

A Game-based Approach to Transcribing Images of Text

Khalil Dahab, Anja Belz

School of Computing, Mathematical and Information Sciences
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
kad12@brighton.ac.uk, asb@brighton.ac.uk

Abstract

We present a methodology that takes as input scanned documents of typed or hand-written text, and produces transcriptions of the text as output. Instead of using OCR technology, the methodology is game-based and produces such transcriptions as a by-product. The approach is intended particularly for languages for which language technology and resources are scarce and reliable OCR technology may not exist. It can be used in place of OCR for transcribing individual documents, or to create corpora of paired images and transcriptions required to train OCR tools. We present Minefield, a prototype implementation of the approach which is currently collecting Arabic transcriptions.

1. Introduction

Creating language resources is expensive and time-consuming, and this forms a bottleneck in the development of language technology, for less-studied non-European languages in particular.

The recent internet phenomenon of crowd-sourcing offers a cost-effective and potentially fast way of overcoming such language resource acquisition bottlenecks. The form of crowd-sourcing pioneered by von Ahn (2006) and dubbed by him ‘Games with a Purpose’ (GWAP) in particular has captured the imagination of HLT researchers, with a number of research projects on this topic currently under way, including AnaWiki (Poesio et al., 2008) and On-toGame (part of the FP7 Incentives project) (Siorpaes and Hepp, 2008).

In a GWAP, the idea is that players simply play the game for those reasons that make people spend vast amounts of time playing games,¹ including entertainment effect and competitive elements. It is only as a by-product that new resources that are useful, e.g. to researchers, are created; in von Ahn’s first GWAP, for example, the resource was a database of labelled images.

The work we present in this paper has the twofold aims to provide a tool that can, as is, (i) substitute for OCR technology in the case of languages for which reliable OCR technology does not exist, and, over a period of time, (ii) be used to create the language resources necessary to create OCR technology for less-resourced languages. We are currently focusing on creating Arabic language resources; Arabic has been identified by the NEMLAR and MEDAR projects as particularly in need of language resource development, with resources for OCR technology among the most urgent (Maegaard et al., 2009).

In the following section, we describe the basic infrastructure that underlies our approach: segmenting an input image file into lines and line segments, and matching typed

transcriptions to the line segments. Section 3. describes how the transcriptions themselves are obtained in a game where two players team up to find a way through a minefield, and Section 4. briefly describes our current prototype implementation.

2. Transcription Method

In this section we describe the transcription process independently of its implementation as a game. The input to this process is a set of scanned or other photographic images of text, either typed or hand-written. Each image is interpreted as a page and segmented into lines (where our current algorithm scans the bitmap image to locate rows of blank pixels which are taken to be line separators). Each line is then segmented into fixed-width segments (where the width is a variable parameter). Figure 1 shows a line of Arabic text segmented in this fashion.

Transcriptions are then obtained (in a game-context, as described in the following section) for pairs of adjacent segments. Each segment pair overlaps the next segment pair by one segment, so that transcriptions are obtained for segment pairs $[S_1S_2]$, $[S_2S_3]$, $[S_3S_4]$, ..., and so on. The overlaps are intended to counteract problems arising from segment boundaries cutting through words and letters, which may make transcription of letters adjacent to a boundary difficult. Transcribing overlapping segments ensures that each letter is seen in its entirety at least once.

A transcription for a given segment is not accepted until a certain number and proportion of transcribers agree on it. The level of agreement is a variable parameter that can be set to achieve a higher or lower likelihood of outputs being correct, as desired. This parameter is useful for controlling the cost of transcriptions.

Once transcriptions at the desired likelihood of accuracy have been obtained for all segments in a document, they are reassembled in the correct order to form the output of the transcription process. In order to do this, the maximal overlaps between the transcriptions of segment pairs are determined as shown in Figure 2. The matching procedure finds the largest overlap between the suffix of transcription

¹According to a recent report by the Entertainment Software Association, every day more than 200 million hours are spent playing computer and video games, see http://www.theesa.com/facts/gamer_data.php.



Figure 1: Example of a segmented line.

$T_{i,i+1}$ of segment pair $[S_i, S_{i+1}]$ and the prefix of transcription $T_{i+1,i+2}$ of segment pair $[S_{i+1}, S_{i+2}]$.

3. Game Design

Rather than paying transcribers to provide transcriptions for segment pairs, these are produced as a by-product of playing a game called Minefield.

Minefield is a two-player game where the scenario is that one player, Player A, is stuck in a minefield and does not know where the mines are. The aim is for Player A to find her way out of the minefield, but taking a step in the wrong direction may result in stepping on a mine. At each turn, Player A finds encrypted messages that will, once decrypted, tell her which way is safe to go. However, Player A does not have the means for decoding the messages, and must instead pass them to Player B, whose character is in headquarters remote from the minefield and has a code-cracking machine that can translate the encrypted messages into directions such as *right*, *left*, *forward*.

The encrypted messages that Player A sees in the minefield are in fact segment pairs (as described in the preceding section). The only way Player A can pass the encrypted message to Player B in headquarters is by copying (transcribing) them on the keyboard. Player B sees a copy of the encrypted message on their screen, along with two other ones. The copy of the encrypted message has a safe direction to take written under it, the other two have directions written under them that are potentially unsafe. If Player A types a correct transcription, this will enable Player B to identify the correct message and pass back to Player A the associated safe direction.

Figure 3 shows example screen shots of what Player A sees when passing the encrypted message to Player B (on the left), and what Player B sees when choosing the decryption (on the right). At each turn, the encrypted messages are selected randomly from the current pool of segment pairs to be transcribed.

3.1. Player Motivation

In order for a game to be successful it needs to be perceived as fun, but also as challenging. Challenge is recognised as a fundamental characteristic of successful games (Sweetser and Wyeth, 2005). We have incorporated a range of features to keep players motivated. A time limit can be set within which a mine has to be passed. Pairs of players are awarded points every time they successfully avoid a mine, and Minefield keeps track of the total point tally. Tables of top scores and player ranks show points collected by each player over different periods of time.

Something we are planning to incorporate in the future is additional financial incentivisation where top players are

periodically rewarded with vouchers.

3.2. Ensuring Accuracy

Minefield employs a number of strategies to ensure output accuracy. The first is the basic design of the game which has been called an inversion-problem game design (von Ahn and Dabbish, 2008): instead of agreeing on outputs (as is the case in many other GWAPs), one player has to enable the other other to identify what the original input was. In Minefield, the only way this can be achieved is by correctly typing the encrypted message out on the keyboard.²

A standard GWAP strategy we use is initially letting players play only with segment pairs for which the correct transcription is known, and testing their performance. Such checks are repeated periodically at later stages to ensure player accuracy levels are not dropping. Furthermore, players are paired up at random, so are unlikely to know who they are playing against, hence collusion is difficult.

4. Implementation of Minefield

A prototype version of the design described in the previous section has been implemented and can be viewed at <http://www.minefield-project.info>. The following resources and tools were used to create it: SQL Server Express 2005 for the Database Server; SubSonic 3.0 and Linq for the Data Access Layer; ASP.NET, C# and .NET Web Services for the Game Logic Layer; ASP.NET Controls, HTML/CSS and jQuery for the Presentation Layer; and Flash SWF/ActionScript for the Game Client. A Flash template from TemplateMonster (<http://www.templatemonster.com>) was used as a starting point for interface design.

While the design and implementation are language-independent, the prototype implementation currently has Arabic documents for transcription. Input of Arabic characters from a Latin keyboard is possible, as can be seen on the left of Figure 3.

Using the Minefield tool for different languages would require very few changes in the tool itself. The line segmentation algorithm would have to be adapted to account for the orientation of lines (horizontal vs. vertical) and the direction of text (left to right vs. right to left). Other than that, the tool is not sensitive to different languages, as the input is an image rather than text.

²Players could not simply type the first part of a segment, because they cannot see what the alternative text snippets are that their partner is looking at.

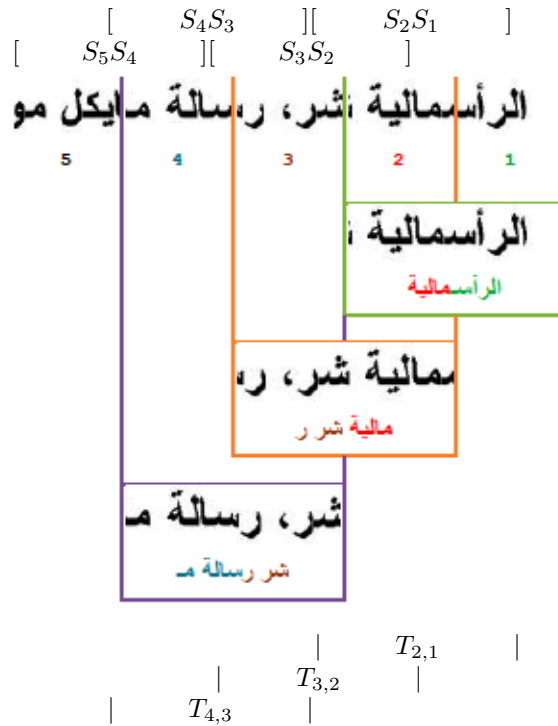


Figure 2: Illustration of how transcriptions of segment pairs are aligned to form output transcription.



Figure 3: Minefield screenshots: what Player A sees when passing the encrypted message to Player B (on the left), and what Player B sees when choosing the decryption (on the right).

5. Related Research

The approach most similar to the core idea in the research presented in this paper is the Recaptcha methodology.³ In Recaptcha, words that were not accurately recognised by OCR technology are used as Captchas⁴ in websites' verification procedures. This approach is currently being used by Google to digitise the New York Times archives.

However, Recaptcha currently deals only with English text. If other languages were to be added, a mechanism would be required for allocating Captchas to websites in the correct language. Furthermore, the approach requires a large number of websites in the given language to ex-

ist already which moreover have to be willing to purchase Captcha verification services. This is not likely to be the case for less-resourced languages, and consequently a transcription project with Recaptcha is not viable for such languages.

In contrast, a game-based approach as in Minefield does not require a large number of websites in a given language to exist already. Minefield makes it possible to target native speakers of a given language more directly, either through a small number of existing websites, or outside of the Web.

6. Concluding Remarks

In this paper we have presented a game-based approach to transcribing scanned or other photographic images of text (hand-written or typed). The method can be used to obtain transcriptions of individual documents or to create large-scale resources for training OCR tools. Our method was

³Originally developed by von Ahn and others at Carnegie Mellon University, now owned by Google (<http://recaptcha.net/learnmore.html>).

⁴Captchas normally are distorted images of words which a user has to type out before accessing certain content on a website.

created particularly with less-resourced languages such as Arabic in mind.

We have implemented a prototype version of Minefield, currently collecting transcriptions of Arabic documents. We have not yet implemented a fully automatic line segmentation procedure that can cope with mixed-content pages (this is left for future work). We are currently completing the testing and user evaluation phase and are planning to deploy the system for large-scale language resource creation and one-off document transcription services in the near future.

7. References

- Bente Maegaard, Mustafa Yaseen, Steven Krauwer, and Khalid Choukri. 2009. Cooperation roadmap. Technical report, The MEDAR Project. <http://www.medar.info/MEDAR-roadmap.pdf>.
- Massimo Poesio, Udo Kruschwitz, and Jon Chamberlain. 2008. ANAWIKI: Creating anaphorically annotated resources through web cooperation. In Nicoletta Calzolari et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.
- Katharina Siorpaes and M. Hepp. 2008. Games with a purpose for the semantic web. *IEEE intelligent systems*.
- Penelope Sweetser and Peta Wyeth. 2005. Gameflow: a model for evaluating player enjoyment in games. *Comput. Entertain.*, 3(3):3–3.
- Luis von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.
- Luis von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.