Literature Survey

-Ishita Chaudhary
-Abdul Mannan
-Jason Carvalho

The focus of our project extends to two domains, Text Recognition from an Image, commonly known as OCR, and Automatic Question Generation.

Pratik Madhukar Manwatkar ; Shashank H. Yadav [1] have created a Text-Recognition System consisting of four modules.The first and most important part of this is segmentation, which is to extract characters. This is done by using the Line detection and Character detection algorithm. The text recognition part is done using feature extraction and classification is done here by using the Kohonen neural network.

One of the most efficient and highly developed OCR's available is Tesseract OCR, which was initially developed at HP in the 1980's, but sponsored by Google since 2006. In 'An Overview of the Tesseract OCR Engine' [2], describes the unique model used which first stored the outlines of components of the connected-component labelling. The outlines were then collected into blobs, and assigning the filtered blobs into lines. The model could also detect curved baselines. Where it finds fixed pitch text, Tesseract chops the words into characters using the pitch, and disables the chopper and associator on fixed-pitch words for the word recognition step. It bought forward the idea that the features in the unknown need not be the same as the features in the training data.

B.Gatos [3] describes a three step method for recognising text taken from historical documents, without prior information about the language. It consists of a pre-processing stage where documents are converted into binary images, a segmentation technique that extracts the characters, the creation of a database by the extracted characters and a recognition stage where the database is used for converting any document into text file.
They have combined several note-worthy approaches from the field of OCR such as binarization and segmentation and more general pattern recognition techniques such as image enhancement and k-clustering. For the typed data, the overall recognition rate, calculated by finding Levenshtein distance was 83.66%.

Unlike previous approaches to text-detection, TextBoxes++ [4] detects text by directly predicting word bounding boxes with quadrilaterals via a single neural network that is end-to-end trainable. Here they combine TextBoxes++ with CRNN, a text recognition module. The approach of using recognition results to further refine the detection results due to the semantic-level awareness of recognised text hasn't been done before. TextBoxes++ outperforms competing methods in terms of text localisation accuracy and runtime.

The Second Part of our project involves Automatic-Question-Generation (AQG). Automatic question generation has come out as a promising area of research in the field of Natural Language Processing (NLP) due to its wide applications in Educational Technology. Two major parts of this process is Sentence Extraction and Keyword Selection.

TextRank [5] is a graph-based ranking model mainly utilised for keyword and sentence extraction. This model basically decides the importance of a vertex in a graph based on a 'voting' or 'recommendation' system. For sentence extraction, this vertex is a sentence. A text unit recommends other related text units, and the strength of the recommendation is recursively computed based on the importance of the units making the recommendation. If a sentence has similarities (calculated by using the number of common tokens of the sentences) to another sentence, it gets upvoted by that sentence. Hence, the sentences are ranked in the order of their votes. Evaluation using the rouge method and based on guidelines by DUC evaluators showed that TextRank succeeded in identifying the most important sentences. The element that makes TextRank unique is that it is unsupervised, and only depends on the provided text to produce the key sentences.

An interesting technique developed was to generate fill-in-the-blank questions on informative sentences in the text data by Das and Majumder [6]. The method identifies and selects these sentences based on Part-of-Speech tags and other rules. It does not follow a template based method. They also generate hints to nudge the user in the direction of the answer. The questions which were answered by identifying domain specific words from the text (an ontology approach) were removed from the review. The accuracy of system for identifying informative, factual sentences was 92.367%.

'Ontology-Based Multiple Choice Question Generation', Alsubait [7] explains that the most difficult part of MCQ generation is generating distractors. They describe an ontology based approach to generate distractors. They have developed a 'Similarity Measurer', inspired by Jaccard's similarity coefficient, to generate questions of varying difficulty by varying the similarity between the key and distractors. This also ensures that the distractors created belong to the domain of the generated question. Two ontologies have been used for testing, and a question was deemed too difficult for students if it was answered correctly by less than 30 % of the students and was too easy if answered by more than 90 % of the students. External reviewers also evaluated if the generated questions were useful or not, with 92% and 96% declared as useful by at least one reviewer of each of the two ontologies used.

Manish Agarwal [8] presented an automatic, rule based open-cloze question generation (OCQG) system of generating questions of varying difficulties. This approach consisted of selecting informative sentences and identifying keywords in these sentences. The dataset used was News reports on Cricket matches. 22 questions (10+12) were generated and evaluated by three different evaluators. The overall accuracy of the system was 3.15 (Eval-1), 3.14 (Eval-2) and 3.26 (Eval-3) out of 4.

[9]This paper uses a mix of syntax and semantic based approach to generate wh- type questions from both simple and complex sentences. It makes use of the Stanford Tagger for POS tagging. It then uses NER over the sentence for identifying the Subject, Object, etc. Questions on complex sentences are determined by the discourse connective(conjunctions). It also extracts the auxiliary verb (if present) to help figure out the tense of the sentence, which is important in proper question generation.

Rapid Automatic Keyword Extraction (RAKE)[9] is an unsupervised method for extracting keywords from text data. Unlike most of the work done in this area, RAKE is domain-independent. RAKE uses stop words and phrase delimiters to partition the document text into candidate keywords. Word Co-occurrence is identified within these candidate keywords. After every candidate keyword is identified and the graph of word co-occurrences is made, a score is calculated for each candidate keyword and defined as the sum of its member word scores. The degree and frequency of word vertices in the graph are used to determine the word scores. This model selects one third of the top scoring words in the graph as keywords. RAKE effectively extracts keywords and outperforms the current state of the art models in terms of precision, efficiency, and simplicity. As the number of words in the document increases, RAKE outperforms TextRank. Comparison between them on the Inspec testing set showed it was about 6 times faster than TextRank.

Citations

[1] P. M. Manwatkar and S. H. Yadav, "Text recognition from images," 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2015, pp. 1-6, doi: 10.1109/ICIIECS.2015.7193210.

[2] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.

[3] G. Vamvakas, B. Gatos, N. Stamatopoulos and S. J. Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents," 2008 The Eighth IAPR International Workshop on Document Analysis Systems, Nara, 2008, pp. 525-532, doi: 10.1109/DAS.2008.73.

[4]M. Liao, B. Shi and X. Bai, "TextBoxes++: A Single-Shot Oriented Scene Text Detector," in IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3676-3690, Aug. 2018, doi: 10.1109/TIP.2018.2825107.

[5]Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[6] Das, B., Majumder, M. Factual open cloze question generation for assessment of learner's knowledge. *Int J Educ Technol High Educ* **14,** 24 (2017).

[7] Alsubait, T., Parsia, B. & Sattler, U. Ontology-Based Multiple Choice Question Generation. *Künstl Intell* **30,** 183–188 (2016)

[8] Agarwal, M (2012). Cloze and open cloze question generation systems and their evaluation guidelines. Master's thesis, International Institute of Information Technology, Hyderabad.

[9]Automatic Question Generation from Paragraph Dhaval Swali,Jay Palan, Ishita Shah-International Journal of Advance Engineering and Research Development Volume 3, Issue 12, December -2016

[9]Rose, Stuart & Engel, Dave & Cramer, Nick & Cowley, Wendy. (2010). Automatic Keyword Extraction from Individual Documents. 10.1002/9780470689646.ch1.