Αναγνώριση γλώσσας ηλεκτρονικού κειμένου

Αθανάσιος Πρωτόπαπας

Α. Ανασκόπηση

1. Εισαγωγή

Με την εξέλιξη των συστημάτων γλωσσικής επεξεργασίας και επικοινωνίας προκύπτει η ανάγκη αυτόματης κατηγοριοποίησης κειμένων προς επεξεργασία ανάλογα με τη γλώσσα στην οποία είναι γραμμένα. Παρόμοια ανάγκη υπάρχει και για την αυτόματη αναγνώριση γλώσσας ομιλίας σε πραγματικό χρόνο (π.χ. για αυτόματη επακόλουθη αναγνώριση ομιλίας ή ακόμα και για τηλεφωνική σύνδεση με κάποιον που μιλάει τη γλώσσα του ατόμου που πραγματοποιεί μια κλήση), όμως το πρόβλημα αυτό είναι εντελώς διαφορετικό και πολύ δυσκολότερο και δεν αναφέρεται στην παρούσα έκθεση. Το πρόβλημα αναγνώρισης της γλώσσας ενός γραπτού κειμένου διαθέσιμου σε ηλεκτρονική μορφή είναι πολύ απλούστερο διότι μπορεί να αναχθεί σε στατιστικό πρόβλημα ενός προκαθορισμένου συνόλου στοιχείων (χαρακτήρων και συνδυασμών τους) χωρίς τα προβλήματα κανονικοποίησης και κατηγοριοποίησης που προκύπτουν για τον ήχο της ομιλίας ή την εικόνα ενός κειμένου από σαρωτή.

Ουσιαστικά το σύνηθες ζητούμενο είναι, για δεδομένη αλληλουχία οκταψήφιων δυαδικών χαρακτήρων (extended ASCII) να επιλεγεί το γλωσσικό μοντέλο (από Ν διαθέσιμα) που είναι βέλτιστα συμβατό με την αλληλουχία βάσει προκαθορισμένης νόρμας (distance metric). Πρόκειται δηλαδή για πρόβλημα ταξινόμησης (κατηγοριοποίησης) και όχι αμιγώς αναγνώρισης διότι στην πράξη αλλά και από σχεδίαση δεν υπάρχει δυνατότητα απόρριψης ενός κειμένου ως «άγνωστης γλώσσας». Αυτό το σημείο αποκτά μεγάλη σημασία στην υλοποίηση και ειδικά σε συγκεκριμένες εφαρμογές διότι καθιστά αναγκαία την πρόβλεψη (και μοντελοποίηση) όλων των γλωσσών που είναι πιθανό να προκύψουν ακόμα και αν δεν υπάρχει ανάγκη ή δυνατότητα περαιτέρω επεξεργασίας παρά για ένα μικρό υποσύνολό τους. Παρά ταύτα, χρησιμοποιείται ο συνήθης όρος «αναγνώριση γλώσσας», με την επιφύλαξη αυτή όσον αφορά στο δόκιμο της χρήσης του.

Όπως και στην πλειονότητα εφαρμογών ταξινόμησης, έτσι και για την περίπτωση της αναγνώρισης γλώσσας υπάρχει η συνήθης ανάγκη συμβιβασμού της αποθηκευτικής ανάγκης με την απαιτούμενη ακρίβεια αναγνώρισης σύντομου κειμένου (ή, ισοδύναμα, με το ελάχιστο απαιτούμενο μήκος κειμένου για δεδομένη ακρίβεια αναγνώρισης). Προφανώς η ακρίβεια αναγνώρισης γλώσσας ενός κειμένου θα τείνει στο μηδέν όταν το μήκος του κειμένου τείνει στο μηδέν. Ούτως ή άλλως για κείμενα ενός ή λίγων μόνο χαρακτήρων δεν τίθεται καν ζήτημα. Η μέγιστη δυνατή απαίτηση προκύπτει πρακτικά για το ελάχιστο μήκος κειμένου ίσο με μια λέξη (το οποίο εξαρτάται από τη γλώσσα!), ενώ η απαίτηση αναγνώρισης από μια παράγραφο λίγων εκατοντάδων χαρακτήρων είναι

ασυμπτωτικά ισοδύναμη με την θεωρητικά ελάχιστη απαίτηση αναγνώρισης γλώσσας κειμένου πολύ μεγάλου μήκους.

Στο πεδίο του μεγέθους των απαιτουμένων γλωσσικών μοντέλων, η οριακή περίπτωση μέγιστης αποθηκευτικής απαίτησης περιλαμβάνει ένα πλήρες λεξιλόγιο για κάθε γλώσσα ενώ η ελάχιστη περίπτωση θα τείνει προς την συγκράτηση ενός μικρού αριθμού συνδυασμών χαρακτήρων (σε ομάδες 1–N χαρακτήρων που μπορεί να σχηματίζουν ολόκληρες λέξεις ή όχι), ενδεχομένως μαζί με στατιστικές πληροφορίες για τη συχνότητα εμφάνισής τους ή για περιορισμούς περικειμένου. Το θεωρητικό μέγιστο αποθήκευσης και χρήσης ενός πλήρους λεξιλογίου αποτελεί και τη βάση αναφοράς για την αποτελεσματικότητα κάθε χρησιμοποιούμενου μοντέλου ενώ δεν είναι πρακτικά υλοποιήσιμη η χρήση του. Η χρήση αυτής της βάσης αναφοράς καθιστά προφανή την εγγενή αδυναμία εκπλήρωσης του στόχου αναγνώρισης όταν μια αλληλουχία χαρακτήρων μπορεί να προκύψει σε περισσότερες από μία «γνωστές» γλώσσες. Δηλαδή η τέλεια επίδοση κάθε δυνατού συστήματος αναγνώρισης γλώσσας είναι εγγενώς μικρότερη από 100%.

Για την κάλυψη ασαφών περιπτώσεων, για τη μείωση του μεγέθους του λεξιλογίου χωρίς μείωση της αξιοπιστίας, αλλά κυρίως για τον ορθό χειρισμό πολυγλωσσικών κειμένων, αναπτύσσονται σύγχρονες μέθοδοι αναγνώρισης γλώσσας φράσεων με χρήση γραμματικών-μορφολογικών μοντέλων γλώσσας και όχι μόνο λεξιλογίων. Πρόκειται για εξελιγμένα (και περίπλοκα) συστήματα γλωσσολογικής ανάλυσης (parsing) τα οποία μπορούν να αντιμετωπίσουν με επιτυχία περιπτώσεις στις οποίες οι «παραδοσιακές» μέθοδοι στατιστικής αλληλουχιών χαρακτήρων είναι απλά ανεπαρκείς. Λόγω της πολυπλοκότητας και της ιδιαιτερότητάς τους, τα συστήματα αυτά δεν εξετάζονται στην παρούσα έκθεση.

2. Μέθοδοι – Περιγραφή

Υπάρχουν αρκετές μέθοδοι στη βιβλιογραφία αλλά και σε online συστήματα επίδειξης στον παγκόσμιο ιστό. Ένας βασικός διαχωρισμός αφορά στη χρήση μονάδων αυθαίρετου ή σταθερού μήκους, καθώς και στη χρήση μονάδων-λέξεων ή αλληλουχιών χαρακτήρων που δεν σχηματίζουν απαραίτητα λέξεις. Οι υπάρχουσες μέθοδοι διαφέρουν επίσης και στη διάσταση της πολυπλοκότητας, από απλή αναγνώριση μονάδων-«κλειδιών» ή καταμέτρηση συχνότητας εμφάνισης αλληλουχιών μέχρι HMM.

2.1. Γράμματα (χαρακτήρες)

Πολλές γλώσσες μπορούν να διακριθούν με βάση μόνο τους χαρακτήρες που χρησιμοποιούνται στη γραφή τους. Π.χ. για τα αγγλικά χρησιμοποιούνται μόνο τα 26 γράμματα του λατινικού αλφαβήτου ενώ για τα γερμανικά χρησιμοποιούνται επίσης και τα ä, ö, ü, ß, για τα γαλλικά τα é, è, ê, ô, á, à, ç, για τα ισπανικά τα ñ, á, é, í, ó, ú, ü, κ.λπ. Για τα ελληνικά (και μόνο) χρησιμοποιείται το ελληνικό αλφάβητο ενώ για τις περισσότερες σλαβικές γλώσσες χρησιμοποιείται το κυριλλικό αλφάβητο. Οι

διαφορές αυτές μπορούν να χρησιμοποιηθούν με αρκετή αξιοπιστία στην αναγνώριση της γλώσσας ενός εντύπου, όχι όμως τόσο στην αναγνώριση ηλεκτρονικού κειμένου. Οι λόγοι είναι δύο: κωδικοποίηση και ασυνέπεια.

Συγκεκριμένα, ο τρόπος συμβολισμού των διαφόρων χαρακτήρων εξαρτάται από το πρότυπο κωδικοποίησης χαρακτήρων, σήμερα συνήθως στις 128 θέσεις οκταψήφιων δυαδικών χαρακτήρων πάνω από τα ASCII και ανάλογα με το αν χρησιμοποιείται ISO-8859-1, ISO-8859-2, ή CP1250 (για το λατινικό αλφάβητο), ISO-8859-7 ή CP1253 (για το ελληνικό αλφάβητο), κ.ο.κ. Η κωδικοποίηση συνήθως δεν είναι γνωστή πριν από τη γλώσσα και ανεξάρτητα από αυτή και θα πρέπει στην τυπική περίπτωση να αναγνωριστεί συγχρόνως (ενδεχομένως και για την ορθή απεικόνιση του κειμένου).

Ακόμα και αν υποτεθεί ότι το ζήτημα της κωδικοποίησης με κάποιο τρόπο λύνεται ή παρακάμπτεται, κατ' αρχήν με χρήση μαρκαρισμένων κειμένου και ειδικά στο μέλλον με την ευρεία χρήση του Unicode, πιο σοβαρό επιχείρημα ενάντια στη χρήση της μεθόδου αυτής είναι η ασυνέπεια μεταξύ κειμένων στην εφαρμογή της ορθογραφίας κάθε γλώσσας. Π.χ. για τα γερμανικά, τα γαλλικά και τα ισπανικά είναι πολύ συνηθισμένο να χρησιμοποιούνται μόνο οι απλοί 26 χαρακτήρες χωρίς τόνους ή άλλα διακριτικά (και στα γερμανικά οι ισοδύναμοι τύποι ae, oe, ue, και ss) είτε διότι κάποιο λογισμικό ή υλικό δεν υποστηρίζει την εμφάνιση ή την πληκτρολόγηση (εισαγωγή) άλλων χαρακτήρων, είτε διότι κάποιος χρήστης δεν γνωρίζει ή δεν κάνει την επιπλέον προσπάθεια εισαγωγής των ειδικών χαρακτήρων ή άλλων σημαδιών (τονικών κ.ά.).

Επιπλέον, υπάρχουν διάφοροι τρόποι συμβολισμού ειδικών χαρακτήρων και σημαδιών που χρησιμοποιούνται ως επί το πλείστον σε μηνύματα ηλεκτρονικού ταχυδρομείου και σε άλλες περιπτώσεις όταν οι συνθήκες ιστορικά επιβάλλουν ή ενθαρρύνουν τη χρήση απλού ASCII. Τέτοιοι είναι π.χ. ο συμβολισμός του ό ως 'o, o', /o, ή `o, ο συμβολισμός του ñ ως ~n, n~, ή n#, κ.λπ. Προφανώς η ποικιλία των συμβολισμών αυτών αποκλείει την αξιόπιστη ανίχνευση των στοιχείων εκείνων που χαρακτηρίζουν κάθε γλώσσα.

Ένας ακόμα λόγος που καθιστά ασύμφορη τη χρήση της μεθόδου αυτής είναι και η προβληματική επεκτασιμότητά της, διότι δεν μπορεί να είναι εκ των προτέρων γνωστό αν υπάρχει κατάλληλο σύνολο χαρακτήρων που να εμφανίζεται με υψηλή συχνότητα και να χαρακτηρίζει επαρκώς κάθε γλώσσα η οποία μπορεί να χρειαστεί να ενσωματωθεί στο σύστημα μελλοντικά.

2.2. Συχνές λέξεις

Μια άλλη μέθοδος, αρκετά παλιά όπως και η προηγούμενη, βασίζεται στην ύπαρξη ορισμένων μικρών πολύ συχνών λέξεων σε κάθε γλώσσα οι οποίες είναι χαρακτηριστικές για τη γλώσσα αυτή. Τέτοιες είναι συνήθως άρθρα, ο συνδετικός σύνδεσμος, κ.ά. Η μέθοδος αυτή αποτελεί ειδική περίπτωση της μεθόδου Ν-γραμμάτων που περιγράφεται στη συνέχεια, και μάλιστα πρόκειται για ιδιαίτερα δύσκαμπτη και απαιτητική μέθοδο και για το λόγο αυτό δε χρησιμοποιείται εδώ και αρκετά χρόνια.

Συγκεκριμένα, για να μπορεί να λειτουργήσει η μέθοδος αυτή πρέπει να εξασφαλίζεται κάποιο σημαντικό μέγεθος κειμένου ώστε η πιθανότητα να απαντάται κάποια από τις χαρακτηριστικές λέξεις να είναι πολύ μεγάλη. Επίσης θα πρέπει να παρέχεται ρέον και όχι τηλεγραφικό κείμενο διότι στο τελευταίο τέτοιες λέξεις είναι αυτές που συνήθως παραλείπονται. Οι δύο αυτοί περιορισμοί είναι σημαντικοί για ένα σύστημα αναγνώρισης γλώσσας που μπορεί να κληθεί να αναγνωρίσει σύντομα ή συνθηματικά μηνύματα. Το πρόβλημα επιδεινώνεται από το γεγονός ότι η μέθοδος αυτή δεν παρέχει διαβαθμισμένο αποτέλεσμα αλλά αποτυγχάνει εντελώς όταν δεν υπάρχουν στο κείμενο κάποιες από τις χαρακτηριστικές λέξεις. Δεν μπορεί δηλαδή, για κάποιο σύντομο κείμενο, να δώσει μια μεγάλη πιθανότητα αυτό να είναι γραμμένο σε κάποια συγκεκριμένη γλώσσα και να δώσει τη δυνατότητα αξιολόγησης μιας πιθανοτικής επιλογής. Τέλος, όπως και στην προηγούμενη μέθοδο, η επεκτασιμότητα της μεθόδου αυτής είναι αμφίβολη διότι δεν μπορεί να προεξοφληθεί ότι για κάθε γλώσσα που μπορεί να χρειαστεί στο μέλλον θα υπάρχουν κατάλληλες αρκετά συχνές λέξεις οι οποίες θα είναι και διαφορετικές από αυτές κάθε άλλης υπάρχουσας γλώσσας.

2.3. Στατιστικές διγραμμάτων (και N-γραμμάτων)

Η μέθοδος αυτή μάλλον πρωτοχρησιμοποιήθηκε από το σύστημα της εταιρείας Automatic Language Processing Systems που κατασκεύασε ο Kenneth Beesley στα τέλη της δεκαετίας του ’80 [4]. Πρόκειται για εκμετάλλευση των στατιστικών χαρακτηριστικών κάθε γλώσσας όσον αφορά στη συχνότητα εμφάνισης ζευγών χαρακτήρων, διότι η ορθογραφική κανονικότητα και συστηματικότητα κάθε γλώσσας είναι εμφανής τόσο στις επιτρεπόμενες αλληλουχίες χαρακτήρων όσο και στη συχνότητα εμφάνισης κάθε αλληλουχίας. Π.χ., κάθε δίψηφο σύμφωνο ή φωνήεν, δίφθογγος, πρόθεμα ή μορφολογική κατάληξη αποτελούν αλληλουχίες χαρακτήρων που εμφανίζονται τακτικά ενώ υπάρχουν σε κάθε γλώσσα άλλες αλληλουχίες χαρακτήρων που απαγορεύονται ή εμφανίζονται σχετικά σπάνια. Η αρχή αυτή γενικεύεται παρακάτω σε αλληλουχίες N χαρακτήρων ενώ εδώ αναφερόμαστε ειδικά σε ζεύγη (N=2) χαρακτήρων.

Για να λειτουργήσει η μέθοδος αυτή απαιτείται φυσικά κάποιο ευμέγεθες σώμα κειμένου από το οποίο θα εξαχθούν τα στατιστικά χαρακτηριστικά κάθε γλώσσας. Το μέγεθος του κειμένου για κάθε γλώσσα πρέπει να είναι τουλάχιστον της τάξης των 100 Kbytes. Το είδος του κειμένου πρέπει να είναι αντιπροσωπευτικό των κειμένων που θα κληθεί να ταξινομήσει γλωσσικά το σύστημα αν προβλέπεται χρήση σε κείμενα ειδικού τύπου, αλλιώς θα πρέπει το κείμενο να είναι όσο το δυνατόν πιο ουδέτερο και να ποικίλλει λεξιλογικά και υφολογικά ώστε να καλύπτει αντιπροσωπευτικά τη γλώσσα και όχι κάποιο συγκεκριμένο είδος περιεχομένου. Πριν από την επεξεργασία καταμέτρησης απαιτείται προεπεξεργασία «καθαρισμού» του κειμένου από σημεία στίξης και ενοποίηση όλων των λεκτικών συνόρων (white space) σε ένα χαρακτήρα. Στη συνέχεια κατασκευάζονται τα γλωσσικά μοντέλα από τη στατιστική επεξεργασία των σωμάτων κειμένου. Στη συγκεκριμένη περίπτωση καταρτίζεται πίνακας όλων των πιθανών ζευγών χαρακτήρων.και μετράται, για κάθε γλώσσα, η σχετική συχνότητα

εμφάνισης κάθε ζεύγους χαρακτήρων στο αντίστοιχο σώμα κειμένου για τη γλώσσα αυτή. Από τον πίνακα που προκύπτει για κάθε γλώσσα μπορούμε είτε να υπολογίσουμε τη λογαριθμική πιθανότητα εμφάνισης κάθε ζεύγους είτε να ταξινομήσουμε τα ζεύγη χαρακτήρων κατά συχνότητα εμφάνισης. Αυτό εξαρτάται από τον τρόπο με τον οποίο θα γίνεται η αναγνώριση, όπως εξηγείται στη συνέχεια.

Ο τρόπος με τον οποίο γίνεται η επιλογή της γλώσσας από τις γνωστές μπορεί να βασίζεται σε σύγκριση κατανομής, σε πλειοψηφικό σύστημα ή σε συγκεντρωτική πιθανότητα, ως εξής:

2.3.1 Σύγκριση κατανομής

Από τη μέτρηση της σχετικής συχνότητας εμφάνισης διγραμμάτων καταρτίζουμε πίνακα διφωνημάτων ταξινομημένο κατά συχνότητα. Αντικαθιστούμε δηλαδή την αριθμητική πληροφορία της σχετικής συχνότητας με την πληροφορία της διάταξης (θέσης μετά την ταξινόμηση). Η διάταξη αυτή των διγραμμάτων (ή Ν-γραμμάτων για τη γενικότερη περίπτωση) αποτελεί και το γλωσσικό μοντέλο το οποίο αποθηκεύεται και αποτελεί μέρος της μόνιμης βάσης δεδομένων του επιλογέα γλώσσας.

Κάθε κείμενο προς αναγνώριση γλώσσας υπόκειται σε όμοια ανάλυση καταμέτρησης των εμφανιζομένων ζευγών χαρακτήρων και ταξινόμησής τους κατά σχετική συχνότητα εμφάνισης. Στη συνέχεια συγκρίνεται για κάθε γλωσσικό μοντέλο η σειρά των διγραμμάτων στο μοντέλο με τη σειρά διγραμμάτων που προέκυψε από την ανάλυση του κειμένου. Η σύγκριση αυτή μπορεί να γίνει είτε υπολογίζοντας για κάθε δίγραμμα τη διαφορά στην απόλυτη θέση κατάταξης μεταξύ μοντέλου και κειμένου είτε με πιο περίπλοκες μεθόδους όπως ο υπολογισμός κάποιου μη παραμετρικού δείκτη συνάφειας (π.χ. συνάφεια διάταξης κατά Spearman, δείκτης τ κατά Kendall, κ.ά.).

Η γλώσσα στης οποίας το μοντέλο η διάταξη διγραμμάτων είναι πιο όμοια με τη διάταξη διγραμμάτων στο κείμενο προς αναγνώριση επιλέγεται ως γλώσσα του κειμένου.

Παρατήρηση: Όπως και σε όλες τις στατιστικές μεθόδους, η επιλογή γλώσσας είναι σχετική και μπορεί να γίνει μόνο συγκριτικά μεταξύ γλωσσών. Δεν υπάρχει αξιόπιστος τρόπος απόρριψης όλων των γλωσσών με εξαγωγή απόφασης «άγνωστης γλώσσας», ειδικά για κείμενα μικρού μεγέθους.

2.3.2 Πλειοψηφική επιλογή

Το γλωσσικό μοντέλο για κάθε γλώσσα αποτελείται από το σύνολο των σχετικών συχνοτήτων (πιθανοτήτων) εμφάνισης όλων των ζευγών χαρακτήρων (διγραμμάτων). Οι τιμές αυτές αποθηκεύονται και αποτελούν μέρος της μόνιμης βάσης δεδομένων του επιλογέα γλώσσας.

Κάθε κείμενο προς αναγνώριση γλώσσας αναλύεται κατ' αρχήν σε λέξεις (απορρίπτοντας σημεία στίξης και ενοποιώντας διαφορετικούς χαρακτήρες διαστήματος) και στη συνέχεια κάθε λέξη αναλύεται σε διγράμματα (ή σε Ν-γράμματα στη γενικότερη περίπτωση) χωρίς να παραλείπονται και τα δύο «ζεύγη» με το χαρακτήρα διαστήματος στην αρχή και στο τέλος της λέξης. Στη συνέχεια για

κάθε λέξη πολλαπλασιάζονται μεταξύ τους οι σχετικές συχνότητες εμφάνισης όλων των διγραμμάτων που περιέχει και αυτό γίνεται με βάση κάθε ένα γλωσσικό μοντέλο ξεχωριστά. Το αποτέλεσμα είναι ανάλογο της σχετικής πιθανότητας της λέξη αυτή να προέρχεται από κάθε αντίστοιχη γλώσσα. Η γλώσσα από της οποίας το μοντέλο προκύπτει η μεγαλύτερη «πιθανότητα» για την κάθε λέξη θεωρείται ότι είναι η γλώσσα στην οποία είναι γραμμένη η λέξη αυτή, η οποία πλέον συνιστά μια «ψήφο» για τη συγκεκριμένη γλώσσα.

Επειδή για όλες τις γλώσσες χρησιμοποιείται ο ίδιος αριθμός πολλαπλασιαστέων δεν είναι απαραίτητο να κανονικοποιηθεί το αποτέλεσμα και να υπολογιστεί ακριβώς η πιθανότητα βάσει του τύπου του Bayes. Αν κάποιο ζεύγος χαρακτήρων δεν εμφανίζεται ποτέ σε μια γλώσσα σύμφωνα με το αντίστοιχο γλωσσικό μοντέλο τότε η αντίστοιχη σχετική συχνότητα είναι μηδέν και το γινόμενο για τη λέξη μηδενίζεται, οπότε και αποκλείεται από την «ψηφοφορία» η γλώσσα αυτή.

Τελικώς γλώσσα του κειμένου θεωρείται η γλώσσα εκείνη με τις περισσότερες «ψήφους», με το μεγαλύτερο δηλαδή αριθμό λέξεων που να μεγιστοποιούν τη συγκεντρωτική πιθανότητα των διγραμμάτων τους χρησιμοποιώντας το αντίστοιχο γλωσσικό μοντέλο. Ας σημειωθεί ότι ο υπολογισμός αυτός περιέχει μια σημαντική μη-γραμμικότητα στον υπολογισμό διότι κάθε λέξη τελικά συνεισφέρει με το ίδιο βάρος στην τελική απόφαση της γλώσσας του κειμένου ανεξάρτητα από το από πόσα και πόσο συχνά διγράμματα αποτελείται. Εναλλακτικά, η μέθοδος συγκεντρωτικής πιθανότητας που ακολουθεί διαφέρει ακριβώς στο σημείο αυτό διότι βασίζεται στη συγκεντρωτική πιθανότητα όλων των διγραμμάτων του κειμένου με ίδιο βάρος (αντί όλων των λέξεων με το ίδιο βάρος ανεξαρτήτως συχνότητας και πλήθους διγραμμάτων).

2.3.3 Συγκεντρωτική πιθανότητα

Αυτή είναι και η μέθοδος που είναι υλοποιημένη και λειτουργεί στο ΙΕΛ σήμερα.

Από τη μέτρηση της σχετικής συχνότητας εμφάνισης διγραμμάτων στο σώμα κειμένου υπολογίζουμε και αποθηκεύουμε για κάθε δίγραμμα (ή Ν-γραμμα στη γενικότερη περίπτωση) το λογάριθμο της πιθανότητας εμφάνισής του. Οι τιμές αυτές λογαριθμικής συχνότητας για το σώμα κειμένου κάθε γλώσσας αποτελούν και το γλωσσικό μοντέλο για τη γλώσσα αυτό το οποίο αποθηκεύεται και αποτελεί μέρος της μόνιμης βάσης δεδομένων του επιλογέα γλώσσας.

Κάθε κείμενο προς αναγνώριση γλώσσας αναλύεται σε διγράμματα (ή σε Ν-γράμματα στη γενικότερη περίπτωση), αφού απορριφθούν τα σημεία στίξης και ενοποιηθούν οι διαφορετικοί χαρακτήρες διαστήματος και χωρίς να παραλείπονται και τα δύο «ζεύγη» με το χαρακτήρα διαστήματος στην αρχή και στο τέλος κάθε λέξης. Στη συνέχεια αθροίζονται όλες οι λογαριθμικές συχνότητες εμφάνισης όλων των διγραμμάτων που περιέχει και αυτό γίνεται με βάση κάθε ένα γλωσσικό μοντέλο ξεχωριστά. Το αποτέλεσμα είναι ανάλογο της σχετικής πιθανότητας του κειμένου συνολικά να είναι γραμμένο στην αντίστοιχη γλώσσα, βάσει της στατιστικής των διγραμμάτων. Η γλώσσα από της

οποίας το μοντέλο προκύπτει η μεγαλύτερη «πιθανότητα» για το κείμενο θεωρείται ότι είναι η γλώσσα στην οποία είναι γραμμένο το κείμενο.

Ο λόγος που χρησιμοποιούνται λογάριθμοι και όχι απευθείας οι σχετικές συχνότητες, παρ' ό,τι πρόκειται για υπολογισμό συγκεντρωτικής πιθανότητας, είναι για να αποφευχθεί η υποχείλιση του υπολογισμών μετά από πολλούς πολλαπλασιασμούς μικρών αριθμών (ειδικά για ένα κείμενο σχετικά μεγάλο όπου κατά πάσα πιθανότητα θα εμφανίζονται και διγράμματα με σχετική συχνότητα εμφάνισης πολύ μικρότερη του 1.0). Το άθροισμα λογαρίθμων ισοδυναμεί με πολλαπλασιασμό των σχετικών συχνοτήτων και εφόσον μας ενδιαφέρει μόνο η σύγκριση των αποτελεσμάτων για κάθε γλωσσικό μοντέλο μας αρκεί η μονοτονικότητα της λογαριθμικής συνάρτησης.

Η μέθοδος αυτή υπολογισμού και σύγκρισης είναι παρόμοια με την προηγούμενη με τη διαφορά ότι εδώ δεν παρεμβάλλεται η μη-γραμμικότητα της αποκοπής κατά λέξη κι έτσι κάθε δίγραμμα συνεισφέρει με ίδιο βάρος στον τελικό υπολογισμό. Έτσι μια λέξη με διγράμματα μεγάλης συχνότητας (σε μια γλώσσα) μπορεί να αποβεί πιο σημαντική για την τελική κατάταξη του κειμένου από μια άλλη λέξη με διγράμματα χαμηλότερης συχνότητας, ενώ στην προηγούμενη περίπτωση αυτό δεν μπορεί να συμβεί.

Οι προαναφερθείσες μέθοδοι για ζεύγη χαρακτήρων (διγράμματα) μπορούν να εφαρμοστούν και για απλούς χαρακτήρες αλλά και για μεγαλύτερα τμήματα κειμένου, π.χ. τριγράμματα, τετραγράμματα, κ.ο.κ. Από δοκιμές μου αλλά και από τη βιβλιογραφία προκύπτει ότι η χρήση απλών γραμμάτων δεν είναι αξιόπιστη. Επίσης προκύπτει ότι δεν χρησιμοποιούνται αλληλουχίες μήκους μεγαλύτερου από τρία (δηλαδή τριγράμματα) – δεν παρατηρήθηκε καμία αναφορά σε χρήση τετραγραμμάτων ή αλληλουχιών μεγαλύτερου μεγέθους – εκτός από την περίπτωση μικτών N-γραμμάτων που αναφέρεται στη συνέχεια, στην οποία περιλαμβάνονται λίγα μόνο N-γράμματα μήκους μεγαλύτερου του 3 (τα πολύ συχνά).

Δεν υπάρχει στη βιβλιογραφία συγκριτική μελέτη μεταξύ μεθόδων που να διαφέρουν αποκλειστικά στο μήκος των τμημάτων-μονάδων που χρησιμοποιούνται. Κάθε ερευνητής προτείνει και υλοποιεί μια δικιά του μέθοδο και τη συγκρίνει με μια ή δύο άλλες το πολύ (ή και με καμία, παραθέτοντας απλώς αποτελέσματα για τη μέθοδό του). Οι διαφορές μεταξύ των κειμένων που χρησιμοποιούνται σε κάθε μελέτη (τόσο για τη σύνταξη των γλωσσικών μοντέλων όσο και για τη μέτρηση της απόδοσης) καθιστούν αδύνατη την ουσιαστική σύγκριση μεταξύ μεθόδων.

Ας αναφερθεί εδώ και μια συγκριτική μελέτη μεθόδων αναγνώρισης γλώσσας [3], κατά την οποία συγκρίθηκε η μέθοδος των συχνών λέξεων με τη μέθοδο της συγκεντρωτικής πιθανότητας τριγραμμάτων. Τα αποτελέσματα έδειξαν τεράστια υπεροχή της μεθόδου τριγραμμάτων για σύντομα κείμενα (1-2 και 3-5 λέξεις) αλλά ισοδυναμία των μεθόδων για μεγαλύτερα κείμενα (6 λέξεις και πάνω). Για κείμενα άνω των 50 λέξεων και οι δύο μέθοδοι είχαν 100% επιτυχία μεταξύ 10 ευρωπαϊκών γλωσσών.

2.4. Διάταξη μικτών N-γραμμάτων

Μια μέθοδος όμοια με αυτές της οικογένειας των N-γραμμάτων που περιγράφηκε στην ενότητα 2.3 έχει παρουσιαστεί πιο πρόσφατα [2], η οποία διαφέρει κυρίως στο ότι χρησιμοποιεί αδιακρίτως και συγχρόνως αλληλουχίες χαρακτήρων διαφόρων μεγεθών με μόνο κριτήριο τη συχνότητα εμφάνισής τους. Συγκεκριμένα, καταμετρούνται οι απλοί χαρακτήρες (1-γράμματα), τα ζεύγη χαρακτήρων (διγράμματα), οι τριάδες χαρακτήρων (τριγράμματα) και ενδεχομένως και μεγαλύτερες αλληλουχίες (τετράδες, πεντάδες, κ.ο.κ). Στη βιβλιογραφία αναφέρεται χρήση N-γραμμάτων για N από 1 έως 5. Στη συνέχεια ταξινομούνται κατά σχετική συχνότητα εμφάνισης όλα τα N-γράμματα ανεξαρτήτως μεγέθους, έτσι ώστε ένα ζεύγος γραμμάτων π.χ. που εμφανίζεται συχνότερα από όσο συναντάται κάποιο σπάνιο γράμμα μόνο του μπορεί να βρίσκεται ψηλότερα στην κατάταξη. (Προφανώς δεν είναι δυνατόν κάποιο δίγραμμα να είναι συχνότερο από κάποιον από τους δύο χαρακτήρες που το απαρτίζουν!)

Πρέπει να καταστεί σαφές ότι δεν είναι δυνατόν να καταρτιστεί πίνακας όλων των N-γραμμάτων για N>2 εφόσον κάτι τέτοιο θα ήταν και απαγορευτικό από πλευράς αποθηκευτικών απαιτήσεων αλλά και ανούσιο διότι οι περισσότεροι συνδυασμοί για N>3 δε θα εμφανίζονται ποτέ ή σχεδόν ποτέ. Έτσι γίνεται εναλλακτικά καταγραφή και καταμέτρηση συγχρόνως των συνδυασμών που εμφανίζονται στο σώμα κειμένου κάθε γλώσσας, για κάθε N.

Στη συνέχεια διατηρούνται τα k συχνότερα N-γράμματα και αποτελούν, ως διατεταγμένο σύνολο, το γλωσσικό μοντέλο για κάθε γλώσσα από το σώμα κειμένου της οποίας έχουν υπολογιστεί. Το k μπορεί να κυμαίνεται πρακτικά μεταξύ 100 και 500 στοιχείων. Παρομοίως με την προαναφερθείσα μέθοδο σύγκρισης κατανομής, κάθε κείμενο προς αναγνώριση γλώσσας υφίσταται όμοια επεξεργασία (καταγραφή, καταμέτρηση, και ταξινόμηση των N-γραμμάτων που περιέχει, για N μεταξύ δύο προκαθορισμένων τιμών) με αποτέλεσμα ένα διατεταγμένο σύνολο N-γραμμάτων. Το σύνολο αυτό συγκρίνεται με το αντίστοιχο κάθε γλωσσικού μοντέλου με τη διαφορά ότι επειδή δεν υπάρχουν κατ' ανάγκη τα ίδια μέλη στα δύο σύνολα (και προφανώς δεν είναι αναμενόμενο να υπάρχουν εφόσον βασιζόμαστε στο ότι οι συχνότερες αλληλουχίες χαρακτήρων διαφέρουν μεταξύ γλωσσών) θέτουμε μια αυθαίρετη μεγάλη τιμή «εκτός συνόλου» η οποία αντιστοιχεί στη διαφορά κατάταξης ενός στοιχείου που υπάρχει μόνο σε ένα από τα δύο σύνολα. Τελικά ως γλώσσα του κειμένου επιλέγεται η γλώσσα εκείνη το μοντέλο της οποίας είναι πιο όμοιο με το σύνολο συνηθέστερων N-γραμμάτων του κειμένου.

2.5. HMM

Στη βιβλιογραφία έχει αναφερθεί επιτυχής εφαρμογή κρυμμένων Μαρκοβιανών μοντέλων, συγκεκριμένα ενός εργοδικού μοντέλου επτά καταστάσεων, στην αναγνώριση της γλώσσας ηλεκτρονικού κειμένου[5]. Δεδομένης της υψηλής απόδοσης των προαναφερθεισών απλούστερων

στατιστικών μεθόδων πρώτης τάξεως είναι αμφίβολης σκοπιμότητας η χρήση τέτοιων μοντέλων για το πρόβλημα της αναγνώρισης γλώσσας.

3. Επιδόσεις

Για τις μεθόδους που χρησιμοποιούνται σήμερα (εκτός δηλαδή από τις μεθόδους ειδικών χαρακτήρων και συχνών λέξεων) τα αποτελέσματα που παρατηρούνται στη βιβλιογραφία αλλά και η σύμφωνη εμπειρία στο ΙΕΛ δείχνουν ότι η επιτυχία εξαρτάται κατά κύριο λόγο από το μήκος του προς αναγνώριση κειμένου και ξεπερνά χωρίς δυσκολία το 99% για κείμενα μερικών εκατοντάδων χαρακτήρων όταν ζητείται επιλογή μεταξύ λίγων μέχρι λίγων δεκάδων γλωσσών.

Β. Σχεδίαση και υλοποίηση του συστήματος του ΙΕΛ

Στην παρούσα ενότητα περιγράφεται το σύστημα ταξινόμησης κειμένων κατά γλώσσα που αναπτύχθηκε στο ΙΕΛ με μια τροποποιημένη μέθοδο συγκεντρωτικής πιθανότητας διγραμμάτων, το οποίο για την ώρα περιλαμβάνει γλωσσικά μοντέλα για πέντε γλώσσες: ελληνικά (δύο κωδικοποιήσεις, σύνολο τέσσερις γραφές), αγγλικά, γερμανικά, γαλλικά και ολλανδικά. Το σύστημα μπορεί πολύ εύκολα να επεκταθεί σε άλλες γλώσσες και κωδικοποιήσεις με μόνη απαίτηση την εκτέλεση του προγράμματος προεπεξεργασίας σε ένα εκτενές σώμα κειμένων για κάθε νέα γλώσσα. Το σύστημα μπορεί να λειτουργήσει αυτόνομα ή στη μορφή δυναμικά συνδεόμενης βιβλιοθήκης (DLL). Η λειτουργία του συστήματος έχει δοκιμαστεί εκτενώς σε διάφορες συνθήκες και για διάφορα μήκη ακολουθιών χαρακτήρων και η απόδοσή του είναι εφάμιλλη των ήδη υπαρχόντων στη βιβλιογραφία και στην αγορά. Τα αποτελέσματα των δοκιμασιών παρουσιάζονται σε επόμενη ενότητα.

4. Μέθοδος ταξινόμησης

Η υλοποιημένη μέθοδος βασίζεται στη συγκεντρωτική πιθανότητα διγραμμάτων, υπολογισμένη επί του συνόλου της διαθέσιμης ακολουθίας χαρακτήρων (χωρίς περιορισμούς κατά λέξη ή φράση-πρόταση).

4.1. Πιθανότητες

Με την υπόθεση ότι κάθε γλώσσα είναι εκ προοιμίου ισοπίθανη για δεδομένο διαθέσιμο κείμενο, η σχετική πιθανότητα καθεμιάς δεδομένου του κειμένου ισούται με την πιθανότητα του κειμένου δεδομένου του αντίστοιχου γλωσσικού μοντέλου και δεν απαιτείται μετατροπή των πιθανοτήτων βάσει του νόμου του Bayes. Η πιθανότητα του κειμένου για δεδομένο γλωσσικό μοντέλο υπολογίζεται πολλαπλασιάζοντας την πιθανότητα κάθε διγράμματος που περιλαμβάνεται (ή, για μοντέλο απλών γραμμάτων, την πιθανότητα κάθε ενός γράμματος). Όμως από το σημείο αυτό στην πράξη εγκαταλείπεται η ορθή έννοια της πιθανότητας και πολλές υλοποιήσεις περνούν στη χρήση κατασκευασμάτων βασισμένων σε πιθανότητες που όμως δεν ακολουθούν τα θεωρητικά μοντέλα.

Αυτό συμβαίνει διότι τα διγράμματα δεν είναι μεταξύ τους ανεξάρτητα αλλά, εφόσον χρησιμοποιούνται όλα, υπάρχει σημαντική εξάρτηση μεταξύ διαδοχικών διγραμμάτων λόγω του υποχρεωτικά κοινού γράμματος. Το πρόβλημα αντιμετωπίζεται μερικώς με την κανονικοποίηση των πιθανοτήτων εμφάνισης κάθε διγράμματος ως προς το πρώτο γράμμα του, παράγοντας έτσι μια δεσμευμένη πιθανότητα η οποία μπορεί με αντίστοιχη κανονικοποίηση στην εφαρμογή του μοντέλου να αντιστοιχιστεί στην κατάλληλη ποσότητα. Από διάφορες δοκιμές που έγιναν όσον αφορά στη χρήση των πιθανοτήτων, με και χωρίς κανονικοποίηση, προέκυψε ότι η ορθή ή όχι εφαρμογή δεν επηρεάζει το αποτέλεσμα, δηλαδή την τελική επίδοση του συστήματος. Συνεπώς στην τελική εφαρμογή δεν έχει ληφθεί μέριμνα για την αποκανονικοποίηση και αποσυσχέτιση των διγραμματικών πιθανοτήτων στον υπολογισμό της συγκεντρωτικής ποσότητας, με μόνη συνέπεια την επιφύλαξη για τη χρήση του όρου «συγκεντρωτική πιθανότητα». Η παρατήρηση αυτή ισχύει για όλες τις κατωτέρω αναφορές σε συγκεντρωτική πιθανότητα, καθώς και για τους πίνακες συχνοτήτων και λογαριθμικών πιθανοτήτων στο παράρτημα.

4.2. Ακολουθίες χαρακτήρων και περιορισμοί

Το σύστημα επεξεργάζεται προς ταξινόμηση οποιαδήποτε ακολουθία χαρακτήρων οκτώ ψηφίων, χωρίς εγγενή περιορισμό μήκους πέρα από τη διαθέσιμη μνήμη του υπολογιστή που μπορεί να διατεθεί σε έναν πίνακα χαρακτήρων. Στα γλωσσικά μοντέλα περιέχονται οι κανονικοποιημένες συχνότητες εμφάνισης όλων των διγραμμάτων μεταξύ των επιτρεπομένων χαρακτήρων. Σε αυτούς περιλαμβάνονται όλα τα γράμματα του λατινικού αλφαβήτου, κεφαλαία και πεζά (θέσεις ASCII από 65 έως 90 και από 97 έως 122), ένας χαρακτήρας γενικευμένου διαστήματος (κατά σύμβαση εδώ ίσος με το διάστημα ASCII 32, καθώς και οι χαρακτήρες extended ASCII από τη θέση 192 και μέχρι τη θέση 254, στους οποίους περιλαμβάνονται οι ελληνικοί χαρακτήρες στην κωδικοποίηση ISO-8859-7, τονισμένοι και άτονοι, καθώς και χαρακτήρες με διακριτικά σημεία (τόνοι, umlaut, κ.ά.) που χρησιμοποιούνται σε άλλες ευρωπαϊκές γλώσσες. Δεν θεωρήθηκε πρακτικά εφικτός ο κατακερματισμός του συνόλου extended ASCII διότι το ποιοι χαρακτήρες είναι κάθε φορά επιτρεπτοί εξαρτάται από τη γλώσσα, ενώ για τους υπολογισμούς είναι στη γενική περίπτωση απαραίτητη η ένωση των συνόλων των χαρακτήρων που χρησιμοποιούνται σε όλες τις γλώσσες.

Ας σημειωθεί ότι δεν έχουν ληφθεί καθόλου υπόψη σημεία στίξης ακόμα και αν ορισμένα απ' αυτά (π.χ. η απόστροφος) ίσως θα μπορούσαν να παράσχουν χρήσιμη πληροφορία. Δεδομένης της πολύ υψηλής επίδοσης του συστήματος ως έχει και της σχετικά χαμηλής συχνότητας εμφάνισης των ενδεχομένως χρήσιμων σημείων στίξης, δε θεωρήθηκε σκόπιμο να προστεθούν περαιτέρω έλεγχοι και θέσεις χαρακτήρων για τη χρήση τους.

5. Δομή και λειτουργία του προγράμματος

Το πρόγραμμα ταξινόμησης κειμένων ανά γλώσσα λειτουργεί αφού πρώτα έχουν κατασκευαστεί γλωσσικά μοντέλα για όλες τις «γνωστές» στο σύστημα γλώσσες. Τα μοντέλα αυτά μπορούν να αποθηκευτούν σε αρχεία, τα οποία να διαβάζονται από το πρόγραμμα ταξινόμησης κάθε φορά που αυτό εκτελείται, ή μπορούν να ενσωματωθούν στο πρόγραμμα ταξινόμησης κατά τη μεταγλώττιση. Η πρώτη μέθοδος έχει το πλεονέκτημα της ευελιξίας και επεκτασιμότητας διότι πρόσθετα γλωσσικά μοντέλα μπορούν να παραχθούν μετά τη μεταγλώττιση και να καταστούν αμέσως ενεργά για χρήση με την επόμενη εκτέλεση του προγράμματος. Ένα σύντομο αρχείο, το οποίο μπορεί εύκολα να ενημερωθεί, περιέχει ένα κατάλογο των γνωστών γλωσσών και των αντίστοιχων αρχείων γλωσσικών μοντέλων. Η επιλογή αυτή έχει και το πλεονέκτημα του μικρού μεγέθους του εκτελέσιμου αρχείου και της συμβατότητας των γλωσσικών μοντέλων μεταξύ υπολογιστικών συστημάτων. Τέλος ακόμα και τα αρχεία γλωσσικών μοντέλων καταλαμβάνουν με την επιλογή αυτή λιγότερο χώρο διότι είναι απλό να αποθηκευτούν συμπιεσμένα μετά τον υπολογισμό τους.

Το κύριο μειονέκτημα της ευέλικτης αυτής μεθόδου είναι ότι τα γλωσσικά μοντέλα πρέπει να φορτώνονται από το δίσκο κάθε φορά που εκτελείται το πρόγραμμα, κάτι που μπορεί να επιμηκύνει σημαντικά το χρόνο εκτέλεσης, ειδικά σε περίπτωση πολλάκις επαναλαμβανόμενης εκτέλεσης. Αντιθέτως, η δεύτερη εναλλακτική λύση (αυτή της ενσωμάτωσης των γλωσσικών μοντέλων στο πρόγραμμα ταξινόμησης κατά τη μεταγλώττιση) προσφέρει υψηλότερη ταχύτητα εκτέλεσης, λόγω της μείωσης στις απαιτούμενες λειτουργίες I/O, κάτι που καθίσταται ιδιαίτερα εμφανές και σημαντικό σε επαναλαμβανόμενες αιτήσεις ταξινόμησης όπως στις δοκιμασίες του προγράμματος που περιγράφονται στη συνέχεια. Και οι δύο επιλογές έχουν υλοποιηθεί και λειτουργούν στο σύστημα του ΙΕΛ, έτσι οι διάφορες εφαρμογές μπορούν να χρησιμοποιήσουν όποια από τις δύο εξυπηρετεί καλύτερα τις ιδιαίτερες προτεραιότητες κάθε περίπτωσης χρήσης.

5.1. Προεπεξεργασία σωμάτων κειμένων

Η κατασκευή των γλωσσικών μοντέλων γίνεται με στατιστική ανάλυση σωμάτων κειμένων από κάθε γλώσσα. Στη βιβλιογραφία αναφέρονται ικανοποιητικά αποτελέσματα ακόμα με ιδιαιτέρως σύντομα κείμενα για την εξαγωγή των γλωσσικών μοντέλων, της τάξης λίγων χιλιάδων χαρακτήρων. Στην παρούσα υλοποίηση χρησιμοποιήθηκαν μεγαλύτερα σώματα, αρκετών εκατοντάδων χιλιάδων χαρακτήρων για κάθε γλώσσα, με σκοπό την πιο αξιόπιστη στατιστική ανάλυση και την εξαγωγή έγκυρων στοιχείων. Τα σώματα κειμένων που χρησιμοποιήθηκαν προήλθαν από τον παγκόσμιο ιστό και περιλαμβάνουν κείμενα ελληνικά (757 kb), αγγλικά (2299 kb), γαλλικά (1881 kb), γερμανικά (1215 kb) και ολλανδικά (594 kb).

Από τα κείμενα αυτά εξήχθησαν στατιστικά στοιχεία για τη συχνότητα εμφάνισης γραμμάτων και διγραμμάτων σε κάθε γλώσσα (υπολογίζοντας μόνο τις προαναφερθείσες θέσεις κωδικοποίησης). Οι συχνότητες εμφάνισης γραμμάτων και οι κανονικοποιημένες (δεσμευμένες) συχνότητες των

διγραμμάτων αποθηκεύτηκαν λογαριθμικά σε αρχεία μεγέθους 54 kb για κάθε γλώσσα (χωρίς συμπίεση). Δηλαδή ο απαιτούμενος χώρος για κάθε ένα γλωσσικό μοντέλο κρίνεται ιδιαίτερα μικρός.

Η προσθήκη άλλων γλωσσικών μοντέλων είναι ιδιαίτερα απλή, με την επεξεργασία του αντίστοιχου σώματος κειμένων και την αποθήκευση του αντίστοιχου αρχείου συχνοτήτων, καθώς και την ενημέρωση του αρχείου γλωσσικών μοντέλων.

Πρέπει να σημειωθεί εδώ η ειδική μέριμνα για τους διαφορετικούς τρόπους γραφής των ελληνικών. Το αρχικό σώμα κειμένου των ελληνικών, το οποίο ήταν εξ' ολοκλήρου σε κωδικοποίηση ISO-8859-7 με κανονική ορθογραφία, μετατράπηκε σε τρεις διαφορετικές μορφές γραφής ελληνικών με χαρακτήρες ASCII. Για τη μετατροπή ελήφθησαν υπόψη οι συνήθεις χρήσεις ελληνικών στο ηλεκτρονικό ταχυδρομείο για τη γραφή των ελληνικών χαρακτήρων. Αυθαίρετα ορίστηκαν, με βάση το αρχικό σύνολο χαρακτήρων

elot="άέήίόύώϊϋΆΈΉΊΌΎΏΪΫαβγδεζηθικλμνξοπρσςτυφχψωΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΓΧΨΩ;
"

οι εξής τρεις αντιστοιχίες (ενώ με τον ίδιο τρόπο πολύ εύκολα μπορούν να οριστούν και άλλες):

asc1="aeniouwiuAEHIOYWIYabgdezn9iklmv3oprsstufxywABGDEZH0IKLMN3OPRSTYFXUW;";

asc2="aehioywiyAEHIOYWIYabgdezh*iklmn*oprsstyfx*wABGDEZH*IKLMN*OPRSTYFX*W?";

asc3: "aehioyviyAEHIOYVIYabgdezhuiklmnjoprswtyfxcvABGDEZHUIKLMNJOPRSTYFXCV;";

Έτσι δημιουργήθηκαν τρία νέα σώματα κειμένων με το ίδιο περιεχόμενο αλλά διαφορετική γραφή («κωδικοποίηση»), από κάθε ένα από τα οποία προέκυψε ένα ξεχωριστό γλωσσικό μοντέλο, με στόχο την αναγνώριση της ελληνικής γλώσσας όταν αυτή παρουσιάζεται γραμμένη με μια από τις ιδιόμορφες συμβάσεις που συναντώνται συχνά σε επικοινωνίες ηλεκτρονικού ταχυδρομείου. Οι κωδικές ονομασίες elot, asc1, asc2 και asc3 χρησιμοποιούνται εφεξής για να υποδηλώσουν τα τεχνητά αυτά σώματα κειμένων και τα αντίστοιχα γλωσσικά μοντέλα για «ASCII ελληνικά». Για ένα πλήρες σύστημα ίσως θα ήταν σκόπιμο να συμπεριληφθεί και ένα γλωσσικό μοντέλο ελληνικών χαρακτήρων αλλά χωρίς τόνους, μια και δυστυχώς είναι αρκετά σύνηθες το φαινόμενο, ειδικά σε ηλεκτρονικά κείμενα, να παραλείπονται εντελώς τα τονικά σημάδια.

Έτσι συνολικά στο τρέχον σύστημα υπάρχει δυνατότητα ταξινόμησης σε πέντε γλώσσες μέ βάση οκτώ συνολικά γλωσσικά μοντέλα.

5.2. Προεπεξεργασία κειμένου

Το προς ταξινόμηση κείμενο (ακολουθία χαρακτήρων) υφίσταται την ίδια προεπεξεργασία όπως και τα σώματα κειμένων που χρησιμοποιήθηκαν για την κατασκευή των γλωσσικών μοντέλων. Δηλαδή αφαιρούνται όλα τα σημεία στίξης και μετατρέπονται σε ένα χαρακτήρα γενικευμένου διαστήματος, μαζί με όλους τους χαρακτήρες «white space» (όπως tab, carriage return κλπ.). Με τον τρόπο αυτό

μένουν μόνο λέξεις που χωρίζονται μεταξύ τους από γενικευμένο διάστημα, το οποίο έχει ουσιαστικά το ρόλο του συνόρου για το διαχωρισμό των λέξεων. Ο χαρακτήρας αυτός προστίθεται επίσης στο τέλος της ακολουθίας χαρακτήρων για να σημειωθεί και εκεί το προφανές τέλος λέξης. Στη συνέχεια η προεπεξεργασμένη ακολουθία χαρακτήρων παραδίδεται στο σύστημα ταξινόμησης ώστε να αναλυθεί στατιστικά και να εκτιμηθεί η συμβατότητα με κάθε υπάρχον γλωσσικό μοντέλο.

5.3. Κριτήριο ταξινόμησης

Όπως και κατά την επεξεργασία του σώματος κειμένων, η προς γλωσσική ταξινόμηση ακολουθία χαρακτήρων υφίσταται στατιστική ανάλυση ως προς το περιεχόμενό της σε χαρακτήρες και ζεύγη αυτών (διγράμματα). Επί του παρόντος η καταμέτρηση χαρακτήρων δε χρησιμοποιείται για την ταξινόμηση ούτε για κανονικοποίηση διότι σε δοκιμαστικές εκτελέσεις του προγράμματος δε φάνηκε να προσθέτει κάποιο πλεονέκτημα. Επίσης σε δοκιμές κατέστη σαφές ότι η συχνότητα απλών γραμμάτων μόνη δεν επαρκεί για ικανοποιητική απόδοση του συστήματος, όπως ήταν άλλωστε αναμενόμενο. Έτσι η ταξινόμηση γίνεται μόνο βάσει διγραμμάτων.

Συγκεκριμένα, για κάθε γλωσσικό μοντέλο, κάθε ζεύγος χαρακτήρων στην προς ταξινόμηση ακολουθία συνεισφέρει σε ένα τρέχον συγκεντρωτικό άθροισμα τη λογαριθμική δεσμευμένη πιθανότητα εμφάνισής του (του ζεύγους) στο αντίστοιχο σώμα κειμένου. Το άθροισμα των λογαρίθμων ισοδυναμεί με πολλαπλασιασμό των πιθανοτήτων, έτσι το τελικό αποτέλεσμα για κάθε γλωσσικό μοντέλο καλείται συγκεντρωτική πιθανότητα της ακολουθίας χαρακτήρων δεδομένου του γλωσσικού αυτού μοντέλου. Όπως προαναφέρθηκε δεν πρόκειται για πραγματική πιθανότητα λόγω της αλληλεξάρτησης μεταξύ διγραμμάτων και λόγω της σχετικής (δεσμευμένης) πιθανότητας κάθε διγράμματος που χρησιμοποιείται. Όμως η μαθηματική αυτή επιφύλαξη δεν έχει καμία αρνητική συνέπεια στην απόδοση του συστήματος, όπως έγινε φανερό μετά από πλήθος δοκιμών με διαφόρους δείκτες και παραμέτρους.

Αφού υπολογιστεί η συγκεντρωτική πιθανότητα της προς ταξινόμηση ακολουθίας χαρακτήρων για κάθε ένα γλωσσικό μοντέλο και κανονικοποιηθεί ως προς το μήκος της ακολουθίας, συγκρίνονται τα αποτελέσματα όλων των γλωσσικών μοντέλων και επιστρέφεται ο κωδικός του γλωσσικού μοντέλου βάσει του οποίου η υπολογίστηκε η μέγιστη συγκεντρωτική πιθανότητα. Το γλωσσικό μοντέλο αυτό θεωρείται και το πλέον πιθανό για τη συγκεκριμένη ακολουθία και η αντίστοιχη γλώσσα επιλέγεται ως η πιο πιθανή γλώσσα του κειμένου (της ακολουθίας χαρακτήρων). Ο κώδικας που επιτελεί τις ανωτέρω λειτουργίες είναι εξαιρετικά απλός και αποτελεσματικός, και δεν αναμένεται να καθυστερήσει σημαντικά οποιαδήποτε εφαρμογή τον χρησιμοποιήσει (με τις επιφυλάξεις για επαναλαμβανόμενες εκτελέσεις στην περίπτωση αποθήκευσης των γλωσσικών μοντέλων σε ξεχωριστά αρχεία).

5.4. Διαθεσιμότητα και συμβατότητα

Το σύστημα ταξινόμησης ακολουθιών χαρακτήρων βάσει γλωσσικών μοντέλων έχει υλοποιηθεί σε περιβάλλον Borland C++ Builder 3 και τρέχει σε λειτουργικό σύστημα Windows 98. Δεν αναμένεται καμία δυσκολία στην προσαρμογή του σε Windows 95, Windows Millennium Edition, Windows 2000, ή Windows NT, με ενδεχόμενη ανάγκη επαναμεταγλώττισης στο αντίστοιχο σύστημα Builder σε κάποιες περιπτώσεις. Δεδομένου ότι όλος ο κώδικας είναι απλή standard C και χρησιμοποιούνται απλά αρχεία κειμένου για τα σώματα κειμένων, η μετατροπή για χρήση σε οποιοδήποτε άλλο σύστημα, π.χ. Unix, δεν θα πρέπει να παρουσιάσει καμία δυσκολία πέραν της ανάγκης μεταγλώττισης. Ούτως ή άλλως δε γίνεται χρήση γραφικών λειτουργιών του Builder παρά μόνο για μία εφαρμογή επίδειξης που χρησιμοποιεί το σύστημα ταξινόμησης για ακολουθίες χαρακτήρων εισαγόμενες από το χρήστη μέσω πληκτρολογίου σε ένα μικρό παράθυρο. Η εφαρμογή αυτή φυσικά δεν αποτελεί μέρος του κυρίως συστήματος και δεν είναι απαραίτητη για τη χρήση του.

Συνοπτικά, η δημιουργία δυναμικά συνδεομένων βιβλιοθηκών για Windows είναι δεδομένη, ενώ πολύ απλά μπορεί να επιτευχθεί το αντίστοιχο αποτέλεσμα και σε άλλα συστήματα.

6. Απόδοση του συστήματος

Βασικό κριτήριο απόδοσης ενός συστήματος ταξινόμησης είναι φυσικά το ποσοστό επιτυχίας για ταξινόμηση αγνώστων κειμένων στο σύνολο των «γνωστών» γλωσσών (δηλαδή των υπαρχόντων γλωσσικών μοντέλων). Από το πλήθος των πειραμάτων και δοκιμαστικών εφαρμογών που πραγματοποιήθηκαν στο ΙΕΛ κατά τη διάρκεια της ανάπτυξης του συστήματος παρατίθενται στο παράρτημα για λόγους οικονομίας χώρου αποτελέσματα από δύο μόνο δοκιμασίες με το σύστημα που τελικά επελέγη. Συνοπτικά οι επιδόσεις του συστήματος είναι τουλάχιστον ικανοποιητικές, εφάμιλλες ή καλύτερες συγκριτικά με τα αναφερόμενα στη βιβλιογραφία, στο βαθμό που μια τέτοια σύγκριση είναι εφικτή. Οι συγκρίσεις είναι δύσκολες διότι διαφέρουν πολλές παράμετροι μεταξύ συστημάτων και μεταξύ δοκιμασιών, αλλά και γιατί τα εμπορικά συστήματα δεν αναφέρουν συστηματικές μετρήσεις των επιδόσεών τους ενώ εστιάζουν κυρίως στο πλήθος των γλωσσών που μπορούν να «αναγνωριστούν». Στην παρούσα εφαρμογή το πλήθος γλωσσών δεν περιορίζεται από τη σχεδίαση του συστήματος, ούτε και αναμένεται να έχει σημαντική επίδραση στις επιδόσεις διότι ήδη συμπεριλαμβάνονται γλώσσες αρκετά συγγενικές. Πάντως εκείνο που μας ενδιαφέρει κυρίως είναι το όσο δυνατόν χαμηλότερο ποσοστό σφαλμάτων για ένα μικρό πλήθος γλωσσών, κείμενα στις οποίες αναμένεται να συναντηθούν.

6.1. Μέθοδος δοκιμασιών

Για την πραγματοποίηση των δοκιμασιών κατατμήθηκαν τα σώματα κειμένων σε διάφορα μήκη αλληλουχιών ώστε να ελεγχθεί η λειτουργία του συστήματος με περιορισμένη πληροφορία. Ας σημειωθεί ότι τα κείμενα δεν υπέστησαν καμία επεξεργασία «καθαρισμού» και έτσι περιλαμβάνουν

αριθμούς, αρκτικόλεξα, ονόματα και πιθανώς λέξεις από άλλες γλώσσες, οι οποίες δυσχεραίνουν την ακριβή ταξινόμηση διότι δεν προβλέπονται από το γλωσσικό μοντέλο αλλά και γιατί αν περιλαμβάνουν ανενεργούς χαρακτήρες μειώνουν ουσιαστικά το χρήσιμο μήκος της ακολουθίας. Αυτό αποτελεί συνειδητή επιλογή για τον έλεγχο της λειτουργίας του συστήματος σε πραγματικές συνθήκες, διότι στην προτεινόμενη χρήση του δεν πρόκειται να αντιμετωπιστούν μόνο προεπεξεργασμένα ή άλλως ελεγμένα κείμενα. Ας σημειωθεί όμως ότι για το λόγο αυτό η κατωτέρω παρουσιαζόμενη επίδοση δεν είναι μια βέλτιστη «εργαστηριακή» επίδοση της οποίας θα ανεμένετο επιδείνωση κάτω από συνθήκες πραγματικής χρήσης.

Για να είναι πάντα δυνατή η χρήση της πληροφορίας τέλους λέξης, όπως αναμένεται και σε πραγματικές συνθήκες χρήσης, δεν ήταν δυνατό να κατατμηθούν τα κείμενα σε αλληλουχίες με επακριβώς προκαθορισμένο μήκος. Η διαδικασία που χρησιμοποιήθηκε ήταν ο περιορισμός του ελάχιστου μήκους, πέρα από το οποίο η αλληλουχία έληγε όποτε έληγε η λέξη κατά την οποία επετεύχθη το ελάχιστο μήκος. Έτσι το πραγματικό μήκος των αλληλουχιών αναμένεται να είναι κατά μέσο όρο μακρύτερο από το ελάχιστο μήκος κατά μισή λέξη, όπως και είναι φανερό από τη μελέτη των στοιχείων του παραρτήματος (ενώ φυσικά το μήκος λέξης εξαρτάται από τη γλώσσα).

6.2. Δυνατότητα απόρριψης άγνωστης γλώσσας

Κάθε τέτοια αλληλουχία ταξινομήθηκε από το σύστημα και κατεγράφησαν οι συγκεντρωτικές πιθανότητες όλων των γλωσσικών μοντέλων. Στους πίνακες του παραρτήματος φαίνεται το εύρος τιμών των συγκεντρωτικών πιθανοτήτων που ελήφθησαν ως βέλτιστες, δηλαδή οδήγησαν σε αντίστοιχη ταξινόμηση. Είναι ιδιαιτέρως σημαντικό να σημειωθεί ότι το εύρος αυτό είναι ιδιαίτερα μεγάλο και δε φαίνεται καμία κατάλληλη τιμή αποκοπής η οποία να μπορεί αξιόπιστα να χρησιμοποιηθεί για την απόρριψη αλληλουχιών χαρακτήρων ως «αγνώστου γλώσσας». Σχετικές δοκιμές κατέδειξαν αποφασιστικά ότι η προσπάθεια απόρριψης είναι μάταιη και οδηγεί σε σημαντική επιδείνωση της επίδοσης του συστήματος στις γνωστές γλώσσες με την ανεπιθύμητη απόρριψη σημαντικού ποσοστού κειμένων γνωστής γλώσσας.

Στα πλαίσια των δοκιμών αυτών ο ενδιαφερόμενος μπορεί να μελετήσει τα αποτελέσματα της δοκιμής που παρατίθεται στο παράρτημα, κατά την οποία δεν συμπεριλαμβάνεται το γλωσσικό μοντέλο για τα ολλανδικά και συνεπώς τα ολλανδικά κείμενα κατατάσσονται υποχρεωτικά σε άλλη γλωσσική κατηγορία, με συχνότερη τη γερμανική η οποία είναι και η πιο παρόμοια. Όπως φαίνεται από τους πίνακες, το εύρος τιμών επιλογής των (λανθασμένων) γλωσσικών μοντέλων δε διαφέρει σημαντικά από αυτά των δοκιμών με κείμενα των άλλων (δηλ. γνωστών) γλωσσών.

6.3. Επιδόσεις και εξάρτηση από μήκος αλληλουχίας

Από τη μελέτη των ενδεικτικών πινάκων του παραρτήματος είναι φανερό ότι για τέσσερις γλώσσες (επτά γλωσσικά μοντέλα λόγω των διαφόρων τρόπων γραφής των ελληνικών) το ποσοστό σφάλματος

πέφτει κάτω από 1% για κείμενο οποιασδήποτε γλώσσας που ξεπερνά σε μήκος τους 35 χαρακτήρες περίπου, ενώ για κείμενο 200 χαρακτήρων το ποσοστό σφάλματος γίνεται αμελητέο (της τάξης του 1‰ ή χαμηλότερο). Η επίδοση αυτή είναι ιδιαίτερα καλή αν αναλογιστεί κανείς ότι για ένα σύνηθες μήκος μιας σειράς κειμένου των 80 χαρακτήρων χρειάζεται μισή μόνο σειρά (4–6 λέξεις, ανάλογα και με τη γλώσσα) για επίτευξη 99% επιτυχίας, ενώ με τρεις σειρές η επιτυχία είναι ουσιαστικά εξασφαλισμένη (ένα λάθος σε 50 σελίδες κειμένου).

Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη της επίδοσης του συστήματος σε σύντομες αλληλουχίες χαρακτήρων, οι οποίες μπορεί να περιέχουν μία ή δύο λέξεις. Τέτοιες περιπτώσεις μπορούν να προκύψουν σε αναγνώριση σύντομων μηνυμάτων, καταλόγων επιλογών προγραμμάτων, ή επιλογών-δεικτών. Στην περίπτωση αυτή η επίδοση του συστήματος είναι σημαντικά χαμηλότερη, όπως άλλωστε και όλων των άλλων υπαρχόντων συστημάτων. Για μέσο πραγματικό μήκος αλληλουχίας γύρω στους 10 χαρακτήρες το ποσοστό σφάλματος κυμαίνεται γύρω στο 10% για τις συγγενικές γλώσσες των αγγλικών, γερμανικών και γαλλικών, ενώ παραμένει ιδιαίτερα υψηλό στην περίπτωση των ελληνικών είτε αυτά είναι γραμμένα με πραγματικούς ελληνικούς χαρακτήρες είτε όχι. Ας σημειωθεί εδώ ότι η αναγνώριση ενός κειμένου ως ASCII-ελληνικού αλλά με λίγο διαφορετική αντιστοιχία ελληνικών-λατινικών δεν αποτελεί λειτουργικό σφάλμα του συστήματος (παρ΄ ό,τι καταγράφεται ξεχωριστά στους πίνακες σύγχυσης) διότι εκείνο που μας ενδιαφέρει είναι η αναγνώριση της γλώσσας, η οποία είναι επιτυχής.

Δεδομένης της συγγένειας μεταξύ αγγλικών, γαλλικών και γερμανικών η σχετικά υψηλή σύγχυση μεταξύ τους για πολύ σύντομες αλληλουχίες χαρακτήρων δεν αποτελεί έκπληξη. Εδώ θα μπορούσε να γίνει ίσως κάποια βελτίωση με συνολική βελτιστοποίηση με εκ των προτέρων πιθανότητες αλλά αυτό θα εξαρτηθεί από την τελική εφαρμογή στην οποία θα χρησιμοποιηθεί το σύστημα και το σύνολο κειμένων το οποίο θα κληθεί να επεξεργαστεί και ταξινομήσει.

7. Συμπέρασμα

Συμπερασματικά μπορεί να υποστηριχθεί ότι η επίδοση του συστήματος είναι ιδιαίτερα ικανοποιητική, τόσο για σύντομες αλληλουχίες χαρακτήρων οι οποίες αποτελούν δοκιμασία υψηλής δυσκολίας, όσο και για αλληλουχίες μεγαλύτερου μήκους, στις οποίες το σφάλμα γίνεται αμελητέο. Το σύστημα κρίνεται όχι απλώς χρησιμοποιήσιμο αλλά συγκρίσιμο με άλλα που αναφέρονται στη βιβλιογραφία ή διατίθενται ως εμπορικά προϊόντα, ενώ μικροβελτιώσεις ίσως είναι δυνατές για πολύ συγκεκριμένες εφαρμογές.

Βιβλιογραφία και διαδικτυακές παραπομπές

[1] Yeshwant K. Muthusamy & A. Lawrence Spitz (1996). Automatic language identification (Section 8.7 of "Survey of the State of the Art in Human Language Technology" edited by R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue) at http://cslu.cse.ogi.edu/HLTsurvey/ch8node9.html.

[2] William B. Cavnar & John M. Trenkle (1994). N-gram-based text categorization. In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV: UNLV Publications/Reprographics, pp. 161-175.

[3] Gregory Grefenstette (1995). Comparing Two Language Identification Schemes. In Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT'95), Rome, Italy.

[4] Kenneth R. Beesley (1988). Language Identifier: A computer program for automatic natural-language identification of on-line text. In Languages at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association, 12-16 October 1988, pp. 47-54.

Available online at http://www.xrce.xerox.com/publis/mltt/mltt-99-01.ps.

[5] Yoshio Ueda and Seiichi Nakagawa (1990). Prediction for phoneme/syllable/word-category and identification of language using HMM. In Proceedings of the 1990 International Conference on Spoken Language Processing, Vol. 2 (pp. 1209-1212). Kobe, Japan.

[6] H. K. Kwan and K. Hirose (1997). Use of Recurrent Network for Unknown Language Rejection in Language Identification System. In Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 97). Rhodes, Greece.

Εταιρείες και συστήματα:

[a] Alis Technologies Inc., language technology solutions at http://www.alis.com/castil/index.html; system for identification of language and character encoding at http://www.alis.com/castil/silc/.

[b] Stochastic Language Identifier by Doug Beeferman at http://www.link.cs.cmu.edu/dougb/ident-doc.html.

[c] TextCat Language Guesser by Gertjan van Noord, implementing the Cavnar & Trenkle algorithm, at http://odur.let.rug.nl/~vannoord/TextCat/.

[d] Automatic Language Identification Bibliography, maintained by Diamantino Caseiro, at http://speech.inesc.pt/~dcaseiro/html/bibliografia.html.

[e] Xerox Research Centre Europe online Language Identifier demo, at http://www.rxrc.xerox.com/research/mltt/tools/guesser.html.