OCR and Automated Translation for the Navigation of non-English Handsets: A Feasibility Study with Arabic

Jennifer Biggs and Michael Broughton
Defence Science and Technology Organisation
Edinburgh, South Australia
{firstname.lastname}@dsto.defence.gov.au

Abstract

In forensics, mobile phones or handsets store potentially valuable information such as *Contact* lists, *SMS Messages*, or possibly *emails* and *Calendar* appointments. However, navigating to this content on non-English configured handsets, when the operator is untrained in the language, becomes a difficult task. We discuss a feasibility study that explored the performance of optical character recognition (OCR) systems against Arabic menus on handset LCD screens. Further, a method of automated spell correction and translation is explored considering fully automated or user-interactive workflow options. A capability technology demonstrator for non-English handset navigation was implemented based on outcomes of these studies, providing a platform for investigating workflow and usability.

1 Introduction

Some cellular exploitation tools support imaging of the handset display after the operator has navigated the handset menus to the content of interest. Such tools may support any handset type. However, navigating to this content on handsets configured for languages other than English (LOTE) is challenging for operators not trained in the language.

We undertook several feasibility studies to investigate the navigation of LOTE handsets for CELLEX purposes. The studies investigated the merits of: 1) applying Commercial-Off-The-Shelf (COTS) Optical Character Recognition (OCR) tools to photographed displays of handsets; and 2) combining LOTE OCR outputs with a method of automated translation.

1.1 OCR accuracy

COTS OCR systems are typically optimised for recognition of text at resolutions in excess of 100 dots per inch (dpi), such as scans of printed documents, newspapers or magazines, advertising accuracy rates of up to 97%. Batawi and Abulnaja (2012) report accuracy rates of between 94.8% and 97.8% for a selection of printed newspaper and magazine Arabic texts. Recognition of non-degraded printed pages may still require identification of optimal image pre-processing options (Yale University, 2008a; 2008b). Recognition accuracy for degraded documents may be expected to be significantly decreased (Herceg et al. 2005).

To utilise a COTS OCR application within a larger system architecture or workflow, where images do not meet application parameters, additional image pre-processing can be applied. Chang et al. (2007) and Chang et al. (2009) used Sakhr Automatic Reader v8.0 on photographed images of text.

1.2 Automated translation of OCR outputs

When applying Machine Translation (MT) processing to OCR output text, OCR errors are compounded. Chang et al. (2009) combined Sakhr Automatic Reader v8.0 with a statistical MT system known as PanDoRA (Zhang and Vogel, 2007), noting that word errors introduced by OCR were amplified during MT processing. For example, in translation of generated images of text from the Basic Travel Expression Corpus (BTEC) (Eck & Hori, 2005), the BLEU score of image text translation without errors was 43.12, while a 10.7% word recognition error rate severely drops the BLEU score to 28.56 (Chang et al., 2007; Chang et al. 2009).

Jennifer Biggs and Michael Broughton. 2014. OCR and Automated Translation for the Navigation of non-English Handsets: A Feasibility Study with Arabic. In *Proceedings of Australasian Language Technology Association Workshop*, pages 108−112.

2 Evaluating OCR Accuracy

The aim of this first study was to determine the feasibility of recognising Arabic text within photographed images of monochrome LCD handset displays by utilising COTS OCR applications.

A late 2003 model handset; a Nokia 2300 lv99, was selected for its backlit monochrome display of 96 x 65 pixels, with an ability to display 4 lines of text in either English and Arabic user interface languages. Image capture was performed using Samsung L200 10.2MP digital camera on a stand fixing orientation and distance with default camera settings. Two COTS OCR systems were selected for recognition of Arabic script. Each COTS system supports a range of either automated or manually determined image pre-processing and recognition settings and either automated or manual text area identification. The COTS systems will be referred to as COTS 1 and COTS 2 only.

2.1 Method

To match photographed images with image parameters expected by the COTS OCR systems, image pre-processing was performed. Images were manually cropped to the handset display area and scaled using cubic interpolation such that text heights were between supported font sizes of 6 – 20 pixels. Observation of binarised images produced by importing the cropped and scaled images into the COTS applications showed significant character erosion and background speckling. Therefore images were manually binarised using a colour threshold function in a raster graphics editor.

For the purposes of the study, manual zoning omitted screen formatting areas such as images or horizontal or vertical bars. In the case of automated zoning, OCR output lines were manually aligned with reference text lines and additional lines from non-text areas were omitted. However, additional OCR outputs from non-text symbols along the same y-axis from a ground truth text area were included.

A number of image pre-processing and recognition settings were applied per COTS OCR system in each of the Arabic and English image text recognition tasks. Accuracy was measured by line, word and character using edit distance.

A test corpus of 259 handset display images (118 Arabic and 141 English) was produced by photographing the Nokia 2300 handset during navigation of menus in both English and Arabic user interface language settings. Four font sizes were observed in the Nokia 2300 display in both English and Arabic interface languages.

Ground truth text for each image was generated containing 407 lines of Arabic and 474 lines of English in one of four font sizes.

2.2 Results

Accuracy results for each COTS system at selected levels of automation are given for Arabic and English in Table 1. Settings used for recognition of English are shown in the shaded rows. Character, word and line accuracy for recognition of English was significantly higher than equivalent settings for recognition of Arabic, except for COTS 2 where automatic settings were applied. In this case, the system output only Arabic script.

The optimal settings for the COTS 1 system provided significantly greater word and line recognition accuracy than COTS 2, although character recognition accuracy was not proportionally higher. This effect was caused by comparative distribution of recognition errors; COTS 1 system errors were clustered in groups more often than those of COTS 2.

COTS system		Character	Word	Line
1	COTS1-A4-1	75.3	43.8	34.1
	COTS1 E1-1	91.8	81.9	78.1
	COTS1-A4 Autosettings	74.4	43.2	32.9
	COTS1 E1 Autosettings	90.7	81.1	77.5
	COTS1-Autozone A1	57.5	23.7	19.2
	COTS Autozone E1	75.4	47	41.9
	COTS1 Autozone Autosettings	45.6	10.6	1.7
2	COTS2 A2-3	70.5	23.7	12.3
	COTS2 E1-3	85.6	75	74
	COTS2 Auto settings Arabic	63	11.7	7.1
	COTS2 Auto settings English	1.5	0.3	0
	COTS2 Auto zone A2	33.5	11.8	2.4
	COTS2 Autozone Autosettings	30.1	6.9	1.7

Table 1: Recognition accuracy of Arabic and English script for increasing levels of automation

3 Translation of OCR outputs

The aim of this second study was to determine the feasibility of applying automated translation to OCR output text recognised from photographed images of a Nokia 2300 handset menu LCD display. Additionally, the study aimed to identify appropriate methods for correction of recognition errors within OCR output text prior to automated translation.

The study utilised automated translation via bilingual dictionary lookup, and compares two methods for error correction of the OCR output text where an exact match is not found in the bilingual dictionary. Each error correction method generates a list of candidate matches, and is measured as fully automated, or with user-interactive selection of a correct match from the candidate list.

3.1 Method

Optimal recognition outputs as described in section 2 from the 118 images of the Arabic portion of the Nokia 2300 handset image corpus for each COTS OCR system were used.

Error correction was performed on each line of OCR output text in each of two sets of 118 text files. Error correction used spell checking based on Levenshtein string distance (or edit distance) to measure text against the spell checking dictionary. Two approaches to error correction were utilised: firstly each OCR recognition line was not tokenised, and secondly whitespace based tokenisation was performed to obtain unigram tokens from each OCR recognition line. The spell checking dictionary contained both tokenised and un-tokenised forms from the Nokia 2300 ground truth text corpus.

By comparing the original image and spell corrected text within an application interface, a user may be able to select the correct text from within spell correction options. Therefore, line accuracy was measured based on two error correction and automated translation workflow options. Firstly, accuracy of the top ranked spell checking match was measured. Secondly, line accuracy was measured where the correct recognition term was found within the top five ranked matches during spell correction.

The first error correction method tokenised each line of OCR output text, and completed word-based automated translation via the bilingual dictionary. The second error correction method used a phrase-based lookup approach based on un-tokenised OCR output lines. Error correction is completed using word n-gram segments of handset menu phrases modelled on the word-wrapped lines in handset displays.

The terminology contained within the Nokia 2300 ground truth text corpus was used as the basis for spelling correction dictionary data. Individual words from each of the n-gram phrases were added, and all menu phrases and words were translated.

A deployed application would typically be required to provide general coverage for a variety of handset makes and models. Therefore, a simulated larger corpus was developed using 1,500 terms between 1 - 4 words in length selected from an Arabic glossary of application menu terminology. The first 375 terms of each length within the glossary that did not appear in the Nokia 2300 ground truth text corpus were used. Word n-grams of length 1 – 4 were selected to simulate OCR recognition lines of word wrapped menu phrases on handset displays with varying width and display resolutions. A final corpus size of 1,665 unique n-gram expressions resulted.

3.2 Results

Line accuracy is reported for both n-gram untokenised and tokenised error correction methods. For both spelling correction methods, line accuracy is reported for user interactive and automated error correction. Automated error correction occurs without user interaction where only the top ranked spell checking match is used. User interactive error correction occurs where the correct term exists within the top five ranked spell checking matches.

Figure 1 shows the un-tokenised n-gram OCR recognition line method provided greater line accuracy than tokenised methods for outputs for both COTS systems outputs, regardless of user interaction. User interaction provided line accuracy increases from 85.9% to 91.1% for COTS 1 and from 80.3% to 86.5% for the un-tokenised method, and from 73% to 84.3% for COTS 1 and from 39.3% to 41.7% for COTS 2 for the tokenised method.

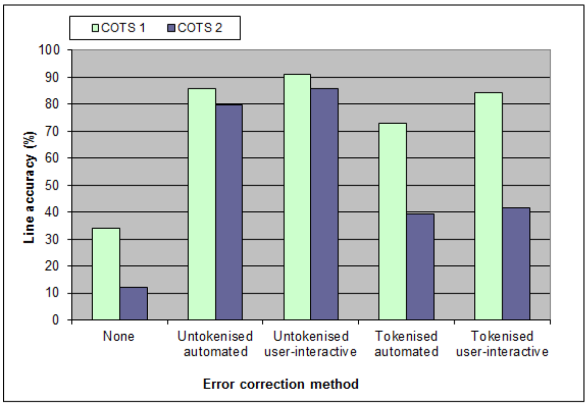

Figure 1: Line accuracy using word tokenised and line based un-tokenised error correction methods

Figure 2 illustrates the overlap between correct lines the two COTS systems following untokenised user-interactive error correction.

81.3% of the recognition zones were correct from both applications, while an additional 9.8% were correct from only COTS 1 recognition outputs and 5.2% were correct from only COTS 2 outputs. 3.6% of recognition zones were not correct by either application.

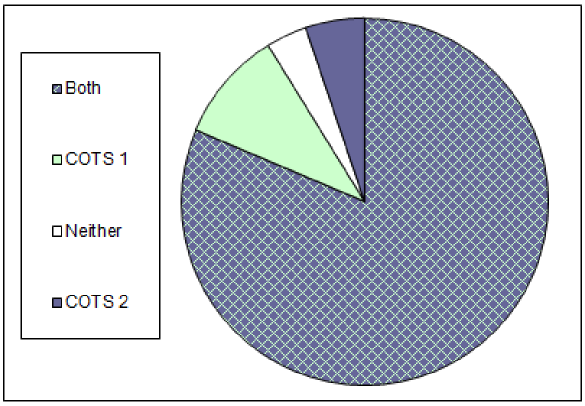

Figure 2: Lines correct following un-tokenised user-interactive error correction

4 Nokia 2300 handset study replication

A demonstrator application was developed, implementing functionalities required to complete all image OCR and translation steps from the studies. The objective in replicating previous studies was to confirm similar results could be achieved using the application and handset menu phrases rather than a simulated corpus based on software application menu phrases. A corpus of an additional 815 Arabic handset menu phrases was collected from user manuals of four handset models and compiled into a corpus suitable for spell checker and bilingual lookup dictionaries as per the method used to create the simulated corpus. This corpus was then used over the 118 Arabic images from the Nokia 2300 image corpus, using manual zone definition and the optimal application settings of the COTS 1 system. Line accuracy of 89.7% was achieved. This was comparable to the optimal line accuracy of 91.1% achieved in previously using a simulated corpus.

5 Discussion

Outcomes from these feasibility studies identified three areas in the workflow as critical to optimising OCR accuracy and overall performance; these were: 1) image interpolation and binarisation transformations; 2) delimitation of text areas in the handset display; and 3) user-interactive error correction. By using error correction on OCR lines, word segmentation errors may be eliminated and n-grams introduced in string distance based error correction.

Evaluation of the OCR and translation workflow considered only the case of a low resolution monochromatic LCD handset display in Arabic. Based on this work, recommendations could be made to improve both overall accuracy and use cases. Performance over a range of handset models, LCD display types, and recognition language should be quantified. Further OCR systems and/or customisation of OCR systems for recognition of specific handset fonts could be evaluated. A multi-OCR engine approach, such as described by Batawi and Abulnaja (2012), could also be considered.

User interactive error correction provided better outcomes than automated error correction for a given error correction approach. As no OCR system can provide 100% accuracy, text verification will be required by comparing recognised script to text in the original image, regardless of whether interactive error correction is completed. Therefore the additional time to complete user interactive error correction at LOTE text processing stages may not be considered prohibitive as the verification task is completed concurrently. However, text verification will present a challenge for those unfamiliar with the writing script, and observations from the use of the demonstrator application indicate that for such users verification is further complicated when the OCR output font differs from the image font (source).

6 Conclusion

Currently, best solutions for mobile device forensics will be either direct data extraction by a COTS solution that supports the given handset, or navigation of LOTE handset menus by a trained linguist. When these options are not available, the described studies and software implementation demonstrated a feasible workflow for navigating non-English handset menu structures by personnel untrained in the language. An outdated handset was selected due to the difference in properties of the font displayed in the low resolution monochrome LED screen to a typical COTS OCR system recognition task. Applying the technique to more current smartphones remains of interest but will also pose additional challenges.

References

Batawi, Yusof A. and Abulnaja, Osama A. (2012) Accuracy Evaluation of Arabic Optical Character Recognition Voting Technique: Experimental Study. *IJECS: International Journal of Electrical & Computer Sciences.* **12** (1) 29-33. ISSN: 2077-1231

Chang, Y., Chen, D., Zhang, Y., Yang, J. (2009) An image-based automatic Arabic translation system. In *Pattern Recognition* **42** (2009) 2127 – 2134.

Chang, Y., Zhang, Y., Vogel, S., Yang, J. (2007) Enhancing Image-based Arabic Document Translation Using a Noisy Channel Correction Model. In: In *Proceedings of MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark.

Eck, M. and Hori, C. (2005) Overview of the IWSLT 2005 evaluation campaign. In: *Proceedings of International Workshop on Spoken Language Translation*, 11-17, Lisbon, Portugal

Herceg, P., Huyck, B.,Van Guilder, L., Kundu, A. (2005). Optimizing OCR Accuracy for Bi-tonal, Noisy Scans of Degraded Arabic Documents. *Visual Information Processing XIV*, edited by Zia-ur Rahman, Robert A. Schowengerdt, *Proceedings of SPIE*, Vol. 5817. pp. 179 Bellingham, WA.

Yale University (2008a) *AMEEL Digitization Manual: Part 9, OCR of Arabic Text with Sakhr*. Updated 2008 [Accessed 15 June 2012] Available from: http://www.library.yale.edu/idp/documentos/OCR_Sakhr.pdf

Yale University (2008b) *AMEEL Digitization Manual: Part 10, OCR of Arabic Text with Verus*. Updated 2008 [Accessed 15 June 2012] Available from: http://www.library.yale.edu/idp/documentos/OCR_Verus.pdf

Zhang, Y., Vogel, S. (2007) PanDoRA: A Large-scale Two-way Statistical Machine Translation System for Hand-held Devices. In: *Proceedings of MT Summit XI*, 10-14 September 2007, Copenhagen, Denmark.