Automatically Detecting and Correcting Input Mode Switching Errors by N-gram Statistics

Seung-Shik Kang, Kookmin University
Contents

1. Input-mode switching bet’n English and Korean
2. Statistical data for Korean and English
   – N-gram(English) and syllable data(Korean)
3. Automatic conversion algorithm with
   – Data structures
   – Error-prone dictionary for exceptional cases
4. Experiments
5. Conclusion
Input-mode Switching

- Bilingual writing system in Korean text
  - Mixture of two character sets in Korean text
    - English alphabet & Korean character → EUC-KR
  - Frequent switching of input-mode
  - Users forget switching the input-mode

- Problem solving methodology
  1. User-friendly switching key: E-K key, shift-space, etc
  2. Change the shape of a cursor or mouse pointer
  3. Automatic switching of input-mode & automatic conversion of input string
Input-mode Switching Errors

- Example 1. Korean string for English

여새ㅡㅁ샻 챔ㅜㅍㄷㄱ냐ㅐㅜ 둔스드
automatic conversion system

- Example 2. English string for Korean

gksdud wkehd wjsghks tltmxpa
한영 자동 전환 시스템
Related Works

- Yoo(1995)
  - detecting English word in Korean mode by scanning 3 letters of a word
  - converted a string into English that violate the syllable construction rules of the Korean language

- Lee(2001)
  - Word-phrase mode algorithm
    - right-to-left manner at the end of the word
  - Predictive mode algorithm
    - left-to-right manner: real-time applications
Types of words in Korean text

- Types of input string
  - a word of Korean syllables only
  - a word of English alphabets only
  - a word of Korean-English mixed string
    - ‘English는’ → English _subject-marker
    - ‘study했다’ → study_past
    - ‘한국IBM은’ → Korea_IBM_subject-marker

- Types of input string
  Type 1. mono-lingual string of $K^+$
  Type 2. mono-lingual string of $E^+$
  Type 3. bilingual string of $E^+K^+$
  Type 4. bilingual string of $K^+E^+$
  Type 5. bilingual string of $K^+E^+K^+$
N-gram Statistics for English

- Trigram/bigram probabilities: dictionary of 68,000 words

- Ex) ‘study’, ‘student’, ‘enter’

WB: {st}, {st}, {en}
WI: {tu, ud}, {tu, ud, de, en}, {nt, te}
WE: {dy}, {nt}, {er}

Freq(‘st’): <2, 0, 0>
Freq(‘tu’): <0, 2, 0>
Freq(‘nt’): <0, 1, 1>

\[ p_e(x) = \frac{\text{freq}(x)}{\sum_j \text{freq}(x_j)} \]
Syllable Statistics for Korean

- Probability for Korean syllable

\[ p_k(s_i) = \frac{freq(s_i)}{\sum_j freq(s_j)} \]

- Raw corpus: 12M words
Data Structure

- char engStr[MAXCHAR]; // English String

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>t</td>
<td>u</td>
<td>d</td>
<td>y</td>
<td>‘\0’</td>
<td></td>
</tr>
</tbody>
</table>

- char korStr[MAXSYLL]; // Korean String

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>ㅏ</td>
<td>서</td>
<td>오</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- int ekMap[MAXCHAR]; // mapping engStr[] → korStr[]

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- char ekMark[MAXCHAR]; // MARK_K, MARK_k, MARK_E, MARK_e
Initialization

Algorithm initialize() {
    int i, j=0;

    if (input_mode == Korean)
        korStr ← input string;
    else engStr ← input string;
    set_ekmap(korStr, engStr, ekMap);
    // initialize korStr or engStr
    // ekMap[]: E–K mapping index
    while (i < strlen(engstr)) {
        i = get_index_syl_start(j);
        j = get_index_syl_end(i);
        syl = get_one_syl(korStr, i, j);
        if (freq(syl) > 0)
            ekmark[i~j] = MARK_k;
        else ekmark[i~j] = MARK_e;
    }
}
Automatic Conversion

• Type of input string by *ekMark[]*
  
  Type 1. A sequence of ‘e’ or ‘E’
    → violate syllable construction rules
  
  Type 2. A sequence of ‘k’
    → need a recalculation of probabilities
  
  Type 3. A combination of ‘e’ and ‘k’
    → need a recalculation of probabilities
Determination of E-K Substring

- Korean probability of a substring

\[ P_K = \sum_i \ln p_k(s_i) \]

- English probability of a substring

\[ P_E = \sum_j \ln p_e(t_j) \]

- Normalization

\[ P_E = P_E \times \frac{n_{\text{syllable}}}{n_{\text{trigram}}} \]
Error-prone Dictionary

- Substrings that frequently cause E-K ambiguity

- Error-prone dictionary
  
  \(<\text{substring}, \text{K/E/I}>\)
  
  - K: ‘substring’ is Korean
  - E: ‘substring’ is English
  - I: depends on the input-mode that user changed
# Experiments

- Brown corpus → English words only
- Newspaper → Korean words only
- Textbook → Korean words with English words
- Mixed words only → E-K mixed words only

<table>
<thead>
<tr>
<th>Test data</th>
<th>No. of words</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown corpus (English)</td>
<td>1,012,930</td>
<td>99.40</td>
</tr>
<tr>
<td>Newspaper (Korean)</td>
<td>98,764</td>
<td>99.93</td>
</tr>
<tr>
<td>Textbook (mixed)</td>
<td>6,274</td>
<td>98.86</td>
</tr>
<tr>
<td>Mixed words only</td>
<td>14,135</td>
<td>98.46</td>
</tr>
</tbody>
</table>
Concluding Remarks

- Statistical method to detecting/correcting input-mode switching errors
- English trigram/bigram & Korean syllable frequency
- Accuracy
  - over 99% for monolingual data
  - about 98.5% E-K mixed data