Improving Hangul to English translation for optical character recognition (OCR)

Emily Park, Jennifer Claudio

Abstract

Real-time translation of languages using camera inputs occasionally results in awkward failures. As a proposed method of assisting such tools for Korean (Hangul) to English translation, an optical assessment method was proposed to help translation algorithms first assess whether the Korean text has been written as English syllables in Korean or in true Korean vocabulary before producing translated phrases. Although the current approach was not viable, future work will implement feedback regarding methods to meaningfully handle the optical data received.

1 Introduction

1.1 Korean alphabetic syllabary

Hangul, the writing system of the Republic of Korea, currently uses an alphabet constituent of 14 consonants and 10 vowels. The Hangul alphabet is described as an alphabetic syllabary, meaning that although alphabet units consist of vowels and consonants working together to depict a sound, letter and syllable combinations have both a vertical and horizontal relationship. This relationship is in contrast to a language such as English, where each alphabetic letter has only a horizontal relationship with the ones that precede or follow it. In Korean, sets of syllables thus produce words which, to a non-Korean speaker, must be converted into semantic units.

Hangul has changed immensely over its history, including historic concern about class differences in the original Korean writing systems to inclusion of the modernized systems. Even more recently, the spread of *Konglish*, words derived from English but used in a Korean context, pose new issues and face new criticisms. Firstly, the linguistic divide between North and South Korea is further emphasized by the divergence of word choice or usage, and secondly transliterations require contextual relevance, as otherwise a homograph may be substituted by the reader. When read by a human, this context can easily be picked up through visual cues such as images associated with the text or with contiguous lines of text, however, an Optical Character Recognition (OCR) reader may only grasp sequential words and less context, hence leading to mistranslation of words.

1.2 Language conversion and accessibility

The relevant forms of language conversion for this situation are transliteration and translation. Transliteration provides a syllabic conversion using characters of another alphabet, whereas translation provides the meaning of a word in a different language.

To couple linguistic and physical accessibility, Optical Character Recognition (OCR) is widely used for recognizing text from scanned documents and converting them to editable data. One method of language-relevant OCR is through Google Translate, though many other programs and platforms exist. As many individuals who have used an OCR language conversion on a food menu may know, awkward translations can occur due to insufficient context for the reader or due to literal translations and loss of figurative speech. Furthermore, the adoption of Konglish presents a problem where a language learner or a speaker who has had less exposure to English may not recognize a word that is actually English that has been transliterated into Korean.

2 Goals

The goal of this project was to create a predictive method for text conversion as an alternative or supplement to sole reliance on counting database references. In doing so, such a predictive method would improve results for Korean language learners and older or more traditional speakers.

Currently, language translators rely on a database of known words, and many include common transliterations and borrowed words. The functionality of any OCR-based language translator therefore depends on the size and integrity of its associated database. While some translators have mentioned AI implementation based on word associations or probability calculations of word linkage, this aspect is beyond the scope of this project.

3 Methods

Fifty common words in English and Korean were both transliterated phonetically and translated across languages to assess preliminary data and feasibility testing. A script was written using Python to determine pixel area represented by the text and the area of its bounding box as determined by the outermost edges of a word's letters. Image samples were taken from different print media, specifically from newspapers, children's books, and advertisements. Each image was fed into the software five times to test reliability, then the ratios of text space to background space were tabulated and calculated. This was done to determine if the ratio retrieved from an OCR could inform the translator as to whether the word was true Korean or transliterated English.

The program imported the *cv2* (a.k.a. OpenCV) Python package for image recognition. The code first

filtered the image into a pure black and white image. Subsequently, the height and width were calculated by detecting the number of pixels that comprised the word. The code looped over every pixel and finally printed the total number of black pixels in the filtered image. The gathered results of black pixels over the area of the bordered word fell between the ratios of 0.3 and 0.5.

Initially, the code used for bordering the text first recognized the words in the image and traced around the detected word to define the border. This was attempted with *pytesseract* (Python Tesseract), which was found unfeasible for recognizing words; this was thus adjusted to use the Google platform to recognize and translate their images.

Upon refining the script to find a bounding edge and verifying that the difference in area between foreground text and background could be determined, word samples were collected and processed.

4 Results and discussion

Although the code generated was able to perform calculations of text versus background, the ratios of foreground to background were not statistically different from each other across transliterated and translated words. This is attributed to the limited number of characters that comprise Hangul, of which fewer than five basic shapes (vertical sticks, horizontal sticks, circles, boxes, and huts). This shape limitation restricts the number of symbols that could be formed. An alternative method of informing a translation platform would be to assess the number of strokes in a word and the number of words that use a given consonant sound. The number of strokes could be viable because many transliterated words result in three syllables, despite being a single syllabic word when pronounced in native Korean.

A secondary issue with the input method included the necessary conversion into black and white. This text then needed to be manually uploaded into the translator, rather than performing in real time in tandem with the OCR itself. This method consequently defeats the purpose of pairing with an OCR.

As expected, font, typeface, and stylizations affected ratios, however, this was determined not to be a contributing factor to the inability to create predictive translation.

5 Conclusions

In the manner approached, using text to background ratios is *not* a viable method of implementing a predictive algorithm without context. Current AI methods used by translators exhibit fairly consistent performance.

6 Acknowledgments

Special thanks to the TEX Users Group and its associated community for their support during our attendance at the annual conference. Additional thanks to Govind and Ganesh Pimpale for their support with generating the Python code for OCR use.

◇ Emily Park, Jennifer Claudio
Oak Grove High School
Science Research Program
San Jose, CA