ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Επανυλοποίηση, βελτίωση, αξιολόγηση και τεκμηρίωση ελληνικού επισημειωτή μερών του λόγου που χρησιμοποιεί μηχανική μάθηση

Κωνσταντίνος Παππάς

Επιβλέπων: Ίων Ανδρουτσόπουλος

Ιούλιος 2008

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1:

ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Στη διάρκεια δύο προηγούμενων εργασιών [Μα05, Χρο06] αναπτύχθηκε ένας επισημειωτής μερών του λόγου (part-of-speech tagger) για ελληνικά κείμενα, ο οποίος χρησιμοποιούσε τον αλγόριθμο μάθησης των *k* κοντινότερων γειτόνων (*k*-nearest neighbors, *k*-NN). Ο επισημειωτής κατέτασσε κάθε λέξη ενός δοθέντος κειμένου στο αντίστοιχο μέρος του λόγου (ρήμα, ουσιαστικό, άρθρο κλπ.). Εναλλακτικά, είχε τη δυνατότητα να κατατάσσει τις λέξεις σε λεπτομερέστερες κατηγορίες, που αντανακλούσαν τον αριθμό, την πτώση και το γένος των ουσιαστικών, τη φωνή, τον αριθμό και το χρόνο των ρημάτων κλπ. (π.χ. ξεχωριστή κατηγορία για τα αρσενικά ουσιαστικά ονομαστικής ενικού, ξεχωριστή για τα αρσενικά ουσιαστικά γενικής ενικού κ.ο.κ.). Ακόμη, ο επισημειωτής ήταν δυνατόν να εκπαιδευθεί με ενεργητική μάθηση, επιλέγοντας δηλαδή ο ίδιος χρήσιμα παραδείγματα εκπαίδευσης (λέξεις) από ένα σώμα κειμένων, τα οποία επισημείωνε κατόπιν ο άνθρωπος-εκπαιδευτής του, αντί ο εκπαιδευτής να σημειώνει με τη σειρά τις λέξεις ενός σώματος κειμένων εκπαίδευσης (παθητική μάθηση). Παρ' όλο που κάποια πειράματα είχαν δείξει ότι η ενεργητική μάθηση υπερτερούσε της παθητικής όταν χρησιμοποιούσαν και οι δύο το ίδιο μικρό σώμα κειμένων εκπαίδευσης, άλλα πειράματα είχαν δείξει ότι η ενεργητική μάθηση αδυνατούσε να εκμεταλλευτεί επαρκώς μεγαλύτερα σώματα εκπαίδευσης, επειδή τα επιλεγόμενα παραδείγματα ήταν πολύ σπάνιες περιπτώσεις. Ακόμη, ο επισημειωτής δεν διέθετε επαρκή τεκμηρίωση και χρησιμοποιούσε μια υλοποίηση του αλγορίθμου *k*-NN άλλης ερευνητικής ομάδας, κάτι που δημιουργούσε προβλήματα διάθεσης του λογισμικού του επισημειωτή.[1]

Στη διάρκεια της παρούσας εργασίας, δημιουργήθηκε μια νέα υλοποίηση του αλγορίθμου *k*-NN, η οποία χρησιμοποιήθηκε κατόπιν ως βάση προκειμένου να επανυλοποιηθεί εξαρχής ο επισημειωτής μερών του λόγου. Η νέα υλοποίηση του επισημειωτή είναι πλήρως τεκμηριωμένη, διαθέτει εύχρηστη διεπαφή χρήστη και παρέχεται ελεύθερα.[2] Δημιουργήθηκαν, επίσης, βελτιωμένα σώματα κειμένων εκπαίδευσης και αξιολόγησης, τα οποία χρησιμοποιήθηκαν σε νέα πειράματα. Μελετήθηκαν, τέλος, πιθανοί τρόποι αντιμετώπισης των προβλημάτων που είχαν παρατηρηθεί στην περίπτωση της ενεργητικής μάθησης.

1.2 Διάρθρωση εργασίας

Συνοπτικά, τα επόμενα κεφάλαια της εργασίας καλύπτουν τα εξής θέματα:

- Το κεφάλαιο 2 αναφέρεται στις μεθόδους μηχανικής μάθησης που χρησιμοποιούνται στην εργασία. Συγκεκριμένα, περιλαμβάνει μια αναλυτική

[1] Ο επισημειωτής χρησιμοποιούσε την πλατφόρμα TiMBL [DaZa03]. Βλ. http://ilk.uvt.nl/timbl/.
[2] Βλ. http://pages.cs.aueb.gr/nlp/software.html.

περιγραφή του αλγορίθμου *k*-NN, καθώς και μια περιγραφή των μέτρων επιλογής παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται στην ενεργητική μάθηση.

- Στο κεφάλαιο 3 περιγράφεται το πρόβλημα της αναγνώρισης μερών του λόγου και εξηγούνται τα διάφορα στάδια επεξεργασίας που χρησιμοποιούνται για τη λύση του με μηχανική μάθηση.
- Στο κεφάλαιο 4 παρουσιάζονται τα δεδομένα εκπαίδευσης και αξιολόγησης που χρησιμοποιήθηκαν στα πειράματα της εργασίας. Επίσης, παρατίθενται τα αποτελέσματα των πειραμάτων και τα συμπεράσματα που προκύπτουν από αυτά.
- Τέλος, στο κεφάλαιο 5 γίνεται μια ανασκόπηση της εργασίας και προτείνονται πιθανές βελτιώσεις και κατευθύνσεις μελλοντικής έρευνας.

1.3 Ευχαριστίες

Αρχικά, ευχαριστώ τον κ. Ίωνα Ανδρουτσόπουλο, που επέβλεψε την πτυχιακή μου εργασία, συμβουλεύοντας και ενθαρρύνοντάς με σε κάθε στάδιό της.

Αναγκαία για την ολοκλήρωση της εργασίας αποδείχθηκε η συμβολή και συνεργασία του Πρόδρομου Μαλακασιώτη, που ήταν παρών κάθε φορά που χρειάστηκε και με βοήθησε τόσο στα σχετικά θεωρητικά ζητήματα όσο και σε θέματα υλοποίησης και διεξαγωγής πειραμάτων. Τον ευχαριστώ πολύ θερμά.

Επίσης, ευχαριστώ το Γιώργο Λουκαρέλλι για το σύστημα διαχωρισμού περιόδων (sentence splitter) που ανέπτυξε και διέθεσε ελεύθερα, καθώς αυτό αποτελεί μια ουσιαστική υπομονάδα του συστήματος.

Τέλος, ευχαριστώ τα μέλη του Εργαστηρίου Επεξεργασίας Πληροφοριών και ιδιαίτερα την Ομάδα Επεξεργασίας Φυσικής Γλώσσας του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών, για τη βοήθεια που μου παρείχαν, εμψυχώνοντάς με και διαθέτοντάς μου τον απαραίτητο εξοπλισμό.

ΚΕΦΑΛΑΙΟ 2:

Ο ΑΛΓΟΡΙΘΜΟΣ ΤΩΝ Κ-ΚΟΝΤΙΝΟΤΕΡΩΝ ΓΕΙΤΟΝΩΝ

2.1 Εισαγωγή

Η Μηχανική Μάθηση [Mit97] είναι ο τομέας της Τεχνητής Νοημοσύνης ο οποίος ασχολείται με την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν σε υπολογιστικά συστήματα να βελτιώνουν τις επιδόσεις τους αξιοποιώντας εμπειρικά δεδομένα του παρελθόντος. Στην παρούσα εργασία χρησιμοποιούνται μέθοδοι Επιβλεπόμενης Μηχανικής Μάθησης [Ko07]. Ο όρος αυτός χρησιμοποιείται σε περιπτώσεις όπου απαιτείται να κατασκευαστεί μια συνάρτηση, δεδομένων παραδειγμάτων ορισμάτων της (εισόδων) και των αντίστοιχων επιθυμητών τιμών της (εξόδων).

Στο πρόβλημα της αναγνώρισης μερών του λόγου, τα ορίσματα της αναζητούμενης συνάρτησης είναι ιδιότητες (attributes) που παρέχουν πληροφορίες για την υπό κατάταξη λέξη και τα συμφραζόμενά της. Οι τιμές που λαμβάνουν οι ιδιότητες στην περίπτωση κάθε λέξης λέγονται χαρακτηριστικά (features) της λέξης και συνήθως παριστάνονται ως διανύσματα. Οι επιθυμητές τιμές (έξοδοι) της συνάρτησης είναι οι κατηγορίες των λέξεων, δηλαδή τα αντίστοιχα μέρη του λόγου ή τα μέρη του λόγου με πρόσθετες πληροφορίες για την πτώση, τον αριθμό κλπ. Ο στόχος μας, δηλαδή, είναι να μάθουμε μια συνάρτηση που να απεικονίζει κάθε λέξη (ακριβέστερα, το διάνυσμα των χαρακτηριστικών της) στη σωστή κατηγορία.

2.2 Αλγόριθμος των *k* κοντινότερων γειτόνων

Ο αλγόριθμος των *k* κοντινότερων γειτόνων (*k*-NN) [Be91] αποτελεί μια από τις απλούστερες μορφές επιβλεπόμενης μάθησης. Κατά το στάδιο της εκπαίδευσης, απλά αποθηκεύει τα διανύσματα χαρακτηριστικών των παραδειγμάτων εκπαίδευσης, μαζί με τις ορθές τους κατηγορίες, οι οποίες παρέχονται από τον άνθρωπο-εκπαιδευτή. Κατόπιν, κατά την κατάταξη νέων περιπτώσεων (λέξεων νέων κειμένων, στην περίπτωση του επισημειωτή της εργασίας), κατασκευάζονται τα διανύσματα χαρακτηριστικών τους και επιλέγονται για κάθε νέα περίπτωση (λέξη) τα *k* κοντινότερα (πιο παρόμοια) διανύσματα εκπαίδευσης (γείτονες), βάσει κάποιου μέτρου απόστασης. Η κάθε νέα περίπτωση κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των ορθών κατηγοριών των *k* επιλεγέντων παραδειγμάτων εκπαίδευσης (κοντινότερων γειτόνων).

Το *k* είναι ένας φυσικός αριθμός και αποτελεί παράμετρο του αλγορίθμου. Στην υλοποίηση του συστήματος, αυτός ο αριθμός προσδιορίζει το πλήθος των κοντινότερων γειτονιών και όχι γειτόνων. Μια γειτονιά ορίζεται ως το σύνολο όλων των παραδειγμάτων εκπαίδευσης που έχουν την ίδια απόσταση από την υπό κατάταξη περίπτωση. Για παράδειγμα, αν $k = 3$ και οι κοντινότερες γειτονιές είναι δύο παραδείγματα εκπαίδευσης σε απόσταση 5 (κοντινότερη γειτονιά), τρία

παραδείγματα εκπαίδευσης σε απόσταση 7 (δεύτερη κοντινότερη γειτονιά) και δύο παραδείγματα εκπαίδευσης σε απόσταση 8 (τρίτη κοντινότερη γειτονιά), θα χρησιμοποιηθούν επτά παραδείγματα εκπαίδευσης ως γείτονες.

2.3 Μέτρα απόστασης

Για τους σκοπούς της εργασίας, θα παρουσιαστούν δύο μόνο μέτρα απόστασης μεταξύ διανυσμάτων χαρακτηριστικών, τα οποία είχαν χρησιμοποιηθεί και στις προηγούμενες εργασίες [Μα05, Χρο06]: το μέτρο επικάλυψης (overlap metric) και το τροποποιημένο μέτρο διαφοράς τιμών (modified value difference metric). Και τα δύο θεωρούν ως απόσταση δύο διανυσμάτων το συνολικό άθροισμα των διαφορών των χαρακτηριστικών τους.

Το μέτρο επικάλυψης ορίζεται από τις δύο παρακάτω σχέσεις. Στις σχέσεις αυτές $\Delta(\vec{X},\vec{Y})$ είναι η απόσταση μεταξύ των διανυσμάτων $\vec{X}$ και $\vec{Y}$, που το καθένα έχει n χαρακτηριστικά (τιμές ιδιοτήτων), $\delta(x_i, y_i)$ είναι η διαφορά των αντιστοίχων χαρακτηριστικών (τιμών της ίδιας ιδιότητας) x_i και y_i, ενώ max_i και min_i είναι η μέγιστη και ελάχιστη τιμή, αντίστοιχα, της i-στής ιδιότητας.

$$\Delta(\vec{X},\vec{Y}) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{2.1}$$

$$\delta(x_i, y_i) = \begin{cases} \left|\dfrac{x_i - y_i}{\max_i - \min_i}\right| & \text{, αν αριθμητικές τιμές, διαφορετικά :} \\ 0 & \text{αν } x_i = y_i \\ 1 & \text{αν } x_i \neq y_i \end{cases} \tag{2.2}$$

Το διαφοροποιημένο μέτρο διαφοράς τιμών προκύπτει εάν αντικαταστήσουμε τη σχέση (2.2) με την παρακάτω, όπου m το πλήθος των κατηγοριών.

$$\delta(x_i, y_i) = \sum_{j=1}^{m} | P(C_j \mid x_i) - P(C_j \mid y_i) | \tag{2.3}$$

όπου οι πιθανότητες $P(C_j \mid x_i)$ και $P(C_j \mid y_i)$ εκτιμώνται από τα δεδομένα εκπαίδευσης.

2.4 Αναλογία κέρδους ιδιοτήτων

Το πληροφοριακό κέρδος (information gain) μετράει πόσο χρήσιμη είναι μια ιδιότητα κατά την πρόβλεψη της κατηγορίας μιας υπό κατάταξη περίπτωσης (λέξης). Μετράει, δηλαδή, τη μείωση της αβεβαιότητας της κατηγορίας που προκαλεί η γνώση της τιμής της ιδιότητας. Η αβεβαιότητα της κατηγορίας C μετριέται ως η εντροπία $H(C)$, που ορίζεται από τον τύπο (2.4). Αν μάθουμε ότι η τιμή της ιδιότητας X είναι x, η εντροπία γίνεται $H(C|X=x)$, όπως στον τύπο (2.5).

$$H\left(C\right) = -\sum_{c} P\left(C = c\right) \log_2 P\left(C = c\right) \tag{2.4}$$

$$H\left(C \mid X = x\right) = -\sum_{c} P\left(C = c \mid X = x\right) \log_2 P\left(C = c \mid X = x\right) \tag{2.5}$$

Γενικότερα, η γνώση της τιμής της ιδιότητας *X* οδηγεί σε αναμενόμενη εντροπία *H(C|X)*, που για ιδιότητες με πεπερασμένο σύνολο δυνατών τιμών υπολογίζεται από τον τύπο (2.6). Το πληροφοριακό κέρδος που προσφέρει η ιδιότητα *X* είναι η αναμενόμενη μείωση της εντροπίας που επιφέρει η γνώση της τιμής της *X*. Υπολογίζεται από τον τύπο (2.7).

$$H(C \mid X) = \sum_{x} P(X = x) \cdot H(C \mid X = x) \tag{2.6}$$

$$IG(C, X) = H(C) - \sum_{x} P(X = x) \cdot H(C \mid X = x) \tag{2.7}$$

Μια κανονικοποιημένη μορφή του πληροφοριακού κέρδους που χρησιμοποιείται συχνά στον *k*-NN είναι η αναλογία κέρδους (gain ratio), η οποία υπολογίζεται ως ο λόγος του πληροφοριακού κέρδους της ιδιότητας διά την εντροπία της, όπως στον τύπο (2.8). Η εντροπία *H(X)* μια ιδιότητας υπολογίζεται από τον τύπο (2.9). Η αναλογία κέρδους «τιμωρεί» ιδιότητες με πολλές ισοπίθανες δυνατές τιμές (άρα και μεγάλη εντροπία), προκειμένου να «τιμωρήσει» ιδιότητες όπως τα κλειδιά, που μπορεί να «προβλέπουν» μεν τις κατηγορίες των παραδειγμάτων εκπαίδευσης, αλλά δεν προσφέρουν χρήσιμες πληροφορίες σε νέες περιπτώσεις.

$$W_i = \frac{IG(C, X)}{H(X)} \tag{2.8}$$

$$H(X) = -\sum_{x} P\left(X = x\right) \log_2 P\left(X = x\right) \tag{2.9}$$

Κατά τον υπολογισμό της απόστασης δύο διανυσμάτων, οι διαφορές τιμών κάθε ιδιότητας μπορούν να πολλαπλασιαστούν με την αναλογία κέρδους της ιδιότητας, ώστε να δίνεται μεγαλύτερο βάρος στις διαφορές τιμών σημαντικών ιδιοτήτων. Στην περίπτωση αυτή η σχέση (2.1) της προηγούμενης ενότητας τροποποιείται ως εξής, όπου x_i και y_i οι τιμές της ιδιότητας X_i στα δύο διανύσματα:

$$\Delta(\vec{\mathrm{X}}, \vec{\mathrm{Y}}) = \sum_{i=1}^{n} IG(C, X_i) \cdot \delta(x_i, y_i) \tag{2.10}$$

2.5 Ζύγισμα γειτόνων βάσει αποστάσεως

Μία συχνά χρησιμοποιούμενη βελτίωση του *k*-NN είναι κάθε ένας από τους *k* γείτονες να έχει βάρος ψήφου w_i αντιστρόφως ανάλογο της απόστασης d_i του γείτονα από την υπό κατάταξη περίπτωση, ώστε η κατάταξη να γίνεται λαμβάνοντας περισσότερο υπόψη τους πιο παρόμοιους γείτονες [Du76].

$$w_i = \frac{1}{\varepsilon + d_j} \tag{2.11}$$

Στον παρονομαστή της (2.8) προστίθεται μια μικρή θετική σταθερά ε, για να αποφευχθεί η διαίρεση με το μηδέν όταν το d_j είναι μηδέν.

Κάθε νέα περίπτωση (λέξη) κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των *k* κοντινότερων γειτόνων της (πιο παρόμοια παραδείγματα εκπαίδευσης), ζυγίζοντας τις ψήφους των γειτόνων με τα βάρη της (2.8). Εάν ισοψηφίσουν δύο κατηγορίες, επιλέγεται μία από αυτές στην τύχη, με πιθανότητα ανάλογη της συχνότητας εμφάνισης της κατηγορίας σε όλα τα παραδείγματα εκπαίδευσης.

2.6 Ενεργητική μάθηση

Κατά την ενεργητική μάθηση, ο ίδιος ο ταξινομητής επιλέγει παραδείγματα εκπαίδευσης από μια δεξαμενή παραδειγμάτων των οποίων δεν είναι γνωστές οι ορθές κατηγορίες και ζητά από τον άνθρωπο-εκπαιδευτή του να τα κατατάξει. Στην περίπτωσή μας, επιλέγονται λέξεις από ένα μη επισημειωμένο σώμα κειμένων. Αν το μέτρο επιλογής παραδειγμάτων είναι κατάλληλο, ενδέχεται ο ταξινομητής να επιτυγχάνει υψηλότερο ποσοστό ορθότητας (accuracy) σε σχέση με το ποσοστό που επιτυγχάνει (για ίσο αριθμό παραδειγμάτων) όταν τα παραδείγματα επιλέγονται τυχαία ή με τη σειρά που εμφανίζονται στη δεξαμενή (π.χ. αν ο άνθρωπος-επισημειωτής κατατάσσει με τη σειρά τις λέξεις του σώματος εκπαίδευσης). Η ενεργητική μάθηση μπορεί, επομένως να οδηγήσει στα ίδια ή καλύτερα ποσοστά ορθότητας με λιγότερα παραδείγματα εκπαίδευσης, άρα με λιγότερη χειρωνακτική κατάταξη παραδειγμάτων. Στην περίπτωση του *k*-NN, η μείωση των παραδειγμάτων εκπαίδευσης μπορεί να οδηγήσει επίσης σε σημαντική μείωση της απαιτούμενης μνήμης και αύξηση της ταχύτητας κατάταξης.

Στο σύστημα της παρούσας εργασίας χρησιμοποιούνται δύο μέτρα επιλογής παραδειγμάτων κατά την ενεργητική μάθηση, τα οποία προέρχονται από τις προηγούμενες εργασίες [Μα05, Χρο06]. Το πρώτο μέτρο ορίζεται από τη σχέση (2.14), όπου $\vec{x}$ μία λέξη προς κατάταξη (για την ακρίβεια, το διάνυσμα των χαρακτηριστικών της). Το μέτρο υπολογίζει ουσιαστικά την αβεβαιότητα (εντροπία) που έχει ο ταξινομητής για την κατηγορία *C* της λέξης, δοθέντος του διανύσματός της $\vec{x}$ και των ήδη επιλεγμένων (και επισημειωμένων) παραδειγμάτων εκπαίδευσης *T*. Για την ακρίβεια, χρησιμοποιείται η κανονικοποιημένη μορφή εντροπίας του τύπου (2.12), όπου |*C*| ο συνολικός αριθμός των κατηγοριών που εμφανίζονται στους *k* κοντινότερους γείτονες. Οι πιθανότητες $P(C = c \mid \vec{x}, T)$ εκτιμώνται με τον τύπο

(2.13), όπου $V(c \mid \vec{x}, T)$ είναι οι (ζυγισμένοι κατά απόσταση) ψήφοι των k κοντινότερων γειτόνων που ανήκουν στην κατηγορία c. Ο τύπος (2.14) δίνει αρνητικές τιμές στα υποψήφια παραδείγματα εκπαίδευσης με μηδενική εντροπία, ώστε να προτιμώνται παραδείγματα με μη μηδενική εντροπία, και μάλιστα δίνει πιο αρνητικές τιμές (τα αξιολογεί ως χειρότερα) στα υποψήφια παραδείγματα που βρίσκονται κοντύτερα σε ήδη επιλεγμένα παραδείγματα ανεξαρτήτως κατηγορίας (μεγαλύτερο άθροισμα ζυγισμένων κατά απόσταση ψήφων όλων των κατηγοριών).

$$H_n\left(C \mid \vec{x}, T\right) = -\frac{\sum_c P\left(C = c \mid \vec{x}, T\right) \log_2 P\left(C = c \mid \vec{x}, T\right)}{\log\left(\mid C \mid\right)} \tag{2.12}$$

$$P\left(C = c \mid \vec{x}, T\right) = \frac{V(c \mid \vec{x}, T)}{\sum_{c'} V(c' \mid \vec{x}, T)} \tag{2.13}$$

$$W(\vec{x}, T) = \begin{cases} H_n\left(C \mid \vec{x}, T\right), & \text{αν } H_n\left(C \mid \vec{x}, T\right) \neq 0 \\ -\sum_c V(c \mid \vec{x}, T), & \text{αν } H_n\left(C \mid \vec{x}, T\right) = 0 \end{cases} \tag{2.14}$$

Το δεύτερο μέτρο είναι παρόμοιο, αλλά διαιρεί την κανονικοποιημένη εντροπία με το άθροισμα των (ζυγισμένων κατά απόσταση) ψήφων των κατηγοριών των k κοντινότερων γειτόνων, ώστε να προτιμώνται παραδείγματα εκπαίδευσης για τα οποία είναι αβέβαιος ο ταξινομητής αλλά και δεν βρίσκονται κοντά σε πολλά ήδη επιλεγμένα παραδείγματα.

$$W(\vec{x}, T) = \begin{cases} \dfrac{H_n\left(C \mid \vec{x}, T\right)}{\log\left(\sum_{c'} V(c \mid \vec{x}, T)\right)}, & \text{αν } H_n\left(C \mid \vec{x}, T\right) \neq 0 \\ -\sum_{c'} V(c \mid \vec{x}, T), & \text{αν } H_n\left(C \mid \vec{x}, T\right) = 0 \end{cases} \tag{2.15}$$

ΚΕΦΑΛΑΙΟ 3:

ΑΝΑΓΝΩΡΙΣΗ ΜΕΡΩΝ ΤΟΥ ΛΟΓΟΥ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΤΩΝ Κ-ΚΟΝΤΙΝΟΤΕΡΩΝ ΓΕΙΤΟΝΩΝ

3.1 Διαχωρισμός λεκτικών μονάδων

Δεδομένου ενός συνόλου κατηγοριών (ετικετών, tags) που παριστάνουν τα μέρη του λόγου (ή και λεπτομερέστερες πληροφορίες, όπως ο αριθμός, η πτώση και το γένος των ουσιαστικών), το πρόβλημα της αναγνώρισης των μερών του λόγου έγκειται στην κατάταξη κάθε λεκτικής μονάδας (token) ενός δοσμένου κειμένου στη σωστή κατηγορία. Αυτό προϋποθέτει το διαχωρισμό του κειμένου σε λεκτικές μονάδες.

Στο σύστημα της εργασίας, ο διαχωρισμός των λεκτικών μονάδων ξεκινά με τον εντοπισμό και επισημείωση των τελειών που σηματοδοτούν τέλη περιόδων. Χρησιμοποιείται ένας προϋπάρχων διαχωριστής περιόδων [Λου05], ο οποίος διαχωρίζει τις τελείες που σηματοδοτούν τέλη περιόδων από τις υπόλοιπες, χρησιμοποιώντας και αυτός μηχανική μάθηση. Ο διαχωριστής περιόδων πετυχαίνει ποσοστό επιτυχίας κοντά στο 98%.

Κατόπιν, εξετάζονται ένας προς έναν όλοι οι χαρακτήρες του κειμένου από την αρχή ως το τέλος. Κατά τη διάρκεια αυτού του περάσματος, αναγνωρίζονται οι λεκτικές μονάδες, ως εξής:

- Ο χαρακτήρας ‘,’ αποτελεί ξεχωριστή λεκτική μονάδα, εάν ακολουθείται από κενό, χαρακτήρα tab ή χαρακτήρα αλλαγής γραμμής.
- Ο χαρακτήρας ‘.’ αποτελεί ξεχωριστή λεκτική μονάδα, εάν έχει σημειωθεί ως τέλος περιόδου από το διαχωριστή περιόδων.
- Οι χαρακτήρες ‘!’ και ‘;’ αποτελούν ξεχωριστές λεκτικές μονάδες. Η εύρεση τέτοιων χαρακτήρων σηματοδοτεί και τέλος περιόδου.
- Οι χαρακτήρες ‘(’, ‘)’, ‘”’, ‘%’, ‘<’, ‘>’, ‘«’,’»’, ‘{’, ‘}’, ‘[’, ‘]’, ‘/’, ‘\’, ‘*’, ‘$’, ‘:’, ‘€’, ‘·’, ‘-’ και ‘+’ αποτελούν ξεχωριστές λεκτικές μονάδες.
- Κατά τα άλλα, τα κενά, οι χαρακτήρες tab και οι χαρακτήρες αλλαγής γραμμής θεωρούνται διαχωριστές λεκτικών μονάδων.

3.2 Ιδιότητες διανυσμάτων

Μετά το διαχωρισμό του κειμένου σε λεκτικές μονάδες, κατασκευάζεται για κάθε μία ένα διάνυσμα χαρακτηριστικών (features, τιμές ιδιοτήτων). Οι ιδιότητες είναι οι ίδιες με αυτές των προηγούμενων εργασιών ([Μα05], [Χρο06]), δηλαδή οι ακόλουθες.

1. Κατάληξη (οι τρεις τελευταίοι χαρακτήρες ή ολόκληρη η λεκτική μονάδα, αν αυτή έχει μήκος μικρότερο ή ίσο του τρία).

2. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων).

3. Ύπαρξη αποστρόφου στη λεκτική μονάδα (1 εάν υπάρχει, 0 διαφορετικά).

4. Ύπαρξη αριθμητικού ψηφίου στη λεκτική μονάδα (1 εάν υπάρχει, 0 διαφορετικά).

5. Ύπαρξη τελείας στη λεκτική μονάδα (1 εάν υπάρχει, 0 διαφορετικά).

6. Ύπαρξη κόμματος στη λεκτική μονάδα (1 εάν υπάρχει, 0 διαφορετικά).

7. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα (1 εάν υπάρχει, 0 διαφορετικά).

8. Η ετικέτα αμφισημίας (βλ. παρακάτω) της λεκτικής μονάδας.

9. Η κατάληξη της επόμενης λεκτικής μονάδας.

10. Η ετικέτα αμφισημίας της επόμενης λεκτικής μονάδας.

11. Η κατάληξη της προηγούμενης λεκτικής μονάδας.

12. Η προηγούμενη λεκτική μονάδα.

13. Η ετικέτα αμφισημίας της προηγούμενης λεκτικής μονάδας.

14. Η λεκτική μονάδα πριν την προηγούμενη λεκτική μονάδα.

15. Η κατάληξη της λεκτικής μονάδας πριν την προηγούμενη λεκτική μονάδα.

Η ετικέτα αμφισημίας (ambivalence tag - ambitag) [DaZa03] μιας λεκτικής μονάδας είναι μια συμβολοσειρά που περιέχει τις κατηγορίες (τα ονόματα των κατηγοριών) που μπορούν να βρεθούν για αυτή τη λεκτική μονάδα στο σύνολο εκπαίδευσης χωρισμένες με παύλες. Εάν η λεκτική μονάδα δεν έχει ακόμη προστεθεί στο σύνολο εκπαίδευσης, τότε η ετικέτα αμφισημίας είναι μια συμβολοσειρά που περιέχει τις κατηγορίες που έχουν βρεθεί για την κατάληξή της χωρισμένες με παύλες. Εάν δεν υπάρχει κάποια κατηγορία ούτε για την κατάληξη στο σύνολο εκπαίδευσης, τότε η ετικέτα αμφισημίας παίρνει την τιμή «unknown». Οι κατηγορίες της λεκτικής μονάδας ή της κατάληξής της εμφανίζονται στην ετικέτα αμφισημίας ταξινομημένες με βάση τη συχνότητα με την οποία έχουν βρεθεί στο σύνολο εκπαίδευσης. Εάν περισσότερες της μίας κατηγορίες εμφανίζονται με την ίδια συχνότητα, η ταξινόμηση γίνεται αλφαβητικά.

3.3 Σύνολα ετικετών

Στο σύστημα της εργασίας, το σύνολο των κατηγοριών (ετικετών) ορίζεται με δύο διαφορετικούς τρόπους, τον βασικό και τον εκτεταμένο. Περισσότερες πληροφορίες για την προέλευση των δύο συνόλων ετικετών παρέχονται στις προηγούμενες εργασίες [Μα05, Χρο06].

Το βασικό σύνολο περιέχει δώδεκα ετικέτες: άρθρο, ρήμα, σημείο στίξης, επίθετο, επίρρημα, σύνδεσμος, ουσιαστικό, αριθμητικό, μόριο, πρόθεση, πρόθεμα και άλλο. Το εκτεταμένο σύνολο περιέχει εκατόν εβδομήντα ετικέτες. Το σύνολο αυτό περιγράφεται αναλυτικά στο παράρτημα Ι, όπου δίνεται μια αναπαράσταση των ετικετών σε XML.

3.4 Κανόνες αυτόματης κατάταξης

Μια πρώτη λύση που ίσως θα πρότεινε κανείς για το πρόβλημα της αναγνώρισης μερών του λόγου είναι η κατασκευή ενός λεξικού που θα περιείχε όλες τις πιθανές λεκτικές μονάδες και τις αντίστοιχες κατηγορίες τους. Μια τέτοια προσέγγιση όμως αποτυγχάνει, επειδή δεν μπορούμε να αποφανθούμε πάντα για το μέρος του λόγου μιας λεκτικής μονάδας χωρίς να εξετάσουμε τα συμφραζόμενά της. Για παράδειγμα, στην πρόταση «Η τσάντα αυτή είναι της κυρίας.» η λεκτική μονάδα «της» είναι άρθρο, ενώ στην πρόταση «Η τσάντα είναι δικιά της.» είναι αντωνυμία. Ωστόσο, σε ορισμένες περιπτώσεις μπορούμε όντως να αναγνωρίσουμε μέρη του λόγου γνωρίζοντας μονάχα την αντίστοιχη λέξη. Για παράδειγμα, η λεκτική μονάδα «!» πρέπει να καταταγεί ως σημείο στίξης, ανεξάρτητα από τα συμφραζόμενά της.

Στο παρόν σύστημα χρησιμοποιείται ένας συνδυασμός των δύο προσεγγίσεων. Έτσι, μια λεκτική μονάδα δίνεται προς κατάταξη στον ταξινομητή *k*-NN μόνο εφόσον δεν μπορεί να καταταγεί βάσει ενός λεξικού που δημιουργείται στη διάρκεια της εκπαίδευσης. Το λεξικό περιλαμβάνει κατηγορίες λέξεων που δεν είναι δυνατόν να εμφανιστούν με περισσότερες από μία ετικέτες. Οι κατηγορίες αυτές είναι (βλ. και παράρτημα Ι) οι ακόλουθες. Σημειώνεται ότι εάν επιλεγεί το βασικό σύνολο ετικετών, τότε χρησιμοποιούνται μόνο οι τρείς τελευταίες.

- article/definite/nominative/masculine/singular
- article/definite/nominative/feminine/singular
- article/prepositional/genitive/feminine/singular
- article/prepositional/accusative/feminine/singular
- article/prepositional/accusative/feminine/plural
- article/prepositional/accusative/masculine/plural
- pronoun/inflectionless
- other/abbreviation
- other/foreign_word
- other/symbol
- other/other
- punctuation
- conjunction
- particle

Για κάθε μία από τις παραπάνω κατηγορίες, δημιουργείται κατά την εκπαίδευση μια λίστα. Για να περιληφθεί μια λεκτική μονάδα εκπαίδευσης σε κάποια λίστα, πρέπει να μην είναι όλοι οι χαρακτήρες της κεφαλαίοι. Με τον έλεγχο αυτό αποφεύγουμε την λανθασμένη κατάταξη ακρωνύμων. Για παράδειγμα, ακόμα και εάν η λεκτική μονάδα «ΚΑΙ» βρεθεί ως σύνδεσμος στο σώμα εκπαίδευσης, δεν αποκλείεται η ίδια

να βρεθεί σε άλλα κείμενα ως ακρώνυμο (π.χ. σε έναν τίτλο). Επιπλέον, για αντίστοιχους λόγους, για να περιληφθεί μια λεκτική μονάδα στην λίστα ξένων λέξεων πρέπει όλοι οι χαρακτήρες της να είναι πεζοί.

Μετά τους παραπάνω ελέγχους, ελέγχεται εάν όλοι οι χαρακτήρες της λεκτικής μονάδας είναι αριθμητικά ψηφία. Αν είναι, τότε η λεκτική μονάδα κατατάσσεται άμεσα ως αριθμητικό. Τέλος, εάν η λεκτική μονάδα περιλαμβάνει τουλάχιστον ένα λατινικό χαρακτήρα, κατατάσσεται ως ξένη λέξη.

ΚΕΦΑΛΑΙΟ 4:

ΔΕΔΟΜΕΝΑ, ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Επιλογή και επεξεργασία δεδομένων

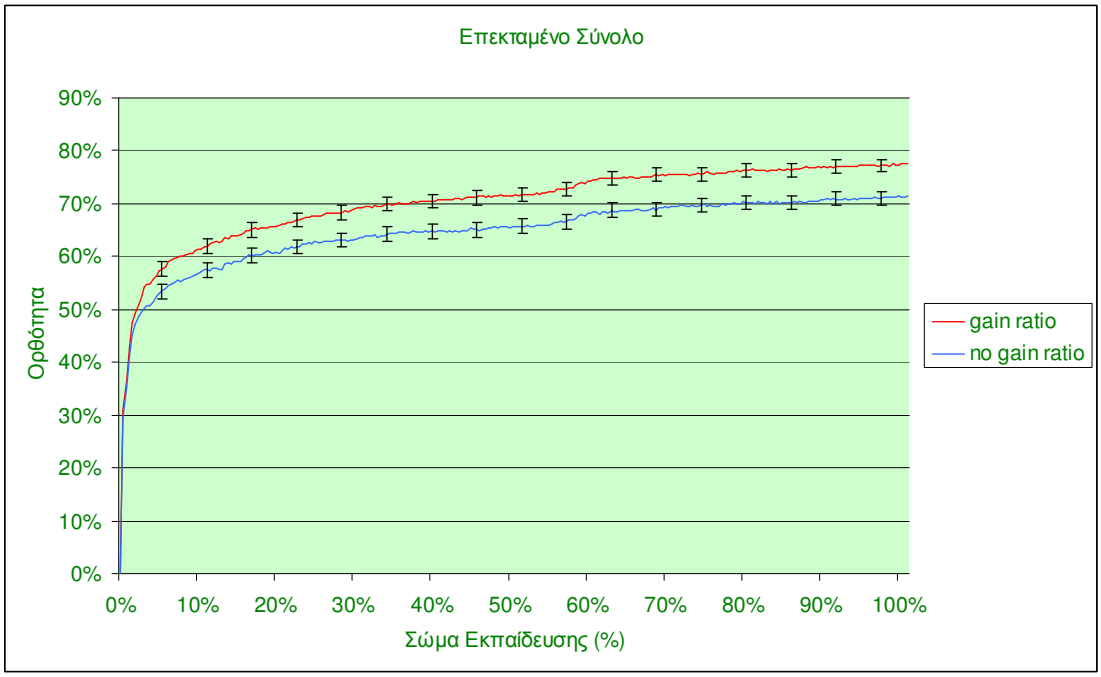
Για την ανάπτυξη του συστήματος υπήρχαν από προηγούμενες εργασίες διαθέσιμα χιλιάδες ειδησεογραφικά κείμενα από τις εφημερίδες «ΤΑ ΝΕΑ» και «ΒΗΜΑ». Από αυτά 46 χρησιμοποιήθηκαν για την κατασκευή ενός σώματος κειμένων ελέγχου 7878 λεκτικών μονάδων και 160 για την κατασκευή ενός σώματος κειμένων εκπαίδευσης 23675 λεκτικών μονάδων.

Τα σώματα κειμένων εκπαίδευσης και ελέγχου ανανεώθηκαν, σε σχέση με αυτά των προηγούμενων εργασιών, με σκοπό να καλυφθεί όσο το δυνατόν μεγαλύτερο φάσμα θεματικών ενοτήτων και αρθρογράφων. Σε αντίθετη περίπτωση, το σύστημα θα «μάθαινε» τις ιδιαιτερότητες λίγων θεματικών ενοτήτων και θα εξειδικευόταν στον ιδιαίτερο τρόπο γραφής συγκεκριμένων αρθρογράφων. Στην παρούσα εργασία επιλέχθηκαν από κάθε ειδησεογραφικό κείμενο οι πρώτες 150 μόνο λέξεις και όσες επιπλέον χρειαζόταν για να ολοκληρωθεί η τελευταία πρόταση. Τα νέα κείμενα επισημειώθηκαν χειρωνακτικά χρησιμοποιώντας το εργαλείο επισημείωσης του συστήματος.

4.2 Ρυθμίσεις του ταξινομητή

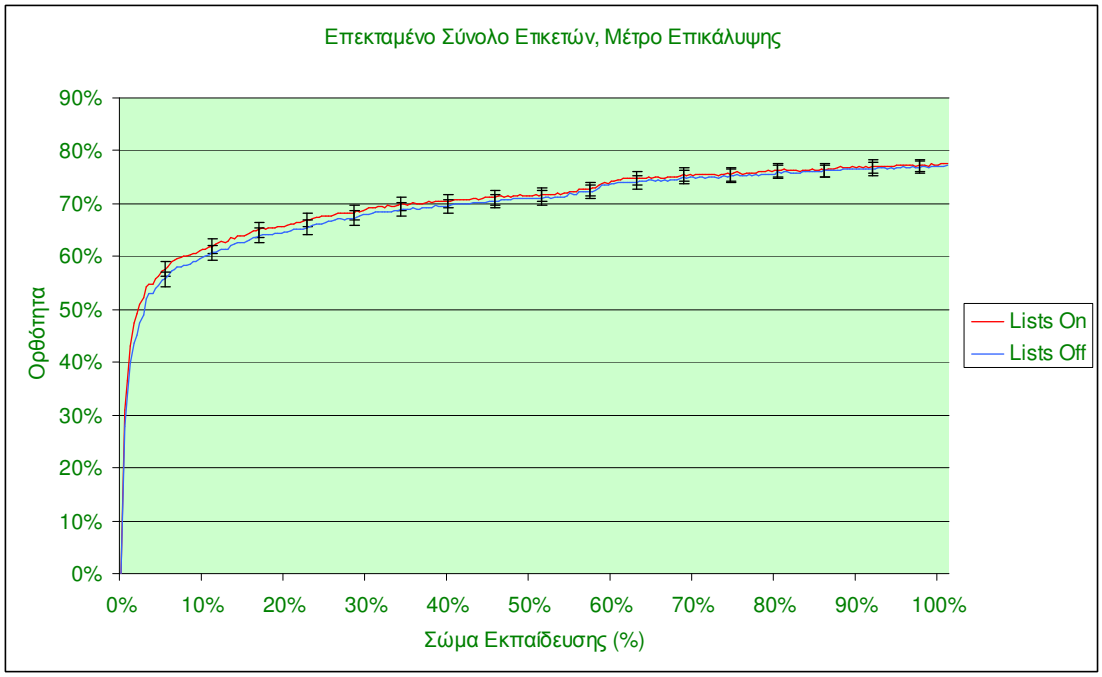
Οι συνιστώσες του συστήματος που αξιολογήθηκαν πειραματικά είναι: η επίδοση του αλγορίθμου k-NN για διάφορες τιμές της παραμέτρου k, η χρήση ή όχι της αναλογίας κέρδους κατά τον υπολογισμό της απόστασης δύο διανυσμάτων, η ενεργοποίηση ή μη των κανόνων αυτόματης κατάταξης και η χρήση του μέτρου επικάλυψης ή του τροποποιημένου μέτρου διαφοράς τιμών.

Προκαταρκτικά πειράματα έδειξαν ότι οι επιδόσεις του ταξινομητή δεν επηρεάζονται σημαντικά από την τιμή του k, ίσως επειδή χρησιμοποιήσαμε σε όλα τα πειράματα ζύγισμα των ψήφων των γειτόνων βάσει των αποστάσεών τους από την υπό κατάταξη περίπτωση. Έτσι, ως προεπιλογή τίθεται η απλούστερη τιμή, δηλαδή k =1. Επίσης, το ζύγισμα με βάση την αναλογία κέρδους φαίνεται να βελτιώνει τις επιδόσεις (βλ. παρακάτω διάγραμμα), ενώ ταυτόχρονα μειώνει το χρόνο κατάταξης. Η μείωση του χρόνου κατάταξης δικαιολογείται, επειδή χωρίς ζύγισμα παίρνουμε πολλά διανύσματα σε ίδιες αποστάσεις, οπότε προκύπτουν πολλοί γείτονες στις κοντινότερες γειτονιές. Αυτό αυξάνει το κόστος εύρεσης πλειοψηφούσας κατηγορίας. Από την άλλη, η αναλογία κέρδους μίας ιδιότητας υπολογίζεται μία μόνο φορά στο στάδιο της εκπαίδευσης οπότε δεν επιβαρύνει το στάδιο κατάταξης. Ως προεπιλογή τίθεται η χρήση αναλογίας κέρδους.

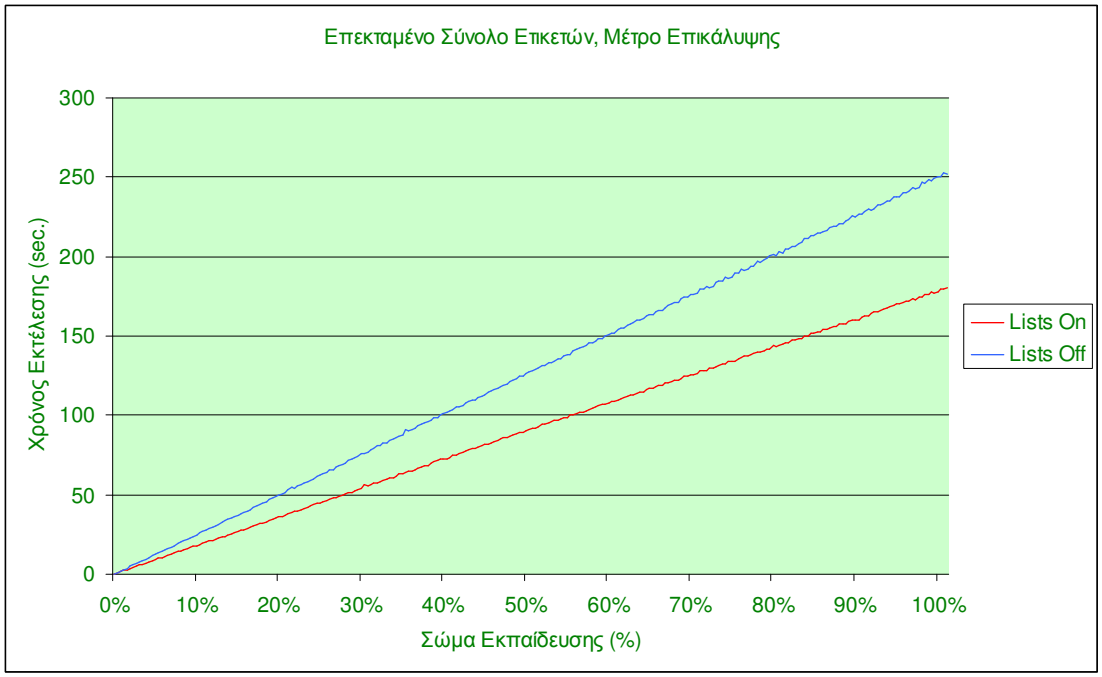


Σχήμα 4.1 Βελτίωση με χρήση αναλογίας κέρδους.

Οι κανόνες αυτόματης κατάταξης γενικά επιδρούν μόνο θετικά στη διαδικασία κατάταξης, επειδή απαντούν με απόλυτη βεβαιότητα για το μέρος του λόγου κάθε λέξης που περιλαμβάνεται στις λίστες τους. Έτσι περιορίζουν τα πιθανά σφάλματα του ταξινομητή *k*-NN μόνο στις υπόλοιπες λέξεις. Πειραματικά φάνηκε ότι, αν και η επίδοση γενικά βελτιώνεται με τη χρήση αυτών των κανόνων, όταν το σώμα εκπαίδευσης είναι αρκετά μεγάλο οι λέξεις που κατατάσσονται από τους κανόνες θα μπορούσαν να καταταγούν εξίσου ορθά από τον ταξινομητή *k*-NN. Εντούτοις, η διαδικασία ολοκληρώνεται αισθητά πιο γρήγορα όταν χρησιμοποιούνται οι κανόνες, επειδή για την κατάταξη των λεκτικών μονάδων που περιλαμβάνονται στις λίστες τους αποφεύγεται η χρονοβόρα εύρεση των κοντινότερων γειτόνων και η επιλογή της πλειοψηφούσας κατηγορίας.

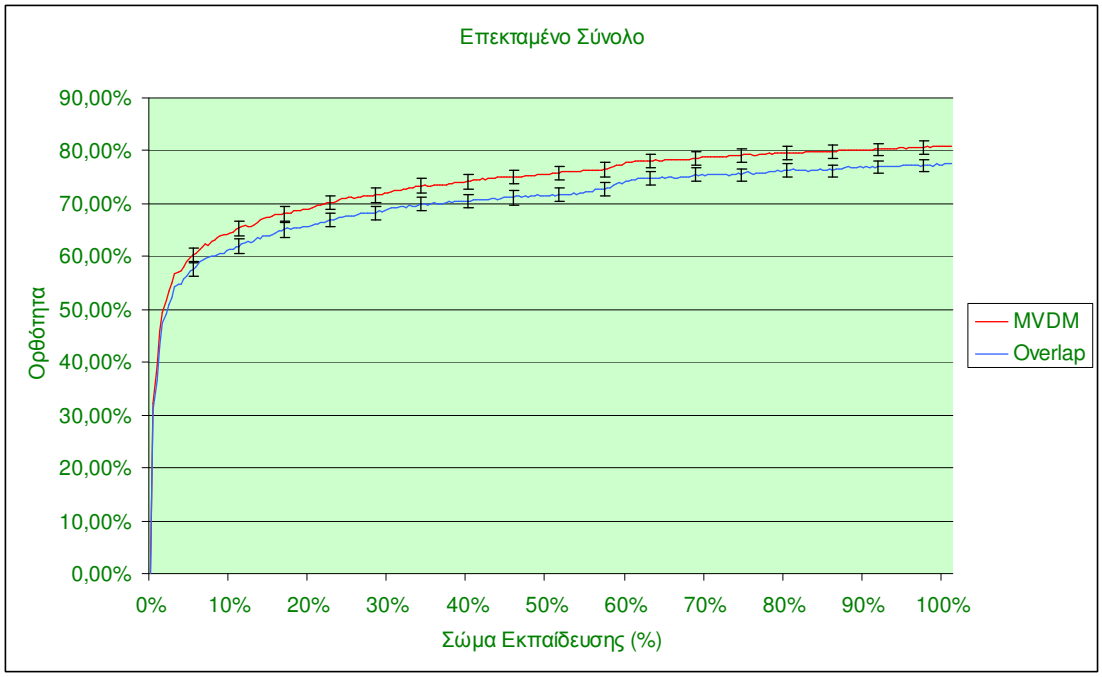


Σχήμα 4.2 Ποσοστό ορθότητας με και χωρίς κανόνες αυτόματης κατάταξης.

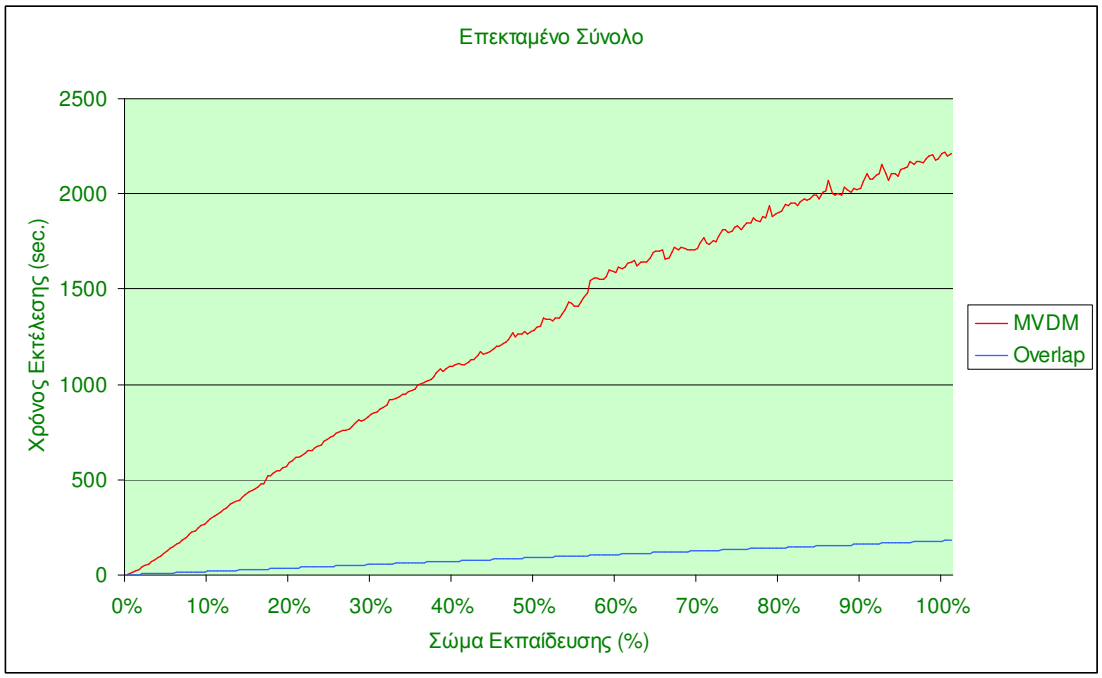


Σχήμα 4.3 Χρόνος κατάταξης με και χωρίς κανόνες αυτόματης κατάταξης.

Τέλος, η χρήση του τροποποιημένου μέτρου διαφοράς τιμών παρόλο που επιτυγχάνει υψηλότερες επιδόσεις σε σχέση με το μέτρο επικάλυψης επιβαρύνει πολύ την εφαρμογή από άποψη χρόνου, όπως φαίνεται στα παρακάτω διαγράμματα.



Σχήμα 4.4 Σύγκριση επίδοσης των δύο μέτρων απόστασης.



Σχήμα 4.5 Σύγκριση χρόνων κατάταξης με τα δύο μέτρα απόστασης.

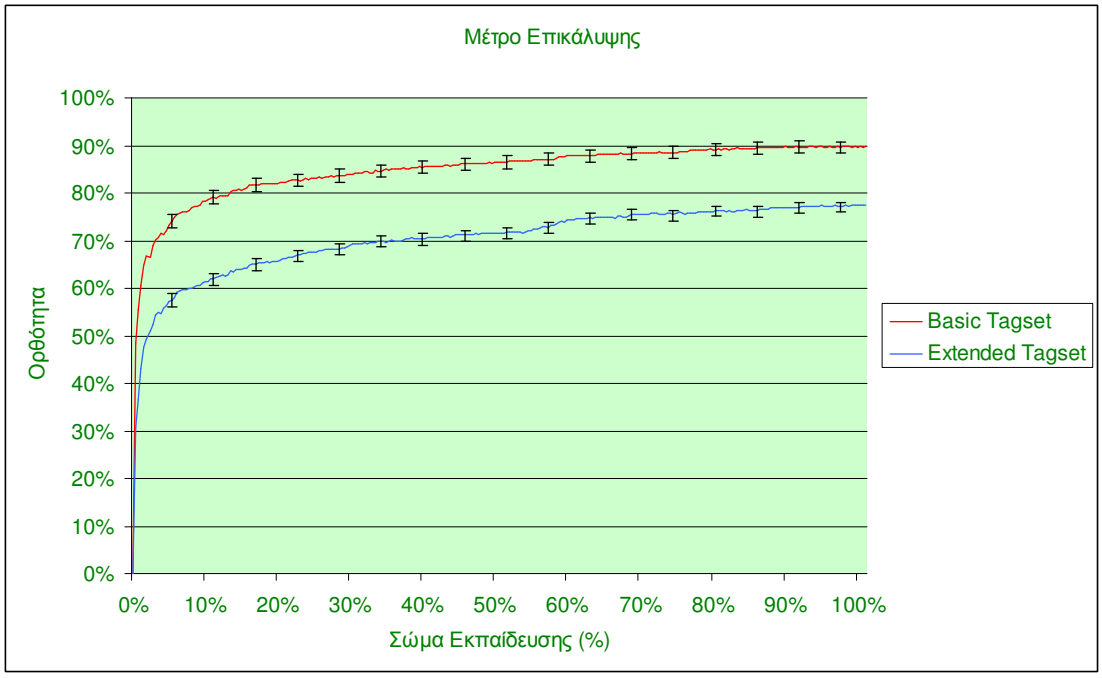
Με βάση τα παραπάνω, επιλέχθηκε ως προεπιλογή το μέτρο επικάλυψης, επειδή κρίθηκε εξαιρετικά μη αποδοτική η καθυστέρηση του διαφοροποιημένου μέτρου.

Στα παραπάνω γραφήματα ποσοστών ορθότητας δίνονται και τα αντίστοιχα διαστήματα εμπιστοσύνης με βαθμό βεβαιότητας 99%.

4.3 Πειράματα με παθητική μάθηση

Τα πειράματα παθητικής μάθησης επιχειρούν να μετρήσουν την ορθότητα που επιτυγχάνει ο ταξινομητής για διάφορα μεγέθη του σώματος εκπαίδευσης. Στα πειράματα που εκτελέστηκαν, το σώμα ελέγχου παρέμενε σταθερό στις 7878 λεκτικές μονάδες. Σε κάθε επανάληψη, το σώμα εκπαίδευσης αυξανόταν κατά 90 λεκτικές μονάδες οι οποίες επιλέγονταν κατά τη σειρά με την οποία εμφανίζονταν στα κείμενα. Ο ταξινομητής εκπαιδευόταν στο νέο σώμα εκπαίδευσης και η ορθότητά του αξιολογούταν εκ νέου. Με τον ίδιο τρόπο εκτελέστηκαν και τα πειράματα της προηγούμενης ενότητας.

Ο ταξινομητής ρυθμίστηκε με βάση τις προεπιλογές που αναφέρθηκαν στην προηγούμενη ενότητα. Οι επιδόσεις του μετρήθηκαν και για τα δύο σύνολα ετικετών, όπως φαίνεται στο παρακάτω διάγραμμα. Είναι φανερό ότι το πρόβλημα αναγνώρισης μερών του λόγου είναι αρκετά πιο εύκολο όταν επιλέγεται το βασικό σύνολο ετικετών.



Σχήμα 4.6 Επιδόσεις ταξινομητή με χρήση παθητικής μάθησης.

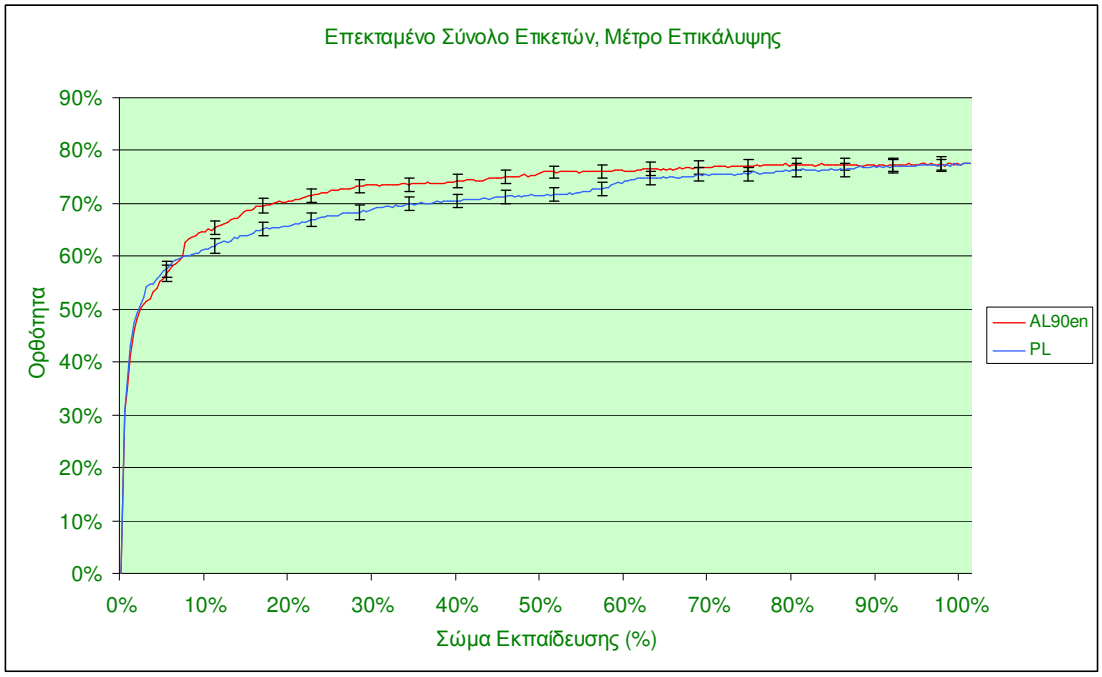
4.4 Πειράματα με ενεργητική μάθηση

Στην ενεργητική μάθηση όλα τα υποψήφια παραδείγματα ταξινομούνται ως προς την χρησιμότητά τους με βάση μέτρα όπως αυτά που περιγράφονται στην ενότητα 2.6. Σε κάθε επανάληψη των πειραμάτων επιλέγεται το χρησιμότερο από τα υποψήφια παραδείγματα, αφαιρείται από το σύνολο υποψήφιων παραδειγμάτων και προστίθεται στο νέο σώμα εκπαίδευσης.

Στα πειράματα που πραγματοποιήθηκαν, για να εξοικονομηθεί χρόνος, σε κάθε επανάληψη επιλέγονταν τα 90 χρησιμότερα παραδείγματα. Χρησιμοποιήθηκε το

μέτρο που περιγράφεται στη σχέση (2.11), το οποίο προηγούμενα πειράματα [Μα05], [Χρο06] έχουν δείξει πως έχει παρόμοια επίδοση με αυτό της σχέσης (2.12). Το σύνολο όλων των υποψήφιων παραδειγμάτων επιλέχθηκε να είναι το ίδιο με αυτό που χρησιμοποιήθηκε και στα πειράματα παθητικής μάθησης, ώστε να μετρήσουμε τα οφέλη της ενεργητικής μάθησης στα ίδια δεδομένα, αλλά και να αποφύγουμε επιπλέον χειρωνακτικές επισημειώσεις. Αυτό έχει ως αποτέλεσμα οι καμπύλες ορθότητας να συγκλίνουν στο σημείο όπου χρησιμοποιείται όλο το σώμα υποψήφιων παραδειγμάτων. Άρα στα παρακάτω αποτελέσματα μας ενδιαφέρει η ταχύτητα με την οποία αυξάνεται η ορθότητα του ταξινομητή συναρτήσει του μεγέθους του σώματος εκπαίδευσης.

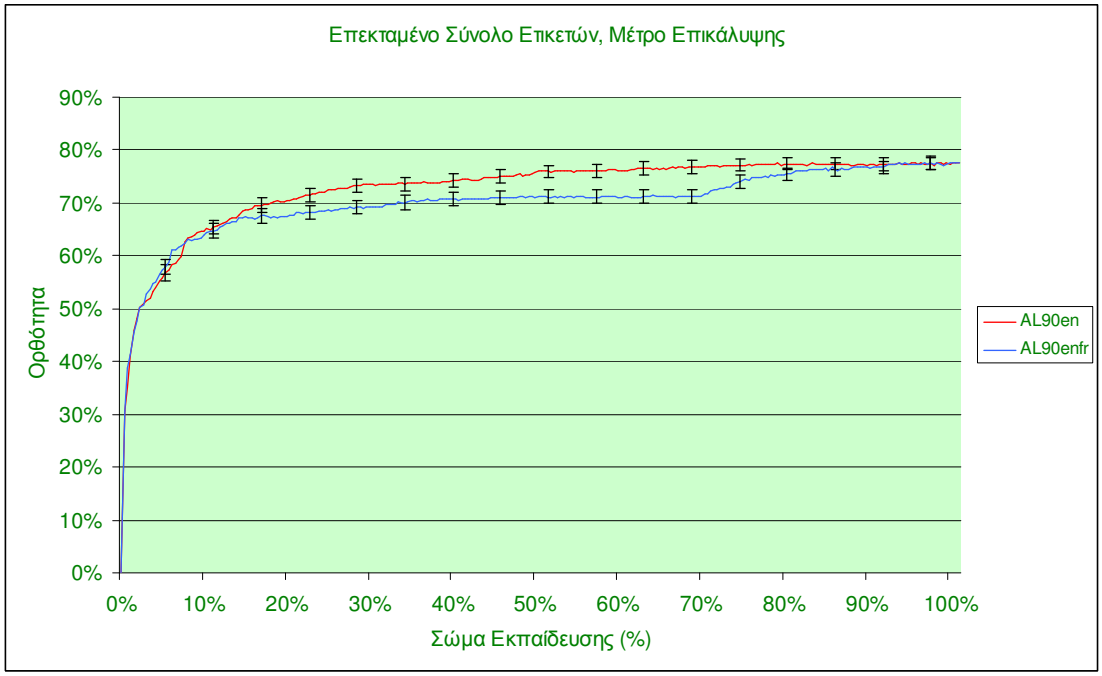
Αρχικά, μετρήθηκε η ορθότητα του ταξινομητή όταν χρησιμοποιείται ενεργητική μάθηση και συγκρίθηκε με τα αντίστοιχα αποτελέσματα που πήραμε στην παθητική μάθηση. Όπως φαίνεται και στο παρακάτω σχήμα, όταν χρησιμοποιείται το μέτρο επιλογής χρησιμότερων παραδειγμάτων της σχέσης (2.11) ο ταξινομητής επιτυγχάνει γενικά υψηλότερα επίπεδα ορθότητας σε σχέση με την παθητική μάθηση.



Σχήμα 4.7 Σύγκριση ενεργητικής και παθητικής μάθησης.

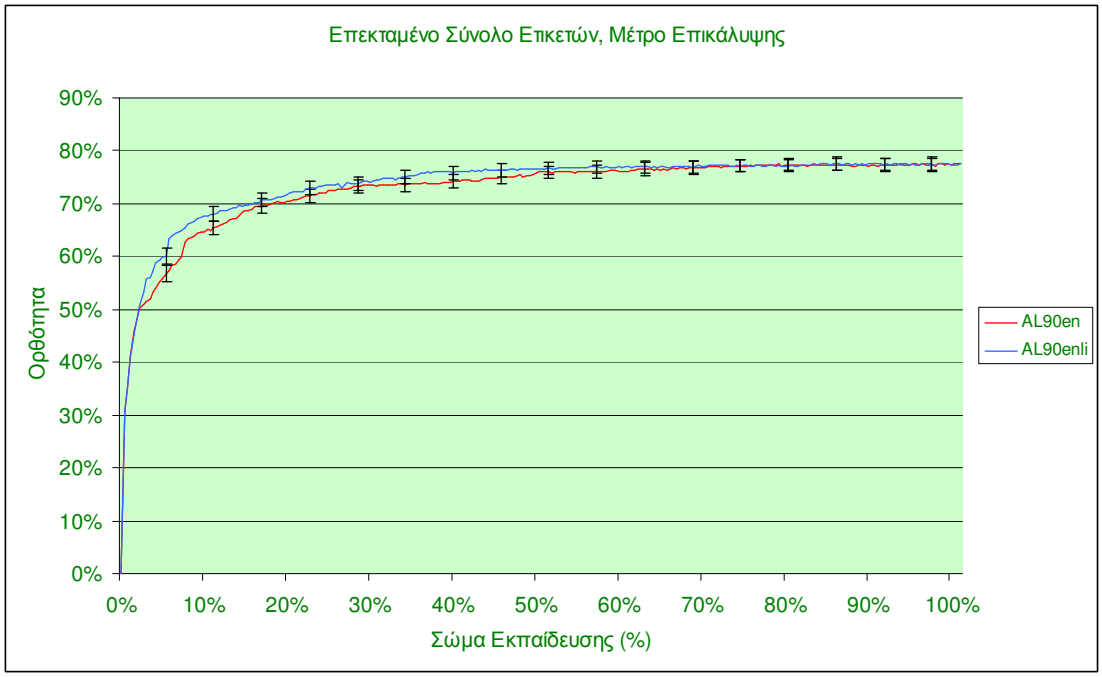
Μια άλλη ιδέα που διερευνήσαμε ήταν κατά την επιλογή παραδειγμάτων εκπαίδευσης στην ενεργητική μάθηση να εξαιρούνται οι σπάνιες περιπτώσεις. Στην περίπτωσή μας, αυτό σημαίνει να αγνοούνται τα υποψήφια διανύσματα εκπαίδευσης που προέρχονται από λέξεις οι οποίες εμφανίζονται λίγες μόνο φορές στο σώμα υποψήφιων παραδειγμάτων εκπαίδευσης. Στο πείραμα που αντιστοιχεί στην κάτω καμπύλη του ακόλουθου διαγράμματος αποκλείστηκαν υποψήφια διανύσματα εκπαίδευσης που προέρχονταν από λέξεις οι οποίες εμφανίζονται μία ή δύο μόνο φορές στο αρχικό σώμα υποψήφιων παραδειγμάτων. Η επιλογή τέτοιων παραδειγμάτων επιτράπηκε μόνο εφόσον είχαν εξαντληθεί όλα τα υπόλοιπα. Παρατηρούμε ότι η ορθότητα γενικά δεν βελτιώνεται. Φαίνεται, δηλαδή, ότι η

εξαίρεση παραδειγμάτων χαμηλής συχνότητας αφαιρεί από το σύνολο υποψήφιων παραδειγμάτων χρήσιμα παραδείγματα.



Σχήμα 4.8 Απόδοση όταν εξαιρούνται παραδείγματα λέξεων συχνότητας ένα και δύο.

Στη συνέχεια, μετρήθηκε η βελτίωση που έχουμε όταν εξαιρούνται από τα υποψήφια παραδείγματα εκπαίδευσης τα διανύσματα που μπορούν να καταταγούν από τους κανόνες αυτόματης κατάταξης. Στο πείραμα του επόμενου διαγράμματος, σε κάθε επανάληψη επιλέγονταν μόνο παραδείγματα τα οποία δεν μπορούσαν να καταταγούν από τους κανόνες αυτόματης κατάταξης, εκτός αν δεν είχαν απομείνει άλλα υποψήφια παραδείγματα. Φαίνεται ότι αυτή η χρήση των κανόνων κατάταξης επιδρά θετικά στα αποτελέσματα της ενεργητικής μάθησης, ιδιαίτερα όταν τα παραδείγματα εκπαίδευσης είναι λίγα, αν και η βελτίωση στο ποσοστό ορθότητας είναι μικρή, όπως και στην περίπτωση της παθητικής μάθησης. Σημειώνεται ότι είναι επόμενο οι καμπύλες να συγκλίνουν στα δεξιά, αφού τότε αναγκαζόμαστε να χρησιμοποιήσουμε παραδείγματα εκπαίδευσης που κατατάσσονται από τους κανόνες, μια που δεν έχουν απομείνει άλλα. Η χρήση των κανόνων μειώνει, επίσης, τις απαιτήσεις σε μνήμη και το χρόνο ταξινόμησης, όπως ήδη αναφέρθηκε στην περίπτωση της παθητικής μάθησης.



Σχήμα 4.9 Χρήση κανόνων αυτόματης κατάταξης στην ενεργητική μάθηση.

4.5 Επιπλέον παρατηρήσεις

Η ιδέα να εξαιρούνται τα υποψήφια διανύσματα εκπαίδευσης που προέρχονται από λέξεις με μικρή συχνότητα εμφάνισης φάνηκε να μη βελτιώνει τα αποτελέσματα. Ωστόσο, στα παραπάνω πειράματα το αρχικό σύνολο υποψήφιων παραδειγμάτων ήταν το ίδιο που χρησιμοποιήθηκε και στην παθητική μάθηση. Ήταν δηλαδή μια συλλογή κειμένων 23675 λεκτικών μονάδων. Σε αυτό το σώμα κειμένων οι λέξεις που εμφανίζονταν με συχνότητα ένα και δύο ήταν στο σύνολό τους 7323. Ήταν, δηλαδή, σχεδόν το ένα τρίτο όλων των λέξεων. Έτσι, περιλαμβάνονταν και πολλές λέξεις που δεν παρουσίαζαν κάποια ιδιορυθμία και ενδεχομένως θα ήταν χρήσιμες στον ταξινομητή. Ενδεικτικά, μερικές από αυτές ήταν οι λέξεις «συνολική», «κατανομή», «νομοσχέδιο», «σύντομη», «αναφορά», «φώτα», «Οδηγός».

Εάν είχε χρησιμοποιηθεί ένα πιο μεγάλο σύνολο κειμένων, για παράδειγμα ένα σύνολο από κείμενα εκατομμυρίων λέξεων, τότε οι λέξεις που θα εμφανίζονταν εκεί με συχνότητα ένα και δύο μόνο θα ήταν αρκετά πιο σπάνιες. Τέτοιες λέξεις ίσως να επιλέγονταν λανθασμένα από το μέτρο επιλογής χρησιμότερων παραδειγμάτων, επειδή θα παρουσίαζαν υψηλή εντροπία για τον ταξινομητή. Επιπλέον, είναι εξίσου πιθανό λόγω της σπανιότητάς τους η εκπαίδευση σε τέτοιες λέξεις να μη βελτίωνε επαρκώς την ορθότητα του ταξινομητή. Τα δύο αυτά προβλήματα είχαν παρατηρηθεί σε προηγούμενη εργασία [Χρο06]. Έτσι, μια μέθοδος όπως αυτή της εξαίρεσης λέξεων χαμηλής συχνότητας κατά την επιλογή παραδειγμάτων είναι πολύ πιθανό να οδηγήσει σε καλύτερα αποτελέσματα όταν χρησιμοποιείται ένα πολύ μεγαλύτερο σώμα κειμένων ως δεξαμενή υποψήφιων παραδειγμάτων. Δυστυχώς, παρ' όλο που είχαμε στη διάθεσή μας μια τέτοια δεξαμενή (τα χιλιάδες άρθρα εφημερίδων που

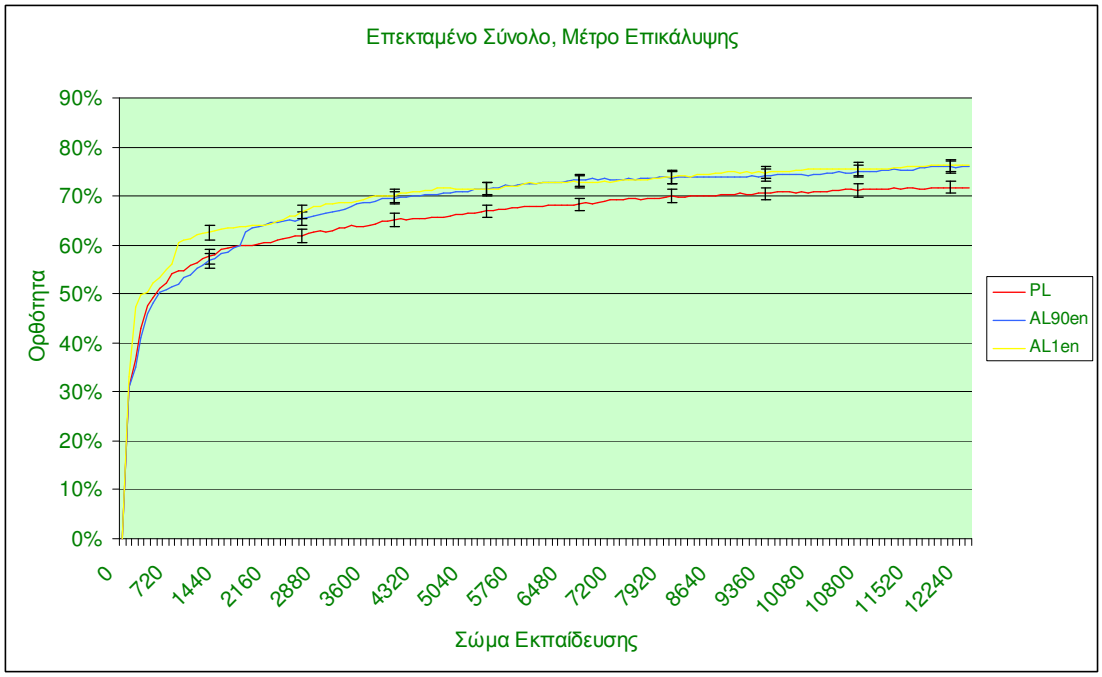
αναφέρθηκαν στην αρχή του κεφαλαίου), περιορισμοί χρόνου δεν επέτρεψαν τη διερεύνηση αυτής της ιδέας.

Επιπλέον, στα παραπάνω πειράματα ενεργητικής μάθησης, μετά την ταξινόμηση όλων των υποψήφιων παραδειγμάτων ως προς την χρησιμότητά τους, επιλέγονταν σε κάθε επανάληψη τα 90 από αυτά, προκειμένου να εξοικονομείται χρόνος (λιγότερες επαναξιολογήσεις των υποψηφίων παραδειγμάτων). Έτσι, όμως, στα 90 χρησιμότερα παραδείγματα που επιλέγονται ανά επανάληψη είναι πιθανό να περιλαμβάνονται πολλά που είναι εξαιρετικά όμοια. Οπότε, μετά την ενσωμάτωση ενός από τα παρόμοια στο σύνολο εκπαίδευσης, τα υπόλοιπα παύουν να είναι χρήσιμα.

Το παραπάνω φαινόμενο αποκτά μεγαλύτερη βαρύτητα όταν το μέγεθος του συνόλου υποψήφιων παραδειγμάτων αυξάνεται. Όσο περισσότερα υποψήφια διανύσματα εκπαίδευσης έχουμε, τόσο πιθανότερο είναι για κάθε ένα από αυτά να μπορούν να βρεθούν παρόμοιά του. Ενδεικτικά, δίνονται τα δέκα χρησιμότερα διανύσματα όπως ταξινομήθηκαν με το μέτρο της σχέσης (2.11) στην πρώτη επανάληψη των πειραμάτων ενεργητικής μάθησης όπου επιλέγονταν 90 παραδείγματα σε κάθε επανάληψη. Επίσης, παρουσιάζεται (επάνω καμπύλη) η βελτίωση που επιτυγχάνεται στα πρώτα στάδια του πειράματος (μέχρι 12330 διανύσματα εκπαίδευσης, που αντιστοιχούν στο 52.08% των συνολικών διανυσμάτων εκπαίδευσης των υπολοίπων πειραμάτων), όταν επιλέγεται ένα μόνο χρησιμότερο παράδειγμα ανά επανάληψη.

```
<ρία, 7,  0, 0, 0, 0, 0, unknown, της, f-5s, ),           ),   Η,   Η,           unknown>
<λών, 5,  0, 0, 0, 0, 0, unknown, της, f-5s, κρατών,      των, -,   -,           unknown>
<τια, 5,  0, 0, 0, 0, 0, unknown, της, f-5s, ότι,         ότι, τα,  τα,          unknown>
<τος, 11, 0, 0, 0, 0, 0, unknown, της, f-5s, του,         του, λου, όλου,        unknown>
<ρες, 5,  0, 0, 0, 0, 0, unknown, της, f-5s, στις,        τις, ρες, περισσότερες, unknown>
<εια, 6,  0, 0, 0, 0, 0, unknown, της, f-5s, τα,          τα,  ικά, στεγαστικά,  unknown>
<ώτα, 4,  0, 0, 0, 0, 0, unknown, της, f-5s, μονοπωλεί,   λεί, τα,  τα,          unknown>
<ένα, 7,  0, 0, 0, 0, 0, unknown, της, f-5s, τα,          τα,  ικά, θεατρικά,    unknown>
<νία, 6,  0, 0, 0, 0, 0, unknown, της, f-5s, χαρακτηρίζει, ζει, η,   η,          unknown>
<»,   1,  0, 0, 0, 0, 0, unknown, της, f-5s, «,           «,   νες, σειρήνες,    unknown>
```

Σχήμα 4.10 Τα δέκα χρησιμότερα διανύσματα της πρώτης επανάληψης.



Σχήμα 4.11 Σύγκριση τριών μεθόδων επιλογής χρησιμότερων παραδειγμάτων.

Στο σχήμα 4.10 φαίνεται πως τα πρώτα διανύσματα είναι όντως πολύ όμοια. Οι 9 στις 15 ιδιότητές τους έχουν την ίδια τιμή. Στο σχήμα 4.11 γίνεται φανερό πως για το συγκεκριμένο σώμα κειμένων των 23675 λέξεων είναι πιο αποδοτικό να επιλεχθούν σειριακά (παθητική μάθηση) τα πρώτα παραδείγματα εκπαίδευσης από το να επιλέγονται τα 90 χρησιμότερα, ενδεχομένως επειδή η κάθε 90-άδα περιλαμβάνει πολλά παρόμοια παραδείγματα. Αντιθέτως, όταν επιλέγεται ένα μόνο παράδειγμα σε κάθε επανάληψη της ενεργητικής μάθησης, η ενεργητική μάθηση επιτυγχάνει εξαρχής καλύτερα αποτελέσματα. Επαναλαμβάνεται ότι αυτή η συμπεριφορά (πολλά παρόμοια παραδείγματα σε κάθε 90-άδα) αναμένεται να εμφανιστεί σε πιο έντονο βαθμό εάν το σύνολο υποψήφιων παραδειγμάτων αυξηθεί σημαντικά. Αξίζει να σημειωθεί πως από ένα σημείο και έπειτα δεν υπάρχει σημαντική διαφορά μεταξύ των περιπτώσεων όπου επιλέγονται ένα ή περισσότερα παραδείγματα ανά επανάληψη.

Το μειονέκτημα του να επιλέγεται σε κάθε επανάληψη ένα μόνο παράδειγμα εκπαίδευσης είναι ότι απαιτούνται περισσότερες επαναλήψεις για να συγκεντρωθεί ο ίδιος αριθμός παραδειγμάτων εκπαίδευσης. Υπενθυμίζεται ότι σε κάθε επανάληψη απαιτείται η επαναξιολόγηση όλων των υποψηφίων παραδειγμάτων εκπαίδευσης, κάτι που είναι εξαιρετικά χρονοβόρο όταν η δεξαμενή υποψηφίων παραδειγμάτων είναι πολύ μεγάλη. Εντούτοις, η βελτίωση στο ποσοστό ορθότητας που επιτυγχάνεται με αυτό τον τρόπο είναι εξαιρετικά σημαντική, ιδιαίτερα στα πρώτα στάδια. Έτσι, στο σύστημα σε κάθε επανάληψη δίνεται η δυνατότητα επιλογής του πλήθους των χρησιμότερων παραδειγμάτων που θα επιλεγούν. Στο μέλλον θα μπορούσε να προστεθεί η δυνατότητα να απορρίπτονται κατά τη συμπλήρωση της δέσμης (batch) επιλεγέντων παραδειγμάτων κάθε επανάληψης τα διανύσματα εκπαίδευσης που είναι πολύ παρόμοια με διανύσματα που έχουν ήδη περιληφθεί στην δέσμη και να αντικαθίστανται από τα αμέσως λιγότερο «χρήσιμα».

ΚΕΦΑΛΑΙΟ 5:

ΕΠΙΛΟΓΟΣ

5.1 Ανασκόπηση

Ο βασικός στόχος της παρούσας εργασίας ήταν η δημιουργία ενός βελτιωμένου και ελεύθερα διαθέσιμου επισημειωτή μερών του λόγου για ελληνικά κείμενα. Ο στόχος αυτός επιτεύχθηκε. Δημιουργήθηκε μια νέα υλοποίηση του αλγορίθμου *k*-NN, η οποία χρησιμοποιήθηκε κατόπιν ως βάση προκειμένου να επανυλοποιηθεί εξαρχής ο επισημειωτής μερών του λόγου που είχε αναπτυχθεί σε προηγούμενες εργασίες. Η νέα υλοποίηση είναι πλήρως τεκμηριωμένη, διαθέτει εύχρηστη διεπαφή χρήστη και παρέχεται ελεύθερα. Δημιουργήθηκαν, επίσης, βελτιωμένα σώματα κειμένων εκπαίδευσης και αξιολόγησης, τα οποία χρησιμοποιήθηκαν σε νέα πειράματα. Μελετήθηκαν, τέλος, πιθανοί τρόποι αντιμετώπισης των προβλημάτων που είχαν παρατηρηθεί στις προηγούμενες εργασίες κατά τη χρήση ενεργητικής μάθησης και προτάθηκαν τρόποι περαιτέρω βελτίωσης του συστήματος.

5.2 Μελλοντικές επεκτάσεις

Πολύ ενδιαφέρον θα ήταν να γίνουν πειράματα ενεργητικής μάθησης με μία πολύ μεγάλη δεξαμενή κειμένων από την οποία θα επιλέγονταν τα παραδείγματα της ενεργητικής μάθησης. Σε αυτό το μεγάλο σώμα κειμένων θα είχε ιδιαίτερο νόημα να διερευνηθεί η απόδοση όταν εξαιρούνται παραδείγματα λέξεων χαμηλής συχνότητας εμφάνισης.

Επιπλέον βελτιώσεις θα εστίαζαν σε μια πιο εκλεπτυσμένη αναθεώρηση του συνόλου ετικετών, για την εξυπηρέτηση πιο εξεζητημένων αναγκών, σε έναν εμπλουτισμό του υπάρχοντος επισημειωμένου σώματος εκπαίδευσης, ενδεχομένως με κείμενα διαφορετικού είδους από τα ειδησεογραφικά, και στην μελέτη ενός πιο αποδοτικού μέτρου επιλογής παραδειγμάτων στην ενεργητική μάθηση. Ως προς το τελευταίο, προτάθηκαν ήδη πιθανές βελτιώσεις στο τέλος του προηγούμενου κεφαλαίου.

ΠΑΡΑΡΤΗΜΑ Ι:

Αναπαράσταση των ετικετών σε XML

Παρακάτω προδιαγράφονται όλες οι XML ετικέτες που περιλαμβάνονται στο εκτεταμένο σύνολο. Παραδείγματα χρήσης των ετικετών περιλαμβάνονται στο συνοδευτικό CD της εργασίας. Σημειώνεται ότι κατά σύμβαση οι παθητικές μετοχές θεωρούνται επίθετα ή ουσιαστικά, ανάλογα με το πώς εμφανίζονται στο κείμενο.

```
tag ::=          'PoS="adjective"' adjectiveTag |
                 'PoS="adverb"' |
                 'PoS="article"' articleTag |
                 'PoS="conjunction"' |
                 'PoS="noun"' nounTag |
                 'PoS="numeral"' |
                 'PoS="particle"' |
                 'PoS="preposition"' |
                 'PoS="pronoun"' pronounTag |
                 'PoS="punctuation"' |
                 'PoS="verb"' verbTag |
                 'PoS="other"' otherTag |

adjectiveTag ::=      genderTag numberTag caseTag

articleTag ::=  'function="def" genderTag numberTag caseTag |
                 'function="prepdef" genderTag numberTag caseTag |
                 'function="indef" genderTag 'number="sg"' caseTag

nounTag ::=           genderTag numberTag caseTag

pronounTag ::=        genderTag? numberTag caseTag |
                      'mode="inflectionless"'

verbTag ::=           tenseTag numberTag voiceTag |
                      infinitiveTag |
                      participleTag

otherTag ::=    'type="abbrev"' |
                 'type="acronym"' |
                 'type="foreign"' |
                 'type="symbol"' |
                 'type="other"'

infinitiveTag ::= 'type="infinitive"' voiceTag

participleTag ::= 'type="participle"'
```

genderTag ::= 'gender="fem"' | 'gender="masc"' | 'gender="neut"'

numberTag ::='number="sg"' | 'number="pl"'

caseTag ::= 'case="nom"' | 'case="gen"' | 'case="acc"' | 'case="voc"'

tenseTag ::= 'tense="present"' | 'tense="past"' | 'tense="future"'

voiceTag ::= 'voice="active"' | 'voice="passive"'

Αναφορές

[Μα05] Π. Μαλακασιώτης, *Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές ενεργητικής μάθησης*. Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[Χρο06] Ι. Χρονάκης, *Επεκτάσεις και περαιτέρω αξιολόγηση συστήματος αναγνώρισης μερών του λόγου για ελληνικά κείμενα*. Πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.

[Mit97] T. Mitchell, *Machine Learning*, Mc-Graw Hill, 1997.

[Ko07] S. Kotsiantis, «Supervised machine learning: a review of classification techniques». *Informatica*, 31:249–268, 2007.

[Be91] B.V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. *IEEE Computer Society Press*, pages 388-397, Los Alamitos, 1991.

[Du76] S. A. Dudani The distance-weighted *k*-nearest neighbour rule. *IEEE Transactions on System, Man, and Cybernetics, volume SMC-6*, σελ. 325-327, 1976.

[Λου05] Γ. Λουκαρέλλι, *Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα*. Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[DaZa03] W. Daelemans, J. Zarvel, K. Van Der Sloot, A. Van Den Bosch. *MBT: Memory-Based Tagger, version 2.0, Reference Guide*, 2003.