Κεφάλαιο 7

Συμπεράσματα - Προοπτικές

7.1 Συμβολή της Διατριβής

Στην διατριβή αυτή παρουσιάστηκε μία προσέγγιση για την αυτόματη ταξινόμηση Νεοελληνικών κειμένων ανάλογα με το είδος και τον συγγραφέα τους. Η διαδικασία αυτή είναι πλήρως αυτοματοποιημένη καθώς τις μετρήσεις των υφολογικών δεικτών, που αναπαριστούν το ύφος του κειμένου, τις παρέχει αυτόματα ο ανιχνευτής ορίων περιόδων και φράσεων. Το εργαλείο αυτό, είναι ένα σύστημα επεξεργασίας φυσικής γλώσσας γενικού σκοπού, που συμβαδίζει με τις σύγχρονες ανάγκες ανάλυσης κειμένου, δηλαδή:

- Δυνατότητα επεξεργασίας μεγάλων όγκων κειμένων πολύ γρήγορα.
- Δυνατότητα ανάλυσης κειμένου χωρίς περιορισμούς.
- Έμφαση σε εργασίες χαμηλού επιπέδου.
- Εύκολη προσαρμογή σε ένα συγκεκριμένο τύπο κειμένου.
- Ελάχιστο υπολογιστικό κόστος.

- Χρήση ελάχιστων πόρων.

Η τελευταία ιδιότητα είναι ίσως η πιο σημαντική. Τα αποτελέσματα ανάλυσης κειμένων χωρίς περιορισμούς αποδεικνύουν ότι είναι δυνατή η γρήγορη και αξιόπιστη ανάλυση κειμένων, με βάση σχετικά απλή πληροφορία και όχι ογκώδη λεξικά και περίπλοκες γραμματικές. Η σύγκριση με μία προσέγγιση που χρησιμοποιεί επιπλέον ένα λεξικό 30.000 λημμάτων έδειξε ότι η μεν ανάκληση βελτιώνεται (ειδικά όσον αφορά τις ΟΦ και τις ΕΦ) η δε ακρίβεια μειώνεται, κυρίως λόγω της αδυναμίας του λεξικού να καλύψει όλες τις δυνατές μορφές μιας λέξης. Αν λάβουμε υπ' όψιν και το χρονικό κόστος που επιφέρει η χρήση λεξικού, τότε οδηγούμαστε στο συμπέρασμα ότι η προσέγγιση ελάχιστων πόρων είναι η καλύτερη λύση για μία εφαρμογή που απαιτεί γρήγορη και αρκετά καλή ανάλυση κειμένων.

Η φιλοσοφία της υλοποίησης των ανιχνευτών ορίων περιόδων και φράσεων συμβαδίζει με τις σύγχρονες τεχνικές στην υπολογιστική γλωσσολογία και μπορεί να συνοψιστεί ως εξής:

- Εμπειρική παρατήρηση των χαρακτηριστικών της Νέας Ελληνικής γλώσσας. Ο ανιχνευτής ορίων περιόδων βασίστηκε στο γεγονός ότι η συντριπτική πλειοψηφία των Νεοελληνικών λέξεων τελειώνει σε συγκεκριμένους χαρακτήρες. Ο μορφολογικός πλούτος της Νεοελληνικής γλώσσας οδήγησε στην εκτίμηση της μορφολογικής πληροφορίας της κάθε λέξης με βάση την κατάληξή της. Επίσης, οι κανόνες ανίχνευσης ορίων φράσεων που χρησιμοποιούνται στα περάσματα ανάλυσης προήλθαν από εμπειρική παρατήρηση και εκμεταλλεύονται την υποχρεωτική χρήση άρθρων, μορίων κτλ.
- Εκμετάλλευση των ήδη διαθέσιμων πόρων. Το λεξικό των κοινών καταλήξεων που χρησιμοποιήθηκε για την εκτίμηση της μορφολογικής περιγραφής της κάθε λέξης, στα πλαίσια του ανιχνευτή ορίων φράσεων, προήλθε από την μετατροπή ενός ήδη υπάρχοντος λεξικού.
- Εφαρμογή μεθόδων μηχανικής εκμάθησης για την αυτόματη εξαγωγή της γλωσσολογικής πληροφορίας. Η εξαγωγή των κανόνων αποσαφήνισης των σημείων στίξης, που χρησιμοποιούνται στα πλαίσια του ανιχνευτή ορίων

περιόδων, έγινε αυτόματα μέσω μιας νέας μεθόδου μηχανικής εκμάθησης, παραλλαγή της EBM.

Η σύγκριση της προτεινόμενης μεθοδολογίας μηχανικής εκμάθησης με την παραδοσιακή EBM έδειξε ότι η μέθοδός μας συμπεριφέρεται καλύτερα όσον αφορά την ακρίβεια αποσαφήνισης αλλά και το χρονικό κόστος εκπαίδευσης. Αυτό οφείλεται στο ότι η EBM είναι μία θεωρία ανεξάρτητη-εφαρμογής και αυτό έχει ως συνέπεια να μην μπορεί να εκμεταλλευτεί τα ιδιαίτερα χαρακτηριστικά μιας συγκεκριμένης εφαρμογής. Έτσι, η μέθοδος που προτείνουμε στο κεφάλαιο 2 ταιριάζει πιο πολύ σε εφαρμογές που διακρίνονται για το ιδιαίτερα μικρό πλήθος των δυνατών μετασχηματισμών. Αυτή η νέα μέθοδος μηχανικής εκμάθησης εφαρμόστηκε και στην ανίχνευση ορίων χειρόγραφων χαρακτήρων συνεχόμενης γραφής (handwritten character segmentation), όπου οι δυνατοί μετασχηματισμοί είναι δύο, επιτυγχάνοντας πολύ ικανοποιητικά αποτελέσματα σε σύγκριση με την παραδοσιακή EBM [53].

Οι υφολογικοί δείκτες που εξάγονται με βάση τον ανιχνευτή ορίων περιόδων και φράσεων ανήκουν σε τρία υφομετρικά επίπεδα. Τα επίπεδα δείγματος και φράσης σχετίζονται με την έξοδο των ανιχνευτών ορίων περιόδων και φράσεων αντίστοιχα. Το επίπεδο ανάλυσης είναι ένας εναλλακτικός τρόπος σύλληψης της υφολογικής πληροφορίας και αφορά στον τρόπο με το οποίο αναλύθηκε το κείμενο από τον ανιχνευτή ορίων φράσεων. Αξίζει να σημειωθεί ότι είναι η πρώτη φορά που το σύνολο των υφολογικών δεικτών δεν είναι προκαθορισμένο αλλά εξαρτάται άμεσα από τον τρόπο με τον οποίο γίνεται η μέτρηση των τιμών των παραμέτρων. Σε όλες τις προηγούμενες προσεγγίσεις η διαδικασίες καθορισμού και μέτρησης των τιμών των παραμέτρων ήταν ανεξάρτητες μεταξύ τους. Ο έλεγχος της σημαντικότητας των υφομετρικών επιπέδων, που πραγματοποιήθηκε στα πλαίσια των πειραμάτων αναγνώρισης είδους κειμένου και συγγραφέα, έδειξε ότι το επίπεδο ανάλυσης είναι σαφώς πιο σημαντικό από αυτό της φράσης, ενώ το πιο σημαντικό επίπεδο, με διαφορά, είναι αυτό του δείγματος. Οι παράμετροι του επιπέδου ανάλυσης γίνονται διαθέσιμοι μόνο εφόσον χρησιμοποιείται ο συγκεκριμένος ανιχνευτής ορίων φράσεων. Παρ' όλα αυτά, παρόμοιους υφολογικούς δείκτες μπορούν να δώσουν σχεδόν όλα τα εργαλεία επεξεργασίας φυσικής γλώσσας. Βέβαια, οι δείκτες αυτοί θα

διαφέρουν από εργαλείο σε εργαλείο αφού σχετίζονται με την μέθοδο που γίνεται η ανάλυση.

Ένα άλλο σημείο που διαφοροποιεί το προτεινόμενο σύνολο των υφολογικών δεικτών είναι ότι δεν χρησιμοποιεί κανένα λεξιλογικό δείκτη. Υπενθυμίζεται ότι σχεδόν όλοι οι ερευνητές βασίζουν εξ ολοκλήρου τις μελέτες τους σε τέτοιους δείκτες, παρά τα προβλήματα που έχει αποδειχτεί ότι έχουν (ισχυρή εξάρτηση από το μήκος του κειμένου, αδυναμία χρήσης τους σε οποιοδήποτε σύνολο υφολογικών κατηγοριών κ.ά.). Με τον αποκλεισμό της λεξιλογικής πληροφορίας, το προτεινόμενο σύνολο των υφολογικών δεικτών γίνεται πιο γενικό καθώς ανεξαρτητοποιείται από συγκεκριμένα είδη κειμένων ή συγκεκριμένους συγγραφείς. Ακόμη, γίνεται πιο ανεξάρτητο-γλώσσας.

Οι ίδιοι υφολογικοί δείκτες εφαρμόζονται και για την αναγνώριση είδους κειμένου και για την αναγνώριση συγγραφέα. Τα αποτελέσματα αυτόματης ταξινόμησης γι' αυτές τις δύο περιπτώσεις δείχνουν ότι το προτεινόμενο σύνολο των υφολογικών παραμέτρων είναι ικανό να διακρίνει οποιαδήποτε υφολογικά ομοιογενή κατηγορία. Η απόδοση των προτεινόμενων συστημάτων είναι σημαντικά βελτιωμένη σε σχέση με προηγούμενα συστήματα. Πιο συγκεκριμένα, η απόδοση του συστήματος αναγνώρισης ειδών κειμένων κυμάνθηκε μεταξύ 82-85% για 10 είδη κειμένων την στιγμή που η καλύτερη απόδοση που έχει αναφερθεί ως σήμερα είναι 79% για 6 μόλις είδη κειμένων της Αγγλικής γλώσσας [54]. Επίσης, η ακρίβεια ταξινόμησης του συστήματος αναγνώρισης συγγραφέα έφτασε το 81% την στιγμή που η αντίστοιχη απόδοση της πιο σύγχρονης λεξιλογικής μεθόδου ήταν 74%. Ο συνδυασμός των δύο προσεγγίσεων έδωσε ακόμα πιο ακριβή αποτελέσματα (87%) και αποτελεί την πιο αξιόπιστη λύση σε περιπτώσεις όπου το μέγεθος του σώματος εκπαίδευσης είναι σχετικά μεγάλο (πάνω από 15 κείμενα από κάθε συγγραφέα).

Στα αρχικά πειράματα χρησιμοποιήθηκαν 10 κείμενα από κάθε είδος κειμένου/συγγραφέα για εκπαίδευση. Ο αριθμός αυτός έχει προταθεί ως ικανός να αναπαραστήσει επαρκώς τα υφολογικά χαρακτηριστικά μιας υφολογικής κατηγορίας [9, 10]. Ωστόσο, η αύξηση του μεγέθους του σώματος εκπαίδευσης βελτιώνει την ακρίβεια, ειδικά στην περίπτωση της αναγνώρισης συγγραφέα, αν και η βελτίωση αυτή δεν είναι γραμμική. Πρέπει να επισημανθεί ότι σε πάρα πολλές εφαρμογές υπάρχουν διαθέσιμα ελάχιστα κείμενα για εκπαίδευση. Όσον αφορά την αναγνώριση

συγγραφέα, η μέθοδός μας υπερτερεί τόσο της λεξιλογικής μεθόδου όσο και του συνδυασμού των δύο προσεγγίσεων έχοντας ως βάση μόνο 10 κείμενα από κάθε συγγραφέα ως σώμα εκπαίδευσης.

Ένας άλλος πολύ σημαντικός παράγοντας που επηρεάζει την αξιοπιστία της ταξινόμησης, ειδικά στην αναγνώριση συγγραφέα, είναι το μήκος του κειμένου. Παρατηρήθηκε ότι οι συγγραφείς με μικρό μέσο μήκος κειμένου (μικρότερο από 1.000 λέξεις) είχαν τα χειρότερα αποτελέσματα ταξινόμησης. Παρόμοια προβλήματα αντιμετώπισε και η λεξιλογική προσέγγιση. Φαίνεται λοιπόν ότι οι 1.000 λέξεις προσφέρουν ένα πολύ αξιόπιστο κάτω όριο για την επίτευξη πολύ υψηλής ακρίβειας ταξινόμησης. Ασφαλώς, υπάρχουν πολλές περιπτώσεις (και τα σώματα κειμένων ανά είδος και ανά συγγραφέα που κατασκευάστηκαν ανήκουν σε αυτές) όπου η θέσπιση ενός τέτοιου ορίου δεν έχει νόημα, καθώς τα περισσότερα κείμενα είναι περιορισμένου μήκους.

Τα κείμενα που χρησιμοποιήθηκαν στα πειράματα προήλθαν από τις σελίδες του Διαδικτύου και δεν υπέστησαν καμία χειρονακτική προεπεξεργασία ή δειγματοληψία. Τα κείμενα αυτά ήταν ήδη σε ηλεκτρονική μορφή και είναι πολύ πιθανόν να περιέχουν διάφορα λάθη. Γίνεται λοιπόν εύκολα αντιληπτό ότι η απόδοση των συστημάτων αναγνώρισης είδους κειμένου και συγγραφέα μπορεί να βελτιωθεί αισθητά εφόσον εφαρμοστεί σε προσεκτικά επιλεγμένα κείμενα που θα έχουν ελεγχθεί για τυχόν λάθη. Ωστόσο, στην πλειοψηφία των εφαρμογών μια τέτοια προεπεξεργασία δεν είναι εφικτή, από πρακτική άποψη.

Για την αυτόματη ταξινόμηση των διανυσμάτων των υφολογικών δεικτών χρησιμοποιήθηκαν δύο τεχνικές της πολυπαραγοντικής στατιστικής: η πολλαπλή παλινδρόμηση και η διαχωριστική ανάλυση. Τα αποτελέσματα που επιτεύχθηκαν ήταν τις περισσότερες φορές παρόμοια, αν και το σφάλμα αναγνώρισης ήταν πιο ομαλά κατανεμημένο με βάση τη διαχωριστική ανάλυση Και οι δύο αυτές τεχνικές χαρακτηρίζονται από τις ελάχιστες απαιτήσεις τους σε υπολογιστικό κόστος τόσο για την εκπαίδευσή τους όσο και για την απόκρισή τους καθώς βασίζονται στον υπολογισμό απλών γραμμικών συναρτήσεων. Επομένως, προσφέρουν μία πολύ καλή λύση για την περίπτωση εφαρμογών που απαιτούν απόκριση σε πραγματικό χρόνο. Επίσης, η απόδοσή τους, όταν χρησιμοποιούνται σχετικά λίγα κείμενα από κάθε υφολογική κατηγορία για εκπαίδευση, έχει αποδειχτεί ότι είναι σημαντικά υψηλότερη

σε σχέση με άλλες μεθόδους, όπως τα νευρωνικά δίκτυα και οι ταξινομητές Bayes [102]. Ωστόσο, η αναλυτική σύγκριση των τεχνικών κατηγοριοποίησης και η επιλογή της βέλτιστης μεθόδου ξεφεύγουν από τα πλαίσια αυτής της διατριβής.

7.2 Προοπτικές

Τα συστήματα που παρουσιάστηκαν στην παρούσα διατριβή μπορούν να χρησιμοποιηθούν σε πλειάδα εφαρμογών. Όσον αφορά το εργαλείο ανίχνευσης περιόδων και φράσεων θα μπορούσε να χρησιμοποιηθεί ως προεπεξεργαστής σε σχεδόν όλες τις εφαρμογές της επεξεργασίας φυσικής γλώσσας και ειδικά σε περιπτώσεις όπου απαιτείται γρήγορη ανάλυση μεγάλων όγκων κειμένων. Ενδεικτικά αναφέρουμε, την ανάκτηση, πληροφορίας, την εξαγωγή πληροφορίας καθώς και την εξαγωγή ορολογίας από κείμενα. Ασφαλώς, οι απαιτήσεις για προεπεξεργασία κειμένου διαφέρουν από εφαρμογή σε εφαρμογή. Ωστόσο, τόσο η ανίχνευση ορίων περιόδων όσο και η ανίχνευση ορίων φράσεων περιλαμβάνονται συνήθως στις προαπαιτούμενες εργασίες πιο περίπλοκων σταδίων επεξεργασίας (π.χ. συντακτική ανάλυση, σημαντική ανάλυση). Αξίζει να σημειωθεί ότι το ολοκληρωμένο σύστημα ανίχνευσης ορίων περιόδων και φράσεων προβλέπεται να χρησιμοποιηθεί στα πλαίσια των εθνικών ερευνητικών προγραμμάτων ΜΙΤΟΣ[1], ΔΗΛΟΣ[2] και ΔΕΙΚΤΗΣ[3] και έχουν ήδη αρχίσει οι απαραίτητες τροποποιήσεις για την επίτευξη όσο το δυνατόν καλύτερων αποτελεσμάτων ανάλυσης για τα εκάστοτε σώματα κειμένων.

Το εργαλείο αναγνώρισης είδους κειμένου θα μπορούσε να βρει εφαρμογή κυρίως σε συστήματα ανάκτησης πληροφορίας. Σε μια τέτοια περίπτωση, η αναγνώριση του είδους ενός κειμένου θα λειτουργούσε ως φίλτρο, έτσι ώστε να προσφέρεται στον χρήστη η καλύτερη δυνατή συλλογή κειμένων που ταιριάζουν με τις απαιτήσεις του. Για παράδειγμα, ένας χρήστης μιας «έξυπνης» μηχανής αναζήτησης θα μπορούσε να ζητήσει επιστημονικά άρθρα για τον ανθρώπινο εγκέφαλο και όχι απλά οτιδήποτε κείμενα περιέχουν τις λέξεις «ανθρώπινος εγκέφαλος» όπως ισχύει στις σύγχρονες μηχανές αναζήτησης.

---

[1] «ΜΙΤΟΣ»: Σύστημα Αναζήτησης και Εξόρυξης Πληροφορίας – Εφαρμογή σε Χρηματοοικονομικές Ειδήσεις (ΕΠΕΤ ΙΙ – νέο ΕΚΒΑΝ 2-1.3-102).
[2] «ΔΗΛΟΣ»: Δίγλωσσο Ηλεκτρονικό Λεξικό Οικονομικών Όρων με Αναφορές Εντάσεως των Όρων σε Κείμενα Συλλογών (ΕΠΕΤ ΙΙ – 98ΓΤ-12).
[3] «ΔΕΙΚΤΗΣ»: Διαλογική Εξαγωγή Πληροφορίας από Ιατρικά Κείμενα Τηλεδιαχειριζόμενων Ηλεκτρονικών Σωμάτων (ΕΠΕΤ ΙΙ – 98ΓΤ-24).

Επίσης, το σύστημα αναγνώρισης είδους κειμένου θα μπορούσε να χρησιμοποιηθεί εύκολα για την γρήγορη ταξινόμηση μεγάλων όγκων κειμένου ανάλογα με την υφολογική τους ομοιογένεια. Σε αυτήν την περίπτωση πρέπει να εφαρμοστεί κάποια τεχνική διάκρισης των πιο βασικών υφολογικών κατηγοριών (clustering) και στην συνέχεια να εκπαιδευτεί το σύστημα αναγνώρισης με βάση κείμενα από την κάθε κατηγορία.

Το σύστημα αναγνώρισης και επιβεβαίωσης συγγραφέα εκτός από την αυτονόητη δυνατότητα εφαρμογής του σε περιπτώσεις προσδιορισμού της πατρότητας ανώνυμων ή αμφισβητούμενων κειμένων, είναι δυνατόν να χρησιμοποιηθεί και για την ανίχνευση υφολογικών ομοιοτήτων ή διαφορών μεταξύ δύο ή περισσότερων συγγραφέων. Ωστόσο, η εξήγηση αυτών των ομοιοτήτων ή των διαφορών είναι αρκετά περίπλοκη υπόθεση καθώς πίσω από την στατιστική ανάλυση του ύφους δεν βρίσκεται κάποια τυπική υφολογική θεωρία. Επομένως, η εξαγωγή τέτοιων συμπερασμάτων θα βασίζεται αναπόφευκτα σε υποκειμενικές κρίσεις.

Αρκετή έρευνα απομένει να γίνει όσον αφορά την διερεύνηση των αλλαγών του ύφους σε ένα κείμενο. Για να γίνει αυτό δυνατό θα πρέπει να οριστεί κάποια *μονάδα κειμένου* (δηλ. ένα ελάχιστο τμήμα κειμένου που αναπαριστά επαρκώς τα υφολογικά χαρακτηριστικά ενός συγγραφέα). Ακόμη, η υφολογική μεταβολή μέσα σε ένα κείμενο δεν θα πρέπει να βασίζεται σε υποκειμενικά κριτήρια αλλά να εξάγεται από κάποια (στατιστική ίσως;) αντικειμενική μεθοδολογία. Από την στιγμή πάντως που επιτευχθεί κάτι τέτοιο, η διαδικασία επιλογής συγκεκριμένων τμημάτων του κειμένου που αναπαριστούν καλύτερα το προσωπικό ύφος του συγγραφέα θα μπορέσει να αυτοματοποιηθεί. Επομένως, και η εξάρτηση της ακρίβειας του συστήματος αναγνώρισης συγγραφέα από το μήκος του κειμένου αλλά και από το μέγεθος του σώματος εκπαίδευσης θα εξασθενούσε σημαντικά.