Language Identification of Code Switching Sentences and Multilingual Sentences of Under-Resourced Languages by Using Multi Structural Word Information

Yin-Lai Yeong, Tien-Ping Tan

School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

Abstract

Language identification (LID) is a process to identify the languages used in a text or speech. Code switching is the switching of a language in a sentence or speech utterance. This paper focuses on LID of words in code switching sentences. Code switching can occur intersentential or intrasentential. The reasons why a writer switches from one language to another due to various reasons and among them are the inability to express opinion in a particular target language, to attract attention, to address different audience, habitual expressions and so on. The difficulty in identifying the languages of each word in a code switching sentence is because the languages have the same character set. In addition, code switching can happen in a sentence as short as a word or as long as a sentence. In this paper, we propose an automatic LID for words in code switching sentences by using multi structural word information (MUSWI) such as grapheme, syllable and word structure and calculate by using n-gram statistical model. The proposed MUSWI approach achieves 96.36% in term of accuracy on the code switching sentences, 99.07% on the multilingual sentences (non-code switching) which are under-resourced and closely related languages.

Index Term: language identification, code switching, n-gram, multilingual, under-resourced languages, closely related languages

1. Introduction

Language identification (LID) is an approach to identify the languages used in a text or a speech. LID is important in many areas of natural language processing. The requirements for LID in different systems may vary. The LID system in an automatic speech recognition system determines the language from the speech spectrum and their context. On the other hand, speech synthesis and machine translation system determine the language from the text itself. The LID system in a speech synthesis system requires that the language of a text is known. For code switching text, the requirement is even higher, where every word needs to be known the languages before it can be synthesized with the right pronunciation rules. LID for code switching text is challenging because the languages use the same character set such as English, Malay, Iban and Indonesian. Besides sharing the same character set, a lot of these languages have many similar morphological structures. In addition, code switching can happen in a sentence as short as a word or as long as a few words.

A speech synthesis system requires that the origin of a word be known. The knowledge of the language of a word is required due to the different pronunciation rules and phoneme sets in different languages. For example, the word “air” in Malay is pronounced as /a e/ or /a e t/, but in English is pronounce as /e r/. However, for a machine translation system, it is not necessary to determine the language of all the words. For example in Malay-English translation, there are no changes in the proper nouns when translating from English to Malay for example “Amerika Syarikat” in Malay instead of America, “Filipina” in Malay instead of Philippines. Our work focuses on LID for sentences in speech synthesis system, especially on code switching text.

Nevertheless, our proposed approach can also be used on documents with a single language. Most of the works in LID attempt to identify only a language in a given sentence or document [1][2][3][4][5][6]. However, when people are multilingual, they are likely to use different languages in their conversation. Thus, some of these approaches may not work for multilingual sentences or documents.

2. Code Switching

A multilingual person may communicate in more than one language. Code switching is the alternate use of two or more languages on a single communicative episode. Code switching can occur in both formal and informal communication. Informal communication, code switching may occur in formal writing, official speeches, news reading or teaching. However, code switching is more commonly used in informal communication compared to formal one such as during chatting among friends, short message service (SMS) and conversation among family members.

Most researchers distinguish two types of code switching: intrasentential and intersentential. Intrasentential code switching is the switching to a word in another language within a sentence, while intersentential code switching the switching to a sentence in another language [7]. Figure 1 illustrates two examples of Malay-English code switching sentences where intrasentential sentence is extracting from the local newspaper Berita Harian Malaysia.

<table>
<thead>
<tr>
<th>1. Intersentential sentence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Tolong simpan buku ini (Malay). Thank you (English).”</td>
</tr>
<tr>
<td>Meaning of sentences: “Please keep this book. Thank you.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Intrasentential sentence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Pembangun permainan (Malay) video popular Candy Crush (English).”</td>
</tr>
<tr>
<td>Meaning of sentence: “Popular video game developer Candy Crush.”</td>
</tr>
</tbody>
</table>

Figure 1: Malay-English code switching sentences.
Malaysia is a multilingual society. Malaysians tend to use more than one language in writing or speaking for example English and Malay, English and Mandarin, Malay and Hokkien, etc. In writing, our estimate from our Utusan Malaysia news corpus shows that on average there are about 15% code switching sentences in the text. Code switching often occurs in writing when there are proper nouns or scientific terms. Writers or readers are more comfortable to use the original terms in English. The amount of code switching in Malay chat or conversation is even higher, with an average of more than 20%.

The most common reason is to overcome the inability to express one’s opinion in the target language. It can also be a social practice where code switching is used to show the social position of the group. Since English is an international language, ability to converse in English in addition to the mother tongue gives a sense of connotation that the person is educated. Thus this gives the incentive for people in Malaysia to often code switch in their daily conversation or writing.

3. Related Works in LID

One of the examples was using n-grams or the statistics of short words frequency to determine the language of the words and word n-gram has shown to be useful in LID task [1][2][3]. Other works in LID [8][9][10] have been reported in order to solve the limitation of existing LID tools such as a very short text, how to correctly handle texts of unknown language and texts comprising of multiple languages especially from the Internet by using LID methods other than n-gram.

The language of an unknown word can also be predicted by calculating the probability of the alphabet sequence for that word in different languages [11]. Ranaivo-Malacon and Ng used alphabet n-gram to identify close languages and we can also model the orthography of the words using alphabet n-gram, and then calculate the probability of an unknown word in different languages. The n-gram model that gives the highest alphabet sequence probability will be selected.

\[
\max(P(a_1, a_2, a_3 \ldots a_n | F_A_L))
\]

where \(F_A\) is the n-gram alphabet model for language L.

4. Proposed Approaches: Multi Structural Word Information (MUSWI)

In this section, we propose an approach that uses multi-structural word information such as grapheme, syllable and word to identify the language in a text. The proposed approach was also tested on multilingual text (non-code switching), which is under resourced. Before the language for each word is identified, a sentence will be segmented into word, syllable and grapheme unit. Table 1 shows an example of Malay-English code switching sentence, and the way the sentence is segmented to word, syllable and grapheme unit.

4.1. LID Using Word Information

In this step, for identifying the language of a sentence, we compute the probability of the word sequence in the sentence. If the language identified is Malay, Iban or those that are likely to have code switching, we will then use MUSWI to identify the language for each word in next processing step. To predict the most probable language, we calculate the probability of the word sequence given the n-gram model of a language. The n-gram model that gives the highest word sequence probability will be selected. The disadvantage of this approach however is it does not work for out of vocabulary (OOV) words.

4.2. LID Using Grapheme Information

To predict the language for each word in a sentence using the grapheme information is similar to the case of using the word information as explained in the previous section. The only difference is that the unit is grapheme.

4.3. LID Using Syllable Structure Information

Before a word is converted to its syllable structure, the word is first converted to grapheme sequence. The grapheme sequence is then merged to become syllables. The largest syllable that can be formed is determined from right to left. The syllables formed from right to left. For example the word “lakukan” after segmented to grapheme is “l.a.k.u.k.a.n”. The syllable “kan” is formed, followed by “ku” and “la”. To predict the language of a word using the syllable information is similar to the case of using grapheme information explained in the earlier section. The only difference is that the unit is syllable.

4.4. Formula for All Approaches

Formula to predict the most probable language by calculated the probability of the word, grapheme and syllable sequence given the n-gram model of a language. The n-gram model that gives the highest word, grapheme and syllable sequence probability will be selected.

\[
\max(P(X_1, X_2, \ldots X_n | F_X_L))
\]

where \(F_X\) is the n-gram word, grapheme and syllable model for language L.

4.5. Interpolation: LID Using Grapheme, Syllable and Word Information

We have discussed three different approaches to predict the language of a word. These approaches use different units to model a word at different levels. To combine the estimation power for these approaches, we propose to use interpolation. The interpolation is carried out using Equation (3) as follows:

\[
\hat{L} = e_1P(G|\Phi_{G,L}) + e_2P(S|\Phi_{S,L}) + e_3P(W|\Phi_{W,L})
\]

where \(e_i\) is the interpolation weight. \(0 \leq e_i \leq 1\) and \(e_1 + e_2 + e_3 = 1\). G is the grapheme sequence of the word and S is the syllable sequence of the word and W is the word sequence of the word. The interpolation weight can be estimated by using a development set. Development set refers to a small amount of training and testing data where the language had been identified. After that, all the combinations of interpolation
weight were tested and a few higher combinations of interpolation weight in percentage of accuracy were selected and used in testing set. Interpolation weight is the weight to combine three set of approaches (grapheme n-gram approach, syllable n-gram approach and word n-gram approach).

5. Experiment & Discussion

In this section, we describe the methodology used to evaluate our proposed LID approach against the existing baseline approaches based on the accuracy to correctly identify the languages of words in a sentence. In testing, we tested the approaches on code switching languages and multilingual sentences. The first test was on Malay-English code switching sentences, followed by traditional LID test, which was performed on sentences either in Malay, Iban or Indonesian.

5.1. Malay-English Code Switching Sentences

We tested the LID approaches using different size of training data from Malay and English, while the size of testing data remains constant. The reason we test the approaches using different amount of data is because we want to know the performance of these approaches when the amount of data and out of vocabulary (OV) words vary. Malay training data was collected from the Malay daily news and English training data was collected from the English Wikipedia database. Then, we trained the models for alphabet n-gram, grapheme n-gram, syllable n-gram and word n-gram by using SRI language model toolkit [13]. We carried out the testing using unigram, bigram and trigram models.

In this test, 900 Malay-English code switching sentences were randomly collected from the daily news of both languages. The testing sentences were unique, and there were no repeated sentences. Code switching sentences were tested using alphabet n-gram approach, grapheme n-gram approach, syllable n-gram approach and word n-gram approach by using different sizes of training data which are “Small” contains 3.5MB (in a range of 15k to 25k sentences) of training data, “Medium” contains 7.0MB (in a range of 30k to 50k sentences) of training data, “Big” contains 43.0MB (in a range of 15k to 25k sentences) of training data and “Large” contains 230MB (more than 700k sentences) of training data. The testing data is verified word by word.

Table 2. Malay-English: The highest accuracy (in %) for each approach in different size of training data.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>No. words</th>
<th>Small</th>
<th>Medium</th>
<th>Big</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet N-gram</td>
<td>1-5</td>
<td>91.97</td>
<td>92.42</td>
<td>92.32</td>
<td>92.07</td>
</tr>
<tr>
<td>Grapheme N-gram</td>
<td>1-5</td>
<td>92.60</td>
<td>92.80</td>
<td>92.61</td>
<td>92.60</td>
</tr>
<tr>
<td>Syllable N-gram</td>
<td>1-5</td>
<td>93.72</td>
<td>94.19</td>
<td>95.23</td>
<td>95.28</td>
</tr>
<tr>
<td>Word N-gram</td>
<td>1-5</td>
<td>90.80</td>
<td>92.65</td>
<td>94.99</td>
<td>95.71</td>
</tr>
<tr>
<td>MUSWI</td>
<td>1-5</td>
<td>95.17</td>
<td>95.55</td>
<td>96.10</td>
<td>96.36</td>
</tr>
</tbody>
</table>

In Table 2, 96.36% was higher than those obtained by using a single approach where the highest value achieved was only 95.71%. Using more training data does not improve the result but instead cause the accuracy to drop due to the limited information provided by alphabet n-gram approach and grapheme n-gram that Malay and English are using the same alphabet. Syllable n-gram approach performs better than the other approaches when the amounts of data are small, medium or big. It is higher than word n-gram by 2.92%, 1.54% and 0.24%, respectively. These results show that the syllable n-gram approach can be used in under-resourced languages when the training data is limited. In addition, combination of three approaches (MUSWI) is able to obtain a better accuracy compared to a single approach. In terms of relative improvement, our proposed approach had increased 15.15% compared with the highest value by using single approach.

5.2. Malay/Indonesia/Iban Sentences (Non-Code Switching Sentences)

Two sets of training data were prepared for testing the multilingual sentences, where we constraint the amount of data used for training. The first set (Test 1) contains approximately 1MB of training data and the second set (Test 2) contains approximately 3MB of training data. Training data for Malay and Indonesian were collected from the Wikipedia database, while the training data for Iban were from the Iban daily news which is Borneo News. For testing, sentences that consist of 5 different lengths were used. Each length category consists of 100 sentences for each language from the same source, and the sentences are unique. Table 3 shows the highest accuracy for each approach in different size of length for Test 1 and Test 2.

Table 3. Malay-English: The highest accuracy (in %) for each approach in different size of length.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>No. words</th>
<th>1-5</th>
<th>6-10</th>
<th>11-15</th>
<th>16-20</th>
<th>21-25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet N-gram</td>
<td>1-5</td>
<td>70.73</td>
<td>79.27</td>
<td>83.20</td>
<td>86.40</td>
<td>89.47</td>
</tr>
<tr>
<td>Grapheme N-gram</td>
<td>1-5</td>
<td>72.33</td>
<td>81.27</td>
<td>84.93</td>
<td>87.13</td>
<td>92.80</td>
</tr>
<tr>
<td>Syllable N-gram</td>
<td>1-5</td>
<td>77.73</td>
<td>87.67</td>
<td>95.07</td>
<td>96.27</td>
<td>98.13</td>
</tr>
<tr>
<td>Word N-gram</td>
<td>1-5</td>
<td>71.00</td>
<td>85.20</td>
<td>94.20</td>
<td>97.07</td>
<td>98.27</td>
</tr>
<tr>
<td>MUSWI</td>
<td>1-5</td>
<td>79.73</td>
<td>90.07</td>
<td>96.00</td>
<td>97.40</td>
<td>98.80</td>
</tr>
</tbody>
</table>

From Table 3, the LID results obtained using MUSWI approach is better than all other approaches for both Test 1 (98.80%) and Test 2 (99.07%) which interpolates grapheme trigram with syllable trigram and word bigram. Trigram models for alphabet n-gram, grapheme n-gram and syllable n-gram give a better result. This is expected because of the limited amount of data for training. For word n-gram, bigram is better. The reason is because the small amount of training data is not sufficient to build a robust word trigram model. Besides MUSWI, results for syllable n-gram are more close to MUSWI for short test sentences. However, for long sentences, word n-gram performs better.

6. Conclusion & Future Work

The proposed MUSWI approach achieves 96.36% in term of accuracy on the code switching sentences and 99.07% on the monolingual sentences. For future work, we are planning to test it on more languages, for example Mandarin and French.

7. Acknowledgements

This project is supported by Research University Grant 1001/PKOMP/817068 from Universiti Sains Malaysia.
8. References


