Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Σύγκριση τεχνολογιών εξατομικευμένης αναγνώρισης βιογραφικών και βιομετρικών χαρακτηριστικών κατά την ταυτοποίηση** **Technologies review for biographic and biometrics recognition in identification process**
Ονοματεπώνυμο Φοιτητή	**Ιωάννης Καλομοίρης**
Πατρώνυμο	**Γεώργιος**
Αριθμός Μητρώου	**ΜΠΠΛ/ 17017**
Επιβλέπων	**Ευάγγελος Σακκόπουλος, Επίκουρος Καθηγητής**

Δεκέμβριος 2021

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Ευθύμιος Αλέπης Αναπληρωτής Καθηγητής	Διονύσιος Σωτηρόπουλος Επίκουρος Καθηγητής	Ευάγγελος Σακκόπουλος Επίκουρος Καθηγητής

Σύνοψη

Η παρούσα διπλωματική εργασία λαμβάνει χώρα σε μια εποχή που η τεχνίτη νοημοσύνη κερδίζει όλο και περισσότερο έδαφος παγκοσμίως. Η παραγωγή κώδικά και εφαρμογών πλησιάζει όλο και περισσότερο σε αυτό που μπορούμε να αναφέρουμε ως «μηχανές που έχουν την δυνατότητα να μαθαίνουν».

Αυτός είναι και ο λόγος για τον οποίο πολλές από αυτές τις εφαρμογές μπορούν να παίζουν ρόλο σε βασικές λειτουργίες τις ανθρώπινης κοινωνίας υποβοηθώντας τομείς της κοινωνικής ζωής.

Η παρούσα διπλωματική εργασία (ΔΕ) φιλοδοξεί να ξεδιπλώσει ένα μικρό μέρος του τομέα της τεχνητής νοημοσύνης και αυτό το μέρος είναι οι τεχνολογίες λογισμικού γύρω από την **Αναγνώριση λεκτικών χαρακτήρων (OCR)** αλλά και την ταυτοποίηση κατά την είσοδο ενός χρήστη σε μια εφαρμογή**.**

Στα πλαίσια της ΔΕ έχει αναπτυχθεί μια εφαρμογή η οποία χρησιμοποίει μια βιβλιοθήκη OCR, το TESSERACT. Στην συνέχεια έχει αναπτυχθεί μια αντίστοιχη που χρησιμοποιεί παρόμοιες τεχνολογίες OCR αλλά αυτή την φορά εδράζεται πάνω στην ήδη τελειοποιημένη σουίτα της εταιρίας Face Tec. Και οι δύο εφαρμογές επιδιώκουν να εκτελέσουν της ίδια διαδικασία, να αναγνωρίσουν ένα τυπικό δημόσιο έγγραφο ταυτοποίησης (κοινή ταυτότητα) και να αποσπάσουν κάποια δεδομένα του ιδιοκτήτη της, με λίγα λόγια να ψηφιοποιήσουν με το πάτημα ενός κουμπιού τα στοιχεία ενός πολίτη. Όμως πριν την παραπάνω περιγραφή προηγείται η παρουσίαση δυο ερευνών σχετικά με την εκπαίδευση και την επεξεργασία των αποτελεσμάτων του OCR.

Στα επόμενα κεφάλαια θα παρουσιαστούν οι δύο έρευνες σαν βάση θεωρίας, στην συνέχεια θα παρουσιαστεί η πρώτη εφαρμογή και οι τεχνολογίες της βιβλιοθήκης Tesseract, μετά θα παρουσιαστούν τα βασικά στοιχεία των τεχνολογιών που εφαρμόζοντα στην σουίτα της Face Tec και η τοπικά εγκατεστημένη εφαρμογή της, σειρά λαμβάνει η σύγκριση των αποτελεσμάτων των δύο εφαρμογών και βεβαίως τα συμπεράσματα που απορρέων από αυτά τα αποτελέσματα. Ιδιαίτερή σημασία όπως είναι φυσικό θα δοθεί στην ακρίβεια της εκτέλεσης της αναγνώρισης λεκτικών χαρακτήρων από κάθε μια εφαρμογή

Abstract

The current thesis project takes place at a time when AI is gaining more and more ground worldwide. The production of code and applications is getting closer to what we can refer to as “machines that have the ability to learn”.

Therefore, many of these applications can play an important role in fundamental functions of human society by assisting areas of social life.

This thesis aims to expand a part of artificial intelligence and this part is the software technologies around **optical recognition (OCR)** but also its role to authentication/ authorization during login in an application.

An application that uses an OCR library, named Tesseract, has been developed. Then a similar one has been developed that uses similar OCR technologies but this time it is based on the already perfected suite of the company FaceTec. Both applications seek to perform the same process, to recognize a standard public identification document and to extract some data of its owner. However, the above description is preceded by the presentation of two studies on training and the processing of OCR results.

The following chapters will present the content of two research as a theory base, then the first application and the technologies of the Tesseract library will be presented, then the basic elements of technologies applied in the FaceTec suite and its locally installed application will be presented. The next step will be the comparison of the results of the two apps and the conclusions that result from these. Importance will naturally be given to the accuracy of the executions results of each application.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Βιβλιογραφία

Κατάλογος Εικόνων:

Κεφάλαιο 1 ο

1 Εισαγωγή

Στην παρούσα μεταπτυχιακή διατριβή γίνεται μια προσπάθεια να καταγραφούν και να παρουσιαστούν πλευρές της τεχνολογία αναγνώρισης βιογραφικών και βιομετρικών χαρακτήρων (OCR). Δεδομένου ότι τα τελευταία χρόνια η εμφάνιση εφαρμογών που σχετίζονται άμεση ή έμμεσα, χρησιμοποιούν κατά κύριο ή μερικό τρόπο την τεχνολογία OCR έχουν πληθύνει σε μεγάλο βαθμό, η αναζήτηση των βέλτιστών δυνατών αποτελεσμάτων αυτής της τεχνολογία ίσως βρίσκεται στο επίκεντρο πολλών ερευνών δείχνοντας ότι ο τομέας του OCR αλλά και της τεχνητής νοημοσύνης εν γένη είναι ένας ζωντανός οργανισμός του αναπτύσσεται.

Η τεχνολογία OCR αφορά την οπτική αναγνώριση τυπωμένου κειμένου από μια μηχανή. Σκεπτόμενοι την σημαντικότητα αυτής της πράξης, μπορούμε να αναλογιστούμε την αξία χρήσης αυτού του επιτεύγματος. Μια σειρά από κείμενα παγκοσμίως, δημόσια έγγραφα ή πνευματικά έργα μπορούν πλέον να ψηφιοποιηθούν (και ήδη συμβαίνει σε μεγάλο βαθμό) με την βοήθεια αυτής της τεχνολογίας. Το γεγονός αυτό χαρίζει νέες διαστάσεις στην δυνατότητα διανομής των έργων αυτών παγκοσμίως αλλά και στην αυτοματοποίηση εργασιών στην δημοσία διοίκηση, παραδείγματος χάρει, ας σκεφτούμε την ευκολία του ανοίγματος ενός λογαριασμού χωρίς την φυσική μας παρουσία σε μια τράπεζα, ή ακόμη πιο σημαντικό την ηλεκτρονική ταυτοποίηση προσωπικών στοιχείων από επίσημα ατομικά έγγραφα. Τα προηγούμενα παραδείγματα αποτελούν κάποιες από τις λίγες αλλά τρανταχτές εφαρμογές της τεχνολογίας OCR. Άλλες μπορεί κάποιος σε προγράμματα μεταφράσεων στα οποία εφαρμόζονται μαζί με την OCR διαδικασία και τεχνολογίες μηχανικής μάθησης.

Όμως, οι παραπάνω εργασίες απαιτούν ακρίβεια. Δηλαδή την αξιόπιστη εξαγωγή ενός εντύπου κείμενου σε ηλεκτρονική μορφή, όπως θα δούμε παρακάτω αυτό αποτελεί μια αρκετά επίπονή και πολύπλοκή εργασία, ενώ πολλές έρευνές εστιάζουν την προσοχή τους ακριβώς σε αυτόν τον τομέα: την περεταίρω επεξεργασία την εξαγόμενων αποτελεσμάτων του OCR προσδοκώντας να τελειοποιήσουμε ακόμα περισσότερο την διαδικασία.

Λαμβάνοντας υπόψιν τα παραπάνω, η παρούσα διατριβή εστιάζει την προσοχή σε σχετικές έρευνες γύρο από τα θέματα της βελτιστοποίησης αποτελεσμάτων και την σύγκριση δυο εφαρμογών που χρησιμοποιούν την τεχνολογία OCR. Η πρώτη δημιουργήθηκε για τις ανάγκες της παρούσας μεταπτυχιακής διατριβής και η δεύτερη είναι η σουίτα εφαρμογών της FaceTec. Και οι δυο εφαρμογές θα χρησιμοποιηθούν ώστε να αποσπάσουν κάποια πεδία από την ελληνική ταυτότητα. Ο λόγος της σύγκρισής μεταξύ των δύο εφαρμογών είναι ο εξής: Η εγκυρότητα των αποτελεσμάτων της επεξεργασίας OCR είναι μια διαδικασία η οποία απαιτεί μια σειρά από ρυθμίσεις ειδικά την αναγνώριση ενός εγγράφου όπως αυτό της ελληνικής ταυτότητας η οποία περιέχει αρκετό οπτικό θόρυβο. Η πρώτη εφαρμογή μπορεί να

χαρακτηριστεί ως ασταθής ως προς την επιτυχία της εξαγωγής αποτελεσμάτων ενώ η δεύτερη ελαχιστοποιεί τα τελικά σφάλματα σε μεγάλο βαθμό.

Στο κεφάλαιο 2 θα παρουσιαστούν δύο έρευνες σχετικά με την μετεπεξεργασία των δεδομένων του OCR, στο κεφάλαιο 3 παρουσιάζεται η πρώτη εφαρμογή OCR της διατριβής μαζί με μια ιστορική αναδρομή και περιγραφή του τρόπου λειτουργίας του Tesseract (βιβλιοθήκη OCR που χρησιμοποιείται από την εφαρμογή), στο κεφάλαιο 4 παρουσιάζεται η εφαρμογή της FaceTec μαζί με την βασική της αρχιτεκτονική, ενώ στο τελευταίο κεφάλαιο, το κεφάλαιο 5, καταγράφοντα κάποιες από τις βασικές παρατηρήσεις από την σύγκριση των δυο εφαρμογών.

Κεφάλαιο 2 ο

2 Σχετικές και προηγούμενες εργασίες

2.1 Γενικά

Σκοπός του κεφάλαιού αυτού είναι να παρουσιαστούν δύο εργασίες, σχετικά πρόσφατες, οι οποίες καταπιάνονται με την επεξεργασία τον αποτελεσμάτων μιας μηχανής οπτικής αναγνώρισης χαρακτήρων (OCR).

Η πρώτη με τίτλο «Έρευνα για τις προσεγγίσεις σχετικά με την προ-OCR διαδικασία» (Survey of Post-OCR Processing Approaches [1]) φιλοδοξεί να αποσαφηνίζει τη σημασία της βελτίωσης της ποιότητας των αποτελεσμάτων της οπτικής αναγνώρισης χαρακτήρων (OCR) μελετώντας τις επιπτώσεις τους στην ανάκτηση πληροφοριών και σε εφαρμογές επεξεργασίας φυσικής γλώσσας. Επιπροσθέτως, ορίζει το πρόβλημα επεξεργασίας των δεδομένων μετά το OCR, απεικονίζει την τυπική ροή της διαδικασίας του και καταγράφει τις πιο σύγχρονες προσεγγίσεις.

Η δεύτερη με τίτλο «Επισκόπηση των προσεγγίσεων βαθιάς μάθησης για την επεξεργασία OCR και την κατανόησης των εγγράφων» (A Survey of Deep Learning Approaches for OCR and Document Understanding [2]) φιλοδοξεί να επανεξετάσει διάφορες τεχνικές για την τεκμηρίωση σχετικά με τα έγγραφα που έχουν συνταχθεί στα αγγλικά και να ενοποιήσει μεθοδολογίες που υπάρχουν ήδη στην βιβλιογραφία θέλοντας να λειτουργήσει ως σημείο εκκίνησης για νέες έρευνες στον τομέα αυτόν.

2.2 Έρευνα για τις προσεγγίσεις σχετικά με την προ-OCR διαδικασία

2.2.1 Εισαγωγή

Δεδομένου ότι η ποσότητα των αναλογικών εγγράφων εξακολουθεί να είναι αρκετά μεγάλη παρά την πρόσφατη ευρεία μεταστροφή τους σε ψηφιακά έγγραφα, έχουν καταβληθεί σημαντικές προσπάθειες για τη μετατροπή των έντυπων υλικών σε ηλεκτρονικά κείμενα με σκοπό την επεξεργασία κειμένου από υπολογιστές, την καλύτερη διατήρηση και την ευκολότερη πρόσβαση. Η διαδικασία μετατροπής (δηλαδή ψηφιοποίηση) περιλαμβάνει την αποτελεσματική σάρωση ή φωτογράφιση εγγράφων, σελίδα προς σελίδα, και τη μετατροπή της εικόνας κάθε σελίδας σε κείμενα αναγνώσιμα από υπολογιστή. Η επιλογή των τεχνικών ψηφιοποίησης βασίζεται σε διάφορους παράγοντες, όπως το μέσο, το εκτυπωμένο αντί για το

χειρόγραφο κείμενο, η γλώσσα Κ.Ο.Κ. Η μετατροπή των ψηφιακών εικόνων σε ηλεκτρονικά κείμενα πραγματοποιείται συχνά δύο προσεγγίσεις: Μη αυτόματη εισαγωγή κειμένου, λογισμικό οπτικής αναγνώρισης χαρακτήρων (OCR) και ημι-αυτόματο.

Η μη αυτόματη πληκτρολόγηση είναι φυσικά πολύ ακριβή και ενδέχεται να σχετίζεται με ορισμένα ζητήματα ασφαλείας κατά την κοινή χρήση πληροφοριών με τρίτους. Η μέθοδος αυτή κοστίζει περίπου 1 ευρώ ανά σελίδα, καθιστώντας την τιμή ψηφιοποίησης ανά βιβλίο περίπου 400, 500 ή ακόμη και 1000 ευρώ. Το λογισμικό OCR είναι μια οικονομική εναλλακτική λύση για τη μη αυτόματη εισαγωγή κειμένου χωρίς σχετικά προβλήματα ασφαλείας. Αυτή η τεχνική προσφέρει καλές τιμές αναγνώρισης και έχει γίνει ένας από τους πιο δημοφιλείς και αποτελεσματικούς τρόπους για τη μετατροπή του εκτυπωμένου κειμένου. Εκτός από το λογισμικό μη αυτόματης εισαγωγής κειμένου και OCR, οι ημιαυτόματες προσεγγίσεις επιτρέπουν τη συνεργασία της μεταγραφής εγγράφων σε χαρτί σε ψηφιακά δεδομένα που στη συνέχεια χρησιμοποιούνται για την εκπαίδευση μοντέλων OCR για αυτόματη δημιουργία μεταγραφών. Ένα από τα εργαλεία μεταγραφής με τη βοήθεια υπολογιστή είναι το Transkribus.

Παρόλο που οι μηχανές OCR βελτιώνονται συνεχώς και μπορούν ήδη να λειτουργήσουν σωστά στο σύγχρονο κείμενο, εξακολουθούν να μην διαθέτουν επαρκή εκπαιδευτικά δεδομένα που αποτελούνται από προηγούμενα έγγραφα και δυσκολεύονται να πετύχουν υψηλή απόδοση σε ιστορικά κείμενα. Η φυσική ποιότητα των αρχικών υλικών, των περίπλοκων διατάξεων, των παλιών γραμματοσειρών προκαλούν σημαντικές δυσκολίες στο τρέχον λογισμικό OCR. Συνεπώς, τα παραγόμενα αποτελέσματα του OCR εξακολουθούν να έχουν θόρυβο και ενδεχομένως να επηρεάσουν τυχόν μεταγενέστερες εφαρμογές που χρησιμοποιούν αυτά τα υλικά κειμένου ως είσοδο.

Πολλές εργασίες έχουν μελετήσει την επίδραση των θορυβωδών εισροών στην **ανάκτηση πληροφοριών (IR)** και **στη γλωσσική επεξεργασία (NLP)**. Η έκθεση των van Strien et al επιβεβαιώνει ότι η χαμηλή ποιότητα του OCR κειμένου επηρεάζει αρνητικά την αναζήτηση πληροφοριών. Ο Chiron et al εκτιμά τον αντίκτυπο των σφαλμάτων οπτικής αναγνώρισης χαρακτήρων στην ψηφιακή βιβλιοθήκη της Gallinica που διαχειρίζεται η εθνική Βιβλιοθήκη της Γαλλίας. Υποδεικνύουν ότι το 7% των κοινών όρων αναζήτησης, για τους οποίους υποβάλλεται ερώτημα τουλάχιστον 35 φορές, ενδέχεται να επηρεάζεται από σφάλματα OCR. Οι επιδόσεις ανάκτησης πληροφοριών παραμένουν καλές για αρκετά υψηλά ποσοστά σφαλμάτων σε μεγάλα κείμενα, ωστόσο μειώνονται δραματικά σε σύντομα κείμενα. Τα αποτελέσματα των Bazzo et al δείχνει ότι οι αισθητές επιπτώσεις αρχίζουν με ποσοστό σφάλματος 5% ανεξάρτητα από το μήκος του κειμένου.

Όσον αφορά το NLP, διάφορες εφαρμογές, π.χ. **αναγνώριση οντοτήτων (NER)**, σήμανση μέρους της ομιλίας (POS), σύνοψη κειμένου, ανίχνευση ορίου προτάσεων, μοντελοποίηση θέματος, η ταξινόμηση κειμένου, η σύνδεση με κατονομασμένη οντότητα κ.ο.κ. επηρεάζονται σοβαρά από τα σφάλματα OCR. Η απόδοση των εργαλείων NER, τα οποία εντοπίζουν τα κατάλληλα ονόματα και τα κατηγοριοποιούν στο σύνολο προκαθορισμένων

κατηγοριών (δηλαδή, πρόσωπο, τοποθεσία, οργάνωση), υποβαθμίζεται σημαντικά μαζί με την αύξηση **του ποσοστού σφάλματος (ER)** της παραγωγής του OCR. Όταν το ποσοστό σφάλματος (WER) του κειμένου αυξάνεται από 0% σε 2.7%, η βαθμολογία F του εργαλείου NER μειώνεται περίπου κατά 3%. Με την υψηλότερη ER, η απόδοση του NER μειώνεται πολύ ταχύτερα, για παράδειγμα, από 90% σε 60% όταν ο WER του κειμένου αυξάνεται από 1% σε 7% ή όταν το ποσοστό σφάλματος χαρακτήρων (CER) αυξάνεται από 8% σε 20%.

Συνολικά, η ποιότητα των αποτελεσμάτων του OCR επηρεάζει ποικιλοτρόπως την ανάκτηση πληροφοριών, καθώς και τις εργασίες NLP. Οι επιδόσεις των εφαρμογών που σχεδιάζονται και εφαρμόζονται με την παραδοχή ότι τα καθαρά δεδομένα συνήθως υποβαθμίζουν τη ποιότητα των κειμένων που προκαλούν θόρυβο. Συνεπώς, είναι σημαντικό να παράγετε καθαρότερη έξοδος.

Έχουν ξεκινήσει πολλές προσπάθειες για την ψηφιοποίηση πολλών εγγράφων που συνιστούν ευρωπαϊκή πολιτιστική κληρονομιά. Ένα μεγάλο μέρος αυτής της κληρονομιάς πρέπει να μετατραπεί σε ψηφιακή μορφή. Επιπλέον, ένα μέρος των ιστορικών κειμένων έχει ήδη υποβληθεί σε επεξεργασία από διάφορους αλγόριθμους OCR, οι οποίοι έχουν χαμηλότερη απόδοση από τους τρέχοντες αλγόριθμους. Ως αποτέλεσμα, πολλές ψηφιοποιημένες συλλογές ιστορικού περιέχομένου εξακολουθούν να έχουν θόρυβο και η ανάπτυξη μηχανών OCR μαζί με τεχνικές μετεπεξεργασίας εξακολουθούν να είναι σε υψηλή ζήτηση. Οι ερευνητές δίνουν προτεραιότητα στην ανάλυση και τη βελτίωση των υφιστάμενων δεδομένων. Στην πραγματικότητα, η ενίσχυση των μοντέλων μετεπεξεργασίας και η ανάλυση των επιπτώσεων των σφαλμάτων οπτικής αναγνώρισης χαρακτήρων (OCR) στα μεταγενέστερα στάδια είναι η πρώτη σύσταση στην ατζέντα για την ψηφιοποίηση ιστορικών και πολυγλωσσικών κειμένων. Ωστόσο, προετοιμάζονται πρότυπα για την επισήμανση της αλήθειας, την αξιολόγηση κ.ο.κ. και καθορίζουν ποιες συλλογές πρέπει να περάσουν πάλι από επεξεργασία OCR. Το ενδιαφέρον της κοινότητας στον τομέα αυτό απεικονίζεται επίσης από τον αριθμό των καταχωρίσεων στους δύο πρόσφατους διαγωνισμούς για τη διόρθωση κειμένου μετά το OCR που διοργανώθηκαν σε συνδυασμό με **τη διεθνή διάσκεψη για την ανάλυση και την αναγνώριση εγγράφων** (ICDAR) το 2017 και το 2019. Η τελευταία έρευνα σχετικά με τις τεχνικές βελτίωσης των αποτελεσμάτων της οπτικής αναγνώρισης χαρακτήρων (OCR) δημοσιεύθηκε το 1997. Αργότερα, υπήρχαν ορισμένες εκθέσεις σχετικά με τις στρατηγικές μετεπεξεργασίας. Στη συνέχεια, είναι απαραίτητο να πραγματοποιηθεί μια νέα έρευνα του τομέα, η οποία θα παρέχει μια επισκόπηση των επίκαιρων προσεγγίσεων για μετά τη επεξεργασία κειμένου με την διενέργεια OCR, αυτός ακριβώς είναι αποτελεί ο κύριος σκοπός αυτού του έργου.

Η έρευνα αυτή μπορεί να είναι επωφελής για διάφορες ομάδες αναγνωστών: Οι μη ειδικοί μπορούν να λάβουν μια συνολική επισκόπηση των προσεγγίσεων επεξεργασίας μετά την OCR. Οι ερευνητές μπορούν να τη χρησιμοποιήσουν ως βάση συζήτησης ή ακόμη και να πάρουν ιδέες για μελλοντική έρευνα. Τέλος, οι προγραμματιστές μπορούν να συμβουλευτούν την έρευνα και να επιλέξουν την κατάλληλη τεχνική, το σύνολο δεδομένων αξιολόγησης ή το κιτ εργαλείων για τη δοκιμή και την ανάπτυξη των προϊόντων τους. Σε αυτό το έργο, κάνουμε τις ακόλουθες συνεισφορές:

1. Η σημασία της βελτίωσης της ποιότητας των αποτελεσμάτων OCR αποσαφηνίζεται με την επανεξέταση των επιπτώσεων του θορυβώδους OCR κειμένου στις μεταγενέστερες εργασίες. Καθορίζουμε το πρόβλημα επεξεργασίας μετά το OCR και καταδύεται η τυπική σειρά μεθόδων που στοχεύουν στη μετεπεξεργασία των αποτελεσμάτων OCR.
2. Στη συνέχεια, πραγματοποιούμε έρευνα σε διάφορες προσεγγίσεις μετεπεξεργασίας, τις κατηγοριοποιούμε σε συγκεκριμένες σύμφωνα με τον βαθμό εξάρτησης από τον άνθρωπο, την έκταση των πληροφοριών που χρησιμοποιούνται και τα εφαρμοζόμενα μοντέλα (π.χ. μοντέλο γλώσσας, μοντέλο αυτόματης μετάφρασης κ.λπ.), αναλύονται τα πλεονεκτήματά τους και τα μειονεκτήματά τους.
3. Δημοφιλείς μετρήσεις αξιολόγησης, περιγράφονται τα προσβάσιμα σύνολα δεδομένων. Επιπλέον, αναφέρονται αρκετές περιπτώσεις ανοικτού κώδικα που παρέχονται από προηγούμενες εργασίες επεξεργασίας μετά το OCR, γλωσσικούς πόρους και χρήσιμα εργαλείων επεξεργασίας.
4. Αναγνωρίζεται η τρέχουσα τάση αυτού του πεδίου και περιγράφονται οι οδηγίες έρευνας για την εργασία επεξεργασίας μετά το OCR.

2.2.2 Ορισμός προβλήματος επεξεργασίας μετά το OCR

Οι προσεγγίσεις μετεπεξεργασίας αποτελούνται από δύο εργασίες: *Ανίχνευση σφαλμάτων* και *διόρθωση σφαλμάτων*. Ο εντοπισμός σφαλμάτων Post-OCR επιτρέπει τον εντοπισμό λανθασμένων διακριτικών από το κείμενο εισαγωγής. Η ανίχνευση σφαλμάτων υποστηρίζει τον ανθρώπινο παράγοντα για τη γρήγορη διόρθωση σφαλμάτων, επιτρέπει επίσης την επισήμανση δεδομένων με θόρυβο για την επανεπεξεργασία τους, εάν είναι απαραίτητο. Επιπλέον, ο εντοπισμός σφαλμάτων δημιουργεί μια λίστα με τα σφάλματα που εντοπίστηκαν ως δεδομένα εισαγωγής της διόρθωσης σφαλμάτων μετά το OCR, η οποία προορίζεται για την επιδιόρθωση μη έγκυρων διακριτικών. Με το που δείνεται ένα σφάλμα, συνήθως, δημιουργείται και βαθμολογείται μια λίστα λέξεων με βάση τις διαθέσιμες πληροφορίες. Στη συνέχεια, επιλέγεται ο υποψήφιος με την υψηλότερη κατάταξη για διόρθωση του σφάλματος σε προσεγγίσεις αυτόματης διόρθωσης ή προτείνονται οι κορυφαίοι n υποψήφιοι στους χρήστες σε ημιαυτόματες προσεγγίσεις.

2.2.3 Τεχνικές επεξεργασίας μετά το OCR

Πρώτα περιγράφεται τη διαδικασία συλλογής σχετικών εγγράφων. Πραγματοποιήθηκε αναζήτηση σε έγγραφα που περιέχουν τις λέξεις “OCR Post*,” “OCR Correct*,” και “OCR Detect*” στο DBLP και στο Google Scholar τον Μάιο του. Τα έγγραφα που προέκυψαν ελέγχθηκαν προσεκτικά εάν παρουσιάζουν μεθόδους μετεπεξεργασίας ή όχι. Στη συνέχεια, τα σχετικά έγγραφα ταξινομήθηκαν σε βασικές ομάδες: Μη αυτόματα, ημιαυτόματα και αυτόματα. Στο επόμενο βήμα, ελέγξαμε όλα τα έγγραφα που αναφέρονται σε ένα από αυτά τα παραπάνω επιλεγμένα έγγραφα και προστέθηκαν τα σχετικά έργα στη συλλογή που μελετήσαμε. Συμπληρώθηκε μια συλλογή με πρόσθετα σχετικά έγγραφα που επιλέγονται μέσω

της DBLP των δημιουργών των επιλεγμένων εγγράφων. Η συλλογή εγγράφων στον τομέα της μετεπεξεργασίας κειμένου OCR κατηγοριοποιείται σε υποκατηγορίες που αναλύονται λεπτομερώς σε αυτήν την ενότητα.

Η βιβλιογραφία σχετικά με την έρευνα επεξεργασίας μετά από το OCR περιλαμβάνει μια πλούσια οικογένεια προσεγγίσεων. Μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες: Μη αυτόματος, ημιαυτόματος και αυτόματος τύπος ανάλογα με το επίπεδο εξάρτησης από τον άνθρωπο. Ωστόσο, υπάρχουν ορισμένες ημιαυτόματες προσεγγίσεις που συχνά συνδυάζονται με αυτόματες, επομένως, εξετάζουμε δύο κύριες κατηγορίες: Μη αυτόματες και (ημιαυτόματες) αυτόματες. Τα χαρακτηριστικά κάθε ομάδας αναλύονται στις ακόλουθες ενότητες.

2.2.3.1 Χειροκίνητη προσέγγιση

Οι προσεγγίσεις μαζικής συλλογής πόρων για τη μετεπεξεργασία των αποτελεσμάτων OCR έχουν δημιουργηθεί για να επωφεληθούν από τη δημόσια προσπάθεια βελτίωσης της ποιότητας των ψηφιοποιημένων κειμένων. Μία από τις πρώτες προσεγγίσεις της πληθοπορισμού που εφαρμόζεται στην επεξεργασία μετά την OCR είναι ένα σύστημα βασισμένο στο διαδίκτυο που ονομάζεται *trove* το 2009. Το σύστημα αυτό έχει αναπτυχθεί από την εθνική Βιβλιοθήκη της Αυστραλίας για τη διόρθωση ιστορικών εφημερίδων της Αυστραλίας. Η προσέγγιση περιλαμβάνει πλήρη άρθρα για τους εθελοντές και επιτρέπει τον καθορισμό κειμένου ανά γραμμή.

Οι Clematide εφαρμόζει μια πλατφόρμα για τη διόρθωση του πλήθους με το όνομα *Kokos* για τη μείωση του ποσοστού σφάλματος των ετησίων βιβλίων του Ελβετικού αλπικού συλλόγου που ψηφιοποιήθηκε από την Abby FineReader. Αυτό το πολύγλωσσο σώμα περιέχει κείμενα του 19ου αιώνα γραμμένα στα γερμανικά και τα γαλλικά. *Ο Kokos* εμφανίζει πλήρη έγγραφα στους χρήστες και τους επιτρέπει να διορθώνουν σφάλματα ανά λέξη. Περισσότεροι από 180,000 χαρακτήρες σε 21,247 σελίδες διορθώθηκαν από εθελοντές σε περίπου 7 μήνες, με ακρίβεια λέξης 99.7%.

Αντί να εμφανίζει πλήρη άρθρα στους χρήστες, ένα άλλο σύστημα (που ονομάζεται Digitalkoot) αναπτύχθηκε από τους Chrons et al χωρίζει τα άρθρα σε ενιαία φράσεις και τα τοποθετεί σε απλά παιχνίδια. Στόχος του μοντέλου παιχνιδιού είναι να προσελκύσει εθελοντές που θα δώσουν το χρόνο τους για τη διόρθωση εσφαλμένων διακριτικών. Η προσοχή που συγκέντρωσε ήταν σχετικά μεγάλη, καθώς 4,768 άτομα έπαιζαν τουλάχιστον ένα παιχνίδι. Το πειραματικό αποτέλεσμα αποκαλύπτει ότι η ποιότητα του διορθωμένου κειμένου είναι πολύ υψηλή με ακρίβεια λέξεων πάνω από 99%. Ωστόσο, το σύστημα αγνοεί το κοντινό περιβάλλον των λέξεων και επιτρέπει στους χρήστες να αλληλεπιδρούν μόνο με μεμονωμένες λέξεις, γεγονός που δημιουργεί αμφιβολίες για το ότι δεν μπορεί να διορθώσει σφάλματα σε πραγματικές συνθήκες.

Το reCAPTCHA είναι ένας άλλος τύπος δημόσιας συνεργασίας μετά τη διόρθωση. **Η πλήρως αυτοματοποιημένη δημόσια δοκιμή κατά τη διάρκεια της δοκιμής για την ενημέρωση των υπολογιστών και των ανθρώπων (CAPTCHA)**, η οποία είναι ένα ευρέως

διαδεδομένων μέτρο ασφαλείας στον Παγκόσμιο ιστό, χρησιμοποιείται για τον περιορισμό της επίθεσης κακόβουλων bot σε ιστότοπους. Οι Von Ahn et al προτείνει να αξιοποιηθούν οι συλλογικές εισροές από τις CAPTCHA για να γίνει η επεξεργασία παλαιών εντύπων. Οι συγγραφείς αναφέρουν ότι το σύστημα reCAPTCHA επιτυγχάνει ακρίβεια 99.1% σε επίπεδο λέξης, ενώ ένα τυπικό OCR στο ίδιο σύνολο τόξων λαμβάνει μόνο το 83.5%. Δυστυχώς, η έκδοση του reCAPTCHA που επιτρέπει την εκ των υστέρων διόρθωση έχει τερματιστεί από τον Μάρτιο του 2018.

Οι προσεγγίσεις συλλογικής επεξεργασίας μετά την OCR αποδεικνύουν τα οφέλη τους με σχετικά υψηλή αποδοτικότητα και είναι οικονομικά αποδοτικές. Επιπλέον, μπορούν εύκολα να εφαρμοστούν σε άλλα έργα ψηφιοποίησης. Ωστόσο, εξαρτώνται σε μεγάλο βαθμό από την εθελοντική εργασία και είναι απαραίτητο να παρακινούνται οι χρήστες. Επιπλέον, οι προσεγγίσεις αυτές αντιμετωπίζουν ορισμένους κινδύνους που προκαλούνται από τους δημόσιους χρήστες, όπως ενδεχόμενους βανδαλισμό κειμένου, λανθασμένες προτάσεις..

2.2.3.2 Ημι-αυτόματες προσεγγίσεις

Αυτός ο τύπος προσέγγισης μπορεί να ταξινομηθεί σε τύπους με μεμονωμένες λέξεις και εξαρτώμενους από το πλαίσιο τύπους με βάση την έκταση των πληροφοριών που χρησιμοποιούνται σε κάθε προσέγγιση. Οι προσεγγίσεις μεμονωμένων λέξεων λαμβάνουν υπόψη μόνο χαρακτηριστικά του ίδιου του διακριτικού στόχου OCR, για παράδειγμα την παρουσία του σε ένα λεξικό, την ομοιότητα μεταξύ του διακριτικού και μιας καταχώρησης λεξικού, τη συχνότητά του, το βαθμό εμπιστοσύνης αναγνώρισης που παρέχεται από το λογισμικό OCR κ.ο.κ. Μέθοδοι που ανήκουν σε αυτό το είδος εντοπίζουν και διορθώνουν κυρίως σφάλματα που δεν είναι σε λέξη. Ωστόσο, οι προσεγγίσεις που εξαρτώνται από το πλαίσιο όχι μόνο λαμβάνουν υπόψη τα χαρακτηριστικά της εστίασης του OCR, αλλά και το περιβάλλον του, για παράδειγμα, το μοντέλο γλώσσας της λέξης, τμήματα της ομιλίας κ.ο.κ. Εξετάζοντας το λεκτικό πλαίσιο, οι προσεγγίσεις αυτού του τύπου είναι σε θέση να χειριστούν τόσο τα μη λεκτικά όσο και τα πραγματικά σφάλματα.

Προσεγγίσεις με μεμονωμένες λέξεις

Προσεγγίσεις αυτού του τύπου τα κείμενα με δυνατότητα εκ των υστέρων διόρθωσης βάσει των χαρακτηριστικών των μεμονωμένων διακριτικών. Ανάλογα με τις εφαρμοζόμενες τεχνικές, μπορούν να ταξινομηθούν σε διάφορες ομάδες, συμπεριλαμβανομένων μεθόδων που βασίζονται στη συγχώνευση εξόδων OCR, λεξικών προσεγγίσεων, μοντέλων σφαλμάτων, γλωσσικών μοντέλων με βάση το θέμα και άλλων μοντέλων. Τα χαρακτηριστικά κάθε ομάδας αναλύονται στις ακόλουθες ενότητες.

(1)Συγχώνευση εξόδων OCR.

Η μία κατεύθυνση εργασίας συνδυάζει πολλαπλές εξόδους OCR. Τα κείμενα με OCR δημιουργούνται πρώτα από πολλούς μηχανισμούς OCR που λειτουργούν στην ίδια είσοδο ή από την ίδια μηχανή OCR σε διαφορετικές ψηφιακές εκδόσεις του αρχικού αρχείου. Τα κείμενα OCR μπορούν επίσης να συλλεχθούν από επαναλήψεις κειμένου στη συλλογή. Στη συνέχεια, εφαρμόζονται αλγόριθμοι ευθυγράμμισης για την ευθυγράμμιση των αποτελεσμάτων OCR. Τέλος, διερευνώνται διάφορες μέθοδοι λήψης αποφάσεων για την επιλογή της τελικής παραγωγής.

Στο πρώτο βήμα, οι προσεγγίσεις αυτού του τύπου λαμβάνουν πολλαπλά αποτελέσματα με διαφορετικούς τρόπους. Ενώ οι Lopresti και Zhou et al [3] χρησιμοποιούν την ίδια μηχανή OCR σε αρκετές σαρώσεις του ίδιου εγγράφου. Αντιθέτως, οι Lin [9], Lund et al [4] και Volk et al [6] χρησιμοποιούν πολλές μηχανές OCR για την ψηφιοποίηση του ίδιου εγγράφου.

Έχουν αναπτυχθεί αρκετές μέθοδοι ευθυγράμμισης για την ευθυγράμμιση πολλαπλών ακολουθιών εξόδου OCR. Οι Lund et al [9] εισάγουν έναν αποτελεσματικό αλγόριθμο για την ευθυγράμμιση των αποτελεσμάτων των πολλαπλών μηχανών οπτικής αναγνώρισης (OCR) χρησιμοποιώντας τον αλγόριθμο A* και στη συνέχεια για την αξιοποίηση των διαφορών μεταξύ τους. Αντί να ευθυγραμμιστούν οι εκδόσεις OCR της ίδιας σάρωσης, η προσέγγιση των Wemhoener et al To [7] επιτρέπει τη δημιουργία μιας ακολουθίας ευθυγράμμισης των αποτελεσμάτων OCR με τις σαρώσεις διαφορετικών αντιγράφων του ίδιου βιβλίου ή των διαφορετικών εκδόσεων του. Al Azawi et al [4, 8] εφαρμόζουν ευθυγράμμιση γραμμής προς σελίδα που ευθυγραμμίζει κάθε γραμμή του 1ου OCR με ολόκληρη τη σελίδα του δεύτερου OCR χρησιμοποιώντας **τους σταθμισμένους μορφοποιητές Finite-State (WFST)**.

Στο τελευταίο βήμα, εφαρμόζονται διάφορες τεχνικές για την επιλογή της καλύτερης ακολουθίας. Οι Lopresti et al [3], Lin [9], Wemhoener et al [7] και Reul et al [10] χρησιμοποιούν την πολιτική ψηφοφορίας, Al Azawi et al [4, 8] χρησιμοποιούν **τη μακροπρόθεσμη βραχυπρόθεσμη μνήμη (LSTM)** [11] για να αποφασίσουν ποια είναι η πιο σχετική παραγωγή. Στη λίστα αποφάσεων εκμάθησης χρησιμοποιούνται διάφορα είδη χαρακτηριστικών (ψηφοφορία, αριθμός, λεξικό, δελτίο και λεξικό), η μέγιστη ταξινόμηση εντροπίας ή οι μέθοδοι τυχαίων πεδίων υπό όρους (CRF) για την επιλογή της καλύτερης δυνατής διόρθωσης από τους Lund et al [12, 4].

Οι προαναφερθείσες προσεγγίσεις τείνουν να χαρακτηρίζονται από χαμηλότερο ποσοστό σφάλματος λέξεων από αυτό που λαμβάνεται με τη χρήση μιας μόνο μηχανής οπτικής αναγνώρισης χαρακτήρων (OCR). Ωστόσο, οι περισσότεροι από αυτές περιορίζουν τις προτάσεις υποψηφίων μόνο στα αποτελέσματα αναγνώρισης των μηχανών OCR. Επιπλέον, οι προσεγγίσεις αυτές δεν λαμβάνουν υπόψη τυχόν συναφείς πληροφορίες, συνεπώς , τα σφάλματα πραγματικών λέξεων δεν μπορούν να διορθωθούν. Επιπλέον, απαιτούν ορισμένες επιπλέον προσπάθειες επεξεργασίας πολλαπλών OCR και την παρουσία της αρχικής εισόδου OCR που δεν είναι πάντα διαθέσιμη.

Για να μετριαστεί το μειονέκτημα της εκτέλεσης OCR πολλές φορές, ορισμένες πρόσφατες μέθοδοι επωφελούνται από επαναλήψεις κειμένου ή μοντέλα k της διασταυρούμενης επικύρωσης με αναδίπλωση k. Ένα έργο των Xu και Smith [13] εντοπίζει

επαναλαμβανόμενα κείμενα στη συλλογή και δημιουργεί μια ευθυγράμμιση πολλαπλών ακολουθιών από αυτά. Η καλύτερη ακολουθία επιλέγεται με πλειοψηφία ή μοντέλο γλώσσας χαρακτήρων. Οι Das et al [14] χρησιμοποιούν επίσης δύο λέξεις στο σώμα. Οι λέξεις ομαδοποιούνται με βάση την ομοιότητα των εικόνων και των χαρακτηριστικών κειμένου. Κάθε παρουσία των ομάδων λέξεων ελέγχεται ως προς την ορθότητά της μέσω ενός λεξικού. Μια ομάδα λέξεων που περιέχει σφάλματα αποδέχεται τη συχνότερη πρόβλεψη ή την ανθρώπινη παρέμβαση ως εκπρόσωπο ολόκληρης της ομάδας. Η προσέγγιση των Reul et al [10] επωφελούνται από τα k μοντέλα πολλαπλής επικύρωσης και ψηφοφορίας με πλειοψηφία. Ειδικότερα, αντί να επιλέξουν το καλύτερο μοντέλο από τα κ πολλαπλάσια διασταυρούμενη επικύρωση και να εφαρμόσουν μόνο αυτό το μοντέλο στα δεδομένα δοκιμής, οι συγγραφείς εξετάζουν όλα τα μοντέλα k στα δεδομένα δοκιμής και επιλέγουν την υψηλότερη ακολουθία ψηφοφορίας ως τελικό αποτέλεσμα.

(2)Λεξικές προσεγγίσεις.

Μια άλλη γραμμή εργασίας είναι οι λεξικές προσεγγίσεις στις οποίες βασίζονται συνήθως Λεξικά και μετρήσεις απόστασης για την επιλογή υποψηφίων για σφάλματα OCR. Παρακάτω περιγράφεται η κάλυψη ενός λεξικού, ο τρόπος κατασκευής ενός δυναμικού λεξικού ή χρήσης διαθέσιμων δεδομένων με ευρετήριο, οι διάφορες μετρήσεις απόστασης καθώς και τα συμπληρωματικά χαρακτηριστικά. Αναπτύσσονται διάφορα δημοφιλή μέτρα, όπως η απόσταση επεξεργασίας Damerau-Levenshtein [15], που είναι ο ελάχιστος αριθμός πράξεων επεξεργασίας ενός χαρακτήρα ή η απόσταση ngram [7] που εξαρτάται από τον αριθμό των κοινών ναρίων μεταξύ δύο συμβολοσειρών. Ο Schulz et al [17] υποδεικνύει ορισμένες βελτιώσεις της απόστασης επεξεργασίας Levenshtein (που δηλώνονται ως απόσταση LV) με βάση τα συχνά μοτίβα σφάλματος. Η μέθοδος τους επιλέγει αποτελεσματικά μικρά σύνολα υποψηφίων με υψηλότερη ανάκληση, ειδικά στο περιβάλλον μεγάλων λεξιλογίου.

Ορισμένα τυπικά παραδείγματα αυτού του τύπου είναι τα Taghva et al [18], Estrella et al [26], Kettunen et al [20], Cappelatti et al [21]. Taghva et al [18] που παρουσιάζουν το σύστημα επεξεργασίας εγγράφων με την ονομασία MAN- ICURE, το οποίο περιλαμβάνει μια μονάδα μετεπεξεργασίας-PPSYS. Αυτή η ενότητα εντοπίζει σφάλματα που εκκρεμούν κυρίως σε λεξικά και τα διορθώνει κατά προσέγγιση, με χρήση σύγχυσης συχνότητας λέξεων και χαρακτήρων. Estrella et al [26] και Kettunen et al [20] εντοπίζει παραπλανητικές μάρκες που βασίζονται στο λεξικό και στη συχνότητα μονογραμματικού κειμένου. Για κάθε μη έγκυρο διακριτικό, δημιουργείται μια λίστα εναλλακτικών βάσει της απόστασης επεξεργασίας και των τυπογραφικών χαρακτηριστικών. Στη συνέχεια, ο πιο συχνός υποψήφιος αντικαθιστά το διακριτικό. Ο Cappelatti et al [21] εφαρμόζει ορθογραφικό έλεγχο για να προτείνει υποψήφιους διόρθωσης και να τους ταξινομεί χρησιμοποιώντας την απόσταση LV ή τις τροποποιημένες Needleman-Wunsch που λαμβάνει υπόψη τη συχνότητα του γράμματος που λαμβάνεται από το σώμα λέξεων.

Λαμβάνοντας υπόψη το γεγονός ότι τα συμβατικά λεξικά δεν περιλαμβάνουν σημαντικό αριθμό λέξεων ενός συγκεκριμένου τομέα, ορισμένες προηγούμενες εργασίες αποσκοπούν στη

μελέτη της επίδρασης της λεξικής κάλυψης σε προσεγγίσεις μετά την εφαρμογή OCR και στην υπόδειξη τεχνικών για τη δυναμική δημιουργία λεξικών εξαρτώμενων από τον τομέα.

Ο Strohmaier et al έχει εκμεταλλευτεί το θεματικό λεξικό για να βελτιώσει τα αποτελέσματα των προσεγγίσεων που ανήκουν στο λεξικό είδος. Δημιουργεί ένα δυναμικό λεξικό συλλέγοντας λεξικά ιστοσελίδων του τομέα εισόδου, τα οποία παρέχουν υψηλότερη κάλυψη από ένα στατικό συμβατικό λεξικό. Ορισμένα δυναμικά λεξικά που δημιουργούνται με αυτόν τον τρόπο εφαρμόζονται στις επόμενες εργασίες τους [22], όπου αναπτύσσουν ένα εργαλείο για τη βελτιστοποίηση της λεξικής μετά τη διόρθωση. Η συσκευή τους περιλαμβάνει δύο φάσεις, (1)την επιλογή παραμέτρων για τη βελτιστοποίηση ενός μοντέλου μετά τη διόρθωση που εφαρμόζεται σε μια μόνο μηχανή OCR, (2) το συνδυασμό των αποτελεσμάτων πολλών μηχανων OCR. Το άρθρο εστιάζει κυρίως στην πρώτη.

Ομοίως, οι Mihov et al [23] δημιουργούν αυτόματα δυναμικά λεξικά ειδικών για κάθε τομέα μέσω της ανάλυσης σχετικών ιστοσελίδων. Κατά την προσέγγισή τους, καταρτίζεται ένα λεξικό ως αυτοματοποίηση LV για να προτείνει υποψηφίους χαρακτήρες προς διόρθωση. Εξασφαλίζουν καλύτερη λεξική κάλυψη από τις προηγούμενες εργασίες [27] χρησιμοποιώντας λεξιλόγιο των 1000 σημαντικότερων σελίδων.

Οι Bassil et al [21], Taghva και Agarwal [3] αξιοποιούν τα μαζικά δεδομένα ευρετηρίου της Google για την παραγωγή OCR μετά την επεξεργασία. Στέλνουν διακριτικά OCR στη μηχανή αναζήτησης Google ως ερωτήματα αναζήτησης. Εάν το ερώτημα περιέχει σφάλματα, τότε ο μηχανισμός αναζήτησης θα προτείνει ορισμένες λέξεις με δυνατότητα αντικατάστασης για ανορθογραφία. Αυτές οι προτάσεις χρησιμοποιούνται ως διορθώσεις για σφάλματα OCR. Και οι δύο αυτές μέθοδοι εξασφαλίζουν υψηλή απόδοση σε μικρά σύνολα δεδομένων. Τα Taghva και Agarwal [3] χρησιμοποιούν επιπλέον την απόσταση LV, τη μεγαλύτερη κοινή ακολουθία, την τιμή mA-TRx σύγχυσης OCR για να επιλέξουν έναν κατάλληλο υποψήφιο χαρακτήρα. Η μέθοδος που χρησιμοποιούν είναι σε θέση να διορθώσουν περισσότερα σφάλματα από τη χρήση μόνο των προτάσεων της Google με βελτίωση περίπου 16.6%. Επιπλέον, η ανάλυσή τους δείχνει ότι η μηχανή αναζήτησης Google δεν προτείνει υποψηφίους, αν το περιβάλλον που βρίσκεται κοντά στο σφάλμα είναι εσφαλμένο.

Πέρα από τα λεξικά και τις μετρήσεις απόστασης, ορισμένες προσεγγίσεις αυτού του τύπου χρησιμοποιούν πληροφορίες που αφορούν τους κανόνες για τη βελτίωση των εκτελέσεών τους. Furrer et al [28] προσδιορίζουν τα σφάλματα OCR που προέρχονται από γοτθικά κείμενα με διάφορους πόρους, π.χ. ένα μεγάλο γερμανικό λεξικό, μια λίστα τοπικών ονομάτων, την εμπιστοσύνη της αναγνώρισης. Οι υποψήφιοι δημιουργούνται από παρόμοιες αντικαταστάσεις χαρακτήρων ή αποστάσεις ngram χαρακτήρων. Το πείραμα σε 35,000 λέξεις δείχνει μια αύξηση της ακρίβειας από 96.72% σε 98.36%. Μαζί με την επεξεργασία της απόστασης και της συχνότητας των λέξεων, Hammarström et al [29] και Jean-Caurant et al [30] θεωρούν ένα ακόμη χαρακτηριστικό: Την ομοιότητα της διανομής (π.χ. Word2Vec [103] ή Glove [31]) για τον εντοπισμό πιθανών ορθογραφικών παραλλαγών κάθε λέξης στο κείμενο GT. Στη συνέχεια, εάν μια παραλλαγή εμφανίζεται στο κείμενο OCR, τότε θα αντικατασταθεί από την αντίστοιχη σωστή λέξη.

Η Reynaert [32, 33] εισάγει μια μη επιτηρούμενη μέθοδο για την επίλυση προβλημάτων λεξικών διακυμάνσεων σε επίπεδο λέξεων [32] ή σε επίπεδο χαρακτήρων [33]. Ο συντάκτης βασίζεται σε μεγάλο εταιρικό υπόβαθρο για τον υπολογισμό των συχνοτήτων λέξεων και τη δημιουργία λεξικού εάν δεν είναι διαθέσιμο. Η μέθοδος εκμεταλλεύεται έναν πίνακα κατακερματισμού και μια συνάρτηση κατακερματισμού για να παράγει μεγάλο αριθμό για την αναγνώριση γραμμών που έχουν τους ίδιους χαρακτήρες (π.χ. τα “παραδείγματα” και “eaxmale” έχουν την ίδια τιμή κατακερματισμού). Αυτή η λειτουργία επιτρέπει την ανάκτηση παρόμοιων λέξεων για μια δεδομένη λέξη (ή λέξη-στόχο) με προσθήκη, αφαίρεση ή και τα δύο με την τιμή κατακερματισμού της. Για παράδειγμα, αφαιρώντας την τιμή κατακερματισμού του χαρακτήρα "y" από τη λέξη "αυτά" θα προκύψει η νέα λέξη "το". Μόλις ανακτήσετε όλες τις παραλλαγές μιας δεδομένης λέξης, εφαρμόζεται η επεξεργασία απόστασης και συχνότητας λέξεων για την επιλογή του υποψηφίου με την καλύτερη αντιστοίχιση.

Ωστόσο, οι λεξικές προσεγγίσεις που εφαρμόζονται εύκολα, έχουν ορισμένες ελλείψεις. Η έλλειψη λεξικών υψηλής κάλυψης είναι το δυσκολότερο πρόβλημα αυτού του τύπου [24], ιδίως με συλλογές ιστορικού των οποίων τα κείμενα δεν ακολουθούν τυπική ορθογραφία. Επιπλέον, οι προσεγγίσεις αυτού του τύπου επικεντρώνονται μόνο σε μεμονωμένες λέξεις, επομένως δεν μπορούν να χειριστούν πραγματικά λάθη.

(3)Μοντέλα σφαλμάτων.

Αρκετές προσεγγίσεις μετεπεξεργασίας OCR επικεντρώνονται σε σφάλματα. Μοντέλα γλώσσας πληκτρολόγησης για τη διόρθωση εσφαλμένων συμβολοσειρών OCR. Τα μοντέλα πρώιμων σφαλμάτων συχνά βασίζονται στις μετρήσεις απόστασης LV που πλησιάζουν όλες τις επεξεργασίες εξίσου και αγνοούν το περιβάλλον. Στο πλαίσιο της ορθογραφίας, οι Kenneth W. Church και William A. Gale [34] αναπτύσσουν ένα μοντέλο σφάλματος που συνδέει τις πιθανότητες με τη εμπλοκή ενός χαρακτήρα. Οι πιθανότητες εισαγωγής και διαγραφής εξαρτώνται από τον προηγούμενο χαρακτήρα. Οι Brill και Moore [35] προτείνουν ένα γενικότερο μοντέλο σφάλματος που επιτρέπει λειτουργίες επεξεργασίας πολλαπλών χαρακτήρων μαζί με τις πιθανότητες που έχουν. Αυτές οι τεχνικές εφαρμόζονται επίσης για τη διόρθωση σφαλμάτων OCR.

Χρησιμοποιώντας λειτουργίες επεξεργασίας πολλαπλών χαρακτήρων, οι Taghva και Stofsky εφαρμόζουν μια δια δραστική μέθοδο μετεπεξεργασίας με το όνομα OCR Spell. Πρώτον, η μέθοδος προσδιορίζει το μεγαλύτερο δυνατό όριο λέξεων αντί απλώς για τη tokening κειμένων που βασίζονται σε λευκούς χώρους. Στη συνέχεια, κάθε λέξη token ελέγχεται ως προς την ορθότητά της μέσω ενός λεξικού-αναζήτησης. Στη συνέχεια, οι υποψήφιοι με δυνατότητα αντικατάστασης για κάθε λανθασμένο διακριτικό δημιουργούνται με βάση τον πίνακα σύγχυσης και την παραλλαγή του, η οποία ενημερώνεται όταν οι χρήστες προσθέτουν μια διόρθωση για ένα σφάλμα. Επιπλέον, ένα ευρετήριο είναι σχεδιασμένο για το εντοπισμό υποψηφίων για λέξεις που περιέχουν μη αναγνωρίσιμο χαρακτήρα (π.χ., ένα σύμβολο περισπωμένης στο χαρτί). Στη συνέχεια, η συνάρτηση Bayesian χρησιμοποιείται για τη

βαθμολογία προτεινόμενων υποψηφίων βάσει των συχνοτήτων συντοπισμού. Τέλος, συνιστάται στους χρήστες η λίστα προτάσεων με κατάταξη.

Οι Kolak και Resnik [36] εστιάζουν στο θορυβώδες μοντέλο του καναλιού και προσαρμόζουν ένα πλαίσιο συντακτικής αναγνώρισης προτύπων για την επίλυση του προβλήματος της επεξεργασίας μετά την OCR. Η απόδοση του μοντέλου σφάλματος TME είναι συγκρίσιμη με την επεξεργασία πολλαπλών χαρακτήρων [37] σε μικρά εκπαιδευτικά δεδομένα, αλλά παρουσιάζει μειωμένη απόδοση σε μεγαλύτερο.

Αντιθέτως, οι Perez-Cortes et al [38] προτείνει στοχαστικό σφάλμα διορθώνοντας την ανάλυση για την επεξεργασία μετά την OCR. Η ιδέα τους είναι να δημιουργήσουν ένα στοχαστικό μηχάνημα πεπτικής κατάστασης, το οποίο θα δέχεται τις συμβολοσειρές στο λεξικό. Εάν μια λέξη γίνει αποδεκτή από το μοντέλο, τότε δεν απαιτείται διόρθωση. Διαφορετικά, εφαρμόζεται ένα μοντέλο σφάλματος που έχει συνταχθεί σε μηχάνημα πεπερασμένου ελέγχου για να προτείνει και να ταξινομεί υποψηφίους χαρακτήρες.

Επεκτείνοντας την προσέγγιση του Perez-Cortes et al [38], δημιουργείται ένα μοντέλο γλώσσας χαρακτήρων, ένα μοντέλο σφάλματος και στη συνέχεια προσθέτει ένα ακόμη μοντέλο που δημιουργήθηκε από την αναγνώριση χαρακτήρων εμπιστευτικά, το οποίο ονομάζεται μοντέλο υπόθεσης. Τρία μοντέλα συγκεντρώνονται ξεχωριστά σε WFSTs και στη συνέχεια συντίθενται στον τελικό υποδοχέα. Ο καλύτερος υποψήφιος επιλέγεται από τη διαδρομή του χαμηλότερου κόστους αυτού του τελικού υποδοχέα. Ωστόσο, η εμπιστοσύνη στην αναγνώριση χαρακτήρων συχνά δεν είναι διαθέσιμη σε ορισμένες ψηφιοποιημένες συντεχνίες, όπως σύνολα δεδομένων που χρησιμοποιούνται σε αυτές τις εργασίες [39], επομένως είναι αδύνατο να εφαρμοστεί πλήρως και να εφαρμοστεί το μοντέλο υπόθεσης. Αυτή η μέθοδος υπερέχει ενός Perez-Cortes et al [38] για τη διόρθωση των ισπανικών ονομασιών σε επίπεδο χαρακτήρων.

Σε μια άλλη εργασία, οι Perez-Cortes et al [40] κατασκευάζουν τα συνδυασμένα γλωσσικά μοντέλα από συναφή πεδία βάσει του WFST. Τρία πεδία, συμπεριλαμβανομένων των επαρχιών, των δήμων και των ταχυδρομικών κωδικών της Ισπανίας, περιέχονται στα μοντέλα τους. Τα πειραματικά αποτελέσματα δείχνουν ότι μειώνουν σημαντικά τα ποσοστά σφαλμάτων με αποδεκτό χρόνο διόρθωσης. Η μέθοδος βελτιώνεται περαιτέρω με το προσαρμόσιμο κατώτατο όριο για την απόρριψη των λιγότερο αξιόπιστων υποθέσεων.

Ομοίως, οι Borovikov et al χρησιμοποιούν δύο μεθόδους μετεπεξεργασίας με βάση το κρυφό μοντέλο Markov και Richter et al [41]. Σε αυτά τα μοντέλα, οι κρυφές καταστάσεις είναι χαρακτήρες λέξεων σε ένα λεξικό, τα παρατηρήσιμα σύμβολα είναι χαρακτήρες OCR. Η πιθανότητα μετάβασης και η αρχική πιθανότητα υπολογίζονται από τα κανονικοποιημένα στατιστικά στοιχεία βιογραφήματος και αρχικών χαρακτήρων από το λεξικό και την πραγματική αλήθεια. Η πιθανότητα εκπομπής υπολογίζεται από πίνακα σύγχυσης.

Η τεχνική των σταθμισμένων μηχανημάτων Finite-State (WFSM) εφαρμόζεται και πάλι στην Kolak και την Resnik [42], όπου οι συγγραφείς χρησιμοποιούν το θορυβώδες μοντέλο καναλιού χρησιμοποιώντας το WFSM για να εκτιμήσουν την πιο πιθανή ακολουθία χαρακτήρων πηγής για μια δεδομένη ακολουθία χαρακτήρων παρατηρούμενων χαρακτήρων

(ή κείμενο OCR). Η μέθοδος τους επιτρέπει την επεξεργασία ενός ή πολλαπλών χαρακτήρων και μπορεί να λειτουργήσει χωρίς πόρο λεξικού.

Οι Reffie και Ringlstetter υποστηρίζουν ότι οι πληροφορίες σχετικά με τις ορθογραφικές παραλλαγές και τα σφάλματα OCR διαδραματίζουν πολύ σημαντικό ρόλο στη βελτίωση της απόδοσης της OCR (π.χ., λεπτομερείς προσεγγίσεις μετά το OCR), καθώς και στην ανάκτηση πληροφοριών σχετικά με ψηφιοποιημένα ιστορικά κείμενα. Ως αποτέλεσμα, προτείνουν μια μη εποπτευόμενη μέθοδο για την αυτόματη ανάλυση ιστορικών εγγράφων OCR για την κατάρτιση προφίλ αυτών των πληροφοριών. Το προφίλ περιέχει γενικές και τοπικές πληροφορίες. Το τοπικό προφίλ παρέχει μια λίστα με πιθανές προτάσεις διόρθωσης για κάθε διακριτικό OCR, χρησιμοποιώντας την αλήθεια εδάφους και τις ιστορικές ορθογραφικές αυξομείωση. Όταν συσσωρεύουν το τοπικό προφίλ, υπολογίζουν το γενικό προφίλ για ένα κείμενο ιστορικού OCR. Αυτό το καθολικό προφίλ περιέχει μια λίστα με τύπους σφαλμάτων OCR και μια λίστα με μοτίβα ιστορικού με τις εκτιμώμενες συχνότητες τους. Τα αποτελέσματά τους αποκαλύπτουν μια ισχυρή συσχέτιση μεταξύ των πραγματικών πληροφοριών στο κείμενο OCR και αυτών στο εκτιμώμενο προφίλ τους. Αυτό το προφίλ μπορεί να εφαρμοστεί σε ψηφιοποιημένα κείμενα που έχουν υποβληθεί σε μετεπεξεργασία στην ίδια γλώσσα ή υποβάλλονται σε επεξεργασία από το ίδιο λογισμικό OCR.

Αυτός ο αυτοματοποιημένος μηχανισμός κατάρτισης προφίλ αποτελεί το σημαντικότερο μέρος ενός δια δραστικού εργαλείου μετεπεξεργασίας OCR, PoCoTo [43]. Κάθε λέξη OCR με το αντίστοιχο τμήμα εικόνας και μια πλήρη προβολή της σελίδας εμφανίζονται παράλληλα με τους χρήστες. Χρησιμοποιώντας τα καθολικά και τοπικά προφίλ, αυτό το εργαλείο επισημαίνει πιθανά εσφαλμένα διακριτικά του κειμένου εισαγωγής OCR, υπολογίζει και προτείνει υποψηφίους διόρθωσης στους χρήστες. Το εργαλείο επιτρέπει επίσης τη μαζική διόρθωση σειρών σφαλμάτων.

Στην επόμενη έκδοση αυτής της μεθόδου, οι Fink et al [44] επιπλέον, εξειδικεύστε τα προφίλ από την τροφοδοσία χρήστη- πίσω από τα βήματα μη αυτόματης διόρθωσης. Επίσης, διευρύνουν το σύνολο των μοτίβων τους με επιπλέον μοντέλα σε έγγραφα προγενέστερων περιόδων. Επιπλέον, εισάγονται μη ερμηνεύσιμα ηλεκτρονικά κλειδιά στο σύνολο των σφαλμάτων, γεγονός που συμβάλλει στη βελτίωση της ανάκλησης της ανίχνευσης σφαλμάτων. Το προφίλ χρησιμοποιείται ξανά σε πρόσφατες εργασίες των Englmeier et al [45]. Παρουσιάζουν το σύστημα A-PoCoTo για πλήρως αυτοματοποιημένη μετεπεξεργασία, στο οποίο οι πληροφορίες εξόδου προφίλ χρησιμοποιούνται ως ένα από τα χαρακτηριστικά για την εκμάθηση μηχανημάτων. Ένα σύστημα combined μεταξύ Του A-PoCoTo και του δια δραστικού, που ονομάζεται A-I-PoCoTo, επιτρέπει στους χρήστες να διορθώσουν εσφαλμένα αποτελέσματα της αυτόματης διόρθωσης ή να επικυρώσουν τις σωστές αποφάσεις.

Σε αντίθεση με άλλες προσεγγίσεις αυτού του τύπου, οι Gerdjikov et al [46] προτείνει μια νέα μέθοδο που ονομάζεται **ενσωμάτωση των γλωσσικών δομών** για τη δημιουργία και την κατάταξη υποψηφίων βάσει δομών λέξεων. Η γενική ιδέα είναι να δείτε πόσο θορυβώδεις λέξεις και λέξεις αναφοράς (λέξεις με θόρυβο ως σφάλματα, λέξεις που αναφέρονται ως λέξεις σε ένα λεξικό) έχουν δημιουργηθεί από άποψη διακριτών διορθώσεων. Μια χαρακτηριστική

επιδιόρθωση είναι είτε ένα πρόθεμα λέξης είτε μια λέξη που δεν διορθώνεται σε τουλάχιστον δύο διακριτά πλαίσια. Υποθέτουν ότι υπάρχει ένα σύνολο ορθογραφικών παραλλαγών που μετασχηματίζει διακριτές ασυμφωνίες λέξεων με θόρυβο σε εκείνες των αντίστοιχων αναφερόμενων λέξεων τους με τρόπο που ταιριάζει καλύτερα με τις δομές και των δύο συνόλων.

(4)Μοντέλα γλώσσας με βάση το θέμα.

Πολλές από τις παραπάνω προσεγγίσεις μετά την OCR βελτιώνουν την ποιότητα της εξόδου OCR συνδυάζοντας το μοντέλο σφάλματος και το μοντέλο γλώσσας του κειμένου. Ωστόσο, η χρησιμοποιούμενη γλώσσα είναι παγκόσμια και ανεξάρτητη από το θέμα του εγγράφου. Με άλλα λόγια, οι πληροφορίες του θέματος σχετικά με το έγγραφο δεν θεωρείται ότι εξαλείφουν τους θορυβώδεις υποψηφίους. Ορισμένες μελέτες προτείνουν στη συνέχεια την εφαρμογή γλωσσικών μοντέλων βάσει θέματος για τη μετεπεξεργασία κειμένων.

Ο Eger et al [47] εξετάζει τέσσερα μοντέλα μετατροπής μεταξύ συμβολοσειρών στο υπόδειγμα διόρθωσης σφαλμάτων: (1) η ακολουθία [48] είναι ένα κοινό μοντέλο ακολουθίας που χρησιμοποιεί πιθανότητες ngram σε ζεύγη υποσειρών των ακολουθιών εισόδου και εξόδου· (2) το DirectTL+ προβάλλει την εργασία ως γραμμή τμηματοποίησης ακολουθίας και επισήμανσης ακολουθίας. Η οδηγία DirecTL+ ενσωματώνει τους κοινούς χαρακτήρες ng στην ετικέτα ακολουθίας· (3)η AliSeTra [44] λειτουργεί με παρόμοιο τρόπο με την DirectTL+. Χρησιμοποιεί το CRF (δηλ. τυχαίο πεδίο υπό όρους) ως αλληλουχία του labeler και αγνοεί τα κοινά γραμμάριο χαρακτήρων· (4) η απόσταση επεξεργασίας βάσει συμφραζομένων θεωρεί το έργο ως σταθμισμένες λειτουργίες επεξεργασίας που μπορούν να εξαρτώνται από το περιεχόμενο εισόδου και εξόδου. Τρία από αυτά τα μοντέλα (sequitur, DirecTL+, AliSeTra) υπερβαίνουν τις γραμμές βάσης, το καλύτερο είναι να επιτυγχάνεται ακρίβεια λέξης περίπου 88.35% ενώ η γραμμή βάσης (δηλ. το θορυβώδες μοντέλο που επιτρέπει την επεξεργασία πολλαπλών χαρακτήρων) λαμβάνει περίπου 84.20%. Ο απλός συνδυασμός (π.χ. ψηφοφορία κατά πλειοψηφία) αυτών των μοντέλων και των γραμμών βάσης επιτυγχάνει ακόμη καλύτερα αποτελέσματα από ό,τι τα μεμονωμένα μοντέλα.

Οι Hämäläinen και Hengchen [49] χρησιμοποιούν ένα μοντέλο NMT επιπέδου χαρακτήρων για την αποκατάσταση σφαλμάτων στα έγγραφα του 18ου αιώνα. Για κάθε σωστή λέξη, ζητούν τις πιο παρόμοιες σημασιολογικές λέξεις του Word2Vec. Στη συνέχεια, οι λέξεις που προκύπτουν κατηγοριοποιούνται ως σωστές λέξεις ή σφάλματα. Στη συνέχεια, ομαδοποιούν κάθε σφάλμα με την πιο παρόμοια σωστή λέξη στη λίστα λέξεων που προκύπτει, βασιζόμενη στην επεξεργασία απόστασης και καταργούν σφάλματα των οποίων οι αποστάσεις επεξεργασίας στις αντίστοιχες σωστές λέξεις είναι μεγαλύτερες από τρεις.

Ομοίως, ο Hakala et al [50] εκπαίδευσε μοντέλα NMT επιπέδων χαρακτήρων χρησιμοποιώντας το OpenNMT και μετατρέψε τα σφάλματα OCR στις αντίστοιχες σωστές λέξεις τους. Ωστόσο, δημιουργούν εκπαιδευτικά δεδομένα που βασίζονται σε αντιγραφή κειμένου στη συλλογή κειμένου OCR. Για κάθε σύμπλεγμα περισσότερων από 20 ακολουθιών, αυτές συμβολίζουν τις ακολουθίες και ομαδοποιούν παρόμοιες λέξεις με βάση την απόσταση LV. Στη

συνέχεια, κάθε ομάδα λέξεων ευθυγραμμίζεται και ο πιο συνηθισμένος χαρακτήρας για κάθε θέση επιλέγεται ως εκπρόσωπός της. Η εμφάνιση αντικατάστασης υπολογίζεται για κάθε χαρακτήρα και οι διανομές συλλέγονται από συχνότητες τέτοιων αντικαταστάσεων από όλα τα συμπλέγματα. Τέλος, δημιουργούν θορυβώδεις συμβολοσειρές με τυχαία διαγραφή, εισαγωγή και αντικατάσταση ενός ή δύο χαρακτήρων από μια δεδομένη συμβολοσειρά.

Προσεγγίσεις ανάλογα με το πλαίσιο. Οι προσεγγίσεις επεξεργασίας μετά την OCR αυτού του είδους δεν αφορούν μόνο τα σφάλματα που δεν αφορούν λέξη αλλά και τις πραγματικές λέξεις εξετάζοντας τα χαρακτηριστικά των μεμονωμένων διακριτικών και των γύρω τους πλαισίων. Ταξινομούνται σε ορισμένες από τις ακόλουθες κατηγορίες: Μοντέλα γλωσσών, μοντέλα εκμάθησης μηχανημάτων βάσει χαρακτηριστικών και μοντέλα ακολουθίας.

Μοντέλα γλωσσών. Αρκετές μέθοδοι μετεπεξεργασίας αξιοποιούν γλωσσικά μοντέλα για τη δημιουργία και την κατάταξη των υποψηφίων διόρθωσης. Τα χωρίζουμε σε δύο κύριες ομάδες ανάλογα με τα μοντέλα γλωσσών που χρησιμοποιούνται, δηλαδή τα μοντέλα γλωσσών που βασίζονται σε στατιστικά και νευρονικά δίκτυα.

Μοντέλα στατιστικής γλώσσας. Τα μοντέλα στατιστικής γλώσσας υπολογίζουν τη κατανομή πιθανοτήτων των ακολουθιών λέξεων, τα οποία προκύπτουν από μετρήσεις συχνότητας με ορισμένες τεχνικές εξομάλυνσης για τη διαχείριση προβλημάτων αραιότητας των δεδομένων. Για λόγους απλότητας, οι μέθοδοι που χρησιμοποιούν τη συχνότητα ngram εξετάζονται επίσης σε αυτήν την ενότητα.

Ο Ringlstetter et al [51] τελειοποιήστε τη στρατηγική ανίχνευσης που χρησιμοποιήθηκε στη προηγούμενη προσέγγιση για την παραγωγή μικρότερων λεξικών με υψηλή κάλυψη. Συγκεκριμένα, η ομοιότητα μεταξύ των ανιχνευμένων σελίδων και του δεδομένου κειμένου εισόδου ελέγχεται με βάση την κανονικοποιημένη απόσταση συνημίτονου. Για κάθε διακριτικό στην είσοδο, επιλέγουν λέξεις των οποίων οι αποστάσεις LV είναι μικρότερο από ένα όριο. Οι προτάσεις διόρθωσης κατατάσσονται επιπλέον κατά συχνότητες γραμμάριο λέξεων από τα ανιχνευμένα δεδομένα.

Οι Poncelas et al [52], Hládek et al [53] και Généreux et al [54] χρησιμοποιούν παρόμοιες τεχνικές για τον εντοπισμό σφαλμάτων και προτείνουν υποψηφίους διόρθωσης. Συγκεκριμένα, εντοπίζουν θορυβώδη ηλεκτρονικά κλειδιά από ένα λεξικό - αναζήτηση και επιλέγουν υποψηφίους με βάση τις αποστάσεις LV μεταξύ ενός δεδομένου σφάλματος και καταχωρήσεων λεξικού. Ωστόσο, κατατάσσουν τις προτάσεις διόρθωσης με διαφορετικούς τρόπους. Ο Hládek et al [55] χρησιμοποιούν HMM με πιθανότητα μετάβασης κατάστασης ως πιθανότητα του μοντέλου της γλώσσας βιογραμμαρίων και πιθανότητα παρατήρησης ως απόσταση εξομάλυνσης για την επιλογή του καλύτερου υποψηφίου. Ο Généreux et al [54] επιλέξτε τον πλέον πιθανό υποψήφιο με το άθροισμα του ακόλουθου χαρακτηριστικού: βάρος σύγχυσης, υποψήφια συχνότητα και συχνότητα βιογραμμαρίων. Αναφέρουν ότι η μέθοδος τους αποδίδει συγκριτικά με το έργο του Hauser [56], το οποίο εφαρμόζει επεξεργασία πολλαπλών χαρακτήρων. Και οι δύο αυτές μέθοδοι επιτυγχάνουν παρόμοια ποσοστά μείωσης σφαλμάτων στα σύνολα δεδομένων που έχουν την ίδια περίοδο δημοσίευσης και την ίδια γραμματοσειρά.

Οι Evershed et al [57] εκμεταλλεύεται τόσο τα λάθη όσο και τα γλωσσικά μοντέλα της λέξης ngram με την μέθοδο μεταποιήσεως. Δημιουργούν προσεκτικά υποψηφίους σε επίπεδο χαρακτήρων χρησιμοποιώντας το μοντέλο σφάλματος και σε επίπεδο λέξης χρησιμοποιώντας “τριγραμμάρια κενού”, το οποίο δίνει την πιθανότητα μιας λέξης να είναι συσκευασμένη στην αριστερή και τη δεξιά γειτονική της. Το μοντέλο σφάλματος χρησιμοποιεί σταθμισμένες επεξεργασίες πολλαπλών χαρακτήρων και μια εκτίμηση βασιζόμενη στη διαδικασία "αντίστροφη OCR" που δημιουργεί γραφικά με χαμηλή ανάλυση για ένα ζεύγος θορυβωδών και σωστών λέξεων και, στη συνέχεια, υπολογίζει μια συσχέτιση bit. Οι προτάσεις κατατάσσονται με βάση το βάρος σύγχυσης και το γράφημα των λέξεων συν το μοντέλο γλώσσας 5 γραμμαρίων.

Η ομάδα ανταγωνισμού του Centro de Estudios de la RAE (που αναφέρεται ως WFST-PostOCR στο διαγωνισμό ICDAR2017 [58], RAE στο ICDAR2019 [59]) συνθέτει μοντέλα πιθανολογικών σφαλμάτων χαρακτήρων στο WFST. Τα μοντέλα γλώσσας γραμμάριου της λέξης και η παραμόρφωση των υποψηφίων που παράγονται από το μοντέλο σφάλματος χρησιμοποιούνται για να αποφασιστεί η καλύτερη εναλλακτική λύση. Η προσέγγιση αυτή επιτυγχάνει την υψηλότερη απόδοση όσον αφορά το έργο ανίχνευσης του διαγωνισμού ICDAR2017 και βελτιώνει την ποιότητα του κειμένου OCR και στους δύο διαγωνισμούς του ICDAR2017 και του ICDAR2019.

Αντί να χρησιμοποιήσουν το συμβατικό μοντέλο σφάλματος, οι Sariev et al Οι [60] και Niklas [61] χρησιμοποιούν το γλωσσικό πρότυπο της λέξης ngram με άλλες τεχνικές δημιουργίας υποψηφίων. Ο Niklas χρησιμοποιεί τον αλγόριθμο κατακερματισμού του αναγραμμένου και μια νέα μέθοδο προσαρμογής OCR για να αναζητήσει τις καλύτερες προτάσεις για εσφαλμένα OCR ηλεκτρονικά κλειδιά. Η προτεινόμενη μέθοδος ταξινομεί πρώτα τους χαρακτήρες σε κλάσεις ισοδυναμίας με βάση τα παρόμοια σχήματα. Οι χαρακτήρες της ίδιας κατηγορίας θα μοιράζονται το ίδιο κλειδί OCR. Το κλειδί του δεδομένου το word χρησιμοποιείται για την ανάκτηση όλων των λέξεων που έχουν το ίδιο κλειδί στο λεξικό.

Ορισμένες μέθοδοι αυτού του είδους βασίζονται στο Google Web ngram corpus [62] για την επιδιόρθωση σφαλμάτων, δηλαδή [9]. Ο Bassil et al [63] εντοπίζει μη λεκτικό σφάλμα με τη χρήση της συχνότητας μονογράμματος της λέξης. Δεύτερον, οι υποψήφιοι παράγονται με μονογραφία και με διγραμμάριο χαρακτήρων. Τέλος, επιλέγουν την καλύτερη εναλλακτική για κάθε εντοπισμένο σφάλμα, βασιζόμενοι στη λέξη συχνότητα 5 γραμμαρίων. Οι Soni et al [64] επικεντρώνονται στη διαχείριση σφαλμάτων τμηματοποίησης μέσω των διαδικτυακών γραμ. της Google 1T. Αποτείνουν εάν ένα διακριτικό πρέπει να κατατμηθεί με βάση την πιθανότητα μονογραφήματος λέξεων βιογραμμαρίων. Στη συνέχεια, εφαρμόζεται η πιθανότητα εμφάνισης μαρκών υψηλότερης τάξης για τη λήψη απόφασης σχετικά με τις μάρκες που είναι πιθανό να εμφανίζονται στο πλαίσιο. Ένα πρόσφατο έργο της Cacho [65] εντοπίζει σφάλματα OCR που βασίζονται στην OCR Spell [66] και προτείνει υποψηφίους για κάθε λανθασμένο διακριτικό βάσει διαγραμμένων λέξεων. Ο υποψήφιος με την υψηλότερη συχνότητα χρησιμοποιείται για την αντικατάσταση του σφάλματος.

Μοντέλα γλωσσών με βάση το νευρονικό δίκτυο. Έκτος από τα μοντέλα στατιστικής γλώσσας, ορισμένα από τις προσεγγίσεις εκμεταλλεύονται γλωσσικά μοντέλα που βασίζονται στο νευρονικό δίκτυο για να δημοσιεύσουν το κείμενο OCR. Μοντέλα νευρονικού δικτύου - γλώσσα εργασίας μαθαίνουν να συσχετίζουν κάθε λέξη με ένα διάνυσμα χαρακτηριστικών συνεχούς τιμής. Συνήθως, κατασκευάζονται και εκπαιδεύονται ως πιθανολογικές εξισώσεις για την πρόβλεψη της κατανομής πιθανοτήτων κατά την επόμενη λέξη, δεδομένου του πλαισίου της [67]. Εκτός από τα μοντέλα σε επίπεδο κειμένου, τα μοντέλα γλώσσας νευρονικού δικτύου μπορούν επίσης να λειτουργήσουν σε επίπεδο χαρακτήρων.

Το μοντέλο γλώσσας αμφίδρομης επικοινωνίας LSTM επιπέδου χαρακτήρων αναπτύσσεται και εφαρμόζεται στα ψηφιοποιημένα γαλλικά κλινικά κείμενα μετά τη διαδικασία από τους D'hondt et al [68]. Με καθαρά κείμενα (δηλ. ψηφιακές συλλογές που δεν περιέχουν σφάλματα), δημιουργούν αυτόματα το εκπαιδευτικό υλικό εφαρμόζοντας τυχαία λειτουργίες επεξεργασίας (δηλ. διαγραφή, εισαγωγή, αντικατάσταση). Σύμφωνα με τα αποτελέσματά τους, τα συστήματα με τις καλύτερες επιδόσεις υπερτερούν της βασικής TICCL. Οι συγγραφείς επισημαίνουν επίσης ότι τα μοντέλα τους δεν είναι καλά στον εντοπισμό σφαλμάτων που σχετίζονται με το όριο λέξεων.

Σε αντίθεση με αυτές τις παραπάνω προσεγγίσεις που λειτουργούν μόνο σε επίπεδο χαρακτήρων, οι Magallon et al [69] συνδύασε τα χαρακτηριστικά των μοντέλων σε επίπεδα χαρακτήρων και λέξεων. Τα χαρακτηριστικά του μοντέλου επιπέδου χαρακτήρων χρησιμοποιούνται ως εισαγωγή του μοντέλου επιπέδου λέξης. Είναι επίσης ένας από τους συμμετέχοντες στον διαγωνισμό ICDAR2017 [70]. Η απόδοσή τους είναι συγκρίσιμη με άλλες προσεγγίσεις στην εργασία ανίχνευσης.

Μοντέλα εκμάθησης μηχανήματος βάσει χαρακτηριστικών. Οι προσεγγίσεις εκμάθησης μηχανήματος μαθαίνουν από διαφορετικά χαρακτηριστικά, ώστε να είναι δυνατή η πιο ισχυρή επιλογή υποψηφίων. Αυτές οι προσεγγίσεις διερευνούν πολλαπλές πηγές για τη δημιουργία υποψηφίων, την εξαγωγή δυνατοτήτων και την κατάταξή τους χρησιμοποιώντας ένα στατιστικό μοντέλο.

Οι Kissos και Dershowitz [71] ελέγξτε κάθε διακριτικό στο έγγραφο εισόδου έναντι ενός λεξικού για σφάλματα από επιλογής. Για κάθε σφάλμα, χρησιμοποιείται ένα μοντέλο σφάλματος που επιτρέπει την επεξεργασία πολλαπλών χαρακτήρων για τη δημιουργία πιθανών υποψηφίων. Στη συνέχεια, κατατάσσουν τους υποψηφίους βάσει ενός μοντέλου παλινδρόμησης σε τέσσερα χαρακτηριστικά: Το βάρος σύγχυσης που προκύπτει από τα εκπαιδευτικά τους δεδομένα και τη συχνότητα των λέξεων. Αυτές οι δυνατότητες και δύο επιπλέον (OCR εμπιστοσύνης, διάρκεια όρων σε ένα OCR χρησιμοποιούνται στο μοντέλο παλινδρόμησης για να καθοριστεί εάν το διακριτικό OCR πρέπει να τοποθετηθεί εκ νέου από τον υποψήφιο με την υψηλότερη κατάταξη. Ωστόσο, η εμπιστοσύνη OCR δεν είναι πάντα διαθέσιμη στις συλλογές κειμένου OCR, επομένως αυτό το μοντέλο δεν μπορεί να εφαρμοστεί πλήρως. Επιπλέον, η προσέγγιση αυτή δεν λαμβάνει υπόψη ένα σημαντικό χαρακτηριστικό που χρησιμοποιείται στη διόρθωση σφαλμάτων [72], το οποίο είναι το σφάλμα. Η ομοιότητα μεταξύ δύο συμβολοσειρών (X , Y) μπορεί να μετρηθεί με βάση την απόσταση LV ή **τη μεγαλύτερη**

κοινή ακολουθία τους (LCS), η οποία είναι μια ακολουθία και *των δύο X* και *Y,* οποιαδήποτε ακολουθία μεγαλύτερη *από το Z* δεν αποτελεί ακολουθία *X* ή *Y* [75].

Η Khirbat [73] κατατάσσει ένα διακριτικό ως εσφαλμένο ή ορθό χρησιμοποιώντας τα ακόλουθα χαρακτηριστικά: παρουσία μη αλφαριθμητικού κειμένου στο διακριτικό, παρουσία του διακριτικού σε λεξικό, εάν η συχνότητα του διακριτικού και το περιβάλλον του στο έγγραφο εισόδου είναι μεγαλύτερη από ένα όριο, και εάν η συχνότητα του βιογραφήματος είναι υψηλότερη από ένα άλλο όριο. Η μέθοδος προτείνει καταχωρίσεις λεξικού των οποίων οι αποστάσεις LV από ένα δεδομένο σφάλμα είναι μικρότερες από ένα όριο ως υποψήφιοι διόρθωσης. Κάθε υποψήφιος βαθμολογείται από έναν αλγόριθμο γενικής βελτιστοποίησης [74] σε ένα σταθμισμένο άθροισμα της απόστασης LV και της LCS. Ο υποψήφιος με την υψηλότερη κατάταξη που παρουσιάζεται στο Google Web 1T ngram corpus προτείνεται να αντικαταστήσει το σφάλμα.

Εμπνευσμένοι από αυτά τα παραπάνω μοντέλα, οι Nguyen et al [76] παράγουν έναν κατάλογο εναλλακτικών λύσεων για κάθε σφάλμα οπτικής αναγνώρισης χαρακτήρων (OCR), με βάση το μοντέλο σφάλματος της επεξεργασίας πολλαπλών χαρακτήρων. Στη συνέχεια, οι υποψήφιοι αυτοί βαθμολογούνται από ένα μοντέλο ανατομής σε σύνολα χαρακτηριστικών (δηλαδή, βάρος σύγχυσης, πιθανότητα θεματικού πλαισίου που δίνεται από το μοντέλο στατιστικής/νευρικής γλώσσας [71]). Τα αποτελέσματα των πειραμάτων δείχνουν ότι αυτή η πολυτμηματική προσέγγιση είναι συγκρίσιμη με εκείνη των συμμετεχόντων στον διαγωνισμό ICDAR2017 [77]. Σε μια άλλη εργασία, οι Nguyen et al [78] επικεντρώνονται κυρίως στον εντοπισμό εσφαλμένων διακριτικών από το κείμενο OCR. Για κάθε διακριτικό στο κείμενο εισαγωγής, οι πιθανές αντικαταστάσεις του δημιουργούνται από τέσσερις πηγές (ή τέσσερα σύνολα δημιουργίας υποψηφίων): μοντέλο σφάλματος χαρακτήρων, περιβάλλον κατά τρία γραμμάρια τοπικών λέξεων, όπου στη συνέχεια επιλέγουν πιθανούς υποψηφίους που εξαρτώνται από τις δύο γειτονικές χώρες στα αριστερά ή τους δεξιούς και τους αριστερούς γείτονές τους, ή τους δύο γείτονές τους στα δεξιά. Η βασική ιδέα αυτού του ανιχνευτή σφαλμάτων είναι ότι ένα διακριτικό OCR πρέπει να αποδειχθεί έγκυρη λέξη μέσω δυαδικής ταξινόμησης με τιμές δυνατοτήτων υπολογισμένες από το σύνολο των υποψηφίων του. Προτείνονται δύο νέα σημαντικά χαρακτηριστικά, συμπεριλαμβανομένης της συχνότητας ενός διακριτικού OCR στα σύνολα δημιουργίας υποψηφίων και μιας τροποποιημένης έκδοσης ενός περίεργου δείκτη. Ο περίεργος δείκτης αντιπροσωπεύει ένα επίπεδο ανύπαρκτων ή σπάνιων n-γραμμαρίων σε μια μάρκα και υπολογίζεται με βάση τις συχνότητες των βιογραμμαρίων/τριγραμμαρίων χαρακτήρων μιας δεδομένης διακριτικού.

Ένα έργο της Cacho [79] εφαρμόζει την OCR Spell [66] για τον εντοπισμό σφαλμάτων και τη δημιουργία επανατοποθετήσεων. Στη συνέχεια, χρησιμοποιείται η ταξινόμηση SVM με πέντε χαρακτηριστικά (δηλ., επεξεργασία απόστασης, σύγχυση βάρους, συχνότητα μονογραμμαρίων και συχνότητες βιογραμμαρίων πίσω/εμπρός) για την επιλογή των καλύτερων εγγενών. Πρόσφατα, οι Nguyen et al [80] επιλέγει επίσης τους πιο σχετικούς υποψηφίους διόρθωσης με βάση κοινά χαρακτηριστικά, δηλαδή την ομοιότητα των συμβολοσειρών, τη συχνότητα των λέξεων, τη συχνότητα των γραμμάριο, και το βάρος της

σύγχυσης. Μετά τον εντοπισμό σφαλμάτων μέσω του λεξικού-αναζήτησης, οι συγγραφείς δημιουργούν και κατατάσσουν τους υποψηφίους κατά ένα

αλγόριθμος στοχαστικής βελτιστοποίησης με αντικειμενική συνάρτηση ως σταθμισμένο συνδυασμό αυτών των παραπάνω χαρακτηριστικών.

Μοντέλα ακολουθίας (Seq2Seq). Ορισμένες προσεγγίσεις θεωρούν τη μετεπεξεργασία OCR ως εργασία αυτόματης μετάφρασης (MT), η οποία μετατρέπει το κείμενο OCR σε διορθωμένο κείμενο στην ίδια γλώσσα. Σε αυτήν την ενότητα, τις ομαδοποιούμε σε σχέση με τα παραδοσιακά και νευρονικά μοντέλα Seq2Seq.

2.2.4 Μετρήσεις αξιολόγησης, σύνολα δεδομένων, γλωσσικοί πόροι

Μετρήσεις

Όσον αφορά την εργασία εντοπισμού σφαλμάτων, ο στόχος είναι να καθοριστεί εάν ένα διακριτικό OCR αναγνωρίστηκε σωστά ή όχι. Με άλλα λόγια, είναι το ίδιο με μια δυαδική ταξινόμηση. Για την αξιολόγηση της απόδοσης των μεθόδων ανίχνευσης σφαλμάτων χρησιμοποιούνται τρία δημοφιλή μετρικά στοιχεία (π.χ. ακρίβεια, ανάκληση, μέση αρμονική τους βαθμολογία F)

Όπου **αληθή θετικά (TP)** ο αριθμός των σφαλμάτων στο σύνολο που ανιχνεύονται με τη μέθοδο. **Ψευδώς θετικά (FP)** ο αριθμός έγκυρων λέξεων στο σύνολο που ανιχνεύονται ως σφάλματα. **Ψευδώς αρνητικά (FN)** ο αριθμός των σφαλμάτων στο σύνολο που δεν ανιχνεύονται.

Όσον αφορά την εργασία διόρθωσης, υπάρχουν ορισμένα κοινά μέτρα αξιολόγησης, συμπεριλαμβανομένου του ποσοστού σφάλματος (ποσοστό σφάλματος χαρακτήρων ως CER, ποσοστό σφάλματος λέξης ως WER), της ακρίβειας (ακρίβεια χαρακτήρων ως CAC, ακρίβεια λέξεων ως WAC), της ακρίβειας, της ανάκλησης, της βαθμολογίας F, **της δίγλωσσης μελέτης αξιολόγησης (BLEU)**.

Η ευρέως χρησιμοποιούμενη μέτρηση είναι η ER που ποσοτικοποιεί τον ελάχιστο αριθμό χαρακτήρων (δηλαδή, εισαγωγές, διαγραφές και αντικαταστάσεις) που απαιτούνται για τη μετατροπή του κειμένου.

Η ακρίβεια (AC) είναι επίσης μια δημοφιλής μέτρηση. Η Reynaert προτείνει τη χρήση των Precision, Recall ή correction rate και F-score για την αξιολόγηση της απόδοσης μιας προσέγγισης μετεπεξεργασίας. Αυτές οι μετρήσεις υπολογίζονται ως η αντίστοιχη εξίσωση (5), στην οποία οι τιμές TP, FP, αληθώς αρνητικά (TN) και FN ορίζονται σύμφωνα με τις τέσσερις πιθανές περιπτώσεις που μπορεί να προκύψουν μετά τη μετεπεξεργασία κειμένου OCR.

Το BLEU είναι ένα μετρικό σύστημα που εφαρμόζεται για την αξιολόγηση της ποιότητας των κειμένων εξόδου στη μηχανική μετάφραση. Σε ορισμένες προσεγγίσεις μετεπεξεργασίας που βασίζονται σε τεχνικές MT, το BLEU χρησιμοποιείται επίσης για την επαλήθευση των ερμηνειών τους, αποφασίζοντας με ποιον τρόπο το μεταφρασμένο κείμενο

Ορισμένα εργαλεία αξιολόγησης αναπτύσσονται από προηγούμενες εργασίες και είναι ελεύθερα προσβάσιμα στην κοινότητα: OcrevalUAtion που αναπτύσσεται από το Κέντρο επιδεξιότητας, το ISRI OCR Evaluation Impoors Toolkit (που δηλώνεται ως εργαλειοθήκη ISRI), και τον ανταγωνισμό για τη διόρθωση κειμένου μετά το OCR.

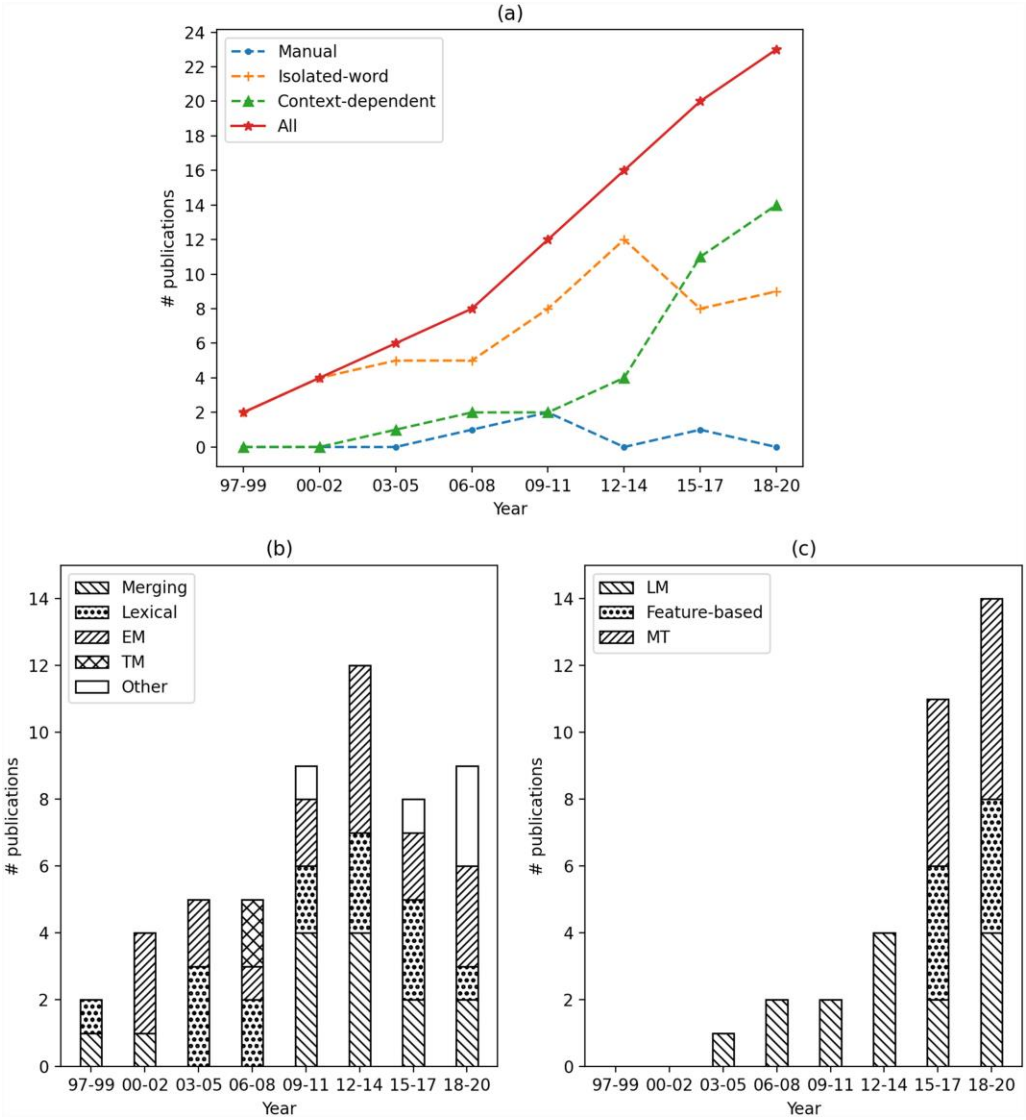
2.2.5 Συζήτηση και συμπεράσματα

Θα συζητήσουμε στη συνέχεια σε αυτήν την ενότητα τις τρέχουσες τάσεις επεξεργασίας μετά το OCR και θα προτείνουμε ορισμένες επεκτάσεις.

Για να έχουμε μια σαφή εικόνα των τάσεων των μεθόδων επεξεργασίας μετά το OCR, πρώτα καταδείξαμε στην εικόνα 4α) τον αριθμό των προσεγγίσεων που εμπίπτουν σε κάθε ομάδα βάσει των ετών δημοσίευσής τους. Φαίνεται ότι ο συνολικός αριθμός των προσεγγίσεων αυξήθηκε σταδιακά από 1997–1999 σε 2015–2017 και στη συνέχεια σε 2018–2020. Έχουν αναπτυχθεί ορισμένες μη αυτόματες προσεγγίσεις συνεργασίας για να επωφεληθούν από τη δημόσια προσπάθεια διόρθωσης σφαλμάτων οπτικής αναγνώρισης χαρακτήρων (OCR), η τελευταία δημοσιεύθηκε το 2016. Παρατηρείται αξιοσημείωτη αλλαγή στη δεσπόζουσα θέση μεταξύ των προσεγγίσεων που εξαρτώνται από την κάθε λέξη και από το πλαίσιο κατά τα τελευταία έτη. Προσεγγίσεις μεμονωμένων λέξεων με συγχώνευση αποτελεσμάτων OCR, εξαρτώμενων από τον τομέα δικαίων, μοντέλων σφαλμάτων κ.ο.κ. ενδέχεται να έχουν ήδη φτάσει στο σημείο κορεσμού, ενώ εξαρτώμενες από το περιβάλλον προσεγγίσεις υποστηρίζονται σε μεγάλο βαθμό από τα υπάρχοντα κιτ εργαλείων και τεχνικές νευρονικού δικτύου συνεχίσουν να ευδοκιμούν. Οι λεπτομερείς στατιστικές για τις μεμονωμένες λέξεις και τις εξαρτώμενες από το περιβάλλον προσεγγίσεις απεικονίζονται στο σχήμα 4β) και γ).

Η επεξεργασία μετά την OCR είναι ένα από τα σημαντικά βήματα για τη μετατροπή των εκτυπωμένων κειμένων σε ηλεκτρονική μορφή, ειδικά για τα ιστορικά έγγραφα στα οποία η απόδοση του λογισμικού OCR μειώνεται σημαντικά λόγω της παλιάς ορθογραφίας, των παρωχημένων γραμματοσειρών (π.χ., Fraktur, Antiqua), των περίπλοκων διατάξεων και της κακής φυσικής ποιότητας του χρησιμοποιημένου υλικού. Ωστόσο, οι προτεινόμενες προσεγγίσεις συχνά αφορούν διαφορετικά σύνολα δεδομένων και χρησιμοποιούν διαφορετικές αξιολογήσεις, επομένως, είναι δύσκολο να συγκριθούν οι επιδόσεις τους.

Οι τρεις πρόσφατοι διαγωνισμοί για τη διόρθωση κειμένου μετά την OCR που διοργανώθηκαν στο ICDAR2017, στο ICDAR2019 και στο ALTA2017 αποτελούν σημεία εκκίνησης για την επίλυση του προβλήματος αυτού, με την αξιολόγηση των προσεγγίσεων που υποβλήθηκαν στο ίδιο σύνολο δεδομένων με τις ίδιες μετρήσεις, ακόμη και με τη χρήση του ίδιου δημοσιευμένου εργαλείου αξιολόγησης.



Εικόνα 1. Πλήθος εγγράφων επεξεργασίας μετά από OCR με βάση τα έτη δημοσίευσής τους.

2.2.6 Προτεινόμενες οδηγίες

Εκτός από τις τεχνικές (ημι-)αυτόματης μετεπεξεργασίας, παρακάτω παρουσιάζονται ορισμένες δευτερεύουσες τεχνικές για την επιλογή μιας κατάλληλης προσέγγισης ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων.

(1) Όταν το σύνολο δεδομένων αποτελείται από καθαρό κείμενο, οι προσεγγίσεις που χρησιμοποιούν τα χαρακτηριστικά του ίδιου του σώματος και τους γλωσσικούς πόρους (π.χ. λεξικές μέθοδοι, μοντέλα σφαλμάτων, μοντέλα στατιστικής γλώσσας κ.λπ.) είναι οι μόνες επιλογές για τον καθαρισμό θορυβώδους κειμένου OCR. Τα περισσότερα από τα προσβάσιμα σύνολα δεδομένων ανήκουν σε αυτήν την περίπτωση.

(2) Όταν το σώμα περιέχει έγγραφα κειμένου μαζί με το χαρακτήρα μπορούμε να χρησιμοποιήσουμε αυτές τις σημαντικές πληροφορίες μαζί με άλλες πηγές για τον εντοπισμό ύποπτων διακριτικών, καθώς και για τη διόρθωσή τους.

(3) Στην καλύτερη περίπτωση, αν οι πρωτότυπες εικόνες είναι επίσης διαθέσιμες και η επανάληψη της διαδικασίας δεν είναι πολύ ακριβή, τότε μπορούμε να εκμεταλλευτούμε τα αποτελέσματα OCR των διαφορετικών μηχανών OCR, καθώς και την εμπιστευτικότητα της αναγνώρισής τους. Εκτός από τις προσεγγίσεις συγχώνευσης των τρεχουσών αποτελεσμάτων OCR, τα μελλοντικά συστήματα μετεπεξεργασίας ενδέχεται να επιτρέπουν την επιλογή της καλύτερης αντικατάστασης από τις ομάδες υποψηφίων που δημιουργούνται από αυτές τις μηχανές OCR και δημιουργούνται από άλλες πηγές (π.χ. μοντέλο σφάλματος, συχνότητα ngram). Επιπλέον, το πλαίσιο θα μπορούσε να χρησιμοποιηθεί για την αντιμετώπιση πραγματικών σφαλμάτων.

(4) Με ένα μικρό σύνολο δεδομένων, μια επιλογή είναι η χρήση συνθετικών δεδομένων για εκπαιδευτικά μοντέλα μετεπεξεργασίας. Προτείνονται ορισμένες τεχνικές για τη δημιουργία τεχνητού υλικού, όπως η τυχαία διαγραφή, ο ρυθμός, και η αντικατάσταση χαρακτήρων από μια δεδομένη λέξη· η απομίμηση ρεαλιστικών σφαλμάτων από κείμενα επανάληψης: Επιλογή μιας εναλλακτικής λύσης από τη λίστα των συχνών αντικαταστάσιμων χαρακτήρων για ένα δεδομένο χαρακτήρα ή χρήση μοντέλου αντίστροφου σφάλματος με την εισαγωγή ngrams.

Εστιάζοντας σε μοντέλα με βάση τους πόρους γλώσσας, τονίζουμε ορισμένες από τις αξιοσημείωτες λειτουργίες τους και τις χρήσιμες συμβουλές τους.

(1) Το μοντέλο σφάλματος $P(s \mid w)$ παίζει ζωτικό ρόλο στην πρόταση και την κατάταξη υποψηφίων χαρακτήρων. Το κεντρικό σημείο αυτού του μοντέλου είναι ο πίνακας σύγχυσης που θα μπορούσε να εκτιμηθεί με την ευθυγράμμιση του OCR κειμένου με τον αντίστοιχο GT. Δεδομένου ότι τα σφάλματα OCR μπορεί να περιλαμβάνουν τμηματοποίηση χαρακτήρων, δηλαδή διαχωρισμό

χαρακτήρων (π.χ. "*το m*" έναντι "*το in*"), συγχώνευση χαρακτήρων (π.χ. "*li*nk" έναντι "*b*nk"), είναι σημαντικό το μοντέλο σφάλματος να λαμβάνει υπόψη τόσο τη μετατροπή ενός όσο και πολλαπλών χαρακτήρων.

(2) Το γλωσσικό μοντέλο *P* (*w*) είναι η πιθανότητα της λέξης *w* στο να υπάρχει στο συμβατικό σώμα. Οι λεξικές προσεγγίσεις αποδεικνύουν ότι τα λεξικά μπορούν να δώσουν πιο σχετικές βαθμολογίες. Εάν η συλλογή δεδομένων περιέχει παλιά έγγραφα, τότε οι μέθοδοι πρέπει να περιλαμβάνουν λεξικά ιστορικών γλωσσών, καθώς οι ιστορικές ορθογραφίες συχνά δεν καταλήγουν σε σύγχρονες συμβάσεις.

(3) Αντί να εστιάζουμε μόνο στο μοντέλο σφάλματος, το μοντέλο γλώσσας, συνιστούμε ανεπιφύλακτα το συνδυασμό αυτών των δύο μοντέλων και άλλων πληροφοριών σε μοντέλα ηλεκτρονικής εκμάθησης που βασίζονται σε χαρακτηριστικά για την επίτευξη μεγαλύτερης απόδοσης.

Λαμβάνοντας υπόψη τις μεθόδους μετά την OCR που βασίζονται στα μοντέλα Seq2Seq, η NMT επιτυγχάνει καλή απόδοση και επίσης κυριαρχεί στα παραδοσιακά μοντέλα Seq2Seq σε ποσότητα, ωστόσο, τα παραδοσιακά μοντέλα Seq2Seq εξακολουθούν να έχουν τα δικά τους πλεονεκτήματα.

2.2.7 Πιθανές επεκτάσεις

Μαζί με τις κατευθυντήριες οδηγίες, προτείνουμε ορισμένες πιθανές οδηγίες για την ανάπτυξη προσεγγίσεων μετεπεξεργασίας.

(1) Οι προσεγγίσεις που βασίζονται στην εκμάθηση μηχανημάτων απαιτούν πολλά δεδομένα για την κατάρτιση ενός επιτυχημένου μοντέλου, εκτός από τις υπάρχουσες μεθόδους παραγωγής συνθετικών δεδομένων, προτείνουμε δύο ακόμη τρόπους για τη δημιουργία τεχνητών υλικών. Μια γνώμη είναι να εφαρμοστεί ένα μοντέλο αυτόματης μετάφρασης με το αρχικό κείμενο που χρησιμοποιείται ως GT και το κείμενο-στόχος που χρησιμοποιείται ως OCR κείμενο για την απόκτηση περισσότερων εκπαιδευτικών δεδομένων. Μια άλλη πρόταση είναι να χρησιμοποιηθεί η εκμάθηση μεταφοράς: Να εκπαιδευθεί ένα μοντέλο OCR και, στη συνέχεια, να επανεκπαιδευτεί σε τρέχοντα δεδομένα για να μάθει τις διανομές σφαλμάτων. Στη συνέχεια, το επανεκπαιδευμένο μοντέλο χρησιμοποιείται για την παραγωγή τεχνητών σφαλμάτων.

(2) Οι προσεγγίσεις που βασίζονται στο νευρονικό δίκτυο προσελκύουν μεγάλη προσοχή. Η απόδοση τους μπορεί να βελτιωθεί εάν χρησιμοποιηθούν από ξωτερικούς πόρους, όπως προεκπαιδευμένοι χαρακτήρες ή γλωσσικά μοντέλα κ.ο.κ. Παρουσιάζονται ορισμένες πιθανές επεκτάσεις στο νευρικό δίκτυο ως εξής:

- Είναι δυνατή η δημιουργία πλαισίων χαρακτήρων για την αποτύπωση της

ομοιότητας σχημάτων χαρακτήρων ή πλαισίων, ειδικά με παλιές γραμματοσειρές, ιστορική ορθογραφία.

- Τα μοντέλα που βασίζονται στο νευρονικό δίκτυο θα μπορούσαν να λειτουργήσουν σε επίπεδο λέξης, εάν υπάρχουν περισσότερα δεδομένα που μπορεί να παραχθούν από αρκετές προαναφερθείσες τεχνικές. Ένα υβριδικό μοντέλο σε επίπεδο χαρακτήρων και κειμένου θα μπορούσε ενδεχομένως να αποφέρει περισσότερα οφέλη για τη μετά OCR επεξεργασία, αν και η αρχική αξιολόγηση [156] δεν φαίνεται να την υποστηρίζει.
- Μία επιλογή είναι να επωφεληθούν από τους εξωτερικούς πόρους σε μοντέλα εκπαίδευσης ή μετά την επεξεργασία των αποτελεσμάτων τους. Μια απλή αλλά αποτελεσματική τεχνική είναι η χρήση αυτών των υλικών για το φιλτράρισμα των αποτελεσμάτων προσεγγίσεων που βασίζονται στο νευρικό δίκτυο ή η χρήση τους ως εισόδου ενσωμάτωσης. Πιο σύνθετη χρήση είναι δυνατή για υψηλότερες αποδόσεις.
- Μια άλλη επέκταση είναι, φυσικά, η σχεδίαση εξειδικευμένων μοντέλων για επεξεργασία μετά την OCR. Τα μοντέλα LM, Seq2Seq που βασίζονται στο νευρονικό δίκτυο ενδέχεται να είναι ορισμένες ενδιαφέρουσες επιλογές.

(3) Μια προσέγγιση επεξεργασίας μετά το OCR μπορεί να λειτουργήσει αποτελεσματικά εάν και οι δύο δευτερεύουσες εργασίες της έχουν καλές επιδόσεις. Οι πολλαπλές προσεγγίσεις εκμεταλλεύονται απλές τεχνικές για τον εντοπισμό σφαλμάτων ή ακόμη και για τον εντοπισμό του καταλόγου των σφαλμάτων, όπως δίνεται. Ωστόσο, αυτό δεν σημαίνει ότι ο εντοπισμός σφαλμάτων δεν είναι σημαντικός, καθώς κανείς δεν μπορεί να διορθώσει σφάλματα χωρίς να γνωρίζει τις θέσεις τους. Όσον αφορά την εργασία διόρθωσης σφαλμάτων, υπάρχουν λίγοι συνδυασμοί μεταξύ των προαναφερθέντων μοντέλων. Προτείνουμε επεκτάσεις με βάση κάθε δευτερεύουσα εργασία ως εξής:

- Θα πρέπει να δοθεί μεγαλύτερη προσοχή στις μεθόδους εντοπισμού σφαλμάτων. Ο μελλοντικός ανιχνευτής μπορεί όχι μόνο να εντοπίσει σφάλματα αλλά και να τα ταξινομήσει σε ομάδες, όπως σφάλματα κατάτμησης (π.χ. “τόπος καταγωγής” έναντι “πόλη προέλευσης”, “ωραία προβολή” έναντι “niceview”), σφάλματα αναγνώρισης (π.χ. “εκεί” έναντι “tberc”) κ.ο.κ.
- Αντί να βασιζόμαστε σε μια ενιαία τεχνική, πιστεύουμε ότι είναι χρήσιμο να συνδυάσουμε υποψηφίους που έχουν δημιουργηθεί από διάφορες τεχνικές: Περίπου αναζήτηση, μοντέλα σφάλματος, γλωσσικά μοντέλα και μοντέλα αυτόματης μετάφρασης. Για την κατάταξη των υποψηφίων, θα μπορούσαν να δοκιμαστούν διάφορες μέθοδοι: Ψηφοφορία, ML βάσει χαρακτηριστικών, κ.ο.κ.

(4) Επιπλέον, είναι απαραίτητο να υπάρχει μια βαθύτερη ανάλυση σχετικά με την απόδοση των προσεγγίσεων Pro-cessing μετά από OCR με κείμενα OCRed που δημιουργούνται από διαφορετικά λογισμικά OCR σε διαφορετικά χαρακτηριστικά των σφαλμάτων OCR, όπως η κατάτμηση χαρακτήρων, η κατάτμηση λέξεων, το μήκος λέξεων κ.ο.κ. Μπορείτε να βρείτε πιο λεπτομερείς

πληροφορίες σχετικά με τους τύπους σφαλμάτων OCR σε ένα

στατιστική έρευνα. Επιπλέον, τα διαδραστικά εργαλεία απεικόνισης και έρευνας που χρησιμοποιούν γλωσσικά μοντέλα μεγάλης κλίμακας θα μπορούσαν να είναι χρήσιμα για την καλύτερη κατανόηση των ψηφιακοποιημένων διδακτορικών, ιδίως των παλαιών. Ένα παράδειγμα είναι η προσέγγιση οπτικοποίησης DICT (έγγραφο στο πλαίσιο του χρόνου της) Το οποίο χρησιμοποιεί γλωσσικά μοντέλα που προέρχονται από το σύνολο δεδομένων Google Books τα οποία αντιστοιχίζονται προσωρινά με την ημερομηνία δημοσίευσης του εγγράφου για την απεικόνιση συχνών/σπάνιων λέξεων, καθώς και νεολογισμών ή παρωχημένων όρων που εμφανίζονται στο έγγραφο προορισμού. Μια τέτοια ανάλυση μπορεί να καταστεί απαραίτητη για τους ερευνητές ή τους προγραμματιστές προκειμένου να επιλέξουν τη σχετική μέθοδο για τα δεδομένα τους.

(5) Από την παρατήρησή μας, υπάρχουν λίγες μέθοδοι μετά την OCR που χειρίζονται εσφαλμένα διακριτικά που περιλαμβάνουν τμηματοποίηση λέξεων. Αυτός ο τύπος σφάλματος αντιπροσωπεύει περίπου το 18% των σφαλμάτων οπτικής αναγνώρισης χαρακτήρων (OCR) και συμβαίνει όταν το λογισμικό οπτικής αναγνώρισης χαρακτήρων (OCR) αναγνωρίζει λανθασμένα τα κενά στα έγγραφα. Πιστεύουμε ότι οι μελλοντικές προσεγγίσεις θα μπορούσαν να επικεντρωθούν περισσότερο στην αντιμετώπιση αυτού του απαιτητικού τύπου σφάλματος.

(6) Όσον αφορά τις μετρήσεις αξιολόγησης, τα ίσα βάρη εκχωρούνται συνήθως σε κάθε λέξη στα δόγματα. Πιστεύουμε ότι θα μπορούσε να έχει μεγαλύτερη σημασία εάν χρησιμοποιούνται διαφορετικά βάρη ανάλογα με τη σημασία των λέξεων, ειδικά όταν εξετάζονται συγκεκριμένες εφαρμογές, όπως η ανάκτηση εγγράφων.

(7) Μεταξύ διαφόρων συνόλων δεδομένων που χρησιμοποιούνται για την αξιολόγηση της απόδοσης 91 εγγράφων, 17 σύνολα δεδομένων είναι ελεύθερα προσβάσιμα. Αποτελούν πολύτιμους πόρους που επιτρέπουν στους ερευνητές να συγκρίνουν τις επιδόσεις τους και να κατανοούν καλύτερα τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων τους. Ωστόσο, ακόμη και με το ίδιο σύνολο δεδομένων, οι διαφορετικοί τρόποι διαχωρισμού των δεδομένων εκπαίδευσης, ανάπτυξης ή δοκιμής θα μπορούσαν επίσης να οδηγήσουν σε δυσκολίες στην αποτελεσματική σύγκριση. Επιπλέον, τα περισσότερα από αυτά τα χρησιμοποιημένα σύνολα δεδομένων δεν διαθέτουν σημαντικές πληροφορίες, όπως ημερομηνία δημοσίευσης εγγράφων, μηχανές OCR που χρησιμοποιήθηκαν για τη δημιουργία κειμένου OCR, εμπιστευτικά στοιχεία αναγνώρισης χαρακτήρων ή λέξεων, πρωτότυπες εικόνες κ.ο.κ. Προτείνουμε ορισμένες προτάσεις για τη δημιουργία συνόλων δεδομένων, συγκεκριμένα, είναι καλύτερο να υπάρχει σαφής διαχωρισμός των τμημάτων δεδομένων (π.χ.

εκπαίδευση, δοκιμές, δεδομένα ανάπτυξης) για δίκαιη σύγκριση. Πιο λεπτομερείς πληροφορίες (π.χ. χρόνος δημοσίευσης, μηχανισμός OCR, εμπιστευτικά αναγνώρισης, αρχικές εικόνες) κάθε συνόλου δεδομένων θα μπορούσαν να είναι χρήσιμες για περαιτέρω μελέτες.

(8) Δεδομένου ότι οι περισσότερες από τις τρέχουσες προσεγγίσεις προορίζονται για τη μετεπεξεργασία αγγλικών κειμένων OCR, είναι απαραίτητο να αναπτυχθούν μέθοδοι μετεπεξεργασίας για άλλες γλώσσες.

2.2.8 Συμπέρασμα

Σε αυτό το έργο, παρέχουμε μια εκτεταμένη έρευνα σχετικά με τις προσεγγίσεις των διορθωμένων εκτυπωμένων κειμένων OCR που εστιάζουν κυρίως στην αγγλική και σε άλλες γλώσσες του λατινικού κειμένου. Πρώτα παρουσιάζετε τον ορισμό της αποστολής και απεικονίζουμε τον τυπικό αγωγό της. Στη συνέχεια, η σημασία της επεξεργασίας μετά την OCR προκύπτει από διάφορες αναλύσεις της επίδρασης των σφαλμάτων OCR στις μεταγενέστερες εργασίες. Οι προσεγγίσεις μετεπεξεργασίας ομαδοποιούνται σε χειροκίνητες και (ημιαυτόματες) αυτόματες προσεγγίσεις, και στη συνέχεια χωρίζονται σε μεμονωμένες λέξεις και εξαρτώμενους από το περιβάλλον τύπους ανάλογα με την έκταση των πληροφοριών που χρησιμοποιούνται. Για κάθε τύπο, οι αντιπροσωπευτικές προσεγγίσεις κατηγοριοποιούνται σε μικρότερες ομάδες, ανάλογα με τις τεχνικές που εφαρμόζονται. Στη συνέχεια, συνοψίζονται με προσοχή τα χαρακτηριστικά κάθε βασικού βήματος και παρουσιάζονται τα πλεονεκτήματα και μειονεκτήματα.

Αναφέρονται δημοφιλείς μετρήσεις αξιολόγησης. Οι μελλοντικές εργασίες σε αυτόν τον τομέα μπορούν να επωφεληθούν από αυτά τα σύνολα δεδομένων και τα υπάρχοντα πειραματικά αποτελέσματα. Ορισμένα από τα διαθέσιμα σύνολα δεδομένων παρέχουν εικόνες ημερήσιας διαμόρφωσης, ώστε οι προσεγγίσεις μετεπεξεργασίας να μπορούν να αναπαρίστανται και να επωφελούνται από τη συγχώνευση διαφορετικών αποτελεσμάτων OCR, καθώς και από την αναγνώριση εμπιστευτικών. Όσον αφορά τις μετρήσεις αξιολόγησης, έχουν εφαρμοστεί μέχρι σήμερα διάφορα μέτρα, π.χ. ποσοστό σφάλματος λέξεων/χαρακτήρων, ποσοστό σφάλματος λέξεων/χαρακτήρων, αποτελεσματικότητα, ακρίβεια, Ανάκληση, F-score. Δεδομένου ότι αυτές οι μετρήσεις ζυγίζουν όλες τις λέξεις ισοδύναμα, μπορεί να μην είναι πάντα εξίσου καλό να εφαρμόζεται για διαφορετικές μεταγενέστερες εργασίες που μπορεί να χρειαστεί να εφαρμόζεται διαφορετική βαρύτητα σε συγκεκριμένους τύπους λέξεων.

Στη συνέχεια, θα παρουσιαστούν συζητήσεις σχετικά με την εξέλιξη των μεθόδων, με βάση την προσημείωση που έχει από το 1997, με την αξιοσημείωτη τάση στην ανάπτυξη προσεγγίσεων ανάλογα με το πλαίσιο κατά τα τελευταία έτη. Ανάλογα με τα χαρακτηριστικά των συνόλων δεδομένων, προτείνεται να εφαρμόζονται διαφορετικές μέθοδοι. Τα μοντέλα νευρονικού δικτύου έχουν αποδείξει τις ικανότητές τους στο εκ των υστέρων σε διορθωμένο κείμενο. Εκτιμάται ότι η μελλοντική έρευνα στον τομέα αυτό θα εξετάσει περαιτέρω την εξέλιξη

σε αυτή την κατεύθυνση. Οι περισσότερες προσεγγίσεις επεξεργάζονται δεδομένα στην αγγλική γλώσσα, απομένουν ακόμη πολλά να γίνουν για τις προσεγγίσεις της μετεπεξεργασίας άλλων γλωσσών.

Οι επικείμενες προσεγγίσεις σε αυτό το έργο δεν θα μπορούσαν μόνο να επικεντρωθούν στη βελτίωση της αποτελεσματικότητας της διόρθωσης, αλλά και στην επεξεργασία μετά την OCR σε άλλες γλώσσες, στις κατά προσέγγιση γλώσσες άλλων κειμένων (π.χ. αραβικά, κινεζικά), στα χειρόγραφα κείμενα και στις προηγμένες τεχνικές του νευρικού δικτύου.

2.3 Επισκόπηση προσεγγίσεων Deep learning για OCR και η κατανόηση εγγράφων

2.3.1 Εισαγωγή

Σε αυτό το έργο, εξετάζουμε το πρόβλημα της κατανόησης εγγράφων για έγγραφα που έχουν συνταχθεί στα Αγγλικά. Εδώ, ο όρος κατανόηση εγγράφου σημαίνει την αυτοματοποιημένη διαδικασία ανάγνωσης, διερμηνείας και εξαγωγής πληροφοριών από το γραπτό κείμενο και εικονογραφημένα σχήματα που περιέχονται στις σελίδες ενός εγγράφου. Από την άποψη των επαγγελματιών της μηχανικής μάθησης, η έρευνα αυτή καλύπτει τις μεθόδους με τις οποίες κατασκευάζουμε μοντέλα για την αυτόματη κατανόηση εγγράφων που αρχικά συντέθηκαν για ανθρώπινη κατανάλωση. Τα μοντέλα κατανόησης εγγράφων μεταφέρουν έγγραφα και τμηματικές σελίδες εγγράφων σε χρήσιμα μέρη (π.χ. περιοχές που αντιστοιχούν σε συγκεκριμένο πίνακα ή ιδιότητα), χρησιμοποιώντας συχνά την οπτική αναγνώριση χαρακτήρων (OCR) με κάποιο επίπεδο ανάλυσης διάταξης εγγράφου. Αυτά τα μοντέλα χρησιμοποιούν αυτές τις πληροφορίες για να κατανοήσουν το περιεχόμενο του εγγράφου γενικά, π.χ. ότι η συγκεκριμένη περιοχή ή το πλαίσιο οριοθέτησης αντιστοιχεί σε μια διεύθυνση. Σε αυτήν την έρευνα, εστιάζουμε σε αυτές τις πτυχές της κατανόησης των εγγράφων σε πιο λεπτομερές επίπεδο και συζητούμε δημοφιλείς μεθόδους για αυτές τις εργασίες. Στόχος μας είναι να συνοψίσουμε τις προσεγγίσεις που υπάρχουν στη σύγχρονη κατανόηση των εγγράφων και να τονίσουμε τις τρέχουσες τάσεις και τους περιορισμούς.

Στην ενότητα 2 , συζητούμε ορισμένα γενικά θέματα στη σύγχρονη NLP και στην κατανόηση των εγγράφων και παρέχουμε ένα πλαίσιο για τη δημιουργία συστημάτων από άκρο σε άκρο της αυτοματοποιημένης κατανόησης εγγράφων. Στη συνέχεια, στην ενότητα 3, εξετάζουμε τις καλύτερες μεθόδους για το OCR που περιλαμβάνει τόσο την ανίχνευση κειμένου όσο και τη μεταγραφή κειμένου. Στην ενότητα 4, έχουμε μια ευρύτερη άποψη του προβλήματος κατανόησης των εγγράφων, παρουσιάζοντας πολλές προσεγγίσεις στην ανάλυση της διάταξης των εγγράφων: Το πρόβλημα του εντοπισμού σχετικών πληροφοριών σε κάθε σελίδα. Στη συνέχεια, συζητούμε δημοφιλείς προσεγγίσεις για την εξαγωγή πληροφοριών (ενότητα 5).

2.3.2 Επεξεργασία και κατανόηση κειμένων

Η επεξεργασία εγγράφων περιλαμβάνει αλγόριθμους βασισμένους σε κανόνες , αλλά με την ευρεία επιτυχία της βαθιάς μάθησης, οι μέθοδοι που βασίζονται στην όραση μέσω υπολογιστή (CV) και στη φυσική γλωσσική επεξεργασία (NLP) έχουν έρθει στο προσκήνιο. Οι εξελίξεις στην ανίχνευση αντικειμένων και στην τμηματοποίηση εικόνων έχουν οδηγήσει σε συστήματα που προσεγγίζουν την ανθρώπινη απόδοση σε μια ποικιλία εργασιών. Ως αποτέλεσμα, οι μέθοδοι αυτές έχουν εφαρμοστεί σε διάφορους άλλους τομείς, συμπεριλαμβανομένων των NLP. Δεδομένου ότι τα έγγραφα μπορούν να αναγνωστούν και να προβληθούν ως μέσο οπτικής πληροφόρησης, πολλοί επαγγελματίες αξιοποιούν επίσης τεχνικές όρασης υπολογιστών και τις χρησιμοποιούν για την ανίχνευση κειμένου και την κατάτμηση των παραστατικών.

Η ευρεία επιτυχία και η δημοτικότητα των μεγάλων προ εκπαιδευμένων γλωσσικών μοντέλων έχουν προκαλέσει τη μετατόπιση της κατανόησης των εγγράφων προς τη χρήση μοντέλων με βάση τη βαθιά μάθηση (Peters et al., 2018, Devlin et al., 2019). Αυτά τα μοντέλα μπορούν να συντονιστούν με ακρίβεια για μια ποικιλία εργασιών και έχουν αντικαταστήσει τους φορείς λέξεων ως de facto πρότυπο προ εκπαίδευσης για εργασίες σε φυσική γλώσσα. Δεδομένου ότι τα επιχειρηματικά κείμενα μπορούν να είναι πολύ πυκνά, απαιτούνται τροποποιήσεις του μοντέλου αρχιτεκτονικής. Η πιο απλή προσέγγιση είναι η περικοπή εγγράφων σε μικρότερες ακολουθίες 512 διακριτικών, έτσι ώστε τα προ εκπαιδευμένα μοντέλα γλωσσών να μπορούν να χρησιμοποιηθούν πιο αποτελεσματικά. Μια άλλη προσέγγιση που είναι δημοφιλής πρόσφατα βασίζεται στη μείωση της πολυπλοκότητας της συνιστώσας του ελέγχου των μοντέλων γλώσσας που βασίζονται σε μετασχηματιστή.

Σύγχρονα συστήματα κατανόησης εγγράφων που υπάρχουν στη βιβλιογραφία ενσωματώνουν πολλές αρχιτεκτονικές βαθέως νευρονικών δικτύων για ανάγνωση και κατανόηση του περιεχομένου ενός εγγράφου. Δεδομένου ότι τα έγγραφα προορίζονται για ανθρώπους και όχι για μηχανήματα, οι επαγγελματίες πρέπει να συνδυάζουν τις αρχιτεκτονικές CV και NLP σε μια ενοποιημένη λύση. Παρόλο που συγκεκριμένες περιπτώσεις χρήσης υπαγορεύουν τις ακριβείς τεχνικές που χρησιμοποιούνται, ένα ολοκληρωμένο σύστημα χρησιμοποιεί:

Μια λειτουργική μονάδα ανάλυσης διάταξης εγγράφου που βασίζεται σε ένα πλάνο, η οποία χωρίζει κάθε σελίδα εγγράφου σε ξεχωριστές περιοχές περιεχομένου. Αυτό το μοντέλο όχι μόνο οριοθετεί τις σχετικές και τις άσχετες περιοχές, αλλά χρησιμεύει επίσης για την κατηγοριοποίηση του τύπου περιεχομένου που προσδιορίζει.

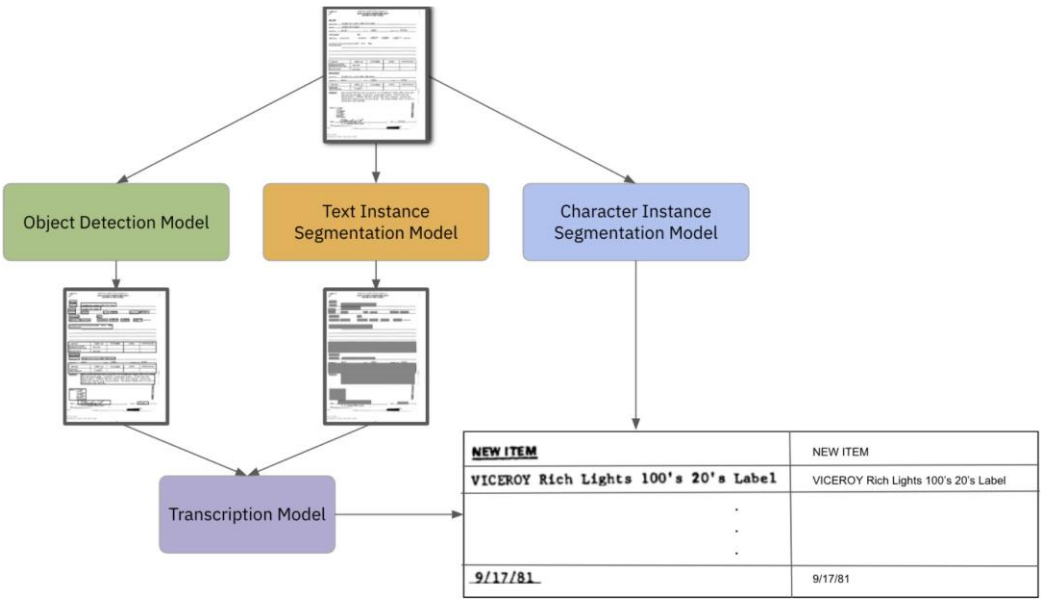
Μοντέλο οπτικής αναγνώρισης χαρακτήρων (OCR), σκοπός του οποίου είναι ο εντοπισμός και η πιστή καταγραφή όλου του κειμένου που υπάρχει στο έγγραφο. Τα μοντέλα OCR που παρεμβάλλονται μεταξύ των μοντέλων CV και NLP μπορούν είτε να χρησιμοποιούν απευθείας ανάλυση διάταξης εγγράφου είτε να επιλύουν το πρόβλημα με ανεξάρτητο τρόπο.

Μοντέλα εξαγωγής πληροφοριών χρησιμοποιούν το αποτέλεσμα της ανάλυσης OCR ή διάταξης εγγράφου για την κατανόηση και τον προσδιορισμό των σχέσεων μεταξύ των πληροφοριών που παρέχονται στο έγγραφο. Τα μοντέλα αυτά, τα οποία συνήθως είναι εξειδικευμένα σε συγκεκριμένο τομέα και εργασία, παρέχουν τη δομή που απαιτείται για την ανάγνωση εγγράφων από μηχανή, παρέχοντας βοηθητικό πρόγραμμα στην κατανόηση των εγγράφων.

Στις παρακάτω ενότητες, επεκτείνουμε αυτές τις έννοιες που αποτελούν μια ολοκληρωμένη λύση κατανόησης εγγράφων.

2.3.3 OCR

Το OCR διαθέτει δύο κύρια στοιχεία: Ανίχνευση κειμένου και μεταγραφή κειμένου. Γενικά, αυτά τα δύο μέρη είναι ξεχωριστά και χρησιμοποιούν διαφορετικά μοντέλα για κάθε εργασία. Στη συνέχεια, συζητούμε τις πιο σύγχρονες μεθόδους για κάθε ένα από αυτά τα στοιχεία και δείχνουμε πώς ένα έγγραφο μπορεί να υποβληθεί σε επεξεργασία μέσω διαφορετικών συστημάτων OCR γενικής χρήσης. Βλ. εικόνα 1 για λεπτομέρειες.



Εικόνα 2: Εδώ εμφανίζεται η γενική διαδικασία OCR

Ανίχνευση κειμένου

Η ανίχνευση κειμένου είναι η εύρεση κειμένου που υπάρχει σε μια σελίδα ή εικόνα. Η είσοδος συχνά αναπαρίσταται από έναν τρισδιάστατο τανυστή, $C\,H\,W$, όπου C είναι ο αριθμός των καναλιών (συχνά τρία, για το κόκκινο, το πράσινο και το μπλε), H είναι το ύψος και W είναι το πλάτος της εικόνας. Η ανίχνευση κειμένου είναι ένα δύσκολο πρόβλημα, επειδή το κείμενο έχει διάφορα σχήματα και προσανατολισμούς και μπορεί συχνά να παραμορφωθεί. Διερευνούμε δύο κοινούς τρόπους με τους οποίους οι ερευνητές θέτουν το πρόβλημα ανίχνευσης κειμένου: Ως εργασία ανίχνευσης αντικειμένων και ως εργασία τμηματοποίησης παρουσίας. Ένα μοντέλο ανίχνευσης κειμένου πρέπει είτε να μάθει να εξάγει συντεταγμένες πλαισίων οριοθέτησης γύρω από κείμενο.

Ανίχνευση κειμένου με ανίχνευση αντικειμένων

Η πρόοδος στη βαθιά μάθηση, ιδίως στην ανίχνευση αντικειμένων και στην σημασιολογική κατάτμηση, έχει οδηγήσει σε αλλαγή στον τρόπο αντιμετώπισης της ανίχνευσης κειμένου. Χρησιμοποιώντας αυτούς τους αποτελεσματικούς ανιχνευτές αντικειμένων από την παραδοσιακή βιβλιογραφία για την όραση σε υπολογιστή, όπως ο Ανιχνευτής πολλαπλών πλαισίων (SSD) μονής λήψης και τα ταχύτερα μοντέλα R-CNN μπορούν να κατασκευαστούν αποδοτικοί ανιχνευτές κειμένου.

Ένα από τα πρώτα έγγραφα που εφαρμόζουν έναν ανιχνευτή με βάση την παλινδρόμηση για το κείμενο είναι το TextBoxes. Προσθέτει μεγάλα προεπιλεγμένα πλαίσια που έχουν μεγάλους λόγους διαστάσεων πλευρών σε SSD, προκειμένου να προσαρμόσουν τον ανιχνευτή αντικειμένων στο κείμενο. Αρκετά έγγραφα βασίζονται σε αυτό για να καταστήσουν τα μοντέλα που βασίζονται σε παλινδρόμηση ανθεκτικά σε προσανατολισμούς. Άλλα έγγραφα έχουν παρόμοια προσέγγιση με το πρόβλημα, αλλά αναπτύσσουν το δικό τους δίκτυο προτάσεων, το οποίο είναι προσαρμοσμένο στο κείμενο και όχι στις φυσικές εικόνες.

Ανίχνευση κειμένου με τμηματοποίηση στιγμιότυπων

Η ανίχνευση κειμένου στα έγγραφα έχει το δικό της μοναδικό σύνολο προκλήσεων: Συγκεκριμένα, το κείμενο είναι συνήθως πυκνό και τα έγγραφα περιέχουν πολύ περισσότερο κείμενο από αυτό που συνήθως υπάρχει στις φυσικές εικόνες. Για την αντιμετώπιση αυτού του προβλήματος πυκνότητας, η ανίχνευση κειμένου μπορεί να τεθεί ως μια εξαιρετικά πυκνή εργασία τμηματοποίησης παρουσίας. Η κατάτμηση στιγμιότυπου είναι η εργασία ταξινόμησης κάθε pixel μιας εικόνας ως συγκεκριμένες, προκαθορισμένες κατηγορίες.

Οι ανιχνευτές κειμένου βάσει τμηματοποίησης λειτουργούν στο επίπεδο εικονοστοιχείων για τον εντοπισμό περιοχών κειμένου. Αυτές οι προβλέψεις ανά pixel χρησιμοποιούνται συχνά για την εκτίμηση πιθανοτήτων περιοχών κειμένου, χαρακτήρων και των σχέσεων τους μεταξύ παρακείμενων χαρακτήρων σε ένα ενοποιημένο πλαίσιο. Οι γιατροί χρησιμοποιούν δημοφιλείς μεθόδους τμηματοποίησης όπως τα πλήρως σύμμορφα δίκτυα (FCN) για την ανίχνευση κειμένου, βελτιώνοντας τα μοντέλα ανίχνευσης αντικειμένων, ειδικά όταν το κείμενο δεν είναι

ευθυγραμμισμένο ή παραμορφωμένο. Αρκετά έγγραφα βασίζονται σε αυτό το θεμέλιο τμηματοποίησης για την παραγωγή περιοχών που οριοθετούν τις λέξεις εξάγοντας περιοχές που οριοθετούν απευθείας από τα αποτελέσματα τμηματοποίησης. Το TextSnake το επεκτείνει περαιτέρω προβλέποντας την περιοχή κειμένου, την κεντρική γραμμή, την κατεύθυνση του κειμένου και την ακτίνα υποψηφίων από ένα FCN. Στη συνέχεια, αυτά τα χαρακτηριστικά συνδυάζονται με έναν αλγόριθμο ραβδώσεων για την εξαγωγή των σημείων του κεντρικού άξονα για την ανακατασκευή του στιγμιότυπου κειμένου.

Ανίχνευση σε επίπεδο λέξεων έναντι χαρακτήρων

Ενώ τα περισσότερα χαρτιά που αναφέρονται παραπάνω προσπαθούν να εντοπίσουν απευθείας λέξεις ή ακόμη και γραμμές λέξεων, ορισμένα χαρτιά υποστηρίζουν ότι ο εντοπισμός σε επίπεδο χαρακτήρων αποτελεί ευκολότερο πρόβλημα από τον γενικό εντοπισμό κειμένου, επειδή οι χαρακτήρες είναι λιγότερο διφορούμενοι από τις γραμμές κειμένου ή τις λέξεις. Το CRAFT χρησιμοποιεί ένα μοντέλο FCN για να εξάγει έναν δισδιάστατο χάρτη θερμότητας Gauss για κάθε χαρακτήρα (Baek et al., 2019). Οι χαρακτήρες που βρίσκονται κοντά μαζί ομαδοποιούνται στη συνέχεια, ομαδοποιούνται σε ένα ορθογώνιο που έχει τη μικρότερη δυνατή περιοχή και εξακολουθεί να περικλείει το σύνολο των χαρακτήρων. Πιο πρόσφατα, οι Ye et al (2020) να συνδυάσει με μεγάλη επιτυχία τα παγκόσμια χαρακτηριστικά, τα χαρακτηριστικά σε επίπεδο λέξεων και τα χαρακτηριστικά σε επίπεδο χαρακτήρων που αποκτώνται με τη χρήση δικτύων προτάσεων περιοχής (Region Subposal Networks, RPN).

Τα περισσότερα από τα μοντέλα που περιγράφονται παραπάνω αναπτύχθηκαν κυρίως για την ανίχνευση της σκηνής κειμένου, αλλά μπορούν εύκολα να προσαρμοστούν στην ανίχνευση κειμένου εγγράφου για το χειρισμό δύσκολων περιπτώσεων, όπως παραμορφωμένο κείμενο. Αναμένουμε λιγότερη παραμόρφωση στα έγγραφα από ό,τι στις φυσικές εικόνες, αλλά τα έγγραφα που έχουν σαρωθεί ανεπαρκώς με συγκεκριμένες γραμματοσειρές μπορεί να εξακολουθούν να δημιουργούν αυτά τα προβλήματα.

Μετεγγραφή κειμένου

Η μεταγραφή κειμένου είναι η εργασία μεταγραφής του κειμένου που υπάρχει σε μια εικόνα. Η εισαγωγή, μια εικόνα, είναι συχνά μια περικοπή που αντιστοιχεί είτε σε χαρακτήρα, λέξη ή ακόλουθία λέξεων και έχει τη διάσταση $C\ H'\ W'$. Ένα μοντέλο μεταγραφής κειμένου πρέπει να μάθει να χρησιμοποιεί αυτήν την περικομμένη εικόνα και να εξάγει μια ακολουθία διακριτικών που ανήκουν σε ένα προκαθορισμένο λεξιλόγιο V . *Το* V συχνά αντιστοιχεί σε ένα σύνολο χαρακτήρων. Για παράδειγμα, για την αναγνώριση των ψηφίων, αυτή είναι η πιο διαισθητική προσέγγιση. Διαφορετικά *, το* V μπορεί επίσης να αντιστοιχεί σε μια σειρά λέξεων, παρόμοια με ένα πρόβλημα γλωσσικής μοντελοποίησης σε επίπεδο λέξης. Και στις δύο περιπτώσεις, το πρόβλημα μπορεί να διαμορφωθεί ως ένα πρόβλημα πολύπλευρής κατάταξης με τον αριθμό των κλάσεων ίσο με το μέγεθος του λεξιλογίου V .

Τα μοντέλα μεταγραφής κειμένου σε επίπεδο λέξης απαιτούν περισσότερα δεδομένα, καθώς ο αριθμός των τάξεων στο πρόβλημα ταξινόμησης πολλαπλών κατηγοριών είναι πολύ μεγαλύτερος από ό,τι στο επίπεδο χαρακτήρων. Αφενός, η πρόβλεψη λέξεων αντί χαρακτήρων μειώνει την πιθανότητα δημιουργίας μικρών τυπογραφικών λαθών (όπως η αντικατάσταση ενός "A" από ένα "o" σε μια λέξη σαν "elephant"). Από την άλλη πλευρά, ο περιορισμός του εαυτού σε λεξιλόγιο επιπέδου λέξης σημαίνει ότι δεν είναι δυνατή η μεταγραφή λέξεων που δεν αποτελούν μέρος αυτού του λεξιλογίου. Αυτό το πρόβλημα δεν υπάρχει σε επίπεδο χαρακτήρων, καθώς ο αριθμός των χαρακτήρων είναι περιορισμένος. Εφόσον γνωρίζουμε τη γλώσσα του εγγράφου, είναι εύκολο να δημιουργηθεί ένα λεξιλόγιο που θα περιέχει όλους τους πιθανούς χαρακτήρες. Οι λέξεις Units (μονάδες) είναι μια βιώσιμη εναλλακτική λύση, καθώς μετριάζουν τα προβλήματα που υπάρχουν τόσο στη μεταγραφή σε επίπεδο λέξης όσο και σε επίπεδο χαρακτήρων.

Τα περισσότερα μοντέλα μεταγραφής κειμένου δανείζονται από τις εξελίξεις στη μοντελοποίηση αλληλουχίας τόσο για κείμενο όσο και για ομιλία και συχνά μπορούν να χρησιμοποιήσουν αυτές τις εξελίξεις ικανοποιητικά με μικρές μόνο προσαρμογές. Κατά συνέπεια, οι επαγγελματίες σπάνια αντιμετωπίζουν άμεσα αυτή την πτυχή σε σχέση με τα άλλα στοιχεία της εργασίας κατανόησης των εγγράφων.

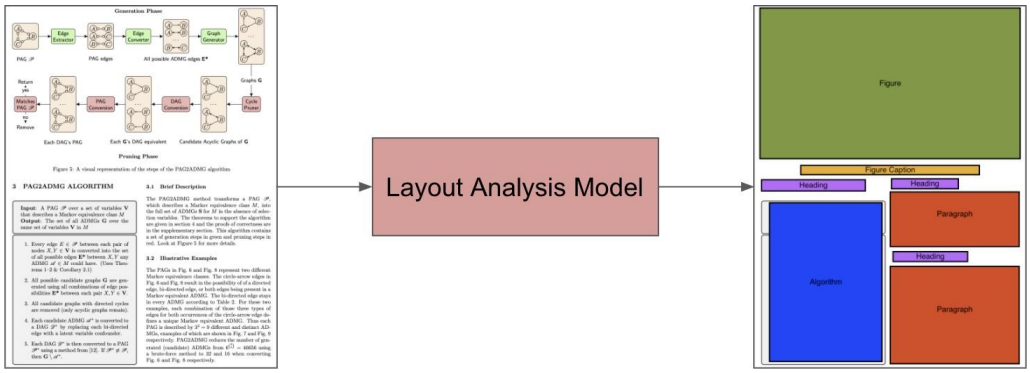
Μοντέλα από άκρο σε άκρο

Οι προσεγγίσεις από άκρο σε άκρο συνδυάζουν την ανίχνευση κειμένου και τη μεταγραφή κειμένου, με σκοπό τη βελτίωση και των δύο στοιχείων από κοινού. Για παράδειγμα, εάν η πιθανότητα της πρόβλεψης κειμένου είναι πολύ μικρή, αυτό σημαίνει ότι το πλαίσιο που εντοπίστηκε είτε δεν εντόπισε ολόκληρη τη λέξη είτε εντόπισε κάτι που δεν είναι κείμενο. Μια προσέγγιση από άκρο σε άκρο μπορεί να είναι πολύ αποτελεσματική σε αυτήν την περίπτωση. Ο συνδυασμός αυτών των δύο μεθόδων είναι αρκετά συχνός και τόσο το Fast Oriented Text Spotting (FOTS) όσο και το TextSpotter με τη ρητή ευθυγράμμιση και προσοχή συνδυάζουν αυτά τα μοντέλα διαδοχικά για την εκπαίδευση από άκρο σε άκρο. Αυτές οι προσεγγίσεις χρησιμοποιούν κοινές λύσεις ως χαρακτηριστικά τόσο για τον εντοπισμό όσο και για την αναγνώριση κειμένου, και εφαρμόζουν μεθόδους για περίπλοκους προσανατολισμούς του κειμένου. Το Mask TextSpotter είναι ένα μοντέλο από άκρο σε άκρο που συνδυάζει δίκτυα προτάσεων περιοχής για πλαίσια οριοθέτησης με τμηματοποίηση κειμένου. Αυτά τα πρόσφατα έργα δείχνουν τη δύναμη των λύσεων OCR από άκρο σε άκρο για τη μείωση των σφαλμάτων.

Ωστόσο, η ύπαρξη ξεχωριστών μοντέλων εντοπισμού κειμένου και αναγνώρισης κειμένου προσφέρει μεγαλύτερη ευελιξία. Πρώτον, τα δύο μοντέλα μπορούν να εκπαιδευτούν ξεχωριστά. Στην περίπτωση που μόνο ένα μικρό σύνολο δεδομένων είναι διαθέσιμο για την εκπαίδευση ολόκληρης της ενότητας OCR, αλλά πολλά δεδομένα αναγνώρισης κειμένου είναι εύκολα προσβάσιμα, είναι λογικό να αξιοποιείτε αυτήν τη μεγάλη ποσότητα δεδομένων στην εκπαίδευση του μοντέλου αναγνώρισης. Επιπλέον, με δύο ξεχωριστά μοντέλα, είναι εύκολο να

υπολογίσουμε δύο ξεχωριστά σύνολα μετρήσεων και να κατανοήσουμε καλύτερα πού θα μπορούσε να είναι το σημείο συμφόρησης.

Ως εκ τούτου, τόσο οι προσεγγίσεις δύο μοντέλων όσο και οι προσεγγίσεις από άκρο σε άκρο είναι βιώσιμες. Το κατά πόσον ο ένας είναι καλύτερος από τον άλλο εξαρτάται κυρίως από τα διαθέσιμα δεδομένα και από το τι θέλει να επιτύχει.



Εικόνα 3: Μοντέλο ανάλυσης διάταξης

1.3.4 Ανάλυση διάταξης εγγράφου

Η ανάλυση διάταξης εγγράφου είναι η διαδικασία εντοπισμού και κατηγοριοποίησης των περιοχών ενδιαφέροντος σε μια εικόνα ή μια σαρωμένη εικόνα μιας σελίδας. Σε γενικές γραμμές, οι περισσότερες προσεγγίσεις μπορούν να στηριχθούν στην κατάτμηση σελίδων και στη λογική δομική ανάλυση. Οι μέθοδοι τμηματοποίησης σελίδας εστιάζουν στην εμφάνιση και χρησιμοποιούν οπτικές ενδείξεις για τη διαίρεση σελίδων σε ξεχωριστές περιοχές. Οι συνηθέστερες είναι το κείμενο, οι εικόνες και οι πίνακες. Αντιθέτως, η λογική διαρθρωτική ανάλυση επικεντρώνεται στην παροχή λεπτομερέστερης σημασιολογικής ταξινόμησης για τις περιοχές αυτές, δηλαδή στον προσδιορισμό μιας περιοχής κειμένου που αποτελεί παράγραφο και τη διάκριση της από τίτλο τίτλου τίτλου ή τίτλου εγγράφου.

Τμηματοποίηση για ανάλυση διάταξης

Όταν εφαρμόζεται στο πρόβλημα της ανάλυσης διάταξης σε επιχειρηματικά έγγραφα, π.χ. οι μέθοδοι τμηματοποίησης προβλέπουν ετικέτες ανά pixel για την κατηγοριοποίηση περιοχών

ενδιαφέροντος. Οι μέθοδοι αυτές είναι ευέλικτες και προσαρμόζονται εύκολα στο ευγενικό καθήκον της κατάτμησης σελίδων ή στο ειδικότερο καθήκον της λογικής δομικής ανάλυσης.

Στο Yang et al (2017) οι συγγραφείς περιγράφουν ένα νευρονικό δίκτυο, το οποίο συνδυάζει τόσο κείμενο όσο και οπτικά χαρακτηριστικά σε μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή που ενσωματώνει επίσης ένα μη εποπτευόμενο δίκτυο προεκπαίδευσης. Κατά τη διάρκεια της αναφοράς, η προσέγγισή τους χρησιμοποιεί μια καθοδική σειρά στρωμάτων για την κωδικοποίηση οπτικών πληροφοριών, η οποία τροφοδοτείται σε μια συμμετρική ανοδική σειρά για αποκωδικοποίηση. Σε κάθε διαδοχικό επίπεδο, η παραγόμενη κωδικοποίηση μεταφέρεται επίσης απευθείας στον αντίστοιχο αποκωδικοποιητή, συνενώνοντας τις αναπαραστάσεις δειγματοληψίας προς τα κάτω και προς τα πάνω. Η αρχιτεκτονική αυτή διασφαλίζει ότι κατά τη διαδικασία κωδικοποίησης και αποκωδικοποίησης λαμβάνονται υπόψη οι οπτικές πληροφορίες χαρακτηριστικών σε διαφορετικά επίπεδα ανάλυσης. Στο τελικό στάδιο, παρέχονται μεταφρασμένα κείμενα παράλληλα με την υπολογισμένη οπτική αναπαράσταση.

Αυτή η αρχιτεκτονική κωδικοποίησης-αποκωδικοποίησης, εμπνευσμένη από το U-Net, έχει υιοθετηθεί για την ανάλυση διάταξης εγγράφων σε διάφορες προσεγγίσεις. Στη συνέχεια, η αναπαράσταση υποβάλλεται σε επεξεργασία με ανοδική δειγματοληψία και μικρότερες 1x1 και 3x3 επίπεδα συνέλιξης. Και τα δύο έργα χρησιμοποιούνται για την εκτέλεση ανάλυσης διαρρύθμισης σε ιστορικά έγγραφα και εφημερίδες από πολλές ευρωπαϊκές γλώσσες, αντίστοιχα.

2.3.4 Εξαγωγή πληροφοριών

Στόχος της εξαγωγής πληροφοριών για την κατανόηση των εγγράφων είναι η λήψη εγγράφων με διαφορετικές διατάξεις και η εξαγωγή πληροφοριών σε δομημένη μορφή. Στα παραδείγματα περιλαμβάνονται η κατανόηση απόδειξης για την αναγνώριση ονομάτων, ποσοτήτων και τιμών των στοιχείων, καθώς και η κατανόηση των φορμών για τον προσδιορισμό διαφορετικών ζευγών βασικών τιμών. Η εξαγωγή εγγράφων από ανθρώπους υπερβαίνει την απλή ανάγνωση κειμένου σε μια σελίδα, καθώς συχνά είναι απαραίτητο να μάθετε τις διατάξεις των σελίδων για πλήρη κατανόηση. ΩΣ εκ τούτου, οι πρόσφατες βελτιώσεις διαθέτουν εκτεταμένες στρατηγικές κωδικοποίησης κειμένου για έγγραφα, κωδικοποιώντας επιπρόσθετα δομικές και οπτικές πληροφορίες κειμένου με διάφορους τρόπους.

2-διάστατη ενσωμάτωση

Έχουν προταθεί προσεγγίσεις για την τοποθέτηση ετικετών πολλαπλών αλληλουχιών, οι οποίες επαυξάνουν τις τρέχουσες ονομαζόμενες μεθόδους αναγνώρισης οντότητας (NER)

ενσωματώνοντας χαρακτηριστικά των κυτίων που οριοθετούν την εικόνα 2D και συγχωνεύοντάς τα με στοιχεία επισήμανσης κειμένου για τη δημιουργία μοντέλων που γνωρίζουν ταυτόχρονα τόσο τη θέση στο πλαίσιο όσο και τη θέση στο χώρο κατά την εξαγωγή πληροφοριών. Κατά τη διάρκεια της προ εκπαίδευσης, το κείμενο αποκρύπτεται τυχαία, αλλά διατηρούνται τα δισδιάστατα παραδείγματα τοποθέτησης. Στη συνέχεια, αυτό το μοντέλο μπορεί να ρυθμιστεί λεπτομερώς σε επεξεργασία επόμενου σταδίου. Εναλλακτικά, οι συντεταγμένες του πλαισίου οριοθέτησης μπορούν επίσης να ενσωματωθούν χρησιμοποιώντας το συνημίτονο και ημίτονο, όπως οι μέθοδοι κωδικοποίησης θέσης. Μπορούν επίσης να ενσωματωθούν και άλλα χαρακτηριστικά, όπως ο αριθμός γραμμής ή ακολουθίας. Σε αυτό το σενάριο, το έγγραφο υποβάλλεται σε προ επεξεργασία για την εκχώρηση ενός αριθμού γραμμής σε κάθε μεμονωμένο διακριτικό. Στη συνέχεια, κάθε διακριτικό παραγγέλνεται από τα αριστερά προς τα δεξιά και λαμβάνει μια διαδοχική θέση. Τέλος, τόσο η θέση γραμμής όσο και η θέση αλληλουχίας είναι ενσωματωμένες.

Αν και οι στρατηγικές αυτές έχουν σημειώσει επιτυχία, η αποκλειστική χρήση του αριθμού γραμμής ή του κυρτού πλαισίου μπορεί να είναι παραπλανητική όταν το έγγραφο έχει σαρωθεί σε ανομοιόμορφη επιφάνεια, οδηγώντας σε καμπυλωτό κείμενο.

Εικόνα ενσωμάτωσης

Ο στόχος του μοντέλου είναι να τμηματοποιήσει σημασιολογικά πληροφορίες ή να σχηματίσει πλαίσια που οριοθετούν περιοχές ενδιαφέροντος. Η στρατηγική αυτή συμβάλλει στη διατήρηση της δισδιάστατης διάταξης του εγγράφου και επιτρέπει στα μοντέλα να αξιοποιούν τις δισδιάστατες συσχετίσεις. Η απευθείας ενσωμάτωση πληροφοριών κειμένου στην εικόνα απλοποιεί την εργασία για την κατανόηση των δισδιάστατων σχέσεων κειμένου. Σε αυτές τις περιπτώσεις, εφαρμόζεται μια συνάρτηση κωδικοποίησης σε ένα προτεινόμενο επίπεδο κειμένου (δηλ. χαρακτήρας, διακριτικό, λέξη) για τη δημιουργία μεμονωμένων φορέων ενσωμάτωσης. Αυτά τα ανύσματα μεταφέρονται σε κάθε εικονοστοιχείο που αποτελείται από το πλαίσιο οριοθέτησης που αντιστοιχεί στο ενσωματωμένο κείμενο, δημιουργώντας τελικά μια εικόνα $W \; H \; D$ όπου W είναι το πλάτος, H είναι το ύψος και D είναι η διάσταση ενσωμάτωσης. Οι προτεινόμενες παραλλαγές παρατίθενται ως εξής:

1. *Char Grid*.
2. *To WordGrid* ενσωματώνει μεμονωμένες λέξεις με τη χρήση Word2vec ή FastText.
3. BERT χρησιμοποιείται για τη λήψη φορέων λέξεων-περιεχομένου.
4. *C+BERTgrid*, συνδυάζει διανύσματα χαρακτήρων και συγκεκριμένου περιβάλλοντος.

Κατά τη σύγκριση των μεθόδων πλέγματος, το C+BERT έχει δείξει την καλύτερη απόδοση, πιθανώς λόγω των δημιουργία πλαισίων λέξεων-φορέων σε συνδυασμό με ένα βαθμό

προσαρμοστικότητας σε σφάλματα OCR. Zhao et al (2019) προτείνει μια εναλλακτική προσέγγιση για την άμεση εφαρμογή γραπτών μηνυμάτων στην εικόνα. Ένα πλέγμα προβάλλεται πάνω από την εικόνα και μια λειτουργία χαρτογράφησης εκχωρεί κάθε διακριτικό σε ένα μοναδικό κελί στο πλέγμα. Στη συνέχεια, μάθετε να αντιστοιχίζετε κάθε κελί στο πλέγμα σε μια κατηγορία. Η μέθοδος αυτή μειώνει σημαντικά τις διαστάσεις λόγω του συστήματός του με πλέγμα, διατηρώντας ταυτόχρονα την πλειοψηφία των δισδιάστατων χωρικών σχέσεων.

Αρχεία ως γραφήματα

Το μη δομημένο κείμενο σε έγγραφα μπορεί επίσης να αναπαρασταθεί ως δίκτυο γραφημάτων, όπου οι κόμβοι σε ένα γράφημα αντιπροσωπεύουν διαφορετικά τμήματα κειμένου. Δύο κόμβοι συνδέονται με ένα άκρο εάν είναι συμμετρικά παρακείμενοι, επιτρέποντας την απευθείας μοντελοποίηση της σχέσης μεταξύ των λέξεων. Ένας κωδικοποιητής, όπως ένα BiLSTM, κωδικοποιεί τμήματα κειμένου σε κόμβους. Τα άκρα μπορούν να αναπαρασταθούν ως δυαδική μήτρα ή ως εμπλουτισμένη μήτρα, κωδικοποιούν πρόσθετες οπτικές πληροφορίες όπως η απόσταση μεταξύ τμημάτων ή το σχήμα των κόμβων πηγής και στόχου. Στη συνέχεια, εφαρμόζεται ένα δίκτυο γραφικής παράστασης σε διαφορετικά πεδία υποδοχής με παρόμοιο τρόπο με τις διεσταλμένες εκδοχές για να διασφαλιστεί ότι μπορούν να διδαχθούν τόσο τοπικές όσο και παγκόσμιες πληροφορίες. Στη συνέχεια, η αναπαράσταση μεταβιβάζεται σε έναν αποκωδικοποιητή tagging ακολουθίας.

Τα έγγραφα μπορούν επίσης να αναπαριστούν ως κατευθυνόμενο γράφημα και ως parser χωρικής εξάρτησης. Σε αυτήν την αναπαράσταση, οι κόμβοι αντιπροσωπεύονται από τμήματα κειμένου, αλλά οι κόμβοι πεδίων που υποδηλώνουν τον τύπο κόμβου χρησιμοποιούνται για την προετοιμασία κάθε DAG. Επιπλέον, ορίζονται δύο είδη άκρων:

1. Άκρα που ομαδοποιούν τμήματα που ανήκουν στην ίδια κατηγορία.
2. Άκρα που συνδέουν τις σχέσεις μεταξύ διαφορετικών ομάδων.

Για τη χωρική κωδικοποίηση του κειμένου χρησιμοποιείται ένας μετασχηματιστής με πρόσθετη δισδιάστατη ενσωμάτωση θέσης. Μετά από αυτό, η εργασία γίνεται πρόβλεψη της μήτρας σχέσεων για κάθε τύπο άκρου. Αυτή η μέθοδος μπορεί να αντιπροσωπεύει αυθαίρετα βαθιές ιεραρχίες και μπορεί να εφαρμοστεί σε περίπλοκες διατάξεις εγγράφων.

Πίνακες

Η εξαγωγή δεδομένων σε μορφή πίνακα παραμένει μια απαιτητική πτυχή της εξαγωγής πληροφοριών λόγω της μεγάλης ποικιλίας μορφών και των σύνθετων ιεραρχιών τους. Τα σύνολα

δεδομένων πίνακα έχουν συνήθως πολλαπλές εργασίες προς εκτέλεση. Η πρώτη εργασία είναι ο εντοπισμός πίνακα που περιλαμβάνει τον εντοπισμό του πλαισίου οριοθέτησης που περιέχει τους πίνακες στο έγγραφο. Η επόμενη εργασία είναι η αναγνώριση δομής πίνακα, η οποία απαιτεί την εξαγωγή των πληροφοριών γραμμής, στήλης και κελιού σε κοινή μορφή. Με αυτό μπορεί να γίνει ένα βήμα παραπέρα στην αναγνώριση του πίνακα, πράγμα που απαιτεί κατανόηση τόσο των δομικών πληροφοριών όσο και του περιεχομένου ταξινομώντας τα μέρη στον ίδιο τον πίνακα. Καθώς τα χαρακτηριστικά κειμένου είναι εξίσου σημαντικά για την ορθή εξαγωγή και κατανόηση των πινάκων, έχουν προταθεί πολλές διαφορετικές μέθοδοι για την εκτέλεση αυτής της εργασίας.

Μία τέτοια πρόταση με την ονομασία TableSense εκτελεί ανίχνευση τράπεζας και αναγνώριση δομής. Το TableSense χρησιμοποιεί μια προσέγγιση τριών σταδίων: Τη λειτουργία κυττάρων, την ανίχνευση αντικειμένων με συμβολικά μοντέλα και ένα μηχανισμό δειγματοληψίας ενεργής μάθησης που βασίζεται στην αβεβαιότητα. Η προτεινόμενη αρχιτεκτονική του TableSense για την ανίχνευση τράπεζας είναι σημαντικά καλύτερη από τις παραδοσιακές μεθόδους στην όραση υπολογιστή, όπως YOLO-v3 ή Mask R-CNN.

2.3.6 Συμπεράσματα

Η κατανόηση των εγγράφων είναι ένα φλέγον θέμα στη βιομηχανία και έχει τεράστια χρηματική αξία. Τα περισσότερα έγγραφα είναι ιδιωτικά δεδομένα που αντιστοιχούν σε ιδιωτικές συμβάσεις, τιμολόγια και αρχεία. Ως εκ τούτου, τα ανοικτά διαθέσιμα σύνολα δεδομένων είναι δύσκολο να προέρχονται και δεν έχουν αποτελέσει επίκεντρο της πανεπιστημιακής Κοινότητας σε σχέση με άλλους τομείς εφαρμογής. Η ακαδημαϊκή βιβλιογραφία σχετικά με τις μεθοδολογίες για την αντιμετώπιση της κατανόησης των εγγράφων είναι εξίσου σπάνια σε σχέση με τομείς με πληθώρα διαθέσιμων στο κοινό δεδομένων, όπως η ταξινόμηση και η μετάφραση των εικόνων. Ωστόσο, οι πιο αποτελεσματικές προσεγγίσεις για την κατανόηση των εγγράφων αξιοποιούν τις πρόσφατες εξελίξεις στη μοντελοποίηση του βαθιού νευρονικού δικτύου. Η ολοκληρωμένη κατανόηση των εγγράφων είναι εφικτή με τη δημιουργία ενός ολοκληρωμένου συστήματος που εκτελεί ανάλυση διάταξης, οπτική αναγνώριση χαρακτήρων και εξαγωγή πληροφοριών για το συγκεκριμένο τομέα. Σε αυτήν την έρευνα, προσπαθούμε να παγιώσουμε και να οργανώσουμε τις μεθοδολογίες που υπάρχουν στη βιβλιογραφία, προκειμένου να είναι ένα σημείο αναφοράς τόσο για τους ακαδημαϊκούς όσο και για τους επαγγελματίες που επιθυμούν να διερευνήσουν την κατανόηση των εγγράφων.

Κεφάλαιο 3 ο

3 Tesseract

3.1 Γενικά

Σκοπός του κεφάλαιού αυτού είναι η παρουσίαση της εφαρμογής που δημιουργήθηκε για τις ανάγκες της παρούσης διατριβής και η παρουσίαση του Tesseract. Στην προαναφερθείσα εφαρμογή χρησιμοποιήθηκε για τις ανάγκες εκτέλεσης αναγνώρισης χαρακτήρων η βιβλιοθήκη της Tesseract η οποία είναι διαθέσιμη σε διάφορα λειτουργικά σύστημα (Windows, Linux, Mac OS). Διατίθεται ως ελεύθερο λογισμικό υπό την άδεια Apache έκδοση 2.0 και από το 2006 συντηρείται από την Google. Αποτελεί μια από τις πλέον ακριβείς μηχανές αναγνώρισης οπτικών χαρακτήρων και έχει ευρεία χρήση σε πλήθος εφαρμογών. Το tesseract μπορεί να χρησιμοποιηθεί μέσο περιβάλλοντος κονσόλας, μέσο γραφικού περιβάλλοντος αλλά και μέσο API. Το τελευταίο χρησιμοποιήθηκε για την παρούσα διπλωματική εργασία.

Στο κεφάλαιο αυτό ο αναγνώστης θα διαβάσει για μια σύντομη αναφορά για την ιστορική πορεία του Tesseract, της βασική του αρχιτεκτονική, τα κύρια γνωρίσματά της λειτουργίας του. Στην συνέχεια παραθέτονται βήμα το βήμα μια πρόταση για εκπαίδευση του Tesseract.

Επίσης, σε αυτό το κεφάλαιο αναπτύσσεται με εικόνες οθόνης η δομή της εφαρμογής τόσο στο τμήμα της που αφορά το backend όσο και αυτό που αφορά το Frontend όπως και η ίδια η εφαρμογή κατά της λειτουργία της και την εκτέλεση της αναγνώρισης χαρακτήρων.

3.1.1 Ιστορική Αναδρομή

Η κατασκευή του λογισμικού ξεκίνησε το 1984 ως διδακτορική διατριβή υπό την χορηγία της Hewlett Packard στο Μπρίστολ του Ηνωμένου Βασιλείου. Το 1987 προστέθηκε και δεύτερο άτομο το οποίο συμμετείχε στην ανάπτυξη του λογισμικού στις εγκαταστάσεις της Hewlett Packard στο Γκρήλεϋ του Κολοράντο στις ΗΠΑ και ξεκίνησαν οι πρώτες προσπάθειες για την εμπορική αξιοποίηση του ως λειτουργία για συσκευές σαρωτών. Το 1995 αξιολογήθηκε ως μια από τις κορυφαίες μηχανές οπτικής αναγνώρισης χαρακτήρων ως προς την ακρίβεια ανάγνωσης, το 1996 δημιουργήθηκε έκδοση για Windows, ενώ το 1998 ξαναγράφτηκε τμήμα του παλαιού κώδικα ο οποίος ο οποίος ήταν γραμμένος σε C στην C++. Κατόπιν η ανανέωση συνεχίστηκε εν μέρει έτσι ώστε να είναι δυνατό όλος ο κώδικας να μεταγλωττιστεί με χρήση μεταγλωττιστή C++.

Από το 1999 δεν υπήρξε καμία δραστηριότητα και δραστηριοποίηση και πέρασε τελείως στην αφάνεια, έως το 2005 όταν η Hewlett Packard σε συνεργασία με το πανεπιστήμιο της Νεβάδα, Λάς Βέγκας αποφάσισε να το διαθέσει υπό ελεύθερη άδεια χρήσης, ενώ από το 2006 την συντήρηση και επέκταση του λογισμικού ανέλαβε η Google (ως έκδοση 1) και έλαβε θετική μετέπειτα αποδοχή ως προς την ποιότητα των αποτελεσμάτων του.

3.1.2 Χαρακτηριστικά

Όλες οι εκδόσεις του λογισμικού είναι διαθέσιμες για τα λειτουργικά συστήματα Linux, Windows, Mac OS. Η ποιότητα των αποτελεσμάτων του εξαρτάται σημαντικά από την ποιότητα και ευκρίνεια των εικόνων τα οποία λαμβάνει ως είσοδο, και προβλήματα όπως ημιπεριστραμένες ή χαμηλής φωτεινότητας σελίδες χρειάζονται ξεχωριστή επεξεργασία πριν ξεκινήσει η διαδικασία αναγνώρισης χαρακτήρων.

Το tesseract μπορεί να χρησιμοποιηθεί μέσο περιβάλλοντος κονσόλας, μέσο γραφικού περιβάλλοντος αλλά και μέσο API. Το τελευταίο χρησιμοποιήθηκε για την παρούσα διπλωματική εργασία

3.1.3 Αρχιτεκτονική Tesseract

Γενικά, η επεξεργασία που ακολουθεί το Tesseract παρουσιάζει χαρακτηριστικά ενός αγωγού δεδομένων και αντίστοιχα η επεξεργασία γίνεται βήμα το βήμα. Ωστόσο ακόμη και σήμερα μερικά από τα στάδια αυτής παραμένουν ασυνήθιστα. Το πρώτο βήμα της επεξεργασίας αποτελεί η διασυνδεδεμένη ανάλυση των δεδομένων. Σε αυτή παράγεται ένα περίγραμμα το οποίο και αποθηκεύεται. Αυτή η μέθοδος ανίχνευσης των εμφωλευμένων

περιγραμμάτων έχει ένα σημαντικό πλεονέκτημα: Είναι πιο εύκολη η αναγνώριση των ανεστραμμένων ή των ασπρόμαυρων κείμενων. Να γιατί το Tesseract ήταν η πρώτη OCR μηχανή που είχε την δυνατότητα να αναγνωρίζει ασπρόμαυρους χαρακτήρες.
Σε αυτό το πρώτο στάδιο τα παραγόμενα περιγράμματα στοιχίζονται σε Blobs.

Τα Blobs οργανώνονται σε γραμμές και περιοχές. Οι γραμμές κειμένου χωρίζονται σε λέξεις ανάλογα με το είδος της απόστασης των χαρακτήρων. Δημιουργούνται κελιά χαρακτήρων και οι λέξεις χωρίζονται με τα κενά και κάποιους πιο ασαφείς χώρους.

Στην συνέχεια, σειρά έχει η επεξεργασία διπλής σάρωσης. Στη πρώτη σάρωση αναγνωρίζεται η κάθε λέξη ξεχωριστά με την σειρά που εμφανίζεται. Εάν η λέξη πληρεί κάποιά ποιοτικά κριτήρια ταξινομείται ως τέτοια και αποθηκεύεται στα δεδομένα εκπαίδευσης (test data). Έτσι το Tesseract έχει την δυνατότητα να αναγνωρίζει τις ποιοτικές απεικονίσεις σε επόμενες λέξεις στην συνέχεια της επεξεργασίας.

3.1.4 Εύρεση γραμμών και λέξεων

Ο λογάριθμος εύρεσής γραμμών έχει σχεδιαστεί έτσι ώστε να μπορεί να αναγνωριστεί μια σειρά λέξεων χωρίς να πρέπει να αποσυντεθεί η σελίδα έτσι βελτιώνεται το τελικό αποτέλεσμα. Ένα από τα βασικά κομμάτια αυτής της διαδικασίας είναι το φιλτράρισμα των Blobs και η κατασκευή γραμμών. Υποθέτοντας ότι η ανάλυση για την διάταξη της σελίδας μας προσφέρει περιοχές κειμένου με ομοιόμορφο μέγεθος κειμένου το φίλτρο ύψους μπορεί να ελαχιστοποιήσει κάθετα και οριζόντια από πάνω την απόσταση των γραμμών με τα γράμματα.

Τελικό βήμα αποτελεί η συγχώνευση των Blobs που επικαλύπτονται. Σε αυτή την περίπτωση το Tesseract τοποθετεί διακριτικά σημάδια και μια κατάλληλη βάση στους μη καθαρούς χαρακτήρες.

3.1.5 Αναγνώριση λέξεων

Βασικό για την επίτευξη της αναγνώρισης των λέξεων είναι να ανιχνευτεί αν μια λέξη μπορεί να αποσυντεθεί στους χαρακτήρες που την αποτελούν. Έτσι η διάσπαση των λέξεων ιεραρχείται πρώτη.

3.1.6 Διόρθωση χαρακτήρων

Όταν η ποιότητα μιας λέξης δεν είναι ικανοποιητική, το Tesseract προσπαθεί να βελτιώσει το αποτέλεσμα. Αυτό συμβαίνει αποσπώντας ένα Blob που συγκεντρώνει την μικρότερη εμπιστοσύνη από

τον ταξινομήτη των χαρακτήρων(character classifier). Στην συνέχεια επιλέγονται υποψήφια σημεία κοπής από διάφορες κορυφές και εκτελείται ο διαχωρισμός του ενός χαρακτήρα από τον άλλο. Αυτή η διαδικασία εκτελείται με σειρά προτεραιότητας και αν κατά την διάρκεια της κοπής δεν βεβαιωθεί η ποιότητα του χαρακτήρα τότε αυτή αναιρείται.

3.2 Ταξινόμηση στατικών χαρακτήρων

Η βασική ιδέα σε αυτό το στάδιο είναι ότι οι προβληματικοί χαρακτήρες δεν χρειάζεται να ταυτίζονται πλήρως με τα δεδομένα εκπαίδευσης για να ταξινομηθούν. Κάθε χαρακτήρας ταξινομείται ξεχωριστά κατά προσέγγιση με ένα χαρακτήρα από τα δεδομένα εκπαίδευσης, αυτό γίνεται με την βοήθεια ενός πολυγωνικού περιγράμματος για κάθε χαρακτήρα.

Η ταξινόμηση χαρακτήρων πραγματοποιείται σε δύο βήματα. Κάθε χαρακτηριστικό παίρνει από ένα τρισδιάστατο πίνακα αναζήτησης ένα διάνυσμα που ταιριάζει κατά προσέγγιση. Τα διανύσματα με την καλύτερη ποιότητα συγκροτούνται σε μια λίστα. Κάθε χαρακτηριστικό ενός αγνώστου χαρακτήρα που προσεγγίζει ένα πρωτότυπο διάνυσμα μιας δοσμένης κλάσης η οποία ενδέχεται να ταιριάζει, στην συνέχεια υπολογίζεται η πραγματική ομοιότητα μεταξύ τους. Κάθε κλάση πρωτότυπου χαρακτήρα αναπαρίσταται από μια λογική έκφραση , κάθε τέτοιος όρους ονομάζεται configuration, έτσι η διαδικασία υπολογισμού της απόστασης αποθηκεύει στοιχεία για την συνολική ομοιότητα για κάθε χαρακτήρα σε κάθε configuration, όπως επίσης και σε κάθε πρωτότυπο. Ο καλύτερος δυνατή απόσταση η οποία υπολογίζεται από το άθροισμα των χαρακτηριστικών και των πρωτότυπων είναι η συνολικά το καλύτερο configuration της κλάσης.

Ενδιαφέρον παρουσιάζει η λειτουργία των δεδομένων εκπαίδευσης (training data). Αν και ο ταξινομιτής είναι σε θέση να αναγνωρίζει εύκολα τους θολούς χαρακτήρες δεν μπορεί να εκπαιδευτεί για τους επόμενους.

3.3 Λεκτική ανάλυση

Το Tesseract εκτελεί μια βραχύβια λεκτική ανάλυση. Όποτε ο μηχανισμός αναγνώρισης λέξεων εντοπίζει ένα νέο απόσπασμα, ο μηχανισμός λεκτικής ανάλυσης διαλέγει την πιο κοντινή λέξη με βάση τα παρακάτω κριτήρια: Αν είναι συχνή λέξη, κορυφαία λέξη λεξιλογίου, κορυφαία αριθμητική λέξη, κορυφαία λέξη με κεφαλαίους χαρακτήρες, κορυφαία λέξη με μικρούς χαρακτήρες. Η τελική απόφαση για το δοσμένος απόσπασμα είναι απλά η λέξη με την μικρότερη απόσταση.

Λέξεις από διαφορετικά αποσπάσματα μπορεί να διαφέρουν στον αριθμό των χαρακτήρων που περιέχουν. Είναι δύσκολο να συγκριθούν άμεσα αυτές οι λέξεις. Το πρόβλημα

αυτό λύνεται με την παράγωγή δύο αριθμών που βοηθούν την ταξινόμηση χαρακτήρων. Ο πρώτος ονομάζεται confidence και ο δεύτερος rating.

3.4 Εκπαίδευση Tesseract

Η εκπαίδευση του Tesseract είναι μια αναγκαία διαδικασία η οποία μπορεί να βελτιστοποιήσει κατά πολύ τα αποτελέσματα της οπτικής αναγνώρισης λεκτικών χαρακτήρων. Για την επίτευξη της εκπαίδευσης μπορούν να χρησιμοποιηθούν μια σειρά από προγράμματα που διευκολύνουν την διαδικασία αυτή. Στην παρούσα εργασία θα παρουσιαστούν τα βασικά βήματα με το jTessBoxEditor.

Παρόλο που το Tesseract έχει προ εγκατεστημένες αρκετές γλώσσες, είναι αναγκαίο να εκπαιδεύσουμε το πρόγραμμα για την συγκεκριμένη μορφή κειμένου που πρόκειται να αντιμετωπίσει, για αυτό υπάρχουν δύο τρόποι. Ο πρώτος είναι να συλλεχθούν εικόνες από το μορφή κειμένου που θα αναγνωριστεί από το Tesseract και ο δεύτερος είναι να γραφτεί σε ένα οποιοδήποτε πρόγραμμα συγγραφής κειμένου (λ.χ. Word) η μορφή κειμένου που πρόκειται να αναγνωριστεί και στη συνέχεια να τραβηχτούν στιγμιότυπα οθόνης (print Screens) από αυτό.

Γενικά, το βασικό περίγραμμα της εκπαίδευσης του Tesseract είναι το εξής:

1) Συγχώνευση δεδομένων εκπαίδευσης στο αρχείο .tiff με την βοήθεια του jTessBoxEditor.
2) Δημιουργία εκπαιδευτικών σημάνσεων, με την κατασκευή αρχείων .box που περιέχουν προβλέψεις για το Tesseract από το .tiff αρχείο και τυχών διορθώσεις ακρίβειας.
3) Εκπαίδευση Tesseract.

Συγκεκριμένα:

Βήμα 1ο : <u>Συγχώνευση δεδομένων εκπαίδευσης</u>

Μετά την συλλογή κάποιων δεδομένων, ανοίγουμε το jTessBoxEditor ώστε να κατασκευάζουμε το αρχείο .tiff, εισάγουμε τις εικόνες της μορφής κειμένου επιλέγουμε όνομα στο αρχείο και πατάμε οκ.

Βήμα 2ο : <u>Δημιουργία εκπαιδευτικών σημάνσεων</u>

Ανοίγουμε το terminal, οδηγούμαστε μέσω αυτού στο αρχείο που περιέχει τις εικόνες και το αρχεία .tiff . Το jTessBoxEditor μπορεί να μας βοηθήσει αρκετά στην διαδικασία αυτή χωρίς να χρειάζεται να επισημάνουμε κάθε μια εικόνα ξεχωριστά. Τρέχουμε την εντολή:

```
tesseract --psm 6 --oem 3 font_name.font.exp0.tif font_name.font.exp0 makebox
```

Η εμφάνιση του psm υποδηλώνει την «πρόθεση» για μοντέλο τμηματοποίησης. Δηλαδή, το πως το Tesseract θα χειριστεί την εικόνα, αν θέλουμε το Tesseract να «βλέπει» κάθε λέξη ως οντότητα τότε πρέπει να προσθέσουμε στην παραπάνω εντολή psm 8.

Χρησιμοποιώντας την παραπάνω εντολή, το παραγόμενο αποτέλεσμα είναι η δημιουργία πλαισίων και προβλέψεων βελτιστοποίησης στις δοσμένες εικόνες τα οποία αποθηκεύονται σε .box αρχεία. Τέλος εάν ανοίξουμε το jTessBoxEditor και ανοίξουμε το .tiff αρχείο θα παρατηρήσουμε ότι κάθε εικόνα έχει τα πλαίσια που προαναφέρθηκαν. Βεβαίως, αν θέλουμε το τελικό αποτέλεσμα της αναγνώρισης κειμένου να είναι καλύτερο θα πρέπει να διορθώσουμε ένα-ένα τα παραγόμενα πλαίσια στις εικόνες.

Βήμα 3ο : Εκπαίδευση Tesseract

Αφού έχουμε δημιουργήσει τα αρχεία .tiff και .box πρέπει να δημιουργήσουμε ένα νέο αρχείο text το οποίο θα περιέχει:

```
font 0 0 0 0 0
```

Το αποθηκεύουμε στο ίδιο φάκελο με τα υπόλοιπα αρχεία, συνεπώς η εικόνα του φακέλου αυτού θα είναι:

Τρέχουμε τις παρακάτω εντολές:

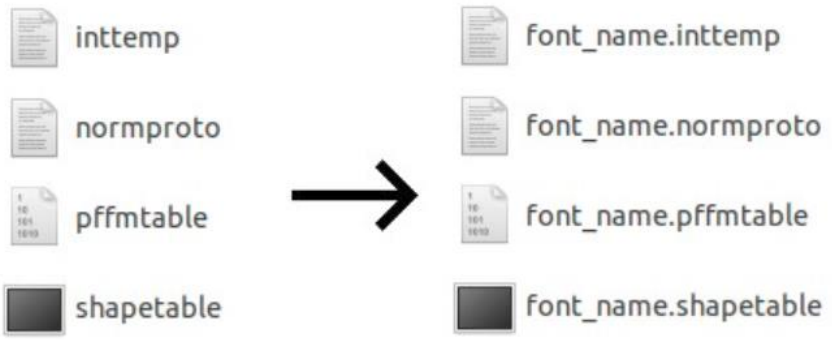
```
# Create a .tr file (training file)
tesseract num.font.exp0.tif font_name.font.exp0 nobatch box.train


# Create a unicharset file
unicharset_extractor font_name.font.exp0.box


# Create a shapetable file
shapeclustering -F font_properties -U unicharset -O
font_name.unicharset font_name.font.exp0.tr


# Create a pffmtable, intemp file
mftraining -F font_properties -U unicharset -O font_name.unicharset
font_name.font.exp0.trecho Clustering..


# Create a normproto file
cntraining font_name.font.exp0.tr
```

Σε αυτό το σημείο είναι πιθανό να αντιμετωπίσουμε κάποια σφάλματα, σε αυτή την περίπτωση πρέπει να τρέξουμε τα εκτελέσιμα (για windows χρήστες) tesseract.exe, unicharset_extractor.exe, cntraining.exe . Θα παρατηρήσουμε ότι στο terminal κάποια παραγόμενα.

Αν εικόνες μας περιέχουν όλους τους αναγκαίους χαρακτήρες τότε θα παρατηρήσουμε ότι ο αριθμός σχημάτων είναι ίσος με τον αριθμό των κλάσεων που χρειαζόμαστε. Για παράδειγμα, εάν θέλουμε το Tesseract να είναι σε θέση να αναγνωρίζει σωστά ψηφία αριθμών τότε θα πρέπει ο αριθμός σχημάτων να ισούται με 10 δηλαδή όσα και τα ψηφία των αριθμών.

Αν ο αριθμός των σχημάτων δεν ισούται με τον αριθμό των κλάσεων τότε πρέπει να ξαναγυρίσουμε στην διαδικασία της εκπαίδευσης και να δημιουργήσουμε πιο «καθαρά» δεδομένα εκπαίδευσης.

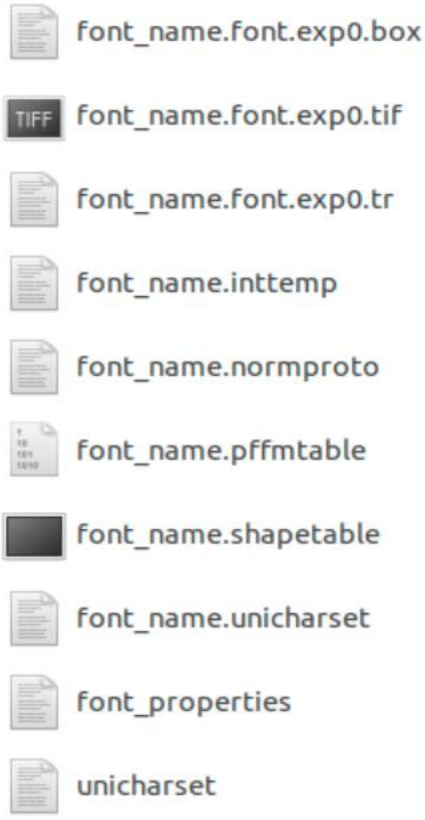
Αλλιώς αν η διαδικασία έχει ολοκληρωθεί ομαλά, θα υπάρχουν τέσσερις παραγόμενοι φάκελοι, shapetable, normproto, intemp και pffmtable. Μετονομάζουμε αυτούς με το πρόθεμα font_name και έχουμε το εξής αποτέλεσμα:



Στην συνέχεια τρέχουμε την εντολή:

```
combine_tessdata font_name.
```

Η νέα εικόνα του φακέλου είναι:



Αντιγράφουμε το αρχείο font_name.trainddata στην τοποθεσία /tessdata.

Συμπεράσματα από την εκπαίδευση:

Η παραπάνω διαδικασία επιλέχθηκε ως η απλούστερη που μπορεί να εφαρμοστεί για την εκπαίδευση αναγνώρισης κειμένου από το Tesseract. Επίσης, αφορούσε την εκπόνηση της αναγνώρισης μιας απλής γραμμής λέξεων σε απλή γραμματοσειρά. Στην δική μας περίπτωση προσπαθούμε να αναγνωρίσουμε τα πεδία μια ταυτότητας η οποία περιέχει πολύ «θόρυβο», δηλαδή αρκετές επικαλύψεις γραμμάτων με εικόνες ή γράμματα παρασκηνίων, πολλές διαφορετικές γραμματοσειρές ταυτόχρονα κτλ. Αυτά τα χαρακτηρίστηκα της ταυτότητας δυσχεραίνουν κατά πολύ την επίτευξη ακρίβειας στο τελικό αποτέλεσμα. Συμπερασματικά, διαφαίνεται πως το Tesseract μπορεί να εκτελέσει με εξαιρετική ταχύτητα την αναγνώριση κειμένου, αλλά η αναγνώριση πιο συνθέτης εικόνας κειμένου απαιτεί άλλου τύπο εκπαίδευσης. Επίσης παρατηρήθηκε ότι για την εκπαίδευση σε απλό κείμενο η αύξηση του πλήθους των εικόνων εκπαίδευσης δεν συνεπάγονταν και καλύτερη ποιότητα κατά αναλογία, ίσως συνέβαινε και το αντίθετο κατά περιπτώσεις.

3.5 Πρόγραμμα που εφαρμόζει την τεχνολογία Tesseract

Παρακάτω θα παρουσιαστεί ένα απλό πρόγραμμα το οποίο χρησιμοποίει την τεχνολογία Tesseract. Η διαδικασία που εκτελεί είναι να απαθανατίσει την ταυτότητα του χρήστη μέσο της κάμερας του υπολογιστή και στην συνέχεια να την επεξεργαστεί και να αποσπάσει κάποια στοιχεία του χρήστη (όπως ο αριθμός ταυτότητας), φυσικά εδώ είναι το σημείο του επεμβαίνει το Tesseract. Επίσης θα παρουσιαστούν με φωτογραφίες τα βασικά σημεία του κώδικα που αναπτύχθηκε όπως και κάποια αποτελέσματα.

Εργαλεία που χρησιμοποιήθηκαν

- Spring Boot java framework (https://spring.io/projects/spring-boot)
- Java 8 (https://www.oracle.com/java/technologies/java8.html)
- Angular 8 (https://angular.io/)
- Nodejs (https://nodejs.org/en/)
- Maven 3.3.6 (https://maven.apache.org/download.cgi)
- IntelliJ (https://www.jetbrains.com/)
- WebStorm (https://www.jetbrains.com/)
- Tesseract 4 (https://github.com/tesseract-ocr/)

Παρουσίαση εφαρμογής

Παρακάτω θα περιγράφουν κάποια βασικά σημεία του κώδικα που παράχθηκε. Όπως περιεγράφηκε συνοπτικά στα εργαλεία, το πρόγραμμα γράφηκε σε γλώσσα Java, και συγκεκριμένα χρησιμοποιήθηκε ένας συγκεκριμένη βιβλιοθήκη/ δομή (framework) της Java, το Spring Boot. Τα τελευταία αποτελούν το πηγαίο κώδικα του server της εφαρμογής. Για τις ανάγκες του frontend χρησιμοποιήθηκε τελευταία έκδοση της angular η οποία αποτελεί μια βιβλιοθήκη/δομή (framework) της JavaScript.

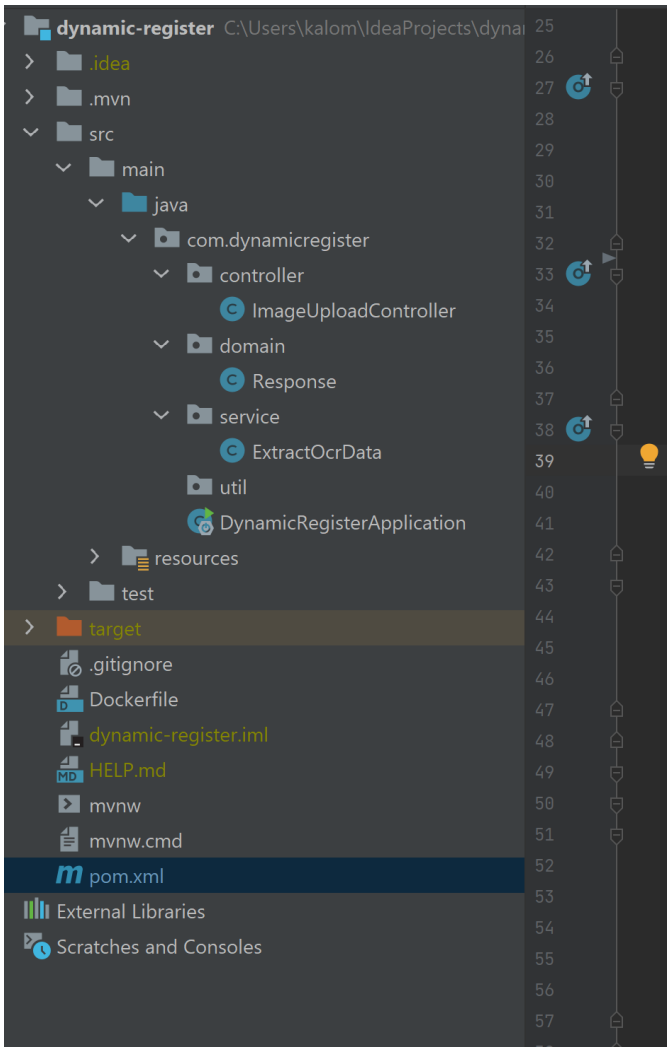
Για τις ανάγκες της διαχείρισης των βιβλιοθηκών, του χτισίματος (Built) στην πλευρά του server, χρησιμοποιήθηκε το maven. Παρακάτω παρουσιάζεται ένα τμήμα από το βασικό αρχείο (pom.xml) για την ρύθμιση του maven.

```
            <optional>true</optional>
        </dependency>
        <dependency>
            <groupId>org.projectlombok</groupId>
            <artifactId>lombok</artifactId>
            <optional>true</optional>
        </dependency>
        <dependency>
            <groupId>org.springframework.boot</groupId>
            <artifactId>spring-boot-starter-test</artifactId>
            <scope>test</scope>
        </dependency>
        <dependency>
            <groupId>net.sourceforge.tess4j</groupId>
            <artifactId>tess4j</artifactId>
            <version>4.4.1</version>
        </dependency>
    </dependencies>
    <dependencyManagement>
        <dependencies>
            <dependency>
                <groupId>com.azure.spring</groupId>
```

Εικόνα 4: Αρχείο POM της εφαρμογής που χρησιμοποίει το Tesseract.

Στην παραπάνω εικόνα παρουσιάζεται το αρχείο pom που περιέχει η εφαρμογή μας. Άξιο παρατήρηση αποτελεί το γεγονός ότι μέσο του maven μπορούμε να έχουμε διαθέσιμο και το Tesseract, είναι ευδιάκριτο στην παραπάνω εικόνα το dependency που περιέχει το tess4j.

Επίσης, η αρχιτεκτονική της εφαρμογής ακολουθεί το MVC design Pattern, παρατίθεται στην παρακάτω εικόνα:



Εικόνα 5: Η δομή των αρχείων της εφαρμογής που χρησιμοποιεί το Tesseract.

Βασικό τμήμα της εφαρμογής όσο αφορά το backend αποτελούν δύο κλάσεις, η ImageUploadController και η ExtractOcrData. Η πρώτη διαχειρίζεται την διασύνδεση του server με το frontend τμήμα της εφαρμογής ενώ η δεύτερη το πιο νευραλγικό της τμήμα, την επεξεργασία με το tesseract της εισερχόμενης εικόνας από το frontend. Παρακάτω παρατίθενται οι αντίστοιχες εικόνες του κώδικα:

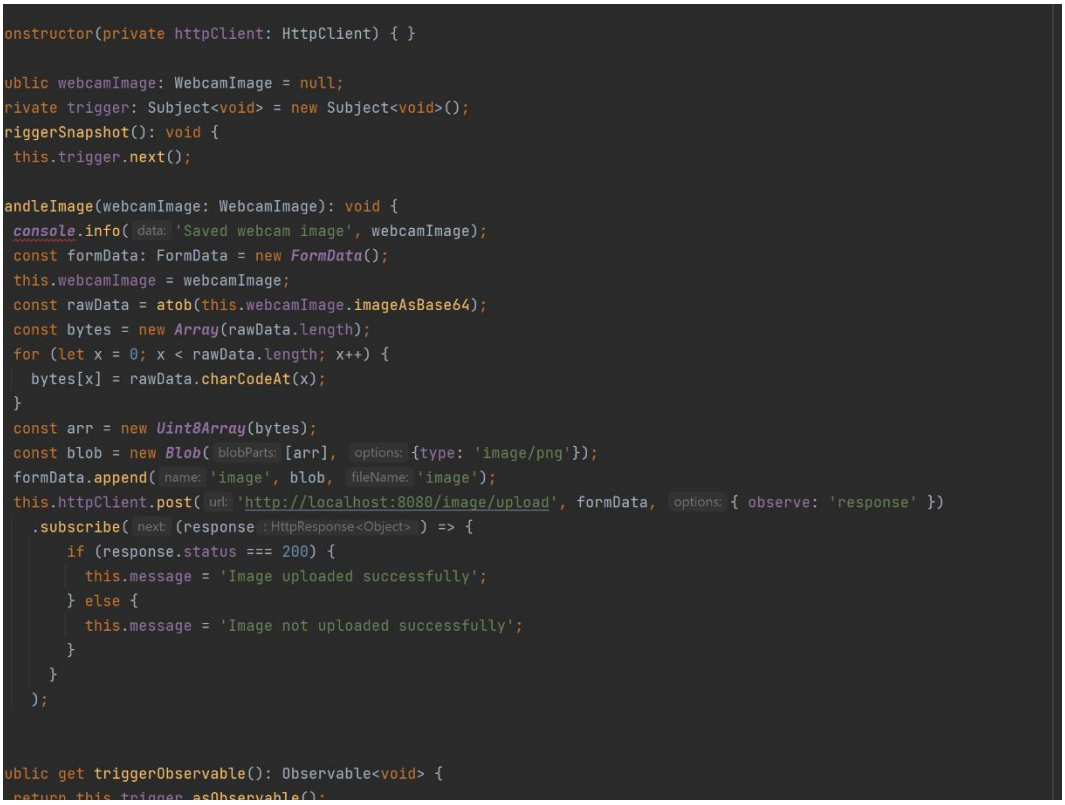
```java
@RestController
@CrossOrigin(origins = "http://localhost:4200")
@RequestMapping(path = "image")
public class ImageUploadController {

    private static final Logger LOGGER = LoggerFactory.getLogger(ImageUploadController.class);

    private final ExtractOcrData extractOcrData;

    public ImageUploadController(ExtractOcrData extractOcrData) { this.extractOcrData = extractOcrData; }

    @PostMapping("/upload")
    public ResponseEntity.BodyBuilder uploadImage(@RequestParam("image") MultipartFile file) throws IOException {
        String strippedText = null;
        try {
            strippedText = extractOcrData.extractTextFromScannedDocument(file);
            LOGGER.info(strippedText);
        } catch (TesseractException e) {
            e.printStackTrace();
        }
        LOGGER.info(strippedText);

        return ResponseEntity.status(HttpStatus.OK);
```

Εικόνα 6: Η κλάση που περιέχει τα end points της εφαρμογής του χρησιμοποιεί το Tesseract.

```
ervice
blic class ExtractOcrData {

  public String extractTextFromScannedDocument(MultipartFile file) throws IOException, TesseractException {

      StringBuilder out = new StringBuilder();

      ITesseract tess = new Tesseract();
      tess.setDatapath("C:\\Program Files\\Tesseract-OCR\\tessdata");
      tess.setLanguage("ell"); // choose your language

      File temp = File.createTempFile( prefix: "tempfile_", suffix: ".png");
      ImageIO.write(ImageIO.read(new ByteArrayInputStream(file.getBytes())), formatName: "png", temp);

      String result = tess.doOCR(temp);

      // Delete temp file
      temp.delete();

      return result;

  }
```

Εικόνα 7: Η κλάση στην οποία εφαρμόζεται το OCR.

Εστιάζοντας στο τμήμα της εφαρμογής που αφορά το frontend, το σημαντικότερο κομμάτι της αποτελεί το αρχείο app.component.ts στο οποίο λαμβάνεται η εικόνα από την κάμερα του υπολογιστή μορφοποιείται και στέλνεται στον server για να γίνει η επεξεργασία με το Tesseract, παρακάτω παρατίθεται η σχετική εικόνα από τον κώδικα:



```
onstructor(private httpClient: HttpClient) { }

ublic webcamImage: WebcamImage = null;
rivate trigger: Subject<void> = new Subject<void>();
riggerSnapshot(): void {
 this.trigger.next();

andleImage(webcamImage: WebcamImage): void {
 console.info( data: 'Saved webcam image', webcamImage);
 const formData: FormData = new FormData();
 this.webcamImage = webcamImage;
 const rawData = atob(this.webcamImage.imageAsBase64);
 const bytes = new Array(rawData.length);
 for (let x = 0; x < rawData.length; x++) {
   bytes[x] = rawData.charCodeAt(x);
 }
 const arr = new Uint8Array(bytes);
 const blob = new Blob( blobParts: [arr], options: {type: 'image/png'});
 formData.append( name: 'image', blob, fileName: 'image');
 this.httpClient.post( url: 'http://localhost:8080/image/upload', formData, options: { observe: 'response' })
   .subscribe( next: (response : HttpResponse<Object> ) => {
       if (response.status === 200) {
         this.message = 'Image uploaded successfully';
       } else {
         this.message = 'Image not uploaded successfully';
       }
     }
   );


ublic get triggerObservable(): Observable<void> {
 return this.trigger.asObservable();
```

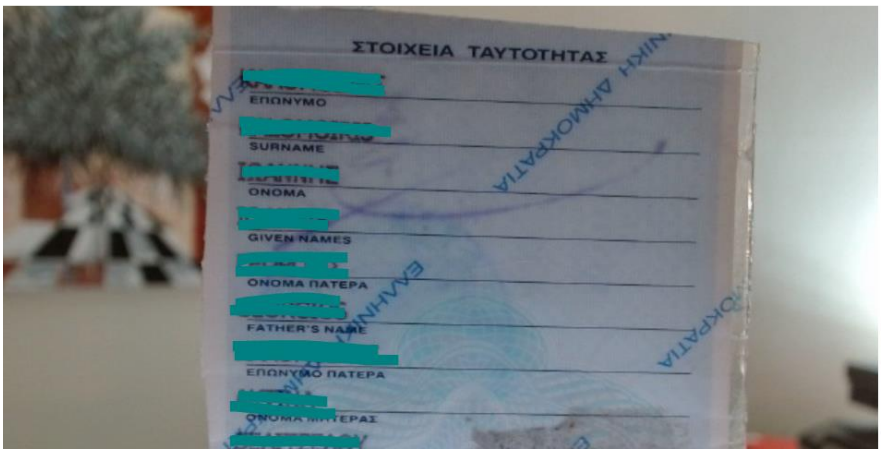
Εικόνα 8: Το service (frontend) που εφαρμόζει τις κλίσεις προς τον server.

Ροή εφαρμογής

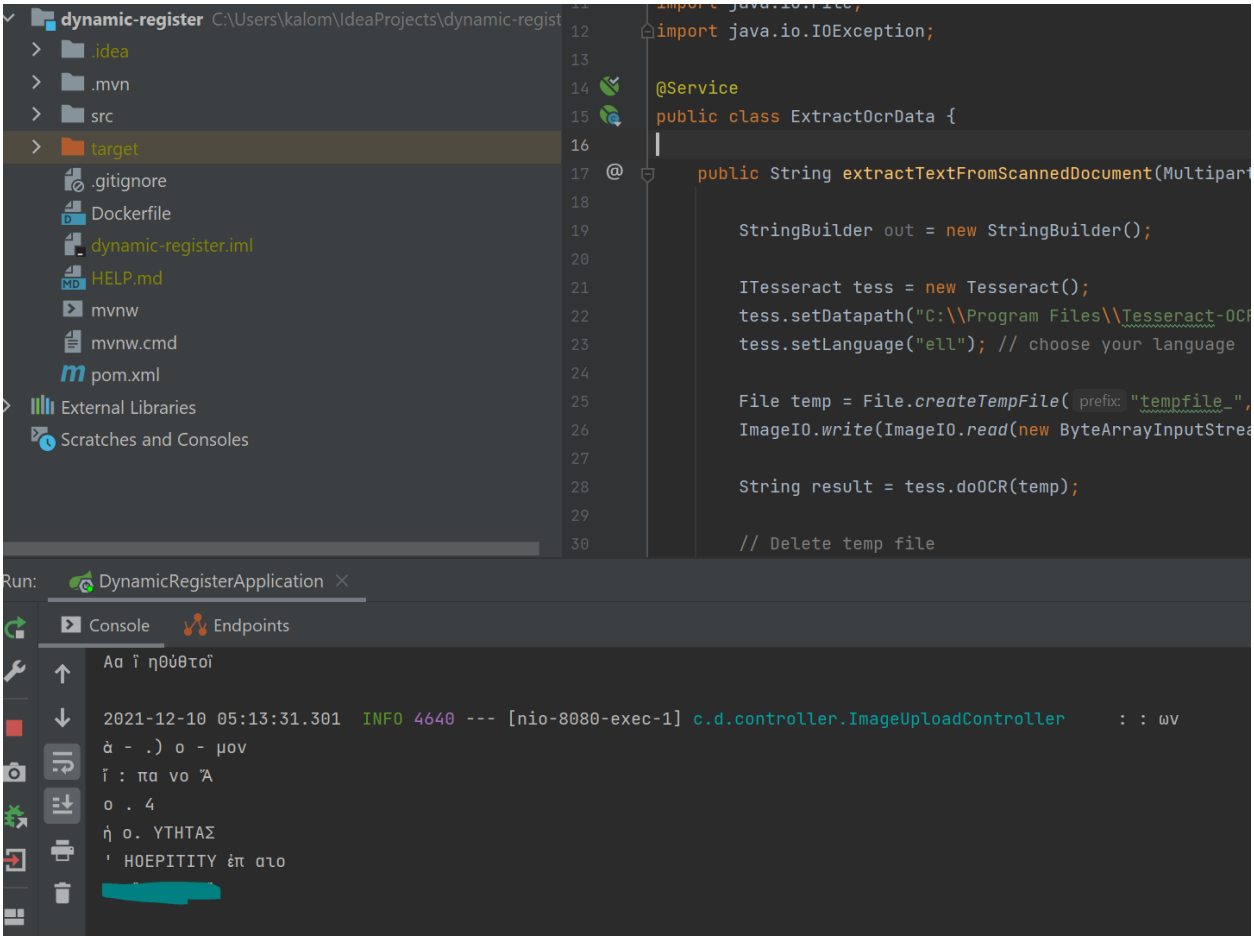
Η εφαρμογή που έχει αναπτυχθεί, όπως έχει ειπωθεί προηγουμένως, με σκοπό ύπαρξης να αποσπάσει κάποια από τα στοιχεία ταυτότητας του χρήστη μέσο μια απλής φωτογραφίας, το αποτέλεσμα της επεξεργασίας από το Tesseract δεν έχει τα αποτελέσματα που επιθυμούμε ή καλύτερα για να έχει τα αποτελέσματα που επιθυμούμε θα χρειάζονταν πολύ παραπάνω προσπάθεια ώστε να ρυθμίσουμε και να εμπλουτίσουν την εφαρμογή με τέτοιο τρόπο ώστε να είναι σε θέση να το κάνει. Στο παρακάτω στιγμιότυπο οθόνης παρατηρούμε πως τα αποτελέσματα του OCR στην κονσόλα είναι αρκετά συγκεχυμένα, παρόλα αυτά, μπορεί να εξαχθεί ο αριθμός ταυτότητας. Παρακάτω παρατίθενται οι σχετικές εικόνες από το UI της εφαρμογής:

Click Here and take the Shot

Here is your image!



Εικόνα 9: Η εφαρμογή κατά την διάρκεια της χρήση της.



Εικόνα 10: Αποτελέσματα της εφαρμογής OCR με Tesseract στην κονσόλα.

Κεφάλαιο 4ο

4 Σουίτα της FaceTec

4.1 Εισαγωγή

Η σουίτα της FaceTec παρέχει μια σειρά από υπηρεσίες οι οποίες σχετίζονται με την βιομετρική κυβερνοασφάλεια. Οι εφαρμογές που αναπτύσσονται γύρο από την σουίτα της Facetec έχουν την δυνατότητα να παρέχουν υπηρεσίες ταυτοποίησης χρηστών τελευταίας τεχνολογίας που περιλαμβάνουν την αναγνώριση ταυτότητας με την χρήση των τεχνολογιών OCR και τρισδιάστατη αναγνώριση χαρακτηριστικών βιομετρικών χαρακτηριστικών.

Οι λύσεις που προσφέρει χρησιμοποιούνται από εκατοντάδες μεγάλες εταιρίες σήμερα. Η FaceTec παρέχει πρόγραμμα μόλις 3MB το οποίο περικλείει τα παραπάνω χαρακτηριστικά αλλά απομακρυσμένες λύσεις (remote/cloud servers) οι οποίες επεκτείνουν κατά πολύ τον τομέα που έχει σχέσης με την διαχείριση της ταυτοποίησης εν γένη

4.2 Τρόπος λειτουργίας

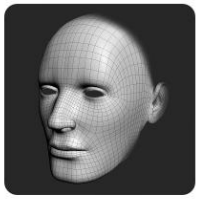
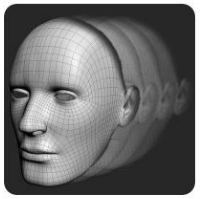
Κατά την διάρκεια της ταυτοποίησης ενός χρήστη, η εφαρμογή χρησιμοποιώντας την κάμερα του κινητού προσπαθεί να αποδείξει ότι η εικόνα που λαμβάνει αποτελείται από τρείς διαστάσεις. Αυτή η συνθέτη διαδικασία παίρνει μόνο δύο δευτερόλεπτα και έρχεται σαν αποτέλεσμα μια πολυσύνθετης επεξεργασίας 100+ βίντεο frames από τις διπλές κάμερες που έτσι και αλλιώς διαθέτουν όλα τα σύγχρονα κινητά τηλέφωνα.

4.3 Εξακρίβωση «ζωντάνιας» (3D liveness)

Όσο οι τεχνολογίες βιομετρικής ταυτοποίησης εξελίσσεται τόσο περισσότερο θα εμφανίζονται τεχνικές εξαπάτησης. Έτσι, είναι πολύ σημαντικό να εξακριβώνεται το αν το πρόσωπο που πρόκειται να αναγνωριστεί είναι πραγματικό. Αυτό είναι εφικτό μόνο εάν εξακριβωθεί ότι είναι ένα 3D πρόσωπο. Αυτό έχει την δυνατότητα να το πράξει η εφαρμογή της FaceTec έχοντας μάλιστα πιστοποιηθεί από τα τεστ NIST/NVLAP level 1 και 2.

4.4 3D ταυτοποίηση & χρόνος

Μάλιστα οι τελευταίες εκδόσεις το πάνε ένα βήμα παρακάτω. Η 2D ταυτοποίηση είναι εύκολο να πέσει θύμα απάτης, για παράδειγμα, κάποιος θα μπορούσε να παρέχει μια φωτογραφεία σε χαρτί αντί το προσώπου του αν η τεχνολογία ταυτοποίησης εδράζεται μόνο σε 2D. Η FaceTec εφαρμόζει μια δισδιάστατη απεικόνιση αλλά κατά την διάρκεια του χρόνου λαμβάνοντας συνεχόμενα βίντεο frames. Έτσι είναι εφικτό να ελεγχθεί αν το πρόσωπο είναι πραγματικό ή όχι. Στην παρακάτω φωτογραφεία συγκρίνονται άλλες τεχνολογίες με αυτή της FaceTec:

TYPE	2D SOFTWARE	3D HARDWARE	FaceTec® 3D SOFTWARE
AXES	X,Y	X,Y,Z	X,Y + TIME
Vendors	Aware, BioID, Daon, FacePhi, Idemia, iProov, ID R&D, etc	Apple Face ID, Google Pixel 4, Intel RealSense®	FaceTec & > 70 Channel Partners Worldwide
Purpose	Face Matching	Unlock Mobile Phones	3D Face Matching
Installed Base	10+ Billion Smart Devices (Android-85% + iOS-14% & Webcams)	Only recent iPhones have Face ID, Pixel 4 = < 1% of market	10+ Billion Smart Devices (Android-85% + iOS-14% & Webcams)
Portable Biometric	Varies	None, re-enroll on each device	Cross-Device & Cross-Platform
Technology	Legacy 2D Matching Software	Hardware: Infrared Camera Array & Neural Network Chip	Software: Real-time Computer Vision + 100% proprietary AI

Interface	Varies	Glance to unlock phone	3D Video Selfie: ~2 Seconds
Racial Bias	Most 2D Algos have racial bias at published FARs	None-Reported	None exhibited in the Lab or Real World usage
Device SDK Info	Varies	No SDK possible, special hardware required	Device SDKs for Android/iOS, web + Server SDK
Liveness Method	Blink, Smile, Turn Head or Flashing Lights, etc	Infrared dots + neural network chip determine if user is 3D	Measures 3D Depth, skin texture, eye reflections, etc
Liveness Strength	Fairly Weak	Fairly Strong	Very Strong
3D Depth Detection	Weak	Very Strong	Very Strong
Intellectual Property	Legacy tech, too old for meaningful patents	20+ infrared related patents acquired in 2013	5 US Patents on 3D process issued, +12 pending globally
FAR/FRR	Varies, but 1/<75,000 at real world usable FRRs	1/1M - No FRR stated	1/12,800,000 FAR @ <1% FRR
Identical Twin Differentiation	Very Weak	"If you have a Twin, use a PIN."	High 1:1 FAR provides Best Possible Twin Differentiation
Liveness Testing Certifications	No, only non-standardized conformances, no camera feed security tested	No Official 3rd Party Testing	Certified Level 1 & 2 Spoof Detection by NIST/NVLAP LAB - Liveness.com

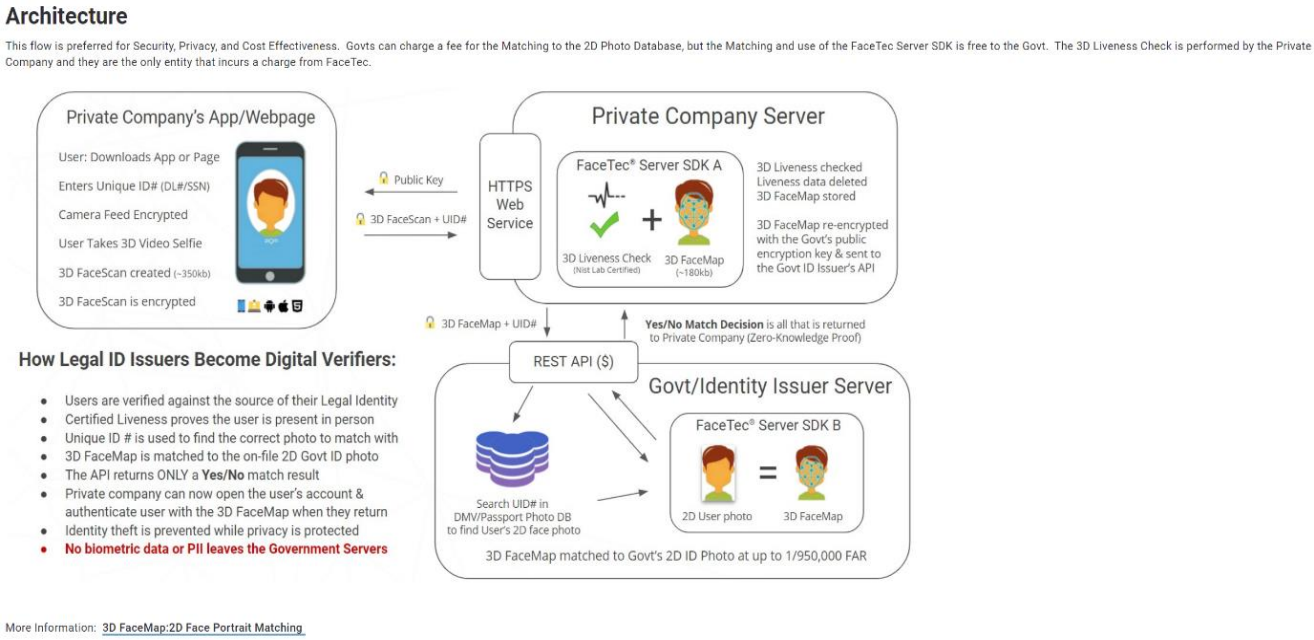
Εικόνα 11: Σύγκριση της σουίτα της FaceTec με άλλες κορυφαίες επιλογές του εμπορίου.

4.5 Αρχιτεκτονική Εφαρμογής FaceTec

Παρακάτω παρουσιάζεται η αρχιτεκτονική της εφαρμογής της FaceTec. Αποτελείται από 3 τμήματα:

1) Το πρώτο είναι το User Interface της εφαρμογής η οποία είναι είτε εγκατεστημένη σε κινητό είτε στον υπολογιστή. Σε αυτό ο χρήστης χρησιμοποιεί την κάμερα του κινητού ή του υπολογιστή του η οποία κρυπτογραφείται και είναι έτοιμη για να αποσταλεί στον server.
2) Το δεύτερο τμήμα είναι η εφαρμογή (web service) που τρέχει σε server (inhouse/cloud) της εκάστοτε εταιρίας/ιδιώτη πελάτη της FaceTec. Σε αυτό εκτελείται η πρώτη νευραλγική διαδικασία ταυτοποίησης του χρήστη, αυτή της επιβεβαίωσης της «ζωντάνιας» του, τέλος δημιουργείται τρισδιάστατη σχεδιάγραμμα προσώπου κρυπτογραφείται εκ νέου και είναι έτοιμο να αποσταλεί σε ένα API το οποίο αποτελεί το τελευταίο βήμα της ταυτοποίησης.
3) Το τρίτο τμήμα αποτελεί από Rest API (Gov/identity Issuer Server) που είναι εγκατεστημένο εκτός του πελάτη. Σε αυτό διασταυρώνεται το τρισδιάστατο σχεδιάγραμμα από το προηγούμενο βήμα με φωτογραφία η οποία είναι

αποθηκευμένη σε βάση δεδομένων του παρόντος Server από την αρχική εγγραφή του χρήστη. Τέλος επιστρέφεται είναι ή όχι ο χρήστης που ισχυρίζεται ότι είναι.

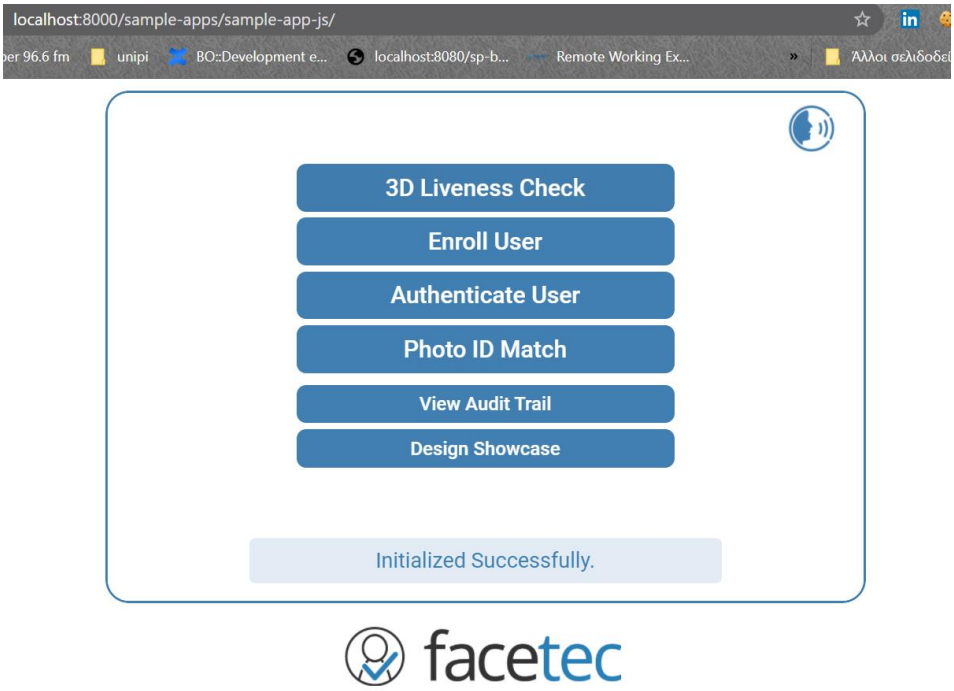


Εικόνα 12: Αρχιτεκτονική της εφαρμογής της FaceTec.

4.5 Τοπικά εγκατεστημένη εφαρμογή FaceTec

Για τις ανάγκες της διπλωματική εργασίας εγκαταστάθηκε τοπικά το demo SDK της FaceTec. Σε αυτό μπορεί να βασιστεί οποιαδήποτε εφαρμογή η οποία απαιτεί authorization/authentication των χρηστών της. Αυτή χωρίζεται σε δύο κομμάτια, το πρώτο αφορά την εφαρμογή που έχει όλες τις λειτουργίες που αφορούν την επαφή με τον ίδιο τον χρήστη που πρόκειται να αξιολογηθεί για την ταυτότητά του και το δεύτερο κομμάτι είναι η εφαρμογή Backoffice (Dashboard) που ο διαχειριστεί έχει πρόσβαση σε όλο τα δεδομένα των χρηστών που έχουν εγγραφεί στο σύστημα.

Παρακάτω παρατίθενται κάποιες εικόνες από την τοπικά εγκατεστημένη εφαρμογή/Dashboard της FaceTec:

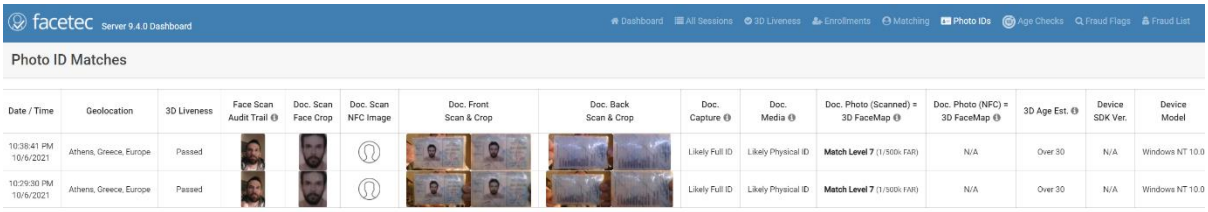


Εικόνα 13: Main menu της εφαρμογής της FaceTec.

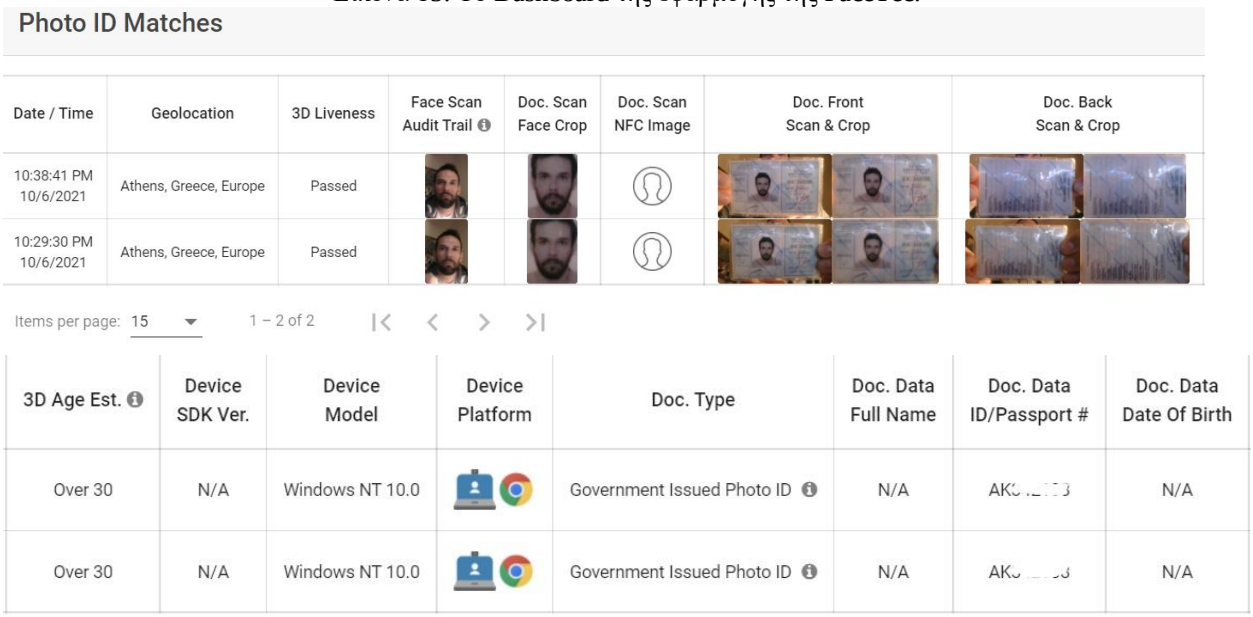
Κεντρικό μενού της FaceTec, σε αυτή ο χρήστης μπορεί να εκτελέσει ταυτοποίηση, να εγγραφεί στο σύστημα, να ελεγχθεί αν είναι πραγματική η εικόνα του (3D liveness) ή να ταυτοποίηση και να εγγράψει στο σύστημα την ταυτότητά του.

Εικόνα 14: Κατά την διάρκεια της ταυτοποίησης χρήστη.

Διαδικασία λήψης φωτογραφία από την εφαρμογή για τις παραπάνω λειτουργίες της εφαρμογής:



Εικόνα 15: Το Dashboard της εφαρμογής της FaceTec.



Εικόνα 16 : Το Dashboard της εφαρμογής.

Το Dashboard της εφαρμογή. Εδώ φαίνονται όλες οι εγγραφές που πραγματοποιήθηκαν. Στην περίπτωσή μας, όπως προαναφέραμε, πραγματοποιήθηκε ταυτοποίηση της ελληνικής ταυτότητας η οποία αποτελεί μια πολύ δύσκολή περίπτωση για την επίτευξη OCR. Παρόλα αυτά τα αποτελέσματά από τις παραπάνω εικόνες δείχνουν ότι η απόσπαση την φωτογραφίας της ταυτότητας έγινε επιτυχώς, εκτίμηση ηλικία αλλά το σημαντικότερο, η απόσπαση του αριθμού της ταυτότητας έγινε με απόλυτη ακρίβεια. Παρόλη την αρκετά μεγάλη ακρίβεια, η εφαρμογή παρόλες τις προσπάθειες φαινόταν να μην μπορεί να εξάγει τα δεδομένα της δεύτερη σελίδα της ταυτότητας. Το τελευταίο θα αποτελέσει και σημείο τριβής/σύγκρισης με την απλή εφαρμογή Tesseract παρακάτω.

Κεφάλαιο 5ο

5 Σύγκριση εφαρμογής Tesseract και faceTec

Στην παρούσα διπλωματική εργασία, όπως προαναφέρθηκε, έχουν αναπτυχθεί δύο εφαρμογές. Η πρώτη είναι αυτή που εδράζεται στην τεχνολογία που προσφέρει το Tessaract και η δεύτερη είναι η εφαρμογή που προσφέρει η εταιρία FaceTec, με το προσφερόμενο JDK περιβάλλον το οποίο εγκαταστάθηκε τοπικά. Στόχος και των δύο εφαρμογών είναι να αναγνωρίσουν, μέσο μιας εικόνας που τους προσφέρεται, πεδία από την ελληνική ταυτότητα.

Σύγκριση:

1) Κόστος:
    a. Για μια εφαρμογή, το Tesseract παρέχεται δωρεάν (Open source)
    b. Η σουίτα της FaceTec δεν είναι δωρεάν. Παρόλα αυτά υπάρχει η δυνατότητα να χρησιμοποιηθεί δωρεάν δοκιμαστικά για ένα περίπου μήνα ένα μέρος των χαρακτηριστικών του.

2) Ευκολία εγκατάστασης:
    a. Μια εφαρμογή tesseract χρειάζεται δουλειά ώστε να είναι σε θέση να παράγει αποτελέσματα.
    b. Η σουίτα της FaceTec μπορεί να εγκατασταθεί αρκετά εύκολα (σχεδόν από τον οποιοδήποτε) και δεν θα χρειαστεί κόπος για να παράξει αποτελέσματα.

3) Εκπαίδευση:
    a. Μια εφαρμογή Tesseract χρειάζεται αρκετό κόπο για την εκπαίδευση της (δες παραπάνω για την εκπαίδευση, την πιο απλή περίπτωση). Η εκπαίδευση μια τέτοιας εφαρμογής ίσως είναι από τα πιο σημαντικά κομμάτια της κατασκευής μια τέτοιας εφαρμογής. Επίσης, σε μια ιδανική περίπτωση στην οποία η εφαρμογή που χρησιμοποιεί Tesseract και είναι μεγάλης έκτασης, χρειάζεται μια ομάδα ανθρώπων μόνο και μόνο για την εκπαίδευση/εισαγωγή δεδομένων.
    b. Η σουίτα της FaceTec δεν χρειάζεται εκπαίδευση. Παρέχει μια μεγάλη λίστα από ήδη εκπαιδευμένων τύπων δημόσιων εγγράφων από δεκάδες χώρες. Στην περίπτωσής της ελληνικής ταυτότητας βεβαίως, παρατηρήθηκε ότι τα αποτελέσματα της πίσω σελίδας δεν αποσπάστηκαν επιτυχώς.

4) Εύρος εργασιών:

a. Η εφαρμογή που αναπτύχθηκε στην παρούσα εργασία, μπορεί μόνο να αποσπάσει μια σειρά πεδίων (όχι και τόσο πετυχημένα πάντα).
b. Η σουίτα της FaceTec μπορεί να εκτελέσει τα παραπάνω αλλά ταυτόχρονα μπορεί να εκτελέσει μια σειρά από άλλες ενέργειες, όπως: αναγνώριση αν τα φωτογραφούμενο πρόσωπο είναι πραγματικό, εκτίμηση ηλικίας, απόσπαση πεδίων δημοσίου εγγράφου. Όλα τα παραπάνω χαρακτηριστικά εντάσσονται σε μια πιο γενικός σχεδιασμό ύπαρξής της σουίτας. Αυτή μπορεί να χρησιμοποιηθεί σαν ένα ολοκληρωμένο σύστημα για της διαχείριση της ταυτοποίησης χρηστών ως κύριο τμήμα login μιας εφαρμογής ή συμπληρωματικό κομμάτι του μηχανισμού login μιας εφαρμογής

5) Αποτελέσματα:
   a. Η εφαρμογή που αναπτύχθηκε στην παρούσα εργασία κατάφερε να αποσπάσει τον αριθμό ταυτότητας, αλλά κατά την διάρκεια των τεστ δεν ήταν λίγες οι φορές που τα αποτελέσματα ήταν λανθασμένα
   b. Η σουίτα της FaceTec κατάφερε να εκτελέσει σχεδόν επιτυχώς της εξαγωγή του αριθμού της ταυτότητας χωρίς λάθη.

Βιβλιογραφία:

[1] Thi Tuyet Hai Nguyen, Adam Jatown και Mickael Coustaty & Antoine Doucet. Έρευνα για τις προσεγγίσεις σχετικά με τον προ-OCR διαδικασία.

[2] Nishant Subramani, Alexandre Matton, Malcolm Greaves και Adrian Lam. Έρευνες για τις προσεγγίσεις σχετικά με το Deep learning για OCR και την κατανόηση κειμένου.

[3] Daniel Lopresti. 2009. Σφάλματα οπτικής αναγνώρισης χαρακτήρων και οι επιπτώσεις τους στην επεξεργασία της φυσικής γλώσσας. *ΕΣΩΤ J.* –151.

[4] William B. Lund και Eric K. Ringger. 2011. Διόρθωση σφαλμάτων με εκπαίδευση εντός τομέα σε πολλά συστήματα OCR. Κατά *τη διάρκεια της διεθνούς διάσκεψης του 2011 για την ανάλυση και την αναγνώριση των εγγράφων*. IEEE, 658–662.

[5] William B. Lund, Daniel D. Walker και Eric K. Ringger. 2011. Προοδευτική ευθυγράμμιση και μεροληπτικό σφάλμα. Κατά *τη διάρκεια της διεθνούς διάσκεψης του 2011 για την ανάλυση και την αναγνώριση των εγγράφων*. IEEE, 764–768. -96

[6] Martin Volk, Torsten Marek και Rico Sennrich. 2010. Μείωση των σφαλμάτων OCR με συνδυασμό δύο συστημάτων OCR. Στο εργαστήριο ECAI 2010 για την τεχνολογία γλωσσών για την πολιτιστική κληρονομιά, τις κοινωνικές επιστήμες και τις ανθρωπιστικές επιστήμες. -161

[7] David Wemhoener, Ismet Zeki Yalniz και R. Manmatha. 2013. Δημιουργία βελτιωμένης έκδοσης με θορυβώδες OCR από πολλές εκδόσεις. Κατά *τη διάρκεια της 12ΗΣ Διεθνούς Διάσκεψης του 2013 για την ανάλυση και την αναγνώριση των εγγράφων*. -163

[8] Mayce Al Azawi, Marcus Liwicki και Thomas M. Breuel. 2015. Συνδυασμός πολλαπλών ευθυγραμμισμένων εξόδων αναγνώρισης με χρήση WFST και LSTM. ΣΤΙΣ *εργασίες της 13ης Διεθνούς Διάσκεψης του 2015 για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'15)*. IEEE, 31–35. -4

[9] Ziafan lean. 2001. Αξιόπιστη λύση OCR για επανεπεξεργασία ψηφιακού περιεχομένου. Στην *αναγνώριση εγγράφων και την ανάκτηση IX*, Τόμος 4670. Διεθνής εταιρεία οπτικών και φωτονικών, 223–231. -87

[10] Christian Reul, Uwe Springmann, Christoph Wick και Frank Puppe. 2018. Βελτίωση της ακρίβειας OCR σε πρώιμα εκτυπωμένα βιβλία, με χρήση γενικής προπόνησης και ψηφοφορίας. Στο *13ο Διεθνές εργαστήριο ανάλυσης εγγράφων της IAPR*(DAS'18). IEEE, 423–428. -129

[11] Sepp Hochreiter και Jürgen Schmidhuber. 1997. Μακροπρόθεσμη μνήμη. *Νευρικός υπολ.* 9, 8 (1997), 1735–1780. -64

[12] William B. Lund, Douglas J. Kennard και Eric K. Ringger. 2013. Συνδυασμός πολλών τιμών τμηματοποίησης τιμής ουδού για τη βελτίωση της εξόδου OCR. Σε *έγγραφο αναγνώρισης και ανάκτησης XX (διαδικασία SPIE)*, τόμος 8658. SPIE, 86580R. -92

[13] Shaobin Xu και David Smith. 2017. Ανάκτηση και συνδυασμός επαναλαμβανόμενων περασμάτων για τη βελτίωση του OCR. ΣΤΙΣ *εργασίες της κοινής διάσκεψης 2017 ACM/IEEE για τις ψηφιακές βιβλιοθήκες (JCDL'17)*. IEEE, 1–4. -167

[14] Deepayan Das, Jerin Philip, Minesh Mathiew και C. V. Jawahar. 2019. Μια οικονομικά αποδοτική προσέγγιση για τη διόρθωση σφαλμάτων OCR σε μεγάλες συλλογές εγγράφων. ΣΤΙΣ *εργασίες της Διεθνούς Διάσκεψης για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'19)*. IEEE, 655–662.

[15] Fred J. Damerau. 1964. Μια τεχνική για τον εντοπισμό και τη διόρθωση των ορθογραφικών λαθών από υπολογιστή. *ΔΙΑΚ. ACM* 7, 3.

[16] Richard C. Angell, George E. Freund και Peter Willett. 1983. Αυτόματη ορθογραφία με χρήση ενός μέτρου ομοιότητας κατά ένα γραμμάριο. *Έγχ. Επεξεργασία. Διαχείριση.* 19, 4 (1983), 255–261 -7

[17] Klaus U. Schulz, Stoyan Mihov και Petar Mitankin. 2007. Γρήγορη επιλογή μικρών και ακριβών συνόλων υποψηφίων από λεξικά για εργασίες διόρθωσης κειμένου. ΣΤΙΣ *εργασίες της 9ης Διεθνούς Διάσκεψης για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'07)*, Τόμος 1. IEEE, 471–475. -142

[18] Kazem Taghva, Allen Credit, Julie Borsack, John Kilburg, Changshi Wu, Και Jeff Gilbreth. 1998. Σύστημα επεξεργασίας εγγράφων. Στο *Document Recognition V*, Τόμος 3305. Διεθνής εταιρεία οπτικών και φωτονικών, 179–184.

[19] Paula Estrella και Pablo Paliza. 2014. Διόρθωση οπτικής αναγνώρισης χαρακτήρων (OCR) των εγγράφων που δημιουργήθηκαν κατά την εθνική διαδικασία επανεπεξεργασίας (reor- ganization) της Αργεντινής. Στα *πλαίσια της 1ης Διεθνούς Διάσκεψης για την ψηφιακή πρόσβαση στη θεματική πολιτιστική κληρονομιά.* 119-123.

[20] Kimmo Kettunen. 2015. Διατήρηση, αλλαγή ή διαγραφή; Δημιουργία ενός χαμηλού πλαισίου οπτικής αναγνώρισης χαρακτήρων (ocr) για μια συλλογή παλαιών φινλανδικών εφημερίδων που έχουν υποστεί ψηφιακή μετατροπή. Στο *πλαίσιο της ιταλικής διάσκεψης ΈΡΕΥΝΑΣ για τις ψηφιακές βιβλιοθήκες*.

[21] Ewerton Cappelatti, Regina De Oliveira Heidrich, Ricardo Oliveira, Cintia Monticelli, Ronaldo Rodrigues, RO - drigo Goulart, Και Eduardo Velho. 2018. Μετά τη διόρθωση σφαλμάτων OCR με χρήση λεγμένων προτάσεων ορθογραφίας που έχουν προκύψει μέσω τροποποιημένου αλγόριθμου που δεν έχει ανάγκη. Στο *πλαίσιο της Διεθνούς Διάσκεψης για την αλληλεπίδραση ανθρώπου-υπολογιστή*. Springer, 3–10.

[22] Christian Strohmaier, Christoph Ringlstetter, Klaus U. Schulz και Stoyan Mihov. 2003. Ένα οπτικό και διαδραστικό εργαλείο για τη βελτιστοποίηση της λεξικής μεταδιόρθωσης των αποτελεσμάτων OCR. *ΣΤΙΣ εργασίες της Διάσκεψης του 2003 σχετικά με την όραση των υπολογιστών και την αναγνώριση των μοτίβων* ,Τόμος 3. IEEE, 32–32.

[23] Stoyan Mihov, Svetla Koeva, Christoph Ringlstetter, Klaus U. Schulz και Christian Strohmaier. 2004. Ακριβής και αποτελεσματική διόρθωση κειμένου με χρήση της αυτόματης λέεβενσσας, των δυναμικών λεξικών Web και των βελτιστοποιημένων μοντέλων διόρθωσης. *Στα πλαίσια του εργαστηρίου για τα διεθνή εργαλεία γλωσσικού ελέγχου και τις τεχνολογίες των γλωσσών (2004)*.

[24] Christian M. Strohmaier, Christoph Ringlstetter, Klaus U. Schulz και Stoyan Mihov. 2003. Λεξική μεταδιόρθωση των αποτελεσμάτων OCR: Το διαδίκτυο ως δυναμικό δευτερεύον λεξικό. Στα *πλαίσια της 7ης Διεθνούς Διάσκεψης για την ανάλυση και την αναγνώριση των εγγράφων.* Citeer, 1133–1137.

[25] Kenneth W. Church και William A. Gale. 1991. Βαθμολόγηση πιθανότητας για ορθογραφία. *Στατ. Υπολ/στή*. 1, 2 (1991), 93–103.

[26] Eric Brill και Robert C. Moore. 2000. Ένα βελτιωμένο μοντέλο σφάλματος για την ορθογραφία του καναλιού με θόρυβο. Στο *πλαίσιο της 38ης ετήσιας συνόδου για την Ένωση για την υπολογιστική γλωσσολογία*.Ένωση για την υπολογιστική γλωσσολογία, 286–293.

[27] Christian M. Strohmaier, Christoph Ringlstetter, Klaus U. Schulz και Stoyan Mihov. 2003. Λεξική μεταδιόρθωση των αποτελεσμάτων OCR: Το διαδίκτυο ως δυναμικό δευτερεύον λεξικό. Στα *πλαίσια της 7ης Διεθνούς Διάσκεψης για την ανάλυση και την αναγνώριση των εγγράφων.* Citeer, 1133–1137.

[28] Lenz Furrer και Martin Volk. 2011. Μείωση σφαλμάτων OCR σε έγγραφα γοτθικού τύπου. Στα πλαίσια του εργαστηρίου για τις γλωσσικές τεχνολογίες για τις ψηφιακές ανθρωπιστικές επιστήμες και την πολιτιστική κληρονομιά. 97-103.

[29] Harald Hammarström, Shafqat Mumtaz Virk και Markus Forsberg. 2017. Το πρόγραμμα OCR της Poor Man μετά τη διόρθωση: Μη διαμελής αναγνώριση της ορθογραφίας σε παραλλαγές που εφαρμόζεται σε συλλογή πολύγλωσσων εγγράφων. Κατά *τη διάρκεια της 2ης Διεθνούς Διάσκεψης για την ψηφιακή πρόσβαση σε πολιτιστική κληρονομιά υπό μορφή κειμένου*. 71-75.

[30] Axel Jean-Caurant, Nouredine Tamani, Vincent Courboulay και Jean-Christophe Burie. 2017. Σειρά με βάση τα στοιχεία της λεξικής για διόρθωση μετά από OCR των κατονομαζόμενων οντοτήτων. ΣΤΙΣ *εργασίες της 14ης Διεθνούς Διάσκεψης για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'17)*, Τόμος 1. IEEE, 1192–1197.

[31] Jeffrey Pennington, Richard Sochre και Christopher Manning. 2014. Γάντι: Καθολικά διανύσματα για την αναπαράσταση λέξεων. Κατά *τη διάρκεια της διάσκεψης του 2014 για τις εμπειρικές μεθόδους επεξεργασίας της φυσικής γλώσσας (EMNLP'14)*. 1532-1543.

[32] Martin Reynaert. 2008. Μη διαλογική OCR μετά τη διόρθωση για έργα ψηφιοποίησης σε κλίμακα Giga. Στα *πλαίσια της Διεθνούς Διάσκεψης για την ευφυή επεξεργασία κειμένου και την υπολογιστική γλωσσολογία*. Springer, 617–630.

[33] Martin W. C. Reynaert. 2011. Σύγχυση χαρακτήρων σε σχέση με την εστίαση της διόρθωσης ορθογραφίας βάσει λέξεων και των παραλλαγών OCR στον εταιρικό κλάδο. *ΕΣΩΤ J. πρωκτικό έγγραφο. Εγγρ.* 14, 2 (2011), 173–187.

[34] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty και Jean-Philippe Moreux. 2017. ICDAR2017 Competition on Post-OCR text correction. ΣΤΙΣ *εργασίες 2017 της 14ης διεθνούς διάσκεψης του IAPR για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'17)*, τόμος 1. IEEE, 1423–1428.

[35] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani και Jean-Philippe Moreux. 2017. Επιπτώσεις των σφαλμάτων οπτικής αναγνώρισης χαρακτήρων (OCR) στη χρήση των ψηφιακών βιβλιοθηκών: Για καλύτερη πρόσβαση στις πληροφορίες. ΣΤΙΣ *εργασίες της κοινής διάσκεψης 2017 ACM/IEEE για τις ψηφιακές βιβλιοθήκες (JCDL'17)*. IEEE, 1–4.

[36] Okan Kolak και Philip Resnik. 2002. Διόρθωση σφάλματος OCR με χρήση μοντέλου καναλιού με θόρυβο. Στο *πλαίσιο της 2ης Διεθνούς Διάσκεψης για την έρευνα στον τομέα της τεχνολογίας των ανθρώπινων γλωσσών*. Morgan Kaufmann Publishers Inc., 257–262.

[37] Eric Brill και Robert C. Moore. 2000. Ένα βελτιωμένο μοντέλο σφάλματος για την ορθογραφία του καναλιού με θόρυβο. Στο *πλαίσιο της 38ης ετήσιας συνόδου για την Ένωση για την υπολογιστική γλωσσολογία*. Ένωση για την υπολογιστική γλωσσολογία, 286–293.

[38] Juan Carlos Perez-Cortes, Juan-Carlos Amengual, Joaquim Arlandis και Rafael Llobet. 2000. Στοχαστικό σφάλμα - Διόρθωση ανάλυσης για μετεπεξεργασία OCR. ΣΤΙΣ *εργασίες της 15ης Διεθνούς Διάσκεψης για την αναγνώριση των προτύπων (ICPR'00)*, Τόμος 4. IEEE, 405–408.

[39] Christophe Rigaud, Antoine Doucet, Mickaël Coustaty και Jean-Philippe Moreux. 2019. Διαγωνισμός ICDAR 2019 για τη διόρθωση κειμένου μετά την OCR. ΣΤΙΣ *εργασίες της Διεθνούς Διάσκεψης του 2019 για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'19)*. IEEE, 1588–1593.

[40] Juan-Carlos Perez-Cortes, Rafael Llobet, J. Ramon Navarro-Cerdan και Joaquim Arlandis. 2010. Χρήση της διασύνδεσης πεδίων για βελτίωση της απόδοσης διόρθωσης σε ένα σύστημα μετεπεξεργασίας ocr με βάση ηχοβολέα. Κατά *τη διάρκεια της 12ης Διεθνούς Διάσκεψης για τα σύνορα, με την αναγνώριση του εγγράφου αυτού*. IEEE, 605–610.

[41] Caitlin Richter, Matthew Wickes, Deniz Beser και Mitch Marcus. 2018. Η επεξεργασία των θορυβωδών δεδομένων οπτικής αναγνώρισης χαρακτήρων (OCR) με χαμηλό πόρο για την ψηφιοποίηση των ιστορικών μεσολόβιο. Στο *πλαίσιο της 11ης διεθνούς διάσκεψης για τους γλωσσικούς πόρους και την αξιολόγηση (LREC'18)*.

[42] Okan Kolak και Philip Resnik. 2005. Μετεπεξεργασία OCR για γλώσσες χαμηλής πυκνότητας. Κατά τη διάρκεια της Διάσκεψης για την τεχνολογία των ανθρώπινων γλωσσών και τις εμπειρικές μεθόδους επεξεργασίας των φυσικών γλωσσών. Σύνδεσμος για Computa- tional linguistics, 867–874.

[43] Thorsten Vobl, Annette Gotscharek, Uli Reffie, Christoph Ringlstetter και Klaus U. Schulz. 2014. PoCoTo-ένα σύστημα ανοικτού κώδικα για αποτελεσματική διαδραστική διόρθωση ιστορικών κειμένων OCRed. Στα *πλαίσια της 1ης Διεθνούς Διάσκεψης για την ψηφιακή πρόσβαση στη θεματική πολιτιστική κληρονομιά*. 57-61.

[44] Florian Freink, Klaus U. Schulz και Uwe Springmann. 2017. Κατάρτιση προφίλ των ιστορικών κειμένων που έχουν αναθεωρηθεί. Κατά *τη διάρκεια της 2ης Διεθνούς Διάσκεψης για την ψηφιακή πρόσβαση σε πολιτιστική κληρονομιά υπό μορφή κειμένου*. 61-66.

[45] Tobias Englmeier, Florian Fink και Klaus U. Schulz. 2019. AI-PoCoTo: Συνδυασμός αυτοματοποιημένου και διαδραστικού ocr

[46] Stefan Gerdjikov, Stoyan Mihov και Vladislav Nengev. 2013. Εξαγωγή των ορθογραφικών παραλλαγών από την υποστήριξη της γλώσσας για διόρθωση κειμένου με θόρυβο. Στα *πλαίσια της 12ης Διεθνούς Διάσκεψης για την ανάλυση και την Ανακτόνωση εγγράφων*. IEEE, 324–328.

[47] Steffen Eger, Tim vor der Brück και Alexander Mehler. 2016. Σύγκριση τεσσάρων μοντέλων μετάφρασης συμβολοσειράς-συμβολοσειράς-επιπέδου χαρακτήρων για τη διόρθωση ορθογραφικού λάθους (OCR). *Πράγα Μπουλλ Μαθηματικά. Μουρούνα.* 105 (2016), 77–100.

[48] Maximilian Bisani και Hermann Ney. 2008. Μοντέλα από κοινού αλληλουχίας για μετατροπή γραφεμέ σε φωνομέ. *Ένα "Speech Commun"* 50, 5 (2008), 434–451.

[49] Mika Hämäläinen και Simon Hengchen. 2019. Από την πρύμνη μέχρι τα έπιπλα: Μια πλήρως αυτόματη μέθοδος NMT και ενσωματώσεως λέξεων για τη μεταδιόρθωση OCR. Κατά *τη διάρκεια της διεθνούς διάσκεψης για τις πρόσφατες προόδους στον τομέα της επεξεργασίας των φυσικών γλωσσών (RANLP'19).* INCOMA Ltd., 431–436.

[50] Kai Hakala, Aleksi Vesanto, Niko Miekka, Tapio Salakoski και Filip Ginter. 2019. Αξιοποίηση επαναλήψεων κειμένου και αποκωδικοποίησης αυτοκωδικοποιητών με OCR μετά τη διόρθωση. *Corr* abs/1906.10907 (2019).

[51] Christoph Ringlstetter, Max Hadersbeck, Klaus U. Schulz και Stoyan Mihov. 2007. Διόρθωση κειμένου με τη χρήση μοντέλων με βάση το domain de- pendent bigram από ανίχνευση μέσω web. Κατά *τη διάρκεια της Διεθνούς ΚΟΙΝΉΣ Διάσκεψης για την τεχνητή νοημοσύνη (IJCAI'07), εργαστήριο για την ανάλυση των μη δομημένων δεδομένων κειμένου με θόρυβο.*

[52] Alberto Poncelas, Mohammad Aboomar, Jan buts, James Hadley και Andy Way. 2020. Ένα εργαλείο για τη διευκόλυνση της χρήσης ocr στα ιστορικά έγγραφα. Στα *πλαίσια του 1ου εργαστηρίου γλωσσικών τεχνολογιών για τις ιστορικές και αρχαίες γλώσσες (LT4HALA'20).* 47-51.

[53] Daniel Hládek, Ján Staš, Stanislav Ondáš, Jozef Juhár και Lászlo Kovács. 2017. Απόσταση συμβολοσειράς εκμάθησης με ομαλό τρόπο για διόρθωση ορθογραφίας OCR. *Εφαρμογές εργαλείων πολυμέσων* 76, 22 (2017), 24549–24567.

[54] Michel Généreux, Egon W. Stemle, Verena Lyding και Lionel Nicolas. 2014. Διόρθωση σφαλμάτων OCR για τα γερμανικά στη γραμματοσειρά fraktur. Κατά *τη διάρκεια της πρώτης Ιταλικής Διάσκεψης για την υπολογιστική γλωσσολογία-IT.* PISA University Press, 186–190.

[55] Daniel Hládek, Ján Staš, Stanislav Ondáš, Jozef Juhár και Lászlo Kovács. 2017. Απόσταση συμβολοσειράς εκμάθησης με ομαλό τρόπο για διόρθωση ορθογραφίας OCR. *Εφαρμογές εργαλείων πολυμέσων* 76, 22 (2017), 24549–24567.

[56] Andreas W. Hauser. 2007. *OCR-Postcorrection of historical texts*. Διατριβή πλοιάρχου. Ludwig-Maximilians-Universität München.

[57] John Evershed και Kent Fitch. 2014. Διόρθωση θορυβώδους OCR: Ο συγκείμενο κτύπος δημιουργεί σύγχυση. Στο *πλαίσιο της 1ης διακρατικής διάσκεψης για την ψηφιακή πρόσβαση σε πολιτιστική κληρονομιά υπό μορφή κειμένου*. ACM, 45–51.

[58] Dana Dannélls και Simon Persson. 2020. Εποπτευόμενη RCO μετά τη διόρθωση ιστορικών κειμένων στη σουηδία: ΠΟΙΟΣ είναι ο ρόλος του συστήματος OCR; ΣΤΙΣ *εργασίες των ψηφιακών ανθρωπιστικών σπουδών στις σκανδιναβικές χώρες, 5η Διάσκεψη* εργασίας CEUR, τόμος 2612. CEUR-WS.org, 24–37.

[59] Christophe Rigaud, Antoine Doucet, Mickaël Coustaty και Jean-Philippe Moreux. 2019. Διαγωνισμός ICDAR 2019 για τη διόρθωση κειμένου μετά την OCR. ΣΤΙΣ *εργασίες της Διεθνούς Διάσκεψης του 2019 για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'19).* IEEE, 1588–1593.

[60] Andrey Sariev, Vladislav Nengev, Stefan Gerdjikov, Petar Mitankin, Hristo Ganchev, Stoyan Mihov και Tinko Tinchev. 2014. Ευέλικτη διόρθωση κειμένου με θόρυβο. Στο *Διεθνές εργαστήριο του 2014 της 11ης IAPR για τα συστήματα ανάλυσης εγγράφων*. IEEE, 31–35.

[61] Κάι Νίκλας. 2010. Μη εποπτευόμενη μετεπεξεργασία σφαλμάτων OCR. *Διατριβή πλοιάρχου. Leibniz Universität Hannover* (2010).

[62] Thorsten Brants και Alex Franz. 2006. Web 1T 5-Gram έκδοση 1 LDC2006T13. Στη *Φιλαδέλφεια: Γλωσσικά δεδομένα con- sortium*. Google Inc

[63] Youssef Bassil και Mohammad Alwani. ΤΊΤΛΟΣ: 2012. Διόρθωση σφαλμάτων με διάκριση περιβάλλοντος OCR βάσει συνόλου δεδομένων 5 γραμμαρίων google web 1T. *Τροπ Ι. ΤΚΣ. Ανάλυση* 50 (2012).

[64] Sandeep Soni, Lauren Klein και Jacob Eisenstein. 2019. Διόρθωση σφαλμάτων λευκαespace σε ψηφιοποιημένα ιστορικά κείμενα. ΣΤΙΣ εργασίες του 3ου κοινού εργαστηρίου SIGHUM σχετικά με την υπολογιστική γλωσσολογία πολιτιστικής κληρονομιάς, τις κοινωνικές επιστήμες, τις ανθρωπιστικές επιστήμες και τη λογοτεχνία. 98-103.

[65] Jorge Ramón Fonseca Cacho, Kazem Taghva και Daniel Alvarez. 2019. Χρησιμοποιήστε το σώμα κώδικα 5 γραμμαρίων του Google web για τη διόρθωση του σφάλματος OCR. Κατά *τη διάρκεια της 16ης Διεθνούς Διάσκεψης για την τεχνολογία των πληροφοριών - νέες γενιές (ITNG'19)*. Springer, 505–511

[66] Kazem Taghva και Eric Stofsky. 2001. OCRSpell: Ένα διαδραστικό σύστημα ορθογραφίας για σφάλματα OCR σε κείμενο.

[67] Yoshua Bengio, Réjean Ducharme, Pascal Vincent και Christian Jauvin. 2003. Ένα μοντέλο νευρικών πιθανολογικών γλωσσών.

[68] EVA D'hondt, Cyril Grouin και Brigitte Grau. 2016. Εντοπισμός και διόρθωση σφάλματος OCR χαμηλών πόρων σε γαλλικά κλινικά κείμενα. *Στα πλαίσια του 7ου Διεθνούς εργαστηρίου για την ανάλυση του κειμένου της Υγείας (Louhi@EMNLP'16)*. Ένωση για την υπολογιστική γλωσσολογία, 61–68.

[69] Thibault Magallon, Frédéric Béchet και Benoît Favre. 2018. Συνδυασμός RNN επιπέδου χαρακτήρων και επιπέδου λέξεων για τον εντοπισμό σφαλμάτων μετά το OCR. Κατά *τη διαδικασία του Actes de la Conférence TALN (CORIA-TALN-RJC'18), τόμος 1*. ATALA, 233–240.

[70] Dana Dannélls και Simon Persson. 2020. Εποπτευόμενη RCO μετά τη διόρθωση ιστορικών κειμένων στη σουηδία: Ποιος είναι ο ρόλος του συστήματος OCR; ΣΤΙΣ *εργασίες των ψηφιακών ανθρωπιστικών σπουδών στις σκανδιναβικές χώρες, 5η Διάσκεψη* εργασίας CEUR, τόμος 2612. CEUR-WS.org, 24–37.

[71] Kissos και Nachum Dershowitz. 2016. Διόρθωση σφάλματος OCR με διόρθωση χαρακτήρων και ταξινόμηση λέξεων βάσει χαρακτηριστικών. Στα *πλαίσια του 12ου Συνεργείου της IAPR για τα συστήματα ανάλυσης εγγράφων (DAS'16)*. IEEE, 198–203.

[72] Aminul Islam και Diana Inkpen. 2009. Διόρθωση ορθογραφίας από πραγματικές λέξεις με χρήση του GOOGLE Web IT: 3 γραμμάρια. Κατά *τη διάρκεια της Διάσκεψης του 2009 για τις εμπειρικές μεθόδους στη μεταποίηση των φυσικών γλωσσών*. 1241-1249.

[73] Gitansh Khirbat. 2017. Διόρθωση κειμένου μετεπεξεργασίας OCR με προσομοίωση ανόπτησης (OPTeCA). Κατά *τη διάρκεια του εργαστηρίου του Συνδέσμου γλωσσικών τεχνολογιών της Αυστραλίας-Ασίας 2017*. 119-123.

[74] Scott Kirkpatrick, C. Daniel Gelatt, και Mario P. Vecchi. 1983. Βελτιστοποίηση με προσομοίωση ανόπτησης. *Science* 220, 4598 (1983), 671–680.

[75] Lloyd Allison και Trevor I. Dix. 1986. Ένας αλγόριθμος δευτερεύουσας ακολουθίας με τη μεγαλύτερη συμβολοσειρά bit. *Ενημερώστε. Επεξεργασία. Λετ.* 23, 5 (1986), 305–310.

[76] Thi-Tuyet-Hai Nguyen, Mickaël Coustaty, Antoine Doucet, Adam Jatowt και Nhu-Van Nguyen. 2018. Προσαρμοστική προσέγγιση επεξεργασίας-απόστασης και παλινδρόμησης για διόρθωση κειμένου μετά από OCR. Στην *ωριμότητα και την καινοτομία στις ψηφιακές βιβλιοθήκες: Εργασίες της 20ης Διεθνούς Διάσκεψης για τις ψηφιακές βιβλιοθήκες Ασίας-Ειρηνικού (ICADL'18) (* Σημειώσεις διαλέξεων στην επιστήμη COM-puter, τόμος 11279. Springer, 278–289.

[77] Dana Dannélls και Simon Persson. 2020. Εποπτευόμενη RCO μετά τη διόρθωση ιστορικών κειμένων στη σουηδία: Ποιος είναι ο ρόλος του συστήματος OCR; ΣΤΙΣ *εργασίες των ψηφιακών ανθρωπιστικών σπουδών στις σκανδιναβικές χώρες, 5η Διάσκεψη* εργασίας CEUR, τόμος 2612. CEUR-WS.org, 24–37.

[78] Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickaël Coustaty, Nhu-Van Nguyen και Antoine Doucet. 2019. Ανίχνευση σφάλματος μετά από OCR με τη δημιουργία ευλογοφανών υποψηφίων. ΣΤΙΣ *εργασίες της Διεθνούς Διάσκεψης του 2019 για την ανάλυση και την αναγνώριση εγγράφων (ICDAR'19)*. IEEE, 876–881.

[79] Jorge Ramón Fonseca Cacho και Kazem Taghva. 2020. Μετεπεξεργασία OCR με χρήση διανυσματικών μηχανημάτων υποστήριξης. Στην *Intel- ligent Computing: Διαδικασίες της διάσκεψης Computing 2020, τόμος 2 (AI'20)*, εξελίξεις στα ευφυή συστήματα και υπολογιστές, τόμος 1229. Springer, 694–713.

[80] Quoc-Dung Nguyen, Duc-Anh Le, Nguyet-Minh Phan και Ivan Zelinka. 2021. Διόρθωση σφάλματος OCR με χρήση μοτίβων διόρθωσης και αυτο-διοργανώνοντας αλγόριθμο μετεγκατάστασης. *Πρωκτικό μοτίβο. Appl. (Εφαρμογή)* 24, 2 (2021), 701–721.