script (includes 15 languages)</td>
<td>21770</td>
<td>25079</td>
<td>28754</td>
<td>30922</td>
<td>32639</td>
</tr>
<tr>
<td>Latin script (includes 28 languages)</td>
<td>34575</td>
<td>41506</td>
<td>49022</td>
<td>53821</td>
<td>57487</td>
</tr>
<tr>
<td>LI in the entire TLI (includes 51 languages)</td>
<td>93786</td>
<td>116769</td>
<td>145253</td>
<td>164850</td>
<td>177285</td>
</tr>
</tbody>
</table>

Table 9. Number of trigrams for different sizes of corpora

<table>
<thead>
<tr>
<th>Group (language number)</th>
<th>Trigram Number</th>
<th>Group (language number)</th>
<th>Trigram Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo–European/Germanic (11)</td>
<td>19207</td>
<td>Indo–European/Balto–Slavic(4)</td>
<td>15248</td>
</tr>
<tr>
<td>Altaic/Turkic (4)</td>
<td>15328</td>
<td>Austronesian/MP (2)</td>
<td>5735</td>
</tr>
<tr>
<td>Indo–European/Italic (6)</td>
<td>12871</td>
<td>Austroasiatic/Vietic (1)</td>
<td>7514</td>
</tr>
</tbody>
</table>

Table 10. Number of trigrams for different language groups of Latin script
each language group, organized the n-grams in reverse order according to their frequencies, and selected the top k n-grams as features. Following feature selection, we used the occurrence of a term in a sentence as the weighting method for both LGI and LI, and calculated the occurrence of each feature in a sentence by converting the sentence into a vector:

$$tf(t_j) = \text{Max}_{j \in L_G} t_{ij}$$  \hspace{1cm} (2)

$$t_{ij} = \frac{\text{Num}(n\_gram_{ij})}{\text{Sum}_j}$$  \hspace{1cm} (3)

4.3.3 Training and Testing
To investigate the effect of corpus size on LGI performance, we used five sizes of sentences in each language by employing bigrams and trigrams. From Tables 11 and 12, we can conclude that an increase in sentence size in the training corpora improved LGI performance in the Maxent classifier. However, once corpus size reached 2,000 sentences in a language, LGI improved only slightly for additional sentences when using both bigram and trigram features. A similar situation occurred with both the NB and the SVM classifiers. We only provide the results for the Maxent classifier owing to limitations of space. To test the effect of LGI on TLI on the same corpus size, for the remainder of this paper, we consider 2,000 sentences in each language for LI.

<table>
<thead>
<tr>
<th>Script</th>
<th>Corpora Size</th>
<th>Feature Size</th>
<th>100</th>
<th>300</th>
<th>500</th>
<th>700</th>
<th>800</th>
<th>1000</th>
<th>1200</th>
<th>2100</th>
<th>2400</th>
<th>3300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>500</td>
<td></td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td></td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td><strong>2000</strong></td>
<td></td>
<td><strong>0.999</strong></td>
<td><strong>0.999</strong></td>
<td><strong>0.999</strong></td>
<td><strong>1.000</strong></td>
<td><strong>0.999</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>500</td>
<td></td>
<td>0.978</td>
<td>0.989</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td>0.979</td>
<td>0.992</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td><strong>1500</strong></td>
<td></td>
<td><strong>0.977</strong></td>
<td><strong>0.991</strong></td>
<td><strong>0.993</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
</tr>
<tr>
<td></td>
<td><strong>2000</strong></td>
<td></td>
<td><strong>0.976</strong></td>
<td><strong>0.992</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td>0.975</td>
<td>0.991</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>Latin</td>
<td>500</td>
<td></td>
<td>0.969</td>
<td>0.987</td>
<td>0.989</td>
<td>0.991</td>
<td>0.991</td>
<td>0.991</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td></td>
<td>0.968</td>
<td>0.989</td>
<td>0.992</td>
<td>0.992</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td><strong>1500</strong></td>
<td></td>
<td><strong>0.969</strong></td>
<td><strong>0.990</strong></td>
<td><strong>0.993</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
</tr>
<tr>
<td></td>
<td><strong>2000</strong></td>
<td></td>
<td><strong>0.970</strong></td>
<td><strong>0.990</strong></td>
<td><strong>0.993</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.994</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.995</strong></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td>0.971</td>
<td>0.991</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 11. LGI F1 score when using bigrams in Maxent
matrix 2. A similar situation occurred in the case of languages using the Cyrillic script in matrix 1, where sentences were often misclassified as belonging to the Slavic language group. As Germanic and Slavic languages are widely used in the world, other language groups contain words in these languages, thus creating noise during LGI. From our analysis, we conclude that our proposed LGI algorithm has a very high efficiency and can be used in LI projects.

<table>
<thead>
<tr>
<th>Script</th>
<th>Corpora Size</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>8000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>500</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>500</td>
<td>0.987</td>
<td>0.990</td>
<td>0.991</td>
<td>0.992</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.988</td>
<td>0.992</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.985</td>
<td>0.990</td>
<td>0.992</td>
<td>0.993</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.986</td>
<td>0.990</td>
<td>0.992</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.985</td>
<td>0.990</td>
<td>0.992</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>Latin</td>
<td>500</td>
<td>0.974</td>
<td>0.983</td>
<td>0.985</td>
<td>0.987</td>
<td>0.991</td>
<td>0.992</td>
<td>0.992</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.976</td>
<td>0.985</td>
<td>0.988</td>
<td>0.990</td>
<td>0.992</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.979</td>
<td>0.986</td>
<td>0.989</td>
<td>0.991</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.980</td>
<td>0.988</td>
<td>0.990</td>
<td>0.992</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.981</td>
<td>0.988</td>
<td>0.990</td>
<td>0.992</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Table 12. LGI F1 score when using trigrams in Maxent

<table>
<thead>
<tr>
<th>Script</th>
<th>Method</th>
<th>Feature Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Arabic</td>
<td>MaxEnt</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.999</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>MaxEnt</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.981</td>
</tr>
<tr>
<td>Latin</td>
<td>MaxEnt</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.981</td>
</tr>
</tbody>
</table>

Table 13. LGI F1 score when using bigrams in the three classification methods
Table 14. LGI F1 score when using trigrams in the three classification methods

<table>
<thead>
<tr>
<th>Script</th>
<th>Method</th>
<th>100</th>
<th>300</th>
<th>500</th>
<th>600</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>5000</th>
<th>10000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>Maxent</td>
<td>0.992</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.986</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.992</td>
<td>0.996</td>
<td>0.998</td>
<td>0.997</td>
<td>0.998</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>Maxent</td>
<td>0.943</td>
<td>0.981</td>
<td>0.989</td>
<td>0.990</td>
<td>0.993</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.920</td>
<td>0.961</td>
<td>0.979</td>
<td>0.982</td>
<td>0.988</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.933</td>
<td>0.975</td>
<td>0.984</td>
<td>0.981</td>
<td>0.988</td>
<td>0.991</td>
<td>0.992</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td>Latin</td>
<td>Maxent</td>
<td>0.929</td>
<td>0.973</td>
<td>0.984</td>
<td>0.988</td>
<td>0.992</td>
<td>0.995</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.911</td>
<td>0.961</td>
<td>0.977</td>
<td>0.981</td>
<td>0.988</td>
<td>0.993</td>
<td>0.994</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.910</td>
<td>0.964</td>
<td>0.980</td>
<td>0.981</td>
<td>0.983</td>
<td>0.991</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Confusion Matrix 1. LGI confusion matrix for languages using the Cyrillic script

```
<table>
<thead>
<tr>
<th></th>
<th>Turkic</th>
<th>Iranian</th>
<th>Mongolic</th>
<th>Slavic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turkic</td>
<td>1194</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Germanic</td>
<td>3</td>
<td>393</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Latin</td>
<td>0</td>
<td>1</td>
<td>199</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1197</td>
</tr>
</tbody>
</table>
```

Confusion Matrix 2. LGI confusion matrix for languages using the Latin script

```
<table>
<thead>
<tr>
<th></th>
<th>Turkic</th>
<th>Italic</th>
<th>MP</th>
<th>Slavic</th>
<th>Vietic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germanic</td>
<td>2192</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Turkic</td>
<td>5</td>
<td>794</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Italic</td>
<td>9</td>
<td>0</td>
<td>1191</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MP</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>399</td>
<td>0</td>
</tr>
<tr>
<td>Slavic</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>798</td>
<td>0</td>
</tr>
<tr>
<td>Vietic</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>199</td>
</tr>
</tbody>
</table>
```

4.4 Language identification (LI)
The preprocessing for LI was identical to that for LGI. To assess the efficiency of the proposed three-stage TLI, we compared three types of TLI tests:

- LI within whole languages in the TLI system without script identification and LGI. This is referred to as “LI in the entire TLI.”

- LI of languages using the same script. Having identified the script of a text, a language is distinguished from other languages using the same script. This is referred to as “LI in the same script.”

- LI within a language group. Following script and language group identification, the language of the text is distinguished from those within the language group. This is referred to as “LI in the LG.”

From Tables 15–19, we observe that increasing feature size can improve the accuracy of identification in the three types of TLI tests, but feature size more than the OFS (the feature size relevant to boldface score in Tables 15–19). LI accuracy improves very little, despite the feature size increasing several times. From the results of the TLI tasks, we can conclude that the SVM required more...
features to reach a high score than Maxent and NB for the three kinds of TLI. Maxent was slightly better than NB for overall LI as well as "LI in the same script," but the "LI in the LG" performance of NB was better than that of Maxent. We only provide the high scores relevant to OFS (boldface and underlined numbers in Tables 21–26) for NB here owing to space restrictions.

<table>
<thead>
<tr>
<th>Method</th>
<th>200</th>
<th>600</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>8000</th>
<th>10000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxent</td>
<td>0.745</td>
<td>0.895</td>
<td>0.927</td>
<td>0.949</td>
<td>0.956</td>
<td><strong>0.959</strong></td>
<td>0.960</td>
<td>0.962</td>
<td>0.964</td>
<td>0.966</td>
<td>0.968</td>
</tr>
<tr>
<td>NB</td>
<td>0.733</td>
<td>0.887</td>
<td>0.919</td>
<td>0.944</td>
<td>0.951</td>
<td>0.955</td>
<td><strong>0.957</strong></td>
<td>0.958</td>
<td>0.960</td>
<td>0.961</td>
<td>0.962</td>
</tr>
<tr>
<td>SVM</td>
<td>0.665</td>
<td>0.855</td>
<td>0.890</td>
<td>0.915</td>
<td>0.924</td>
<td>0.931</td>
<td>0.935</td>
<td><strong>0.939</strong></td>
<td>0.942</td>
<td>0.947</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Table 15. LI F1 score of the entire TLI system

<table>
<thead>
<tr>
<th>Script</th>
<th>Method</th>
<th>200</th>
<th>600</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>9000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxent</td>
<td>0.972</td>
<td>0.985</td>
<td>0.985</td>
<td>0.987</td>
<td><strong>0.988</strong></td>
<td>0.988</td>
<td>0.988</td>
<td>0.988</td>
<td>0.989</td>
<td>0.990</td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>NB</td>
<td>0.967</td>
<td>0.987</td>
<td>0.988</td>
<td><strong>0.990</strong></td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.953</td>
<td>0.978</td>
<td>0.970</td>
<td>0.979</td>
<td>0.983</td>
<td><strong>0.985</strong></td>
<td>0.986</td>
<td>0.987</td>
<td>0.987</td>
<td>0.987</td>
</tr>
<tr>
<td>Maxent</td>
<td>0.894</td>
<td>0.939</td>
<td>0.947</td>
<td>0.954</td>
<td><strong>0.956</strong></td>
<td>0.956</td>
<td>0.956</td>
<td>0.957</td>
<td>0.957</td>
<td>0.957</td>
<td>0.958</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>NB</td>
<td>0.888</td>
<td>0.938</td>
<td>0.946</td>
<td>0.952</td>
<td><strong>0.954</strong></td>
<td>0.955</td>
<td>0.957</td>
<td>0.957</td>
<td>0.957</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.859</td>
<td>0.913</td>
<td>0.925</td>
<td>0.930</td>
<td>0.928</td>
<td>0.937</td>
<td><strong>0.940</strong></td>
<td>0.941</td>
<td>0.945</td>
<td>0.948</td>
</tr>
<tr>
<td>Maxent</td>
<td>0.859</td>
<td>0.934</td>
<td>0.948</td>
<td>0.958</td>
<td><strong>0.964</strong></td>
<td>0.967</td>
<td>0.969</td>
<td>0.969</td>
<td>0.971</td>
<td>0.971</td>
<td>0.972</td>
</tr>
<tr>
<td>Latin</td>
<td>NB</td>
<td>0.888</td>
<td>0.938</td>
<td>0.946</td>
<td>0.952</td>
<td><strong>0.954</strong></td>
<td>0.955</td>
<td>0.957</td>
<td>0.957</td>
<td>0.957</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.797</td>
<td>0.897</td>
<td>0.913</td>
<td>0.929</td>
<td>0.941</td>
<td>0.949</td>
<td><strong>0.954</strong></td>
<td>0.956</td>
<td>0.960</td>
<td>0.961</td>
</tr>
</tbody>
</table>

Table 16. LI F1 score for the same script

<table>
<thead>
<tr>
<th>Language Group</th>
<th>Method</th>
<th>100</th>
<th>400</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>12000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo–Iranian</td>
<td>Maxent</td>
<td>0.967</td>
<td>0.974</td>
<td>0.973</td>
<td>0.974</td>
<td><strong>0.975</strong></td>
<td>0.973</td>
<td>0.974</td>
<td>0.978</td>
<td>0.980</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.969</td>
<td>0.978</td>
<td>0.979</td>
<td><strong>0.980</strong></td>
<td>0.980</td>
<td>0.980</td>
<td>0.981</td>
<td>0.981</td>
<td>0.981</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.960</td>
<td>0.964</td>
<td>0.947</td>
<td>0.953</td>
<td>0.953</td>
<td>0.956</td>
<td>0.960</td>
<td><strong>0.969</strong></td>
<td>0.971</td>
<td>0.971</td>
</tr>
<tr>
<td>Turkic</td>
<td>Maxent</td>
<td>0.987</td>
<td>0.998</td>
<td>0.999</td>
<td><strong>1.000</strong></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.981</td>
<td>0.999</td>
<td>0.999</td>
<td><strong>1.000</strong></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.984</td>
<td>0.999</td>
<td>0.999</td>
<td><strong>1.000</strong></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 17. LI F1 score for language groups using the Arabic script
4.4.1 LI in Latin Script

Our TLI corpus consisted of 28 languages using the Latin script and belonging to six language groups as shown in Table 3. Of these, the Austroasiatic/Vietic language group had only one member, and was identified in the LGI stage; its F1 score on the Maxent method was 0.997, as shown in Table 20, and only one of 200 sentences were misclassified as belonging to the Germanic language group in confusion matrix 2.

From the results of comparison tests for Germanic languages given in Tables 21–23, we can observe that hierarchical LI that narrows the range of LI improved identification accuracy, LI accuracy within a language group was higher than that for language in the same script, and LI accuracy of languages using the same script was higher than the overall TLI. Hierarchical LI to narrow the range of LI can significantly reduce feature size as shown in Tables 21–23. For example, for the Germanic languages (Table 21), the OFSs were 2000, 3000, and 5000 for the three types of comparison tests.

<table>
<thead>
<tr>
<th>Language Group</th>
<th>Method</th>
<th>Feature Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Turkic</td>
<td>Maxent</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.947</td>
</tr>
<tr>
<td>Germanic</td>
<td>Maxent</td>
<td>0.840</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.770</td>
</tr>
<tr>
<td>Italic</td>
<td>Maxent</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.934</td>
</tr>
<tr>
<td>MP</td>
<td>Maxent</td>
<td>0.724</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.625</td>
</tr>
<tr>
<td>Slavic</td>
<td>Maxent</td>
<td>0.940</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.937</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.924</td>
</tr>
</tbody>
</table>

Table 18. LI F1 score for the language groups using the Cyrillic script

Table 19. LI F1 score of language groups using the Latin script
There were some highly similar languages, such as Norwegian Nynorsk (nno) and Norwegian Bokmål (nob) (Table 20), and Malay (msa) and Indonesian (ind) (Table 21). Their identification accuracy values were lower than those of languages using the Latin script as confusion errors often occurred in these cases. For example, we inspected the confusion matrix for one-time naïve Bayes classification when the feature size was 2,000. In matrix 3, eight Norwegian Nynorsk sentences of a total of 200 sentences were misclassified as Norwegian Bokmål, and 13 sentences of Norwegian Bokmål were misclassified as those of Norwegian Nynorsk language. We also investigated the confusion matrix for Malay and Indonesian using one-time naïve Bayes classification. A total of 28 Indonesian sentences out of 200 sentences were misclassified as those of Malay, and 28 Malay sentences were misclassified as those of Indonesian.

<table>
<thead>
<tr>
<th>Script</th>
<th>Language Group</th>
<th>Feature Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Arabic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semitic</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>Indo–Iranian</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>Turkic</td>
<td>1.000</td>
</tr>
<tr>
<td>Cyrillic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indo–Iranian</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>Mongolic</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td>Slavic</td>
<td>0.989</td>
</tr>
<tr>
<td>Latin</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Germanic</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>Turkic</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>Italic</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td>Slavic</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>Vietic</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 20. LGI F1 score for every language group when using bigrams in Maxent.

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>afr</th>
<th>dan</th>
<th>Deu</th>
<th>eng</th>
<th>fao</th>
<th>Fry</th>
<th>isl</th>
<th>nld</th>
<th>nno</th>
<th>nob</th>
<th>Swe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Germanic Languages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.987</td>
<td>0.948</td>
<td>0.993</td>
<td>0.990</td>
<td>0.995</td>
<td>0.973</td>
<td>1.000</td>
<td>0.963</td>
<td>0.933</td>
<td>0.898</td>
<td>0.980</td>
</tr>
<tr>
<td>3000</td>
<td>0.988</td>
<td>0.955</td>
<td>0.995</td>
<td>0.989</td>
<td>0.996</td>
<td>0.975</td>
<td>1.000</td>
<td>0.966</td>
<td>0.941</td>
<td>0.913</td>
<td>0.984</td>
</tr>
<tr>
<td>5000</td>
<td>0.989</td>
<td>0.968</td>
<td>0.996</td>
<td>0.989</td>
<td>0.997</td>
<td>0.975</td>
<td>1.000</td>
<td>0.967</td>
<td>0.948</td>
<td>0.928</td>
<td>0.988</td>
</tr>
<tr>
<td><strong>LI in Latin Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.981</td>
<td>0.933</td>
<td>0.990</td>
<td>0.980</td>
<td>0.991</td>
<td>0.971</td>
<td>1.000</td>
<td>0.957</td>
<td>0.916</td>
<td>0.875</td>
<td>0.970</td>
</tr>
<tr>
<td><strong>3000</strong></td>
<td>0.985</td>
<td>0.943</td>
<td>0.991</td>
<td>0.980</td>
<td>0.994</td>
<td>0.973</td>
<td>1.000</td>
<td>0.961</td>
<td>0.923</td>
<td>0.891</td>
<td>0.976</td>
</tr>
<tr>
<td>5000</td>
<td>0.987</td>
<td>0.956</td>
<td>0.992</td>
<td>0.978</td>
<td>0.995</td>
<td>0.973</td>
<td>1.000</td>
<td>0.963</td>
<td>0.937</td>
<td>0.910</td>
<td>0.980</td>
</tr>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.961</td>
<td>0.889</td>
<td>0.983</td>
<td>0.968</td>
<td>0.985</td>
<td>0.963</td>
<td>1.000</td>
<td>0.940</td>
<td>0.887</td>
<td>0.827</td>
<td>0.952</td>
</tr>
<tr>
<td>3000</td>
<td>0.973</td>
<td>0.917</td>
<td>0.988</td>
<td>0.970</td>
<td>0.990</td>
<td>0.967</td>
<td>1.000</td>
<td>0.951</td>
<td>0.902</td>
<td>0.857</td>
<td>0.960</td>
</tr>
<tr>
<td><strong>5000</strong></td>
<td>0.982</td>
<td>0.933</td>
<td>0.989</td>
<td>0.977</td>
<td>0.992</td>
<td>0.972</td>
<td>1.000</td>
<td>0.959</td>
<td>0.916</td>
<td>0.876</td>
<td>0.969</td>
</tr>
</tbody>
</table>

Table 21. Comparative TLI results for Germanic languages in NB.
### Table 22. Comparative TLI results for Italic and Malayo-Polynesian (MP) languages in NB

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Cat</th>
<th>fra</th>
<th>ita</th>
<th>por</th>
<th>ron</th>
<th>spa</th>
<th>Feature Size</th>
<th>ind</th>
<th>msa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Italic languages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in MP languages</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.987</td>
<td>0.997</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td>0.986</td>
<td>2000</td>
<td>0.856</td>
<td>0.851</td>
</tr>
<tr>
<td>3000</td>
<td>0.988</td>
<td>0.997</td>
<td>0.997</td>
<td>0.996</td>
<td>0.997</td>
<td>0.988</td>
<td>3000</td>
<td>0.863</td>
<td>0.862</td>
</tr>
<tr>
<td>5000</td>
<td>0.989</td>
<td>0.998</td>
<td>0.997</td>
<td>0.996</td>
<td>0.997</td>
<td>0.989</td>
<td>5000</td>
<td>0.862</td>
<td>0.864</td>
</tr>
<tr>
<td><strong>LI in Latin Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in Latin Script</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.981</td>
<td>0.993</td>
<td>0.988</td>
<td>0.990</td>
<td>0.989</td>
<td>0.980</td>
<td>2000</td>
<td>0.798</td>
<td>0.805</td>
</tr>
<tr>
<td>3000</td>
<td>0.984</td>
<td>0.995</td>
<td>0.989</td>
<td>0.993</td>
<td>0.992</td>
<td>0.983</td>
<td>3000</td>
<td>0.812</td>
<td>0.817</td>
</tr>
<tr>
<td>5000</td>
<td>0.987</td>
<td>0.996</td>
<td>0.993</td>
<td>0.995</td>
<td>0.993</td>
<td>0.984</td>
<td>5000</td>
<td>0.830</td>
<td>0.835</td>
</tr>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.974</td>
<td>0.984</td>
<td>0.980</td>
<td>0.984</td>
<td>0.983</td>
<td>0.974</td>
<td>2000</td>
<td>0.740</td>
<td>0.759</td>
</tr>
<tr>
<td>3000</td>
<td>0.979</td>
<td>0.988</td>
<td>0.984</td>
<td>0.988</td>
<td>0.985</td>
<td>0.976</td>
<td>3000</td>
<td>0.767</td>
<td>0.784</td>
</tr>
<tr>
<td>5000</td>
<td>0.983</td>
<td>0.993</td>
<td>0.989</td>
<td>0.992</td>
<td>0.990</td>
<td>0.981</td>
<td>5000</td>
<td>0.799</td>
<td>0.810</td>
</tr>
</tbody>
</table>

### Table 23. Comparative TLI results for Turkic and Slavic languages in NB

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Aze</th>
<th>tuk</th>
<th>tur</th>
<th>uzb</th>
<th>Feature Size</th>
<th>ces</th>
<th>pol</th>
<th>slk</th>
<th>slv</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Turkic languages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in Slavic languages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.984</td>
<td>0.998</td>
<td>0.984</td>
<td>0.994</td>
<td>2000</td>
<td>0.991</td>
<td>0.999</td>
<td>0.990</td>
<td>0.996</td>
</tr>
<tr>
<td>3000</td>
<td>0.987</td>
<td>0.999</td>
<td>0.990</td>
<td>0.996</td>
<td>3000</td>
<td>0.994</td>
<td>0.999</td>
<td>0.993</td>
<td>0.997</td>
</tr>
<tr>
<td>5000</td>
<td>0.988</td>
<td>0.999</td>
<td>0.991</td>
<td>0.996</td>
<td>5000</td>
<td>0.995</td>
<td>0.999</td>
<td>0.993</td>
<td>0.997</td>
</tr>
<tr>
<td><strong>LI in Latin Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in Latin Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.975</td>
<td>0.992</td>
<td>0.966</td>
<td>0.988</td>
<td>2000</td>
<td>0.972</td>
<td>0.994</td>
<td>0.976</td>
<td>0.987</td>
</tr>
<tr>
<td>3000</td>
<td>0.985</td>
<td>0.998</td>
<td>0.983</td>
<td>0.992</td>
<td>3000</td>
<td>0.980</td>
<td>0.997</td>
<td>0.985</td>
<td>0.992</td>
</tr>
<tr>
<td>5000</td>
<td>0.987</td>
<td>0.998</td>
<td>0.987</td>
<td>0.993</td>
<td>5000</td>
<td>0.985</td>
<td>0.998</td>
<td>0.987</td>
<td>0.992</td>
</tr>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.951</td>
<td>0.962</td>
<td>0.903</td>
<td>0.941</td>
<td>2000</td>
<td>0.951</td>
<td>0.987</td>
<td>0.951</td>
<td>0.974</td>
</tr>
<tr>
<td>3000</td>
<td>0.970</td>
<td>0.989</td>
<td>0.932</td>
<td>0.945</td>
<td>3000</td>
<td>0.959</td>
<td>0.993</td>
<td>0.965</td>
<td>0.981</td>
</tr>
<tr>
<td>5000</td>
<td>0.977</td>
<td>0.991</td>
<td>0.937</td>
<td>0.944</td>
<td>5000</td>
<td>0.974</td>
<td>0.995</td>
<td>0.977</td>
<td>0.986</td>
</tr>
</tbody>
</table>

### Confusion Matrix 3. LI confusion matrix for languages using the Latin script

<table>
<thead>
<tr>
<th>language</th>
<th>afr</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fao</th>
<th>fry</th>
<th>isl</th>
<th>nld</th>
<th>nno</th>
<th>nob</th>
<th>swe</th>
</tr>
</thead>
<tbody>
<tr>
<td>afr</td>
<td>199</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>dan</td>
<td>0</td>
<td>186</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>deu</td>
<td>0</td>
<td>0</td>
<td>199</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>199</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>fao</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>fry</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>194</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>isl</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>nld</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>194</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>nno</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>185</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>nob</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>185</td>
</tr>
<tr>
<td>swe</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>193</td>
</tr>
</tbody>
</table>
4.4.2 LI in Cyrillic and Arabic Scripts
Our TLI corpus contained texts in 15 languages using the Cyrillic script belonging to four language groups as shown in Table 5. Of these, the Mongolian language group had only one member: Mongolian. Its F1 score on the Maxent method was 0.993, as shown in Table 20, where only one Mongolian sentence was misclassified as belonging to the Common Turkic languages on the Maxent classifier in confusion matrix 1.

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Bak</th>
<th>chv</th>
<th>kaz</th>
<th>kir</th>
<th>sah</th>
<th>Tat</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Turkic languages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.678</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.993</td>
<td>0.720</td>
</tr>
<tr>
<td>3000</td>
<td>0.690</td>
<td>0.995</td>
<td>0.998</td>
<td>0.996</td>
<td>0.994</td>
<td>0.721</td>
</tr>
<tr>
<td>5000</td>
<td>0.696</td>
<td>0.995</td>
<td>0.998</td>
<td>0.997</td>
<td>0.994</td>
<td>0.719</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>oss</th>
<th>tgk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Iranian languages</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>5000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Bak</th>
<th>chv</th>
<th>kaz</th>
<th>kir</th>
<th>sah</th>
<th>Tat</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Cyrillic Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.672</td>
<td>0.995</td>
<td>0.997</td>
<td>0.993</td>
<td>0.989</td>
<td>0.717</td>
</tr>
<tr>
<td>3000</td>
<td>0.672</td>
<td>0.995</td>
<td>0.998</td>
<td>0.995</td>
<td>0.989</td>
<td>0.722</td>
</tr>
<tr>
<td>5000</td>
<td>0.693</td>
<td>0.995</td>
<td>0.998</td>
<td>0.997</td>
<td>0.989</td>
<td>0.731</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>oss</th>
<th>tgk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Cyrillic Script</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>0.991</td>
<td>0.981</td>
</tr>
<tr>
<td>5000</td>
<td>0.999</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>bel</th>
<th>bul</th>
<th>mkd</th>
<th>Rus</th>
<th>Srp</th>
<th>ukr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>bel</th>
<th>bul</th>
<th>mkd</th>
<th>Rus</th>
<th>Srp</th>
<th>ukr</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 24. Comparative TLI results for Turkic and Iranian languages in NB

Table 25. Comparative TLI results for Slavic languages in NB

<table>
<thead>
<tr>
<th>Language</th>
<th>bak</th>
<th>chv</th>
<th>kaz</th>
<th>Kir</th>
<th>sah</th>
<th>tat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bak</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>Chv</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kaz</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kir</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>199</td>
<td>0</td>
</tr>
<tr>
<td>Tat</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>154</td>
</tr>
</tbody>
</table>

Confusion Matrix 4. LI in Turkic language group.
There were three different language groups of languages using the Arabic script, as shown in Table 4. Of these, the Semitic language group had only one member, Arabic. It was identified during the LGI process for the Arabic script. Its identification F1 score attained optimal accuracy.

The identification accuracy of the languages improved gradually following script identification and language group identification for languages using the Cyrillic and the Arabic scripts. From Tables 24–26, it can be seen that the OFS decreased significantly after hierarchical LI. For the Slavic languages, as can be seen in Table 25, the OFS occurred at 2000, 3000, and 5000 for three types of tests.

Highly similar languages were also present among those using the Cyrillic script; the Bashkir and Tatar languages belong to the Common Turkic language group. Confusion errors often arise between Bashkir and Tatar. For example, in confusion matrix 4 (in a one-time naive Bayes classification when the feature size was 2000), 49 Bashkir (bak) sentences out of 200 sentences were misclassified as belonging to Tatar (tat), and 46 sentences in Tatar out of 200 sentences were misclassified as those in Bashkir.

### 4.5 Time Efficiency Analysis for Hierarchical LI

To evaluate the temporal efficiency of our proposed three-stage TLI, we compared the training and testing times of the three types of TLI tests when the feature size was equal to the OFS. We selected the Maxent classifier with bigrams as the feature for LGI because the Maxent classifier requires fewer features to reach its OFS when using bigrams, as shown in Table 13. We used Python's datetime Toolbox to calculate the training and testing times.

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>fas</th>
<th>kur</th>
<th>pus</th>
<th>Urd</th>
<th>Feature Size</th>
<th>Kaz(ara)</th>
<th>Kir(ara)</th>
<th>uig</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LI in Indo–Iranian</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in Turkic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.963</td>
<td>0.999</td>
<td>0.995</td>
<td>0.963</td>
<td>800</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>2000</td>
<td>0.963</td>
<td>1.000</td>
<td>0.996</td>
<td>0.963</td>
<td>2000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>5000</td>
<td>0.963</td>
<td>1.000</td>
<td>0.996</td>
<td>0.963</td>
<td>5000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>LI in Arabic Script</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in Arabic Script</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.961</td>
<td>0.997</td>
<td>0.993</td>
<td>0.963</td>
<td>800</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>2000</td>
<td>0.963</td>
<td>0.999</td>
<td>0.996</td>
<td>0.963</td>
<td>2000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>5000</td>
<td>0.963</td>
<td>1.000</td>
<td>0.995</td>
<td>0.963</td>
<td>5000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>LI in the entire TLI</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.952</td>
<td>0.994</td>
<td>0.984</td>
<td>0.958</td>
<td>800</td>
<td>0.936</td>
<td>0.939</td>
<td>0.978</td>
</tr>
<tr>
<td>2000</td>
<td>0.960</td>
<td>0.995</td>
<td>0.992</td>
<td>0.961</td>
<td>2000</td>
<td>0.990</td>
<td>0.989</td>
<td>0.994</td>
</tr>
<tr>
<td>5000</td>
<td>0.961</td>
<td>0.998</td>
<td>0.994</td>
<td>0.963</td>
<td>5000</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
</tbody>
</table>

**Table 26. Comparative TLI results for Indo–Iranian and Turkic languages in NB**

<table>
<thead>
<tr>
<th>Method</th>
<th>Arabic script</th>
<th>Cyrillic script</th>
<th>Latin script</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OFS</td>
<td>train</td>
<td>test</td>
</tr>
<tr>
<td>Maxent</td>
<td>700</td>
<td>00:02.9</td>
<td>00:00.1</td>
</tr>
</tbody>
</table>

**Table 27. Training and testing times for LGI for the three scripts**

<table>
<thead>
<tr>
<th>Method</th>
<th>Arabic script</th>
<th>Cyrillic script</th>
<th>Latin script</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OFS</td>
<td>train</td>
<td>test</td>
</tr>
<tr>
<td>Maxent</td>
<td>3000</td>
<td>00:09.5</td>
<td>00:00.3</td>
</tr>
<tr>
<td>NB</td>
<td>2000</td>
<td>00:01.8</td>
<td>00:00.3</td>
</tr>
<tr>
<td>SVM</td>
<td>4000</td>
<td>00:05.5</td>
<td>00:00.4</td>
</tr>
</tbody>
</table>

**Table 28. Training and testing times for LI in the same script**
taken for classification. The unit format is minute: second. The time to make a prediction (testing) was shorter for Maxent. The training time for Maxent was longer than that of the other two classifiers when using NB for LI. Most of the training time was consumed by the LGI process, but the training time for the NB in LI in the LG was shorter than those for the Maxent and the SVM classifiers. The training time for Maxent and SVM in LI in the LG was shorter than those on LI in the script and LI for the overall TLI test.

Comparing the times of the three types of TLI tests, we find that following script identification and LGI, language prediction time (testing time) was significantly shorter for the three classification methods. The testing time in hierarchical language classification was shorter than LI for overall TLI and LI in the same script. Testing in the same script was faster than in overall LI. The testing time of Maxent for LI in the LG was shorter than those of the NB and the SVM classifiers.

### Table 29. Training and testing times for LI in LG for the Arabic script

<table>
<thead>
<tr>
<th>Method</th>
<th>Indo-European</th>
<th>Turkic</th>
<th></th>
<th>Turkic</th>
<th>Iranian</th>
<th>Slavic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OFS train</td>
<td>Test</td>
<td>OFS train</td>
<td>Test</td>
<td>OFS train</td>
<td>Test</td>
<td>OFS train</td>
</tr>
<tr>
<td>Maxent</td>
<td>800 00:00.9</td>
<td>00:00.0</td>
<td>700 00:00.5</td>
<td>00:00.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>700 00:00.3</td>
<td>00:00.0</td>
<td>800 00:00.2</td>
<td>00:00.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2000 00:01.1</td>
<td>00:00.1</td>
<td>700 00:00.3</td>
<td>00:00.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 29. Training and testing times for LI in LG for the Arabian script

<table>
<thead>
<tr>
<th>Method</th>
<th>Turkic</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OFS train</td>
<td>Test</td>
<td>OFS train</td>
</tr>
<tr>
<td>Maxent</td>
<td>1000 00:03.6</td>
<td>00:00.1</td>
<td>900 00:00.3</td>
<td>00:00.0</td>
</tr>
<tr>
<td>NB</td>
<td>2000 00:01.0</td>
<td>00:00.1</td>
<td>300 00:00.1</td>
<td>00:00.0</td>
</tr>
<tr>
<td>SVM</td>
<td>3000 00:03.1</td>
<td>00:00.2</td>
<td>900 00:00.3</td>
<td>00:00.0</td>
</tr>
</tbody>
</table>

### Table 30. Training and testing times for LI in LG for the Cyrillic script

<table>
<thead>
<tr>
<th>Method</th>
<th>OFS train</th>
<th>test</th>
<th>OFS train</th>
<th>test</th>
<th>OFS train</th>
<th>test</th>
<th>OFS train</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Germanic</td>
<td>Italic</td>
<td>MP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maxent</td>
<td>2000 00:14.3</td>
<td>00:00.2</td>
<td>2000 00:04.1</td>
<td>00:00.1</td>
<td>2000 00:00.6</td>
<td>00:00.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>2000 00:02.0</td>
<td>00:00.3</td>
<td>2000 00:01.1</td>
<td>00:00.1</td>
<td>2000 00:01.1</td>
<td>00:00.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>3000 00:05.7</td>
<td>00:00.3</td>
<td>2000 00:01.8</td>
<td>00:00.1</td>
<td>3000 00:01.0</td>
<td>00:00.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 31. Training and testing times for LI in LG for the Latin script

<table>
<thead>
<tr>
<th>Method</th>
<th>OFS train</th>
<th>test</th>
<th>OFS train</th>
<th>test</th>
<th>OFS train</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slavic</td>
<td>Turk</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maxent</td>
<td>2000 00:01.6</td>
<td>00:00.1</td>
<td>1000 00:00.8</td>
<td>00:00.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>2000 00:00.9</td>
<td>00:00.1</td>
<td>2000 00:00.7</td>
<td>00:00.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2000 00:01.1</td>
<td>00:00.1</td>
<td>2000 00:01.2</td>
<td>00:00.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 32. Training and testing times for LI in the three LI tests

<table>
<thead>
<tr>
<th>LI Method</th>
<th>LI in LG train</th>
<th>LI in LG test</th>
<th>LI in script train</th>
<th>LI in script test</th>
<th>LI in the entire TLI train</th>
<th>LI in the entire TLI test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxent</td>
<td>00:52.5</td>
<td>00:01.1</td>
<td>02:12.0</td>
<td>00:01.7</td>
<td>05:25.2</td>
<td>00:02.1</td>
</tr>
<tr>
<td>NB</td>
<td>00:30.7</td>
<td>00:01.4</td>
<td>00:15.6</td>
<td>00:01.9</td>
<td>00:24.4</td>
<td>00:02.6</td>
</tr>
<tr>
<td>SVM</td>
<td>00:39.7</td>
<td>00:01.6</td>
<td>00:53.6</td>
<td>00:02.6</td>
<td>01:15.2</td>
<td>00:02.8</td>
</tr>
</tbody>
</table>

5. Conclusions

Languages are written in different scripts, each of which has its own range in Unicode. Languages in the same language group are similar in their vocabularies and structure. We used these facts to propose a hierarchical short text LI system. When identifying a sentence as belonging to a language, our method identifies its script, the group of languages with the same script, and then the language itself within the language group. Experimental results showed that our method significantly improves LI accuracy, requires shorter training and testing times, and needs a smaller feature size to achieve optimal accuracy than traditional LI systems.

Some highly similar languages were in the same language group. Confusion errors often occurred among them, and reduced their identification accuracy compared to other languages in the same group. Other methods to improve the identification accuracy for highly similar languages need to be studied. Some noisy features were also found in both LGI and LI stages. In future work, we will examine ways to remove noisy features and select the most efficient ones.

Acknowledgment

This work was partially supported by the National Program on Key Basic Research Projects of China (2014CB340506) and the Key Project of the National Natural Science Foundation of China (61331011).

References


[29] https://en.wikipedia.org/wiki/Arabic_script


