ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές ενεργητικής μάθησης* »

Πρόδρομος Μαλακασιώτης

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2005

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

Η διπλωματική αυτή εργασία έχει ως στόχο να μελετήσει τη συμβολή της ενεργητικής μάθησης στο πρόβλημα της αναγνώρισης μερών του λόγου (part of speech tagging), μια περιοχή της επεξεργασίας φυσικής γλώσσας στην οποία οι τεχνικές της ενεργητικής μάθησης δεν έχουν αξιοποιηθεί ως τώρα επαρκώς. Η εργασία επικεντρώθηκε στην αναγνώριση μερών του λόγου σε ελληνικά κείμενα. Χρησιμοποιήθηκαν και προτάθηκαν νέες τεχνικές ενεργητικής μάθησης και μετρήθηκε η αποτελεσματικότητά τους. Πιο συγκεκριμένα, επαναπροσδιορίστηκε η έννοια της ενεργητικής μάθησης, η οποία χωρίστηκε σε τρία στάδια. Σε κάθε στάδιο υλοποιήθηκαν ήδη υπάρχουσες αλλά και νέες τεχνικές και μετρήθηκε πειραματικά η αποτελεσματικότητά τους.

Τα πειράματα περιελάμβαναν χειρωνακτική επισημείωση ενός σώματος ελληνικών ειδησεογραφικών κειμένων συνολικού μεγέθους 20374 λέξεων, διαχωρισμό τους σε σώμα εκπαίδευσης (15300 λέξεις) και σώμα ελέγχου (5074 λέξεις) και εφαρμογή των τριών σταδίων ενεργητικής μάθησης σε συνδυασμό με τον αλγόριθμο κατάταξης των k κοντινότερων γειτόνων. Αν και δεν επιτεύχθηκαν ιδιαίτερα υψηλά ποσοστά ορθότητας (περίπου 80%), τα αποτελέσματα είναι ενθαρρυντικά, καθώς δείχνουν τη θετική συμβολή της ενεργητικής μάθησης, συμπεριλαμβανομένων των νέων τεχνικών που προτείνουμε. Θα πρέπει να σημειωθεί, επίσης, ότι χρησιμοποιήθηκε ένα ιδιαίτερα μεγάλο σύνολο ετικετών για την επισημείωση των λέξεων (112 ετικέτες), κάτι που κάνει το πρόβλημα κατάταξης που είχαμε να αντιμετωπίσουμε ιδιαίτερα δύσκολο.

Τέλος, αξίζει να σημειωθεί ότι οι τεχνικές ενεργητικής μάθησης οι οποίες χρησιμοποιήθηκαν μπορούν εύκολα να εφαρμοστούν και σε άλλα προβλήματα της επεξεργασίας φυσικής γλώσσας.

1 Εισαγωγή

1.1 Αντικείμενο και στόχοι της εργασίας

Με τον όρο «αναγνώριση μερών του λόγου» (part of speech tagging) εννοούμε τη διαδικασία αντιστοίχισης μοναδικής ετικέτας (tag) σε κάθε λέξη ενός συνόλου κειμένων, ώστε η ετικέτα να παριστάνει το μέρος του λόγου στο οποίο ανήκει η λέξη. Η αναγνώριση μερών του λόγου αποτελεί μέρος του ευρύτερου σταδίου της μορφολογικής ανάλυσης κειμένων και χρησιμοποιείται σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας. Είναι μία ενδιαφέρουσα περιοχή τόσο από πρακτικής όσο και από ερευνητικής πλευράς. Ιδιαίτερο ερευνητικό ενδιαφέρον παρουσιάζει η περίπτωση χρήσης τεχνικών μηχανικής μάθησης, ιδιαίτερα ενεργητικής μάθησης, κατά την οποία το ίδιο το σύστημα συμμετέχει στην επιλογή των παραδειγμάτων εκπαίδευσής του. Αξίζει να σημειωθεί ότι η πλειοψηφία των συστημάτων αναγνώρισης μερών του λόγου χρησιμοποιεί ήδη μηχανική μάθηση αλλά οι τεχνικές ενεργητικής μάθησης δεν έχουν ακόμα αξιοποιηθεί επαρκώς στην περιοχή αυτή.

Αν και η δημιουργία ενός πολύ καλού συστήματος αναγνώρισης μερών του λόγου έχει εξέχουσα σημασία, λόγω των πολλών εφαρμογών στις οποίες μπορεί να χρησιμοποιηθεί, ο στόχος της παρούσης εργασίας είναι περισσότερο ερευνητικός. Άλλωστε υπάρχει μία πληθώρα πολύ καλών συστημάτων, κυρίως για τα Αγγλικά, τα οποία ήδη χρησιμοποιούνται. Πιο συγκεκριμένα, η εργασία μελετά το αν και κατά πόσο η ενεργητική μάθηση μπορεί να βελτιώσει τα αποτελέσματα ενός συστήματος αναγνώρισης μερών του λόγου. Η εργασία προτείνει μια ευρύτερη θεώρηση του ρόλου της ενεργητικής μάθησης, όπου ο ρόλος του συστήματος δεν περιορίζεται στον εντοπισμό νέων παραδειγμάτων εκπαίδευσης αλλά περιλαμβάνει και τον εντοπισμό λαθών επισημείωσης στα υπάρχοντα παραδείγματα εκπαίδευσης και τη συμμετοχή του στην ανεύρεση κατάλληλων ιδιοτήτων. Για κάθε ένα από τα τρία αυτά στάδια, η εργασία χρησιμοποιεί υπάρχουσες αλλά και νέες μεθόδους και μελετά πειραματικά την απόδοσή τους.

Αξίζει να σημειωθεί ότι οι μέθοδοι ενεργητικής μάθησης της εργασίας μπορούν να εφαρμοστούν και σε άλλα προβλήματα επεξεργασίας φυσικής γλώσσας. Τέλος, παρ' όλο που η εργασία εστιάζεται στην αναγνώριση μερών του λόγου σε ελληνικά κείμενα, όλες οι τεχνικές που προτείνονται μπορούν εύκολα να εφαρμοστούν και σε άλλες γλώσσες.

1.2 Διάρθρωση της εργασίας

Το υπόλοιπο της εργασίας είναι διαρθρωμένο ως εξής:

Το κεφάλαιο 2 αναφέρεται στη μηχανική μάθηση. Πιο συγκεκριμένα περιέχει εκτεταμένη ανάλυση του αλγορίθμου k-NN, ο οποίος είναι ο αλγόριθμος μηχανικής μάθησης που χρησιμοποιήθηκε

Στο κεφάλαιο 3 περιγράφεται αναλυτικότερα το πρόβλημα της αναγνώρισης μερών του λόγου, ενώ γίνεται και μία επισκόπηση των σημαντικότερων προσεγγίσεων που έχουν προταθεί στην περιοχή.

Το κεφάλαιο 4 περιγράφει τη διαδικασία της ενεργητικής μάθησης και παρουσιάζει τις σημαντικότερες μεθόδους που έχουν προταθεί γι' αυτή, καθώς και τη δική μας προσέγγιση.

Το κεφάλαιο 5 περιγράφει την πειραματική διαδικασία και παρουσιάζει τα αποτελέσματά της.

Τέλος στο κεφάλαιο 6 γίνεται αξιολόγηση της εργασίας και παρουσιάζονται θέματα για πιθανή μελλοντική έρευνα.

1.3 Ευχαριστίες

Αρχικά θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου, κ. Ίωνα Ανδρουτσόπουλο, για τις ουσιαστικές κατευθύνσεις και πολύτιμες συμβουλές που μου έδωσε κατά τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, καθώς και τον κ. Θεόδωρο Καλαμπούκη, που

αποδέχθηκε το ρόλο του δεύτερου αξιολογητή. Επιπλέον, θα ήθελα να ευχαριστήσω το Γιώργο Λουκαρέλλι για τα ειδησεογραφικά κείμενα που μου παρείχε σε μορφή HTML, καθώς και για τις ουσιαστικές συζητήσεις τις οποίες κάναμε πάνω στα θέματα των εργασιών μας. Τέλος, θα ήταν παράλειψη να μην ευχαριστήσω τον Κώστα Στρογγυλό για την εφαρμογή αφαίρεσης ετικετών HTML που μου διέθεσε και για τις πολύτιμες συμβουλές του σε θέματα υλοποίησης και συγγραφής κώδικα.

2 Μηχανική Μάθηση (Machine Learning)

2.1 Εισαγωγή

Η Μηχανική Μάθηση αποτελεί έναν από τους παλαιότερους και σημαντικότερους τομείς έρευνας της Τεχνητής Νοημοσύνης. Στόχος της είναι η δημιουργία συστημάτων που να είναι σε θέση να διδάσκονται από προηγούμενα εμπειρικά δεδομένα, ώστε να εκτελούν την εργασία για την οποία προορίζονται αποτελεσματικότερα. Η διαδικασία εκμάθησης αποτελείται από τα παρακάτω στάδια

- Απόκτηση εμπειρικών δεδομένων από την αλληλεπίδραση με το περιβάλλον.
- Επεξεργασία των δεδομένων, ούτως ώστε να βρεθούν πιθανές γενικεύσεις ή εξειδικεύσεις.
- Χρησιμοποίηση των αποτελεσμάτων της επεξεργασίας και λήψη ανατροφοδότησης από το περιβάλλον, έτσι ώστε να βελτιωθεί περαιτέρω το σύστημα.

2.2 Επιβλεπόμενη Μάθηση (Supervised Learning)

Μία κατηγορία μηχανικής μάθησης είναι η Επιβλεπόμενη Μάθηση. Ένα σύστημα που χρησιμοποιεί επιβλεπόμενη μάθηση αρχικά εκπαιδεύεται σε ένα σύνολο παραδειγμάτων εκπαίδευσης τα οποία συνοδεύονται και από τις κατηγορίες στις οποίες ανήκουν. Για παράδειγμα ένα σύστημα παραγωγής ιατρικών διαγνώσεων θα είχε ως παραδείγματα εκπαίδευσης ιατρικές εξετάσεις συνοδευόμενες από τις ορθές διαγνώσεις. Από την εκπαίδευση προκύπτει ένα μοντέλο των κατηγοριών, το οποίο στη συνέχεια χρησιμοποιείται για να κατατάξει νέες περιπτώσεις των οποίων δεν είναι γνωστή η κατηγορία. Ένα παράδειγμα επιβλεπόμενης μάθησης είναι η μάθηση που χρησιμοποιεί αποθήκευση στη μνήμη (ενότητα 3.2.4). Υπάρχουν πολλοί γνωστοί αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης, όπως ο αλγόριθμος των k κοντινότερων

γειτόνων (k-Nearest Neighbours, k-NN), ο Naïve Bayes, ο ID3 κ.τ.λ. ([Mi97]). Εμείς όμως θα περιοριστούμε στην παρουσίαση των αλγορίθμων που χρησιμοποιούνται στη εργασία.

Ο βασικός τρόπος εκπαίδευσης και αξιολόγησης ενός συστήματος επιβλεπόμενης μάθησης περιγράφεται παρακάτω. Αρχικά πρέπει να οριστούν δύο σύνολα από αντικείμενα: το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Τα αντικείμενα των συνόλων αυτών πρέπει να έχουν καταταγεί χειρωνακτικά σε δύο η περισσότερες κατηγορίες. Στην περίπτωση που εξετάζουμε σε αυτή την εργασία, η τομή των κατηγοριών ανά δύο πρέπει να είναι κενή. Δε μπορεί δηλαδή ένα αντικείμενο να καταταγεί σε περισσότερες από μία κατηγορίες. Επιπλέον πρέπει να οριστεί και ένα σύνολο ιδιοτήτων. Κάθε αντικείμενο των συνόλων εκπαίδευσης και ελέγχου αναπαρίσταται από ένα διάνυσμα, κάθε συντεταγμένη του οποίου αποτελεί την τιμή μίας συγκεκριμένης ιδιότητας. Ο αλγόριθμος εκπαιδεύεται στα αντικείμενα του συνόλου εκπαίδευσης και έτσι δημιουργείται ένα ταξινομητής, η ορθότητα του οποίου αξιολογείται στο σύνολο ελέγχου, συγκρίνοντας τις αποφάσεις του ταξινομητή με τις σωστές κατηγορίες.

2.2.1 Ο αλγόριθμος των k κοντινότερων γειτόνων (k-NN)

2.2.1.1 Η βασική ιδέα

Η βασική ιδέα του k-NN είναι ότι κατά τη διάρκεια της εκπαίδευσης ο αλγόριθμος απλά αποθηκεύει τα διανύσματα των αντικειμένων του συνόλου εκπαίδευσης. Στη συνέχεια για κάθε αντικείμενο του συνόλου ελέγχου υπολογίζεται η απόσταση του διανύσματός του από τα διανύσματα όλων των αντικειμένων του συνόλου εκπαίδευσης. Τέλος επιλέγονται τα k αντικείμενα εκπαίδευσης που έχουν τις μικρότερες αποστάσεις από το εξεταζόμενο αντικείμενο, το οποίο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k γειτόνων.

Σημαντικό ρόλο για την καλύτερη απόδοση του αλγορίθμου παίζει και η επιλογή της τιμής του k. Συνήθως, μία τιμή μεταξύ του 5 και του 10 δίνει πολύ καλά αποτελέσματα για δεδομένα με λίγες διαστάσεις, ενώ μία καλή τεχνική για τον καθορισμό του k είναι η διασταυρωμένη επικύρωση (ενότητα 2.3). Για την εύρεση των k κοντινότερων γειτόνων ενός εξεταζόμενου αντικειμένου πρέπει να ορισθεί κατάλληλα ένα μέτρο απόστασης. Φυσικά ο κάθε χρήστης μπορεί να προσθέσει τα δικά του μέτρα απόστασης, ανάλογα με το είδος του εργασίας που θέλει να επιτελέσει.

Πολύ σημαντικό πλεονέκτημα του k-NN αποτελεί το γεγονός ότι μπορεί να μάθει κάθε είδους συνάρτηση και δεν περιορίζεται, για παράδειγμα, μόνο σε γραμμικούς διαχωριστές. Τα παραπάνω, σε συνδυασμό με την απλότητα του αλγορίθμου, τον καθιστούν έναν από τους πιο σημαντικούς δημοφιλείς αλγορίθμους μηχανικής μάθησης.

2.2.1.2 Μέτρο επικάλυψης (Overlap metric)

Το πιο σύνηθες μέτρο που χρησιμοποιείται για τον προσδιορισμό της απόστασης αντικειμένων είναι το μέτρο επικάλυψης που περιγράφεται από τις σχέσεις 2.1 και 2.2. Στις σχέσεις αυτές $\Delta(\vec{X},\vec{Y})$ είναι η απόσταση μεταξύ των αντικειμένων $\vec{X}$ και $\vec{Y}$, που το καθένα έχει n ιδιότητες, ενώ $\delta(x_i, y_i)$ είναι η απόσταση ανά ιδιότητα. Η απόσταση μεταξύ δύο αντικειμένων είναι απλά το άθροισμα των διαφορών ανάμεσα στις ιδιότητες. Ο αλγόριθμος που χρησιμοποιεί το συγκεκριμένο μέτρο είναι η απλούστερη μορφή του k-NN και είναι γνωστός στους χρήστες του TiMBL (βλ. παράρτημα II) ως IB1 ([AhKi91]).

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

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

2.2.1.3 Τροποποιημένο μέτρο διαφοράς τιμών

Το μέτρο επικάλυψης έχει το μειονέκτημα ότι περιορίζεται σε ακριβές ταίριασμα μεταξύ των τιμών των συμβολικών ιδιοτήτων. Αυτό σημαίνει ότι όλες οι δυνατές τιμές που μπορεί να πάρει μία συμβολική ιδιότητα θεωρείται ότι διαφέρουν το ίδιο μεταξύ τους. Εν τούτοις, υπάρχουν περιπτώσεις όπου αυτό δεν ισχύει. Για το σκοπό αυτό οι Stanfill και Waltz όρισαν ένα μέτρο ([StWa86]) το οποίο βελτιώθηκε περαιτέρω από τους Cost και Salzberg ([CoSa93]). Το μέτρο αυτό ονομάζεται τροποποιημένο μέτρο διαφοράς τιμών (Modified Value Difference Metric (MVDM)) και περιγράφεται από τη σχέση 2.3. Η σχέση αυτή μπορεί να χρησιμοποιηθεί και για αριθμητικές ιδιότητες, αλλά σε πολλές υλοποιήσεις (και στο TiMBL) χρησιμοποιείται για αυτές το μέτρο επικάλυψης (σχέση 2.2)

$$\delta(v_1, v_2) = \sum_{i=1}^{m} | P(C_i \mid v_1) - P(C_i \mid v_2) | \quad (2.3)$$

Στην ουσία το μέτρο αυτό παρέχει έναν τρόπο προσδιορισμού της ομοιότητας των τιμών μίας ιδιότητας εξετάζοντας τη συνύπαρξη των τιμών αυτών με τις κατηγορίες. Δηλαδή προσπαθεί να υπολογίσει την επίπτωση που έχει στην πρόβλεψη της κατηγορίας το γεγονός ότι το ένα αντικείμενο έχει τιμή για την ιδιότητα v_1 ενώ το άλλο v_2.

Όταν χρησιμοποιείται το μέτρο επικάλυψης της προηγούμενης ενότητας με συμβολικές (μη αριθμητικές) ιδιότητες, η απόσταση μετριέται ως το άθροισμα των ιδιοτήτων στις οποίες τα αντικείμενα έχουν διαφορετικές τιμές. Αυτό συχνά κάνει πολλά παραδείγματα εκπαίδευσης να φαίνεται ότι

ισαπέχουν από το υπό κατάταξη αντικείμενο. Εντούτοις, σε κάποια από αυτά τα παραδείγματα εκπαίδευσης οι διαφορές στις τιμές των ιδιοτήτων από το υπό κατάταξη αντικείμενο μπορεί να είναι πιο σημαντικές από ό,τι σε άλλα. Με το MVDM, οι αποστάσεις αντανακλούν καλύτερα τις ουσιώδεις διαφορές μεταξύ των αντικειμένων. Έτσι είναι πιο δύσκολο να βρεθούν πολλά παραδείγματα εκπαίδευσης στην ίδια απόσταση από το υπό κατάταξη αντικείμενο, κάτι που κάνει ευκολότερη την επιλογή των k κοντινότερων γειτόνων.

2.2.1.4 Πληροφοριακό κέρδος

Το μέτρο απόστασης της σχέσης 2.1 δίνει ίση βαρύτητα σε όλες τις ιδιότητες. Αυτή η επιλογή είναι λογική αν όλες οι ιδιότητες είναι εξίσου σημαντικές. Διαφορετικά μπορούμε να ορίσουμε βάρη για τις ιδιότητες, τα οποία θα υποδηλώνουν το πόσο χρήσιμες είναι οι διάφορες ιδιότητες για την πρόβλεψη της κατηγορίας στην οποία πρέπει να καταταγεί ένα νέο αντικείμενο. Η Θεωρία Πληροφορίας μας παρέχει ένα τέτοιου είδους εργαλείο, το πληροφοριακό κέρδος ([Qu86]).

Το πληροφοριακό κέρδος (information gain) εξετάζει κάθε ιδιότητα ξεχωριστά και μετράει πόση πληροφορία συνεισφέρει στη γνώση τη σωστής κατηγορίας. Το πληροφοριακό κέρδος της i-οστής ιδιότητας μετριέται υπολογίζοντας την αναμενόμενη μείωση της αβεβαιότητας για τη σωστή κατηγορία που προκαλεί η γνώση της τιμής της ιδιότητας (σχέση 2.4).

$$w_i = H(C) - \sum_{v \in V_i} P(v) \times H(C \mid v) \tag{2.4}$$

Στη σχέση 2.4 με C συμβολίζουμε το σύνολο των κατηγοριών, με V_i το σύνολο των τιμών της i-οστής ιδιότητας, ενώ $H(C)$ είναι η εντροπία των κατηγοριών και $H(C \mid v)$ η εντροπία των κατηγοριών αν η τιμή της ιδιότητας είναι v (σχέσεις 2.5, 2.6).

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c) \tag{2.5}$$

$$H(C \mid v) = -\sum_{c \in C} P(c \mid v) \log_2 P(c \mid v) \tag{2.6}$$

Οι πιθανότητες υπολογίζονται από τις σχετικές συχνότητες στο σύνολο εκπαίδευσης.

Για ιδιότητες με αριθμητικές τιμές, πρέπει να γίνει ένα ενδιάμεσο βήμα καθώς είναι δύσκολο να εκτιμηθούν οι πιθανότητες για όλες τις δυνατές αριθμητικές τιμές. Για κάθε μία αριθμητική ιδιότητα, τα παραδείγματα εκπαίδευσης τοποθετούνται στον άξονα των πραγματικών αριθμών, σύμφωνα με τις τιμές που έχουν στη συγκεκριμένη ιδιότητα. Στη συνέχεια ο άξονας των πραγματικών αριθμών διαχωρίζεται σε διαστήματα (προεπιλεγμένη τιμή στο TiMBL, 20) κάθε ένα από τα οποία περιέχει τον ίδιο αριθμό παραδειγμάτων εκπαίδευσης (προεπιλεγμένη τιμή στο TiMBL, 1/20 του συνολικού αριθμού των παραδειγμάτων εκπαίδευσης). Στη συνέχεια, τα παραδείγματα εκπαίδευσης σε κάθε ένα από αυτά τα διαστήματα χρησιμοποιούνται στον υπολογισμό του πληροφοριακού κέρδους σαν να έχουν όλα την ίδια τιμή. Αυτός ο διαχωρισμός είναι μόνο προσωρινός και δε χρησιμοποιείται στον υπολογισμό του μέτρου απόστασης.

Ένα από τα μειονεκτήματα του πληροφοριακού κέρδους είναι ότι έχει την τάση να υπερεκτιμά τη σημαντικότητα των ιδιοτήτων που έχουν μεγάλο πλήθος τιμών. Τέτοιου είδους ιδιότητες έχουν πολύ μεγάλο πληροφοριακό κέρδος αλλά δεν προσφέρουν καμία γενίκευση στα νέα αντικείμενα. Έστω για παράδειγμα ότι σε ένα σύστημα εξαγωγής ιατρικών διαγνώσεων έχουμε ως ιδιότητα τον κωδικό του ασθενή. Η τιμή της ιδιότητας αυτής είναι ξεχωριστή για κάθε παράδειγμα εκπαίδευσης και συνεπώς έχει μεγάλο πληροφοριακό κέρδος αφού φαίνεται να προβλέπει πλήρως τη διάγνωση. Στην ουσία όμως ο κωδικός του ασθενή δεν παρέχει καμία σημαντική πληροφορία σε μελλοντικές περιπτώσεις. Το παραπάνω πρόβλημα εμφανίζεται κυρίως όταν χρησιμοποιούμε ιδιότητες που χαρακτηρίζουν μονοσήμαντα τα αντικείμενα

εκπαίδευσης (ιδιότητες κλειδιά). Για την αντιμετώπιση τέτοιων περιπτώσεων ο Quinlan ([Qu93]) εισήγαγε μία κανονικοποιημένη έκδοση του πληροφοριακού κέρδους, η οποία ονομάζεται αναλογία κέρδους (Gain Ratio), και υπολογίζεται ως ο λόγος του πληροφοριακού κέρδους διά την εντροπία των τιμών της ιδιότητας ($si(i)$, σχέσεις 2.6, 2.7).

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C \mid v)}{si(i)} \tag{2.6}$$

$$si(i) = -\sum_{v \in V_i} P(v) \log_2 P(v) \tag{2.7}$$

Οι τιμές που προκύπτουν από τον υπολογισμό είτε του πληροφοριακού κέρδους είτε της αναλογίας κέρδους μπορούν να χρησιμοποιηθούν ως βάρη κατά τον υπολογισμό του μέτρου απόστασης, όπως φαίνεται στη σχέση 2.8. Ο αλγόριθμος k-NN που χρησιμοποιεί αυτό το μέτρο απόστασης ονομάζεται στο TiMBL IB1-IG ([DaBo92]).

$$\Delta(\vec{X}, \vec{Y}) = \sum_{i=1}^{n} w_i \cdot \delta(x_i, y_i) \tag{2.8}$$

Η δυνατότητα του αυτόματου προσδιορισμού των βαρών των ιδιοτήτων μάς επιτρέπει να προσθέτουμε στο σύνολο ιδιοτήτων ιδιότητες που δεν είμαστε βέβαιοι για τη χρησιμότητά τους, καθώς το πληροφοριακό κέρδος θα μετριάσει πολύ τη συνεισφορά τους στη συνολική απόσταση αν δεν είναι χρήσιμες. Εν τούτοις, αυτή δεν είναι απαραίτητα η ενδεδειγμένη στρατηγική, καθώς αν έχουμε πολλές «περιττές» ιδιότητες επιβαρύνονται άσκοπα οι υπολογισμοί των αποστάσεων που απαιτούνται κατά την κατάταξη νέων αντικειμένων.

2.2.1.5 Ζυγισμένη ψήφος με βάση την απόσταση (Distance-weighted class voting)

Η πιο συνηθισμένη μέθοδος ψηφοφορίας των k κοντινότερων γειτόνων για να αποφασίσουν την κατηγορία ενός νέου αντικειμένου είναι η μέθοδος

της πλειοψηφίας, κατά την οποία η ψήφος κάθε γείτονα έχει την ίδια βαρύτητα. Το νέο αντικείμενο κατατάσσεται στην κατηγορία που θα πάρει τις περισσότερες ψήφους.

Μία επέκταση της παραπάνω διαδικασίας ψηφοφορίας είναι η ψηφοφορία κατά την οποία η ψήφος κάθε γείτονα έχει διαφορετική βαρύτητα ανάλογα με το πόσο απέχει ο γείτονας αυτός από το νέο αντικείμενο. Αυτό είναι λογικό καθώς οι κοντινοί γείτονες μοιάζουν περισσότερο με το νέο αντικείμενο από ό,τι οι μακρινοί. Ένα μέτρο βαρύτητας που εξυπηρετεί το σκοπό αυτό είναι το «βάρος αντίστροφης απόστασης» (inverse distance weight) που προτάθηκε από τον Dudani ([Du76]) και περιγράφεται στη σχέση 2.9. Στον παρονομαστή της σχέσης αυτής συνήθως προστίθεται μία σταθερά για να αποφευχθεί η διαίρεση με το μηδέν ([We94]).

$$w_i = \frac{1}{d_j}, \text{αν } d_j \neq 0 \tag{2.9}$$

2.2.1.6 Αντιμετώπιση ισοπαλιών (tie breaking)

Όταν χρησιμοποιείται η μέθοδος της πλειοψηφίας, είναι πολύ πιθανό να υπάρξουν ισοπαλίες μεταξύ δύο ή περισσοτέρων κατηγοριών. Το φαινόμενο αυτό είναι εντονότερο όταν δε χρησιμοποιούνται ζυγισμένες ψήφοι. Για παράδειγμα αν έχουμε ένα σύνολο δέκα κοντινότερων γειτόνων, μπορεί πέντε να ψηφίσουν για την κατηγορία Α και πέντε για την κατηγορία Β. Μία μέθοδος, η οποία χρησιμοποιείται στο TiMBL, για την αντιμετώπιση του προβλήματος είναι η ακόλουθη. Αρχικά, αυξάνεται το k κατά ένα και διενεργείται νέα ψηφοφορία από το καινούριο σύνολο κοντινότερων γειτόνων. Αν η ισοπαλία παραμένει, επιλέγεται από τις υποψήφιες κατηγορίες αυτή με τα περισσότερες εμφανίσεις στο σύνολο εκπαίδευσης. Αν υπάρχουν περισσότερες από μία τέτοιες κατηγορίες, τότε επιλέγεται τυχαία μία από αυτές. Εναλλακτικά, η αντιμετώπιση ισοπαλιών μπορεί να γίνει με τυχαία επιλογή από τις ισόπαλες κατηγορίες, φροντίζοντας, όμως, κάθε κατηγορία να έχει πιθανότητα επιλογής ανάλογη της συχνότητάς της στο σύνολο

εκπαίδευσης. Τέλος στην περίπτωση που έχουμε μόνο δύο κατηγορίες μπορούμε απλά να επιλέξουμε περιττή τιμή για το k, οπότε δε θα υπάρξουν ποτέ ισοπαλίες.

2.3 Διασταυρωμένη επικύρωση

Ένας τρόπος για να ελέγξουμε την αποτελεσματικότητα ενός συστήματος που χρησιμοποιεί μηχανική μάθηση είναι η διασταυρωμένη επικύρωση (cross validation). Η διαδικασία αυτή μπορεί να χρησιμοποιηθεί για οποιονδήποτε αλγόριθμο εκπαίδευσης και η εκτέλεσή της γίνεται ως εξής. Το σύνολο των δεδομένων διαχωρίζεται σε k ισάριθμα τμήματα και ο αλγόριθμος εκπαιδεύεται σε k-1 από αυτά. Η αποτελεσματικότητα του ταξινομητή που δημιουργείται ελέγχεται με το εναπομείναν τμήμα των δεδομένων. Συνολικά διενεργούνται k πειράματα αφήνοντας κάθε φορά διαφορετικό τμήμα έξω από τα δεδομένα εκπαίδευσης. Στο τέλος υπολογίζεται ο μέσος όρος των αποτελεσμάτων που προέκυψαν από τα k πειράματα. Οι πιο δημοφιλείς τιμές για το k είναι το 5 και το 10 οπότε έχουμε αντίστοιχα πενταπλή και δεκαπλή διασταυρωμένη επικύρωση (5-fold, 10-fold cross validation).

3 Αναγνώριση μερών του λόγου

3.1 Εισαγωγή

Ένα από τα σημαντικότερα στάδια της μορφολογικής ανάλυσης κειμένων, είναι η «Αναγνώριση Μερών του Λόγου» (part of speech tagging). Πιο συγκεκριμένα, δεδομένου ενός συνόλου κειμένων και ενός συνόλου ετικετών (κατηγοριών) που παριστάνουν τα μέρη του λόγου (ουσιαστικό, ρήμα, άρθρο, κλπ.), ένα σύστημα αναγνώρισης μερών του λόγου πρέπει να αντιστοιχίσει κάθε λέξη του συνόλου κειμένων σε μία και μόνο κατηγορία του συνόλου ετικετών. Σε πολλές περιπτώσεις οι ετικέτες δεν αντιστοιχούν απλά στα μέρη του λόγου αλλά υποδηλώνουν και διάφορα μορφολογικά χαρακτηριστικά της λέξης, όπως για παράδειγμα το γένος, τον αριθμό, το χρόνο, τη φωνή, κ.τ.λ. (π.χ. αρσενικό ουσιαστικό ενικού αριθμού, ρήμα στον ενεστώτα της ενεργητικής φωνής). Στην περίπτωση αυτή, το σύστημα μπορεί να θεωρηθεί ότι δεν κάνει απλά αναγνώριση μερών του λόγου, αλλά γενικότερα μορφολογική ανάλυση. Παρ' όλα, αυτά συνηθίζεται να χρησιμοποιείται ο όρος «Αναγνώριση Μερών του Λόγου» και για αυτά τα συστήματα και θα ακολουθήσουμε την ίδια πρακτική.

Η σημαντικότητα της συγκεκριμένης περιοχής έγκειται στο γεγονός ότι η μορφολογική πληροφορία που αποδίδεται σε κάθε λέξη ενός κειμένου αποτελεί τη βάση για την περαιτέρω επεξεργασία του. Συνεπώς ένα τέτοιο σύστημα, μορφολογικής ανάλυσης κειμένων, μπορεί να χρησιμοποιηθεί ως τμήμα διαφόρων άλλων συστημάτων επεξεργασίας φυσικής γλώσσας, όπως συντακτικοί αναλυτές, διορθωτές κειμένων, συστήματα αναγνώρισης φωνής κ.α. Εκτός όμως από πρακτικό ενδιαφέρον, η αναγνώριση μερών του λόγου έχει και ερευνητικό ενδιαφέρον, καθώς ενσωματώνοντας πολλές μορφολογικές πληροφορίες στις ετικέτες δημιουργείται ένα πολύ μεγάλο πλήθος κατηγοριών. Έτσι η μορφολογική ανάλυση κειμένων καθίσταται ένα δύσκολο πρόβλημα κατηγοριοποίησης.

Ένα σημαντικό ερώτημα που θα μπορούσε ίσως να προκύψει είναι το γιατί δε χρησιμοποιείται ένα ηλεκτρονικό λεξικό, της εκάστοτε γλώσσας, για τη μορφολογική ανάλυση των λέξεων. Η απάντηση είναι απλή και έγκειται στους περιορισμούς που υπεισέρχονται σε μία τέτοια λύση. Καταρχάς, η κατασκευή ενός ηλεκτρονικού λεξικού είναι μία χρονοβόρα και ιδιαίτερα ακριβή διαδικασία, με συνέπεια τα περισσότερα ηλεκτρονικά λεξικά να μην είναι ελεύθερα διαθέσιμα στο κοινό. Επιπλέον, ένα λεξικό, περιέχοντας πεπερασμένο πλήθος λέξεων, είναι αδύνατο να καλύψει όλες τις πιθανές λέξεις που μπορεί να εμφανιστούν, ιδιαίτερα κύρια ονόματα και νέους τεχνικούς όρους. Τέλος, ένα σύστημα που συμβουλεύεται απλά ένα λεξικό χωρίς να λαμβάνει υπόψη του τα συμφραζόμενα της κάθε λέξης, σε πολλές περιπτώσεις δε μπορεί να αποφασίσει για την ετικέτα που θα αποδώσει σε μία λέξη (π.χ. η λέξη «διατάξεις» εμφανίζεται και ως ρήμα και ως ουσιαστικό).

3.2 Προηγούμενες προσεγγίσεις στην αναγνώριση μερών του λόγου

Όπως έχει ήδη αναφερθεί η αναγνώριση μερών του λόγου αποτελεί απαραίτητο στάδιο σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας. Λογικό είναι λοιπόν να έχουν παρουσιαστεί κατά καιρούς διάφορες προσεγγίσεις για την επίλυση του συγκεκριμένου προβλήματος. Παρακάτω θα παρουσιάσουμε τις σημαντικότερες από αυτές.

3.2.1 Μάθηση στηριζόμενη σε κανόνες μετασχηματισμού οδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)

Πρόκειται για μια γενική μέθοδο μηχανικής μάθησης, η οποία όμως έχει εφαρμοστεί με μεγάλη επιτυχία σε συστήματα αναγνώρισης μερών του λόγου [Br92], [Br93], [Br95b]. Η μέθοδος βασίζεται στην εκμάθηση κανόνων μετασχηματισμού, καθένας από τους οποίους τροποποιεί, στην περίπτωση της αναγνώρισης μερών του λόγου, την ετικέτα της λέξης στην οποία εφαρμόζεται,

εφόσον ικανοποιούνται οι περιορισμοί του. Το σύστημα μαθαίνει τέτοιους κανόνες χρησιμοποιώντας ένα σώμα κειμένων καθώς και ένα σύνολο από μορφότυπους (πρότυπα) κανόνων που έχουν προκαθοριστεί.

Πιο συγκεκριμένα, υπάρχουν δύο ειδών κανόνες: λεκτικοί κανόνες (lexical rules), οι οποίοι προβλέπουν την πιθανότερη ετικέτα για τις άγνωστες λέξεις εξετάζοντας μόνο τις ίδιες τις λέξεις (καταλήξεις, προθέματα, κ.τ.λ.), και κανόνες συμφραζομένων (contextual rules), οι οποίοι εξετάζουν τα συμφραζόμενα των λέξεων.

Οι κανόνες συμφραζομένων είναι της μορφής: *Η ετικέτα$_i$ αντικαθίσταται από την ετικέτα$_j$ αν P*, το οποίο σημαίνει ότι η αρχική ετικέτα που έχει αποδοθεί σε μία λέξη (ετικέτα$_i$) αντικαθίσταται από μία νέα ετικέτα (ετικέτα$_j$) αν τα συμφραζόμενα είναι P. Τα συμφραζόμενα αποτελούνται από την υπό εξέταση λέξη και την ετικέτα που της έχει αποδοθεί καθώς και από τις δύο προηγούμενες λέξεις μαζί με τις ετικέτες που τους έχουν αποδοθεί.

Η εκπαίδευση διενεργείται σε γενικές γραμμές ως εξής. Αρχικά, το σύστημα αναθέτει σε κάθε λέξη του σώματος εκπαίδευσης μία ετικέτα. Στην απλούστερη περίπτωση, ανατίθεται σε κάθε λέξη η συχνότερη ετικέτα της γλώσσας (π.χ. ανατίθεται σε όλες τις λέξεις η ετικέτα του ουσιαστικού). Κατόπιν μετριέται ο αριθμός ορθών ετικετών που προέκυψαν από την παραπάνω διαδικασία, συγκρίνοντάς τις με τις ορθές ετικέτες που έχουν προστεθεί χειρωνακτικά στο σώμα εκπαίδευσης. Στη συνέχεια το σύστημα δοκιμάζει κάθε έναν από τους δυνατούς κανόνες συμφραζομένων ξεχωριστά σε ολόκληρο το σώμα εκπαίδευσης και για κάθε έναν κανόνα μετρά τον αριθμό ορθών ετικετών στον οποίο οδηγεί η εφαρμογή του κανόνα. Ο κανόνας που οδηγεί στη μεγαλύτερη αύξηση του αριθμού των ορθών ετικετών υιοθετείται, δηλαδή οι ετικέτες που έχει αναθέσει το σύστημα στις λέξεις του σώματος εκπαίδευσης θεωρείται πλέον ότι είναι οι αρχικές με τις διορθώσεις του κανόνα που υιοθετήθηκε. Ο κανόνας που υιοθετήθηκε τοποθετείται επίσης στο τέλος μίας λίστας που περιέχει τους υιοθετημένους κανόνες. Η ίδια

διαδικασία επαναλαμβάνεται, δηλαδή υιοθετείται σε κάθε επανάληψη ένας νέος κανόνας, που προστίθεται στη λίστα, μέχρι το σημείο όπου δεν είναι δυνατόν να αυξηθεί ο αριθμός των ορθών ετικετών με την εφαρμογή κανενός κανόνα. Έτσι, δημιουργείται μία λίστα με κανόνες οι οποίοι είναι ταξινομημένοι με τη σειρά που εφαρμόστηκαν στην παραπάνω διαδικασία.

Αφού ολοκληρωθεί η διαδικασία εκπαίδευσης, το σύστημα μπορεί να χρησιμοποιηθεί για να προστεθούν ετικέτες μερών του λόγου σε νέα, μη χειρωνακτικά επισημειωμένα κείμενα.Αρχικά, κάθε λέξη επισημειώνεται με μία ετικέτα όπως και στη διαδικασία εκπαίδευσης. Αν η λέξη δεν υπάρχει στο σώμα εκπαίδευσης (άγνωστη) τότε εφαρμόζονται σε αυτήν οι λεκτικοί κανόνες. Στη συνέχεια εφαρμόζονται οι κανόνες συμφραζομένων που έμαθε το σύστημα με τη σειρά που εφαρμόστηκαν κατά τη διαδικασία εκπαίδευσης.

3.2.2 Κρυφά μοντέλα Markov

Από στατιστικής πλευράς η μορφολογική ανάλυση κειμένων μπορεί να οριστεί ως πρόβλημα μεγιστοποίησης. Έστω $\beta = \{c_1, c_2, ..., c_N\}$ το σύνολο των ετικετών και $v = \{w_1, w_2, ..., w_m\}$ το σύνολο των λέξεων που είναι δυνατόν να εμφανιστούν στα κείμενα. Για μια ακολουθία λέξεων μήκους L, $W = w_1, w_2, ..., w_L$, το ζητούμενο αποτέλεσμα είναι η εύρεση μίας ακολουθίας ετικετών $\hat{C}$ μέγιστης πιθανότητας, δηλαδή:

$$\hat{C} = \arg\max_C P(C \mid W) = \arg\max_C \left(\frac{P(C) \cdot P(W \mid C)}{P(W)} \right), C \in \beta^L \qquad (3.1)$$

Επειδή η πιθανότητα $P(W)$ είναι σταθερή μπορεί να παραληφθεί και το πρόβλημα ανάγεται στο πρόβλημα μεγιστοποίησης του αριθμητή της σχέσης 3.1. Στη σχέση αυτή το γλωσσικό μοντέλο, $P(C)$, αναπαριστά τις πιθανές ακολουθίες ετικετών, ενώ οι λεκτικές πιθανότητες , $P(W \mid C)$, αναπαριστούν τη σχέση ανάμεσα στο λεξιλόγιο και τις ετικέτες.

Για την επίλυση της σχέσης 3.1 πρέπει να γίνουν οι υποθέσεις Markov που περιγράφονται παρακάτω ούτως ώστε να απλοποιηθεί το πρόβλημα.

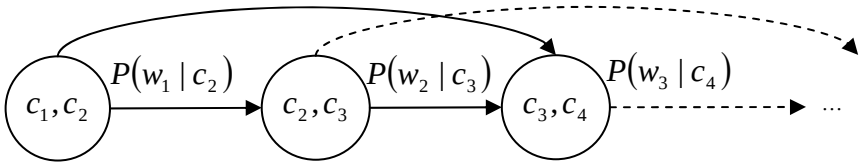
- Έστω $X = (X_1,..., X_L)$ μία ακολουθία τυχαίων μεταβλητών οι οποίες παίρνουν τιμές από κάποιο πεπερασμένο σύνολο $S = \{s_1,..., s_N\}$, το οποίο ονομάζεται σύνολο καταστάσεων. Αν η X έχει τις δύο ακόλουθες ιδιότητες είναι μία αλυσίδα Markov.
  - Περιορισμένος Ορίζοντας (Limited Horizon): $P(X_{l+1} = s_k \mid X_1,..., X_l) = P(X_{l+1} = s_k \mid X_l)$, δηλαδή η τιμή μιας τυχαίας μεταβλητής (διαισθητικά, η κατάσταση στην οποία βρισκόμαστε κατά το χρόνο *l+1*) εξαρτάται μόνο από την τιμή της προηγούμενης τυχαίας μεταβλητής (την κατάσταση στην οποία βρισκόμασταν κατά το χρόνο *l*). Στη γενικότερη περίπτωση η τιμή μίας τυχαίας μεταβλητής μπορεί να εξαρτάται από τις τιμές των m προηγούμενων, οπότε έχουμε μοντέλο Markov m τάξης
  - Ανεξαρτησία χρόνου – Στασιμότητα (Time Invariant – Stationary): $P(X_{l+1} = s_k \mid X_l) = P(X_2 = s_k \mid X_1)$, δηλαδή η παραπάνω εξάρτηση παραμένει η ίδια με την πάροδο του χρόνου.

Θεωρώντας ότι οι ακολουθίες των ετικετών ικανοποιούν τις υποθέσεις Markov δεύτερης τάξης και θεωρώντας ότι η πιθανότητα εμφάνισης μιας λέξης στη θέση i της ακολουθίας W εξαρτάται μόνο από την ετικέτα c_i της αντίστοιχης θέσης το πρόβλημα ανάγεται στο παρακάτω πρόβλημα μεγιστοποίησης:

$$\hat{C} = \arg\max_{c_1 \ldots c_L} \left(\prod_{1 \ldots L} P(c_i \mid c_{i-1}, c_{i-2}) \cdot P(w_i \mid c_i) \right) \tag{3.2}$$

Οι παράγοντες τις παραπάνω σχέσης μπορούν να αναπαρασταθούν ως ένα «Κρυφό μοντέλο Markov» (Hidden Markov model, HMM) ως εξής (εικόνα 3-1):

- Κάθε κόμβος του μοντέλου αποτελείται από ζεύγη ετικετών οι οποίες είναι οι δύο τελευταίες ετικέτες που έχουμε συναντήσει στο κείμενο.
- Οι πιθανότητες συμφραζομένων (contextual probabilities), $P(c_i \mid c_{i-1}, c_{i-2})$, αντιστοιχούν στις πιθανότητες μετάβασης ανάμεσα στους κόμβους (c_{i-2}, c_{i-1}) και (c_{i-1}, c_i).
- Κατά τη μετάβαση από έναν κόμβο σε έναν άλλο εμφανίζεται στην έξοδο μία λέξη w_i με πιθανότητα $P(w_i \mid c_i)$ που εξαρτάται μόνο από την τελευταία ετικέτα που συναντήσαμε και η οποία προκάλεσε τη μετάβαση. Οι πιθανότητες $P(w_i \mid c_i)$ είναι ουσιαστικά οι λεκτικές πιθανότητες.



Εικόνα 3-1

Θεωρούμε ότι η ακολουθία λέξεων W, στις λέξεις της οποίας έχουμε να αναθέσουμε ετικέτες, παράγεται από τη διάσχιση ενός μονοπατιού του κρυφού μοντέλου Markov και το πρόβλημα έγκειται στον εντοπισμό του πιθανότερου μονοπατιού που είναι δυνατόν να έχει παραγάγει την ακολουθία W. Οι ζητούμενες ετικέτες $c_1, ..., c_L$ της σχέσης (3.2) είναι οι ετικέτες που συναντούμε

κατά μήκος του πιθανότερου μονοπατιού. Η εύρεση του πιθανότερου μονοπατιού, δηλαδή τις πιθανότερης ακολουθίας ετικετών για μια ακολουθία λέξεων W, μπορεί να γίνει με δυναμικό προγραμματισμό χρησιμοποιώντας τον αλγόριθμο Viterbi ([Vi67]).

Έχουν προταθεί διάφορες βελτιώσεις της παραπάνω γενικής ιδέας. Ορισμένες από αυτές επιχειρούν να λάβουν υπόψη τους και τις ετικέτες των λέξεων που ακολουθούν την w_i, αντίθετα από τη σχέση (3.2) που λαμβάνει υπόψη της μόνο τις ετικέτες των προηγούμενων δύο λέξεων ([BaMo04]). Επίσης, έχουν προταθεί και τεχνικές με χαλάρωμα των υποθέσεων Markov ([LeTs00]), καθώς και τεχνικές κατά τις οποίες αναπτύσσονται μοντέλα Markov με μεταβλητή μνήμη και συνεπώς καλύτερη αποθηκευτική ικανότητα ([KiRi03]).

3.2.3 Συστήματα μέγιστης εντροπίας

Μια άλλη στατιστική προσέγγιση χρησιμοποιεί την αρχή της Μέγιστης Εντροπίας (Maximum Entropy), η οποία είχε χρησιμοποιηθεί από τον Rosenfeld ([Ro96]) σε προβλήματα μοντελοποίησης της γλώσσας καθώς και σε συστήματα αναγνώρισης προφορικού λόγου. Για την εκτίμηση της πιθανότητας μία ακολουθία ετικετών, $t_1,...,t_n$, να αντιστοιχεί σε μία ακολουθία λέξεων, $w_1,...,w_n$, χρησιμοποιείται η σχέση 3.3. Στη σχέση αυτή με h_i συμβολίζονται τα συμφραζόμενα της λέξης w_i.

$$p(t_1...t_n \mid w_1...w_n) = \prod_{i=1}^{n} p(t_i \mid t_1...t_{i-1}, w_1...w_n) \approx \prod_{i=1}^{n} p(t_i \mid h_i) \tag{3.3}$$

$$H(p) = -\sum_{i=1}^{n} p(t_i \mid h_i) \cdot \log p(t_i \mid h_i) \tag{3.4}$$

Η βασική ιδέα του μοντέλου μέγιστης εντροπίας είναι ότι επιλέγεται η κατανομή πιθανότητας p η οποία έχει τη μέγιστη εντροπία (σχέση 3.4) από όλες τις κατανομές που ικανοποιούν συγκεκριμένους περιορισμούς. Όπως

φαίνεται από τους τύπους όταν η εντροπία της σχέσης 3.4 μεγιστοποιείται, τότε μεγιστοποιείται και η πιθανότητα $p(t_1...t_n \mid w_1...w_n)$ της σχέσης 3.3, ως γινόμενο των πιθανοτήτων $p(t_i \mid h_i)$ που συμμετέχουν στην εντροπία. Οι περιορισμοί προκύπτουν από στατιστικά στοιχεία που εξάγονται από τα δεδομένα εκπαίδευσης. Τα στατιστικά αυτά στοιχεία εκφράζονται ως οι αναμενόμενες τιμές κατάλληλων συναρτήσεων, που αναπαριστούν ιδιότητες των λέξεων οι οποίες εξαρτώνται από τα συμφραζόμενα και τις ετικέτες. Για παράδειγμα αν θέλουμε να περιορίσουμε το μοντέλο έτσι ώστε να επισημειώνει τη λέξη «διατάξεις» ως ρήμα ή ως ουσιαστικό με την ίδια συχνότητα που αυτή έχει επισημειωθεί ως ρήμα ή ως ουσιαστικό στο σώμα εκπαίδευσης, πρέπει να ορίσουμε τις παρακάτω ιδιότητες:

1. $f_1(h,t)=1$ ανν w_i = διατάξεις και t = ρήμα
2. $f_2(h,t)=1$ ανν w_i = διατάξεις και t = ουσιαστικό

Έτσι αν στο σώμα εκπαίδευσης η λέξη «διατάξεις» παρατηρείται επισημειωμένη ως ρήμα ή ως ουσιαστικό με μία αναλογία 7/3 οι παραπάνω ιδιότητες θα ωθήσουν το μοντέλο να επισημειώνει τη λέξη «διατάξεις» στα νέα κείμενα (του σώματος ελέγχου) ως ουσιαστικό ή ως ρήμα με την ίδια αναλογία.

Μία γνωστή προσπάθεια εφαρμογής του μοντέλου μέγιστης εντροπίας στην αναγνώριση μερών του λόγου έγινε από τον Ratnaparkhi ([Ra96]). Στη συγκεκριμένη προσπάθεια, οι ιδιότητες λαμβάνουν υπόψη την τρέχουσα λέξη, τις δύο προηγούμενες και τις δύο επόμενες λέξεις, καθώς και τις ετικέτες των δύο προηγούμενων λέξεων. Επιπλέον αν η υπό εξέταση λέξη είναι μία σπάνια ή μη συνηθισμένη λέξη, στις ιδιότητες περιλαμβάνεται και μορφολογική πληροφορία, όπως για παράδειγμα καταλήξεις, προθέματα, τον αν η λέξη περιέχει αριθμούς, κεφαλαία γράμματα και ειδικά σύμβολα.

3.2.4 Εκμάθηση με αποθήκευση στη μνήμη (memory-based learning)

Η μάθηση με αποθήκευση στη μνήμη είναι ένα είδος επιβλεπόμενης μάθησης που στηρίζεται στην εξαγωγή συμπερασμάτων βάσει της ομοιότητας που παρουσιάζουν νέες περιπτώσεις με γνωστές περιπτώσεις του παρελθόντος ([DaZa96]). Ο γνωστότερος αλγόριθμος αυτού του είδους είναι ο k-NN (ενότητα 2.2.1). Ένα γνωστό σύστημα αναγνώρισης μερών του λόγου που βασίζεται σε μάθηση με αποθήκευση στη μνήμη είναι το MBT ([DaZa96]), το οποίο χρησιμοποιεί το πακέτο TiMBL. Το TiMBL περιέχει υλοποιήσεις διαφόρων παραλλαγών του k-NN και χρησιμοποιήθηκε και στο σύστημα της παρούσας εργασίας. Το MBT χρησιμοποιεί ως προεπιλογή τον αλγόριθμο IB1-IG του TiMBL (ενότητα 2.2.1.4), ενώ παράλληλα παρέχει τη δυνατότητα επιλογής και των άλλων αλγορίθμων που υλοποιούνται από το TiMBL. Στο MBT οι προς επισημείωση λέξεις των νέων κειμένων χωρίζονται σε γνωστές (λέξεις που έχουμε συναντήσει στη διάρκεια της εκπαίδευσης) και άγνωστες. Για τις γνωστές λέξεις, οι ιδιότητες αφορούν μόνο τα συμφραζόμενα, ενώ για τις άγνωστες χρησιμοποιούνται και ιδιότητες που έχουν σχέση με τη μορφολογία της λέξης.

3.2.5 Άλλες προσεγγίσεις

Έκτος από τις παραπάνω προσεγγίσεις έχουν προταθεί και διάφορες σύνθετες τεχνικές που έχουν ως σκοπό να συνδυάσουν τα πλεονεκτήματα των παραπάνω προσεγγίσεων. Έτσι οι Clark, Curran και Osborne ([ClCu03]) προτείνουν ένα σύστημα συνεκπαίδευσης, στο οποίο χρησιμοποιούν δύο υπάρχοντα συστήματα αναγνώρισης μερών του λόγου, τα οποία επανεκπαιδεύουν επαναληπτικά, το ένα με νέα κείμενα που έχουν επισημειωθεί από το άλλο. Επιπλέον, οι Nakagawa, Kudo και Matsumoto ([NaKu02]) προτείνουν ένα σύστημα στο οποίο συνδυάζουν το χαμηλό υπολογιστικό κόστος των HMMs με τη ικανότητα γενίκευσης των μηχανών διανυσμάτων υποστήριξης (Support Vector Machines, [CoVa95], [Bu98], [Va98]). Η βασική ιδέα είναι η χρησιμοποίηση ενός δυαδικού ταξινομητή,

όπως είναι οι μηχανές διανυσμάτων υποστήριξης στη βασική τους μορφή, για την αναθεώρηση των λαθών που έγιναν από το HMM.

3.2.6 Πειράματα για την ελληνική γλώσσα

Αρκετές από τις παραπάνω τεχνικές έχουν εφαρμοστεί και στη ελληνική γλώσσα με αρκετά καλά αποτελέσματα. Πιο συγκεκριμένα οι Δερματάς και Κοκκινάκης έκαναν πειράματα με HMM ([DeKo95]), ενώ οι Ορφανός, Καλλές και λοιποί ([OrKa99]) και οι Ορφανός και Χριστοδουλάκης ([OrCh99]) χρησιμοποίησαν δένδρα αποφάσεων. Τέλος δύο πολύ σημαντικές προσπάθειες που βασίζονται στη μέθοδο TBED (ενότητα 3.2.1) έγιναν από την ερευνητική ομάδα του ΕΚΕΦΕ «Δημόκριτος» ([PePa99]) καθώς και από την ερευνητική ομάδα του Ινστιτούτου Επεξεργασίας Λόγου (ΙΕΛ, [PaPr00]).

4 Αναγνώριση μερών του λόγου με ενεργητική μάθηση

4.1 Ενεργητική Μάθηση

Η Ενεργητική Μάθηση (Active Learning) είναι ένας τομέας της μηχανικής μάθησης στον οποίο ο αλγόριθμος εκπαίδευσης επιλέγει ο ίδιος τα παραδείγματα που πρέπει να επισημειωθούν και να συμπεριληφθούν στο σώμα εκπαίδευσης. Με τον τρόπο αυτό είναι δυνατόν το μέγεθος των δεδομένων που χρειάζονται για την εκπαίδευση ενός αλγορίθμου επιβλεπόμενης μάθησης να μειωθεί σημαντικά. Αυτό είναι πολύ σημαντικό, αν αναλογιστεί κανείς το υψηλό κόστος της επισημείωσης δεδομένων εκπαίδευσης, η οποία συνήθως γίνεται χειρωνακτικά. Έτσι, η ενεργητική μάθηση, αντί να επιλέγει τυχαία παραδείγματα προς επισημείωση για τη δημιουργία το σώματος εκπαίδευσης, προτείνει προς επισημείωση τα παραδείγματα εκείνα που αναμένεται να επιφέρουν το μεγαλύτερο όφελος στον αλγόριθμο εκμάθησης.

Ένα τυπικό σύστημα ενεργητικής μάθησης αποτελείται από τα ακόλουθα στοιχεία, όπως περιγράφονται από τους Tong και Koller ([ToKo00]). Τα διαθέσιμα δεδομένα χωρίζονται σε δύο σύνολα, X και U. Το X αποτελείται από αρχικά λίγα επισημειωμένα παραδείγματα, ενώ το U είναι μία δεξαμενή μη επισημειωμένων παραδειγμάτων. Τέλος, υπάρχει ένας αλγόριθμος εκμάθησης, l, που εκπαιδεύεται στα επισημειωμένα παραδείγματα και μία διαδικασία επερώτησης q. Η διαδικασία q επιλέγει ποια παραδείγματα από το σύνολο U θα επισημειωθούν και θα συμπεριληφθούν στο σύνολο X, το οποίο στη συνέχεια θα χρησιμοποιηθεί για την εκπαίδευση του l. Σε περίπτωση που χρησιμοποιείται παθητική μάθηση (passive learning), τα παραδείγματα διαλέγονται με τυχαίο τρόπο από το U, σε αντίθεση με την ενεργητική μάθηση όπου επιλέγονται τα παραδείγματα που μεγιστοποιούν ένα συγκεκριμένο κριτήριο επιλογής. Ας σημειωθεί ότι στην περίπτωση της ενεργητικής μάθησης η εκπαίδευση στο X ενδέχεται να οδηγήσει σε υψηλότερο ποσοστό ορθής κατάταξης νέων αντικειμένων

(accuracy) από ό,τι η εκπαίδευση σε ολόκληρο το U . Αυτό είναι δυνατόν να συμβεί, επειδή τα επιπλέον παραδείγματα του U που δεν περιλαμβάνονται στο X ενδέχεται να εισάγουν θόρυβο, με αποτέλεσμα να μην εκπαιδεύεται καλά το σύστημα.

Η αποτελεσματικότητα της ενεργητικής μάθησης μετριέται με δύο τρόπους. Ο πιο δημοφιλής είναι η μείωση, σε σχέση με την παθητική μάθηση, του μεγέθους των επισημειωμένων δεδομένων που απαιτούνται για την επίτευξη ενός συγκεκριμένου ποσοστού ορθότητας. Ο δεύτερος είναι η αύξηση του ποσοστού ορθότητας για συγκεκριμένο όγκο επισημειωμένων δεδομένων εκπαίδευσης.

Η ενεργητική μάθηση έχει εφαρμοστεί σε πολλούς τομείς, συμπεριλαμβανομένων προβλημάτων επεξεργασίας φυσικής γλώσσας, με πολύ καλά αποτελέσματα. Για παράδειγμα, έχει χρησιμοποιηθεί σε συστήματα κατηγοριοποίησης κειμένων, αναγνώρισης ονομάτων οντοτήτων, αναγνώρισης φωνής, εξαγωγής πληροφοριών κ.τ.λ. Λόγω της μεγάλης της σημασίας, έχουν προταθεί αρκετές μέθοδοι ενεργητικής μάθησης, οι σημαντικότερες εκ των οποίων περιγράφονται παρακάτω.

4.1.1 Δειγματοληψία βασισμένη στην αβεβαιότητα (Uncertainty based sampling)

Η δειγματοληψία που βασίζεται στην αβεβαιότητα ([LeGa94], [CoGh95]) μετράει το βαθμό βεβαιότητας που έχει ο ταξινομητής (που έχει προκύψει από την εκπαίδευση ως εκείνη τη στιγμή) για τις κατηγορίες μη επισημειωμένων παραδειγμάτων. Ο ταξινομητής αναμένεται να έχει καλύτερα αποτελέσματα αν εκπαιδευτεί σε παραδείγματα για τα οποία έχει μεγάλη αβεβαιότητα. Η μέθοδος αυτή χρειάζεται έναν πιθανοτικό ταξινομητή, ο οποίος αποδίδει στα μη επισημειωμένα παραδείγματα μια πιθανότητα για κάθε δυνατή ετικέτα (κατηγορία). Στη συνέχεια υπολογίζεται η εντροπία της κατανομής των ετικετών για κάθε μη επισημειωμένο παράδειγμα, και

επιλέγεται το παράδειγμα ή τα παραδείγματα με τη μεγαλύτερη εντροπία. Υψηλή τιμή εντροπίας υποδηλώνει μεγάλη αβεβαιότητα ως προς την κατηγορία στην οποία πρέπει να καταταγεί το μη επισημειωμένο παράδειγμα.

4.1.2 Επερώτηση με Επιτροπή (Query by Committee)

Η επερώτηση με επιτροπή ([SeOp92]) είναι μία μέθοδος κατά την οποία μετριέται ο βαθμός συμφωνίας ανάμεσα σε μία επιτροπή ταξινομητών ως προς την ετικέτα η οποία θα αποδοθεί σε ένα μη επισημειωμένο παράδειγμα εκπαίδευσης. Η επιτροπή των ταξινομητών εκπαιδεύεται στα επισημειωμένα παραδείγματα και στη συνέχεια προσπαθεί να ταξινομήσει τα μη επισημειωμένα παραδείγματα. Τα παραδείγματα στα οποία παρουσιάζεται ο μεγαλύτερος βαθμός ασυμφωνίας μεταξύ των ταξινομητών επισημειώνονται, ενσωματώνονται στα δεδομένα εκπαίδευσης και η διαδικασία επαναλαμβάνεται. Ένας τρόπος για να μετρηθεί η ασυμφωνία είναι μετρώντας την εντροπία των ψήφων των ταξινομητών ([ArDa99]). Η βασική υπόθεση της μεθόδου αυτής είναι ότι αν η επιτροπή ταξινομητών δε μπορεί να συμφωνήσει ως προς την κατηγορία στην οποία πρέπει να κατατάξει ένα παράδειγμα, τότε το σώμα εκπαίδευσης δεν περιέχει αρκετά, ή και καθόλου, παρόμοια παραδείγματα, προκαλώντας έτσι αντικρουόμενες αποφάσεις από τους ταξινομητές. Τέλος δεν πρέπει να παραληφθεί ότι η μέθοδος αυτή έχει καλύτερα αποτελέσματα αν η επιτροπή αποτελείται από διαφορετικών ειδών ταξινομητές (π.χ. ταξινομητές που έχουν προκύψει με χρήση διαφορετικών αλγορίθμων μάθησης), των οποίων οι αποφάσεις δε σχετίζονται μεταξύ τους.

4.1.3 Επιλογή παραδειγμάτων κοντά στην επιφάνεια διαχωρισμού

Σε ένα πρόβλημα κατηγοριοποίησης, ο διαχωρισμός των στιγμιοτύπων σε κατηγορίες γίνεται μέσω υπερ-επιφανειών που μαθαίνει ο αλγόριθμος μηχανικής μάθησης. Μία μέθοδος ενεργητικής μάθησης που έχει προταθεί επιλέγει τα παραδείγματα που βρίσκονται κοντά στις διαχωριστικές υπερ-επιφάνειες ή ακόμα και πάνω σε αυτές. Τα παραδείγματα αυτά αντιστοιχούν

σε περιπτώσεις μέγιστης αβεβαιότητας. Η μέθοδος αυτή έχει εφαρμοστεί κυρίως σε μηχανές διανυσμάτων υποστήριξης (ΜΔΥ, SVM) δύο κατηγοριών, όπου η υπερ-επιφάνεια διαχωρισμού είναι ένα υπερ-επίπεδο, πιθανώς σε ένα νέο διανυσματικό χώρο ([Br03], [ScCo00], [ToKo00]). Αρκετές, μάλιστα, προσεγγίσεις επιχειρούν να «εμπλουτίσουν» τη μέθοδο αυτή ούτως ώστε να βελτιώσουν την απόδοσή της. Έτσι, οι Shen, Zhang και λοιποί ([ShZh04]), προτείνουν, επιπλέον, και επιλογή των πιο αντιπροσωπευτικών παραδειγμάτων, μεταξύ εκείνων που βρίσκονται κοντά στο υπερ-επίπεδο διαχωρισμού. Το κατά πόσο αντιπροσωπευτικό είναι ένα παράδειγμα εξαρτάται από το πλήθος των μη επισημειωμένων παραδειγμάτων που είναι παρόμοια με αυτό. Επίσης, αν τα παραδείγματα επιλέγονται κατά δεσμίδες (batches), προτείνουν τα παραδείγματα που εισάγονται στις δεσμίδες να διαφέρουν όσο το δυνατόν περισσότερο μεταξύ τους. Τέλος, οι Nguyen και Smeulders [NgSm04] χρησιμοποιούν ομαδοποίηση (clustering) των μη επισημειωμένων παραδειγμάτων εκπαίδευσης και εκπαιδεύουν έναν ταξινομητή στους αντιπροσώπους (representatives) των ομάδων (clusters) αφού πρώτα τους επισημειώσουν χειρωνακτικά. Στη συνέχεια γίνεται επιλογή των καλύτερων μη επισημειωμένων παραδειγμάτων εκπαίδευσης όπως και στις κλασσικές τεχνικές ενεργητικής μάθησης. Επιλέγονται δηλαδή τα παραδείγματα που βρίσκονται κοντά στην υπερ-επιφάνεια διαχωρισμού. Στη συνέχεια αν χρειαστεί γίνεται εκ νέου ομαδοποίηση των υπολοίπων μη επισημειωμένων παραδειγμάτων εκπαίδευσης και δημιουργούνται ομάδες μικρότερου μεγέθους. Οι νέοι αντιπρόσωποι επιλέγονται και πάλι για επισημείωση. Η διαδικασία αυτή εκτελείται επαναληπτικά μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού.

Στον k-NN οι υπερ-επιφάνειες διαχωρισμού, που δεν είναι απαραιτήτως υπερ-επίπεδα, χωρίζουν τις περιοχές στις οποίες πλειοψηφούν συγκεκριμένες κατηγορίες. Έτσι το πρόβλημα έγκειται και πάλι στην επιλογή υποψηφίων προς επισημείωση παραδειγμάτων που να βρίσκονται κοντά σε υπερ-

επιφάνειες διαχωρισμού. Επιστρέφουμε στον τρόπο επιλογής υποψηφίων παραδειγμάτων του k-NN στην ενότητα 4.2.4.

4.1.4 Κριτήρια τερματισμού

Μέχρι τώρα, έχουμε αναφερθεί σε διάφορες μεθόδους της ενεργητικής μάθησης καθώς και στα οφέλη τα οποία έχουν. Ένα ερώτημα που γεννάται είναι το πότε πρέπει να σταματήσει η διαδικασία της ενεργητικής μάθησης. Με δεδομένο ότι ο κύριος λόγος χρήσης της ενεργητικής μάθησης είναι η μείωση του κόστους επισημείωσης παραδειγμάτων εκπαίδευσης, η ανάγκη εύρεσης ενός καλού κριτηρίου τερματισμού καθίσταται επιτακτική. Εν τούτοις, η εύρεση καλών κριτηρίων τερματισμού δεν είναι εύκολη. Και αυτό γιατί αν επιλέξουμε να σταματήσουμε την ενεργητική μάθηση σε κάποιο συγκεκριμένο αριθμό παραδειγμάτων εκπαίδευσης, κανείς δεν μας εγγυάται ότι έχουμε φτάσει στο καλύτερο δυνατό αποτέλεσμα. Αυτό το πρόβλημα, βέβαια, το αντιμετωπίζει και η παθητική μάθηση. Μία μερική απάντηση στο παραπάνω ερώτημα είναι ότι οποιοσδήποτε ταξινομητής είναι σοβαρά λανθασμένος είναι σχεδόν σίγουρο ότι θα αποκαλυφθεί με μεγάλη πιθανότητα μετά από ένα μικρό αριθμό παραδειγμάτων καθώς θα κάνει μία λανθασμένη πρόβλεψη. Συνεπώς, οποιοσδήποτε ταξινομητής είναι συνεπής με ένα επαρκώς μεγάλο σύνολο δεδομένων εκπαίδευσης είναι απίθανο να είναι σοβαρά λανθασμένος, δηλαδή πρέπει να είναι «πιθανώς περίπου σωστός» (probably approximately correct – PAC, [RaNo04]). Επιπλέον, μία ένδειξη για το ότι πλησιάζουμε στο βέλτιστο σημείο τερματισμού, είναι η οριζοντίωση της καμπύλης εκμάθησης ή η φθίνουσα πορεία της. Η καμπύλη εκμάθησης δείχνει το ποσοστό ορθότητας που επιτυγχάνει ένας ταξινομητής σε νέα δεδομένα ελέγχου (test set) συναρτήσει του μεγέθους των δεδομένων εκπαίδευσης. Αν παρατηρείται οριζοντίωση όσο αυξάνεται ο όγκος των δεδομένων εκπαίδευσης, τότε δεν αποκομίζουμε κάποιο ουσιαστικό όφελος από την προσθήκη νέων δεδομένων εκπαίδευσης. Η φθίνουσα πορεία ενδέχεται να οφείλεται στο ότι τα νέα επιλεγόμενα παραδείγματα εκπαίδευσης εισάγουν θόρυβο ή οδηγούν σε

μεγάλη αύξηση του αριθμού παραδειγμάτων μίας κατηγορίας, με αποτέλεσμα το σύστημα να μην εκπαιδεύεται το ίδιο καλά με πριν (βλ. και [Vl04]).

4.2 Αναθεώρηση της έννοιας της ενεργητικής μάθησης

4.2.1 Εισαγωγή

Η πλειοψηφία των μέχρι τώρα προσεγγίσεων θεωρούν την ενεργητική μάθηση ως τη διαδικασία κατά την οποία το σύστημα επιλέγει και προτείνει σε ένα χρήστη που είναι υπεύθυνος για την εκπαίδευσή του (τον εκπαιδευτή) τα παραδείγματα που πρέπει να επισημειωθούν και να συμπεριληφθούν στο σώμα εκπαίδευσης. Εμείς αναθεωρούμε αυτήν την άποψη και βλέπουμε την ενεργητική μάθηση κάτω από μία γενικότερη σκοπιά. Πιο συγκεκριμένα, θεωρούμε ως ενεργητική μάθηση την ευρύτερη διαδικασία κατά την οποία το σύστημα αλληλεπιδρά με τον επόπτη σε όλους εκείνους τους τομείς που μπορεί να συμβάλουν στην καλύτερη εκπαίδευση του συστήματος. Έτσι διακρίνουμε τρία στάδια ενεργητικής μάθησης. Στο πρώτο, το σύστημα βρίσκει πιθανά λάθη επισημείωσης που οφείλονται στον ανθρώπινο παράγοντα και τα υποδεικνύει στον επόπτη για διόρθωση. Στο δεύτερο, το σύστημα βοηθάει τον επόπτη να προσθέσει ιδιότητες ούτως ώστε να διαχωρίζονται καλύτερα οι κατηγορίες. Τέλος, στο τρίτο στάδιο ενεργητικής μάθησης επιλέγονται τα καλύτερα παραδείγματα προς επισημείωση, τα οποία και ενσωματώνονται στα δεδομένα εκπαίδευσης. Στην ουσία, δηλαδή, το τρίτο στάδιο ενεργητικής μάθησης, σύμφωνα με τη θεώρηση μας, είναι η κλασσική ενεργητική μάθηση όπως εφαρμόζεται στην πλειοψηφία των προηγούμενων προσεγγίσεων. Εδώ θα πρέπει να τονίσουμε ότι, καθώς το πρόβλημα με το οποίο ασχολούμαστε είναι η αναγνώριση μερών του λόγου, και τα τρία στάδια ενεργητικής μάθησης είναι επικεντρωμένα στην περιοχή αυτή. Εν τούτοις, πιστεύουμε ότι μπορούν με επιτυχία να εφαρμοστούν (κυρίως το δεύτερο και το τρίτο) και σε άλλους τομείς της επεξεργασίας φυσικής γλώσσας. Παρακάτω θα επιχειρήσουμε να παρουσιάσουμε αναλυτικότερα τα τρία προαναφερθέντα στάδια.

4.2.2 1ο στάδιο ενεργητικής μάθησης (εύρεση λαθών στην επισημείωση)

Σε κάθε διαδικασία στην οποία υπεισέρχεται ο ανθρώπινος παράγοντας είναι πιθανό να προκύψουν λάθη. Η επισημείωση λέξεων (στην περίπτωσή μας με ετικέτες που αντιστοιχούν σε μέρη του λόγου και άλλες μορφολογικές πληροφορίες) σε κείμενα δε μπορεί να αποτελεί εξαίρεση. Έτσι, οι Dickinson και Meurers πρότειναν μία μέθοδο εύρεσης περιπτώσεων λανθασμένης επισημείωσης λέξεων [DiMe03]. Εμείς χρησιμοποιούμε τη διαδικασία αυτή έτσι ώστε το σύστημα να προτείνει στο χρήστη τις πιθανώς λανθασμένα επισημειωμένες λέξεις. Ο χρήστης στη συνέχεια κρίνει αν οι λέξεις που του προτάθηκαν είναι όντως λανθασμένα επισημειωμένες και αν ναι τις διορθώνει.

4.2.2.1 Βασική ιδέα

Κατά τη διάρκεια της χειρωνακτικής επισημείωσης, ο χρήστης καλείται να επιλέξει από ένα σύνολο ετικετών αυτήν που αντιστοιχεί σε κάθε λέξη. Στην επιλογή αυτή συμβάλλουν τόσο τα χαρακτηριστικά της κάθε λέξης όσο και τα χαρακτηριστικά των λέξεων που την περιβάλλουν, οι οποίες αποτελούν τη γειτονιά της. Μία λέξη είναι αρκετά φυσικό να εμφανίζεται περισσότερες από μία φορές σε ένα κείμενο. Επιπλέον, δεν είναι υποχρεωτικό η λέξη αυτή να έχει πάντα την ίδια ετικέτα. Χαρακτηριστικό παράδειγμα είναι η λέξη «του», η οποία μπορεί να είναι είτε οριστικό αρσενικό άρθρο στη γενική πτώση (**του** καιρού), είτε οριστικό ουδέτερο άρθρο στη γενική πτώση (**του** παιδιού), είτε αδύνατος τύπος προσωπικής αντωνυμίας αρσενικού γένους στη γενική πτώση (**του** έδειξε το δρόμο, εννοώντας ότι τον έδειξε στο γείτονα), είτε, τέλος, αδύνατος τύπος προσωπικής αντωνυμίας ουδέτερου γένους στη γενική πτώση (**του** έδειξε το δρόμο, εννοώντας ότι τον έδειξε σε ένα παιδί). Τέτοιου είδους αμφισημίες υπάρχουν πολλές, επομένως δε μπορούμε να στηριχθούμε απλά στην παρατήρηση ίδιων λέξεων με διαφορετικές ετικέτες για την εύρεση λαθών στην επισημείωση. Χρειαζόμαστε κάποια πληροφορία παραπάνω ούτως ώστε να σιγουρευτούμε ότι έχουμε βρει κάποιο λάθος στην επισημείωση. Αυτή την

πληροφορία μας την παρέχει η γειτονιά της λέξης που εξετάζουμε. Πιο συγκεκριμένα, η εμφάνιση της ίδιας λέξης με διαφορετική ετικέτα, σε δύο ή περισσότερες ίδιες γειτονιές, σηματοδοτεί, με μεγάλη πιθανότητα, την ύπαρξη λάθους.

4.2.2.2 Ο αλγόριθμος

Ο αλγόριθμος που προτείνεται από τους Dickinson και Meurers στηρίζεται στον εντοπισμό και επεξεργασία ακολουθιών λέξεων των κειμένων εκπαίδευσης οι οποίες περιέχουν μία πιθανώς λανθασμένα επισημειωμένη λέξη. Αυτές οι ακολουθίες λέξεων ονομάζονται n-γράμματα διαφοροποίησης (variation n-grams), όπου το n αντιστοιχεί στο μήκος (αριθμός λέξεων) της ακολουθίας. Με τον όρο n-γράμμα διαφοροποίησης εννοούμε μία ακολουθία λέξεων για την οποία υπάρχει μία τουλάχιστον άλλη ακριβώς ίδια με αυτήν ακολουθία λέξεων μέσα στα κείμενα εκπαίδευσης, στην οποία μία από τις λέξεις έχει διαφορετική ετικέτα από την αντίστοιχη λέξη της άλλης ακολουθίας. Προφανώς ένα n-γραμμα περιέχει δύο (n-1)-γράμματα, που προκύπτουν αν αφαιρέσουμε είτε την πρώτη είτε την τελευταία λέξη. Ο αλγόριθμος αρχικά παράγει όλα τα πιθανά n-γράμματα διαφοροποίησης σύμφωνα με την παρακάτω διαδικασία:

1. Εντοπίζουμε όλα τα 1-γράμματα διαφοροποίησης (μεμονωμένες λέξεις) των επισημειωμένων κειμένων εκπαίδευσης και τα αποθηκεύουμε, μαζί με τις θέσεις που κατέχουν στα κείμενα. Δηλαδή εντοπίζουμε ίδιες λέξεις με διαφορετικές ετικέτες και τις αποθηκεύουμε μαζί με τις θέσεις τους.
2. Βασισμένοι στις θέσεις των n-γραμμάτων που αποθηκεύτηκαν τελευταία, επεκτείνουμε κάθε ένα από αυτά και από τις δύο πλευρές, παράγοντας δύο νέα (n+1)-γράμματα, αν αυτό είναι εφικτό. Η επέκταση δεν είναι εφικτή μόνο αν έχουμε φτάσει στην αρχή ή στο τέλος κάποιου κειμένου. Για κάθε ένα από τα (n+1)-γράμματα που προκύπτουν, ελέγχουμε αν υπάρχουν άλλα (n+1)-

γράμματα όμοια με αυτό και αν υπάρχουν τα αποθηκεύουμε όλα, μαζί με τις θέσεις τους.

3. Επιστρέφουμε στο βήμα 2 μέχρι να φτάσουμε σε σημείο που δεν υπάρχουν όμοια n-γράμματα διαφοροποίησης.

Στο βήμα 2 δεν είναι απαραίτητο να ελέγχουμε αν τα όμοια (n+1)-γράμματα περιέχουν ίδιες λέξεις με διαφορετικές ετικέτες, γιατί αυτό το εγγυάται το βήμα 1.

Αφού υπολογιστούν όλα τα n-γράμματα διαφοροποίησης μπορούν να χρησιμοποιηθούν ευριστικές μέθοδοι για την ανακάλυψη πιθανών σφαλμάτων. Συγκεκριμένα, οι Dickinson και Meurers προτείνουν δύο ευριστικές. Αυτή που χρησιμοποιούμε εμείς εξετάζει το μήκος των n-γραμμάτων διαφοροποίησης. Είναι πιο πιθανόν μία διαφορά στην επισημείωση μίας λέξης να είναι λάθος αν η λέξη αυτή περιέχεται με διαφορετικές ετικέτες σε όμοια n-γράμματα μεγάλου μήκους. Επομένως, όσο πιο μεγάλο είναι το μήκος των όμοιων n-γραμμάτων που περιέχουν τη λέξη με διαφορετική ετικέτα, τόσο μεγαλύτερη είναι η πιθανότητα η λέξη αυτή να έχει επισημειωθεί λανθασμένα. Η δεύτερη ευριστική των Dickinson και Meurers θεωρεί ότι όταν η λέξη για την οποία έχουμε διαφορετικές ετικέτες είναι στα άκρα των n-γραμμάτων τότε είναι πολύ πιθανόν να μην υπάρχει λάθος στην επισημείωση ακόμα και αν το n-γράμμα είναι μεγάλο. Αυτό συμβαίνει γιατί για τις λέξεις που βρίσκονται στα άκρα του n-γράμματος χάνεται η πληροφορία των συμφραζομένων που δεν ανήκουν στο n-γράμμα. Ο λόγος για τον οποίο δεν χρησιμοποιήσαμε αυτή την ευριστική είναι ότι το σύστημα θέλουμε απλά να προτείνει τις πιθανόν λανθασμένα επισημειωμένες λέξεις. Δε μας πειράζει ιδιαίτερα αν προταθεί μία λέξη η οποία έχει επισημειωθεί σωστά, καθώς ο χρήστης στη συνέχεια μπορεί εύκολα να κρίνει αν η λέξη αυτή είναι λανθασμένα επισημειωμένη. Επιπλέον, δεν θέλουμε να εξαιρέσουμε τις λέξεις με διαφορετικές ετικέτες που βρίσκονται στα άκρα των n-γραμμάτων, καθώς ενδέχεται κάποιες από αυτές να αποτελούν λάθη επισημείωσης.

4.2.2.3 Παραδείγματα

Για να γίνουν περισσότερο κατανοητά τα παραπάνω παραθέτουμε ορισμένα παραδείγματα, όπως αυτά προέκυψαν κατά την εκτέλεση της παραπάνω διαδικασίας στο σώμα κειμένων το οποίο επισημειώσαμε. Η διαδικασία σταμάτησε όταν φτάσαμε σε μήκος n-γράμματος 14.

Στα παρακάτω παραδείγματα με υπογράμμιση σημειώνεται το n-γράμμα, ενώ δίπλα στη λέξη στην οποία παρατηρήθηκε η διαφορά στην ετικέτα σημειώνεται η ετικέτα (ενότητα 5.2 και παράρτημα Ι). Αρχικά ας ξεκινήσουμε με ένα 4-γράμμα:

1. Γράψτε τον αριθμό των/AtDfMaPlGe αποθηκευτικών χώρων
2. Γράψτε τον αριθμό των/AtDfFePlGe φορολογικών αποθηκών
3. Γράψτε τον αριθμό των/AtDfFePlGe χρήσεων
4. Γράψτε τον αριθμό των/AtDfFePlGe πρόσκαιρων ή άλλων εγκαταστάσεών
5. Γράψτε τον αριθμό των/AtDfFePlGe συνδεδεμένων επιχειρήσεων
6. Γράψτε τον αριθμό των/AtDfFePlGe θέσεων

Όπως παρατηρούμε η λέξη «των» παρατηρήθηκε με δύο διαφορετικές ετικέτες. Όμως και στις έξι περιπτώσεις η επισημείωση είναι σωστή. Το ίδιο γεγονός παρατηρήθηκε στις περισσότερες περιπτώσεις 4-γραμμάτων, ισχυροποιώντας, έτσι, την πεποίθηση ότι μικρό μήκος n-γράμματος συνεπάγεται μικρή πιθανότητα λάθους στην επισημείωση. Όσο μεγάλωνε το μήκος του n-γράμματος, τόσο μεγάλωνε και η πιθανότητα σφάλματος. Έτσι για μήκος 7 συναντήσαμε μοιρασμένες περιπτώσεις ύπαρξης και μη σφάλματος στην επισημείωση. Χαρακτηριστικά είναι τα δύο ακόλουθα παραδείγματα.

1. : σημειώστε « Χ » στο τετραγωνίδιο/NoNeSgAc
2. : σημειώστε « Χ » στο τετραγωνίδιο/AtDfNeSgAc

Παραπάνω η λέξη «τετραγωνίδιο» έχει επισημειωθεί λανθασμένα ως άρθρο στη δεύτερη περίπτωση. Αξίζει να σημειωθεί ότι αν χρησιμοποιούσαμε και τη δεύτερη ευριστική των Dickinson και Meurers (ενότητα 4.2.2.2) το σύστημα δε θα πρότεινε τη συγκεκριμένη λέξη.

1. Για την εξεύρεση των εξωλογιστικών κερδών των/AtDfMaPlGe ελεύθερων επαγγελματιών
2. Για την εξεύρεση των εξωλογιστικών κερδών των/AtDfFePlGe επιχειρήσεων
3. Για την εξεύρεση των εξωλογιστικών κερδών των/AtDfFePlGe επιχειρήσεων
4. Για την εξεύρεση των εξωλογιστικών κερδών των/AtDfFePlGe επιχειρήσεων
5. Για την εξεύρεση των εξωλογιστικών κερδών των/AtDfFePlGe επιχειρήσεων

Αντίθετα, στις παραπάνω περιπτώσεις, παρά τη διαφορετική επισημείωση της λέξης «των», δεν υπάρχει λάθος.

Τέλος, στη μοναδική περίπτωση που παρατηρήθηκε 14-γραμμο η λέξη στην οποία υπήρχε η διαφοροποίηση στην ετικέτα είχε επισημειωθεί λανθασμένα.

1. περιλαμβάνονται στον πίνακα/NoMaSgAc αυτό και στην οποία θα περιλάβετε τις εξής στήλες : Στήλη «
2. περιλαμβάνονται στον πίνακα/NoFeSgAc αυτό και στην οποία θα περιλάβετε τις εξής στήλες : Στήλη «

4.2.3 2ο στάδιο ενεργητικής μάθησης (συμβολή του συστήματος στην εύρεση ιδιοτήτων)

Για την καλύτερη εκπαίδευση του συστήματος είναι επιτακτική η ανάγκη ορισμού ενός συνόλου ιδιοτήτων που θα παρέχει αρκετές

πληροφορίες, ώστε να είναι δυνατόν να διαχωριστούν οι λέξεις διαφορετικών κατηγοριών. Από την άλλη πλευρά, το σύνολο ιδιοτήτων δεν θα πρέπει να περιλαμβάνει περιττές ιδιότητες, δηλαδή ιδιότητες που δεν παρέχουν χρήσιμες πληροφορίες για τις κατηγορίες των λέξεων ή που οι πληροφορίες τους μπορούν να συναχθούν από τις πληροφορίες άλλων ιδιοτήτων. Θα μπορούσε κανείς να ισχυριστεί ότι ο προσδιορισμός του καλύτερου συνόλου ιδιοτήτων θα μπορούσε να γίνει με έναν αλγόριθμο αναζήτησης, όπως για παράδειγμα η αναρρίχηση λόφου (hill climbing), χρησιμοποιώντας ως συνάρτηση αξιολόγησης των υποψηφίων συνόλων ιδιοτήτων το ποσοστό ορθότητας που επιτυγχάνουν σε ένα πείραμα διασταυρωμένης επικύρωση (ενότητα 2.3). Πιο συγκεκριμένα, θα επιλεγόταν τυχαία ένα αρχικό σύνολο ιδιοτήτων, το οποίο θα αποτελούσε την αρχική κατάσταση της αναρρίχησης λόφου, και κάθε μετάβαση θα πρόσθετε ή θα αφαιρούσε μία ιδιότητα οδηγώντας σε μια νέα κατάσταση (σύνολο ιδιοτήτων). Κάθε φορά θα επιλεγόταν η μετάβαση που οδηγεί στην καλύτερη κατάσταση (υψηλότερο ποσοστό ορθότητας στη διασταυρωμένη επικύρωση) μέχρι το σημείο όπου δε θα υπήρχε βελτίωση. Προκειμένου να αποφευχθεί ο κίνδυνος εγκλωβισμού σε τοπικό μέγιστο, η αναζήτηση θα μπορούσε να επαναληφθεί αρκετές φορές, με διαφορετικό σύνολο αρχικών ιδιοτήτων κάθε φορά. Σε μια απλούστερη και πιο διαδεδομένη εκδοχή, η κατασκευή του συνόλου ιδιοτήτων μπορεί να γίνεται ξεκινώντας από το κενό σύνολο και προσθέτοντας διαδοχικά ιδιότητες, από τις καλύτερες προς τις χειρότερες, όπως θα αξιολογούνταν μεμονωμένα με ένα κριτήριο όπως το πληροφοριακό κέρδος.

Υπάρχουν, όμως, προβλήματα μάθησης στα οποία η κατασκευή του συνόλου ιδιοτήτων δεν είναι δυνατόν να γίνει με τις παραπάνω μεθόδους, επειδή είναι αδύνατον να προσδιοριστούν εκ των προτέρων όλες οι δυνατές ιδιότητες. Η αναγνώριση μερών του λόγου εντάσσεται στην κατηγορία αυτών των προβλημάτων, καθώς μπορεί κανείς να σκεφτεί ένα πολύ μεγάλο πλήθος ιδιοτήτων, που αφορούν τόσο τη μορφολογία της ίδιας της λέξης, όσο και τη μορφολογία των λέξεων που την περιβάλλουν και δεν είναι εφικτό να δοθούν

όλες αυτές οι ιδιότητες στο σύστημα εκ των προτέρων ως δυνατές ιδιότητες. Το δεύτερο στάδιο ενεργητικής μάθησης που προτείνουμε έχει ως στόχο να βοηθήσει τον εκπαιδευτή του συστήματος να εντοπίσει τις ιδιότητες που πρέπει να προστεθούν σε ένα αρχικά πολύ μικρό σύνολο ιδιοτήτων. Το ίδιο το σύστημα δεν προτείνει ιδιότητες. Παρέχει, όμως, στον εκπαιδευτή στοιχεία που τον βοηθούν να σκεφτεί εκείνος τις ιδιότητες που πρέπει να προστεθούν. Τέλος, εφόσον έχει εκτελεστεί το πρώτο στάδιο ενεργητικής μάθησης θεωρούμε ότι δεν υπάρχουν λάθη στην επισημείωση των λέξεων. Βέβαια το 1ο στάδιο δεν εγγυάται ότι έχουν βρεθεί όλα τα λάθη επισημείωσης αλλά θεωρούμε ότι το ποσοστό των λαθών είναι τόσο μικρό που δεν επηρεάζει τη συνολική λειτουργία του συστήματος.

4.2.3.1 Βασική ιδέα

Η βασική ιδέα που βρίσκεται πίσω από την πρότασή μας είναι η παρατήρηση ότι αν βρούμε παραδείγματα εκπαίδευσης διαφορετικών κατηγοριών που παριστάνονται από ίδια διανύσματα ιδιοτήτων, τότε οι ιδιότητες οι οποίες έχουμε στο σύνολο ιδιοτήτων δεν παρέχουν αρκετές πληροφορίες για να διαχωριστούν αντικείμενα (στην περίπτωσή μας λέξεις) διαφορετικών κατηγοριών και άρα πρέπει να προστεθούν νέες ιδιότητες. Μάλιστα, μας ενδιαφέρουν διανύσματα τα οποία παρουσιάζονται με μεγάλη συχνότητα στο σώμα εκπαίδευσης και κατανέμονται σχετικά ομοιόμορφα σε διαφορετικές κατηγορίες. Τα διανύσματα θέλουμε να έχουν μεγάλη συχνότητα γιατί έτσι έχουν τη δυνατότητα να επηρεάσουν σημαντικά το ποσοστό ορθότητας. Επιπλέον, αν δεν κατανέμονται σχετικά ομοιόμορφα σε διαφορετικές κατηγορίες αλλά ανήκουν με μεγάλη συχνότητα σε μία κατηγορία, ο ταξινομητής που προκύπτει ενδέχεται να μαθαίνει να τα κατατάσσει πάντα στην πιο συχνή κατηγορία με αποτέλεσμα να επιτυγχάνει υψηλό ποσοστό ορθότητας και συνεπώς να μην αντιμετωπίζει σημαντικό πρόβλημα. Για το λόγο αυτό χρησιμοποιούμε ως μέτρο επιλογής τέτοιων διανυσμάτων δύο κριτήρια: α) την εντροπία της κατηγορίας τους και β) τη συχνότητα εμφάνισης των διανυσμάτων στο σώμα εκπαίδευσης. Πιο

συγκεκριμένα, κατατάσσουμε τα διανύσματα με φθίνουσα σειρά εντροπίας, ενώ σε περιπτώσεις ισοβαθμίας γίνεται κατάταξη με φθίνουσα σειρά συχνότητας. Ένας εναλλακτικός τρόπος για την κατάταξη των διανυσμάτων, είναι να χρησιμοποιηθεί ως μέτρο το γινόμενο των δύο κριτηρίων που προαναφέραμε. Έτσι, αποφεύγονται περιπτώσεις υψηλής κατάταξης διανυσμάτων με μικρό πλήθος αλλά μεγάλη εντροπία. Από την άλλη όμως, ελλοχεύει ο κίνδυνος υψηλής κατάταξης διανυσμάτων με μεγάλο πλήθος και χαμηλή εντροπία. Αυτό το πρόβλημα μπορεί να μετριαστεί χρησιμοποιώντας λογάριθμο στο κριτήριο πλήθους.

4.2.3.2 Παράδειγμα

Παρακάτω παραθέτουμε ένα παράδειγμα για να γίνει περισσότερο κατανοητός ο τρόπος λειτουργίας του 2ου σταδίου ενεργητικής μάθησης.

1. φορολογίας εισοδήματος που περιλαμβάνει τα **καθαρά**/AjNePlAc κέρδη όλης της διαχειριστικής περιόδου
2. Στον πίνακα αυτόν προσδιορίζονται τα **καθαρά**/AjNePlNm κέρδη των επιχειρήσεων που τα
3. και 606 : Γράψτε τα **καθαρά**/AjNePlAc κέρδη από την εκτέλεση των
4. κέρδη » : Γράψτε τα **καθαρά**/AjNePlAc κέρδη από την εκτέλεση του
5. κέρδη » : Γράψτε τα **καθαρά**/AjNePlAc κέρδη από την πώληση του
6. και άλλες εκμεταλλεύσεις τα **καθαρά**/AjNePlNm κέρδη των οποίων υπολογίζονται στις
7. το οικονομικό έτος 2002 τα **καθαρά**/AjNePlNm κέρδη που θα λαμβάνονται για

Στο παράδειγμα αυτό η λέξη «καθαρά» μπορεί να καταταγεί σε δύο κατηγορίες. Είτε είναι ουδέτερο επίθετο σε ονομαστική πληθυντικού, είτε είναι

ουδέτερο επίθετο σε αιτιατική πληθυντικού. Το διάνυσμα των ιδιοτήτων σε όλες τις περιπτώσεις είναι: <αρά, 6, 0, 0, 0, 0, 0, ρδη, τα>. Οι επτά πρώτες ιδιότητες που έχουν επιλεγεί αφορούν τη μορφολογία της λέξης, ενώ οι δύο τελευταίες είναι κατά σειρά η κατάληξη της επόμενης και της προηγούμενης λέξης. Πιο συγκεκριμένα οι εννιά ιδιότητες που έχουμε είναι:

1. Η κατάληξη της λέξης.
2. Το μήκος της λέξης.
3. Αν η λέξη περιέχει απόστροφο.
4. Αν η λέξη περιέχει αριθμό.
5. Αν η λέξη περιέχει τελεία.
6. Αν η λέξη περιέχει κόμμα.
7. Αν η λέξη περιέχει λατινικούς χαρακτήρες.
8. Η κατάληξη της επόμενης λέξης.
9. Η κατάληξη της προηγούμενης λέξης.

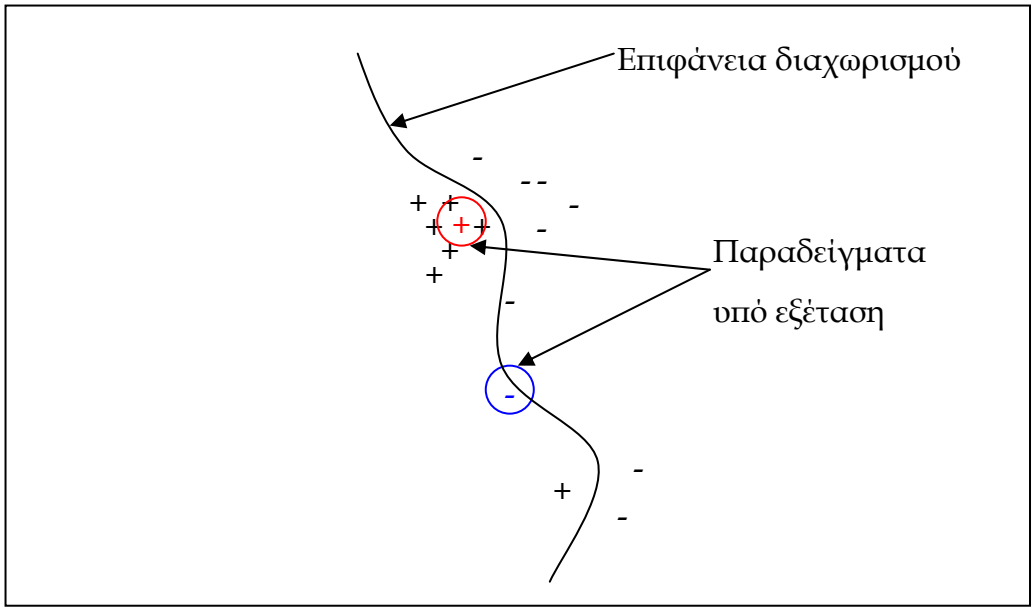
Είναι προφανές ότι δε μπορούν να προκύψουν άλλες «χρήσιμες» ιδιότητες για τη μορφολογία της λέξης καθώς σε κάθε περίπτωση όλες οι ιδιότητες που αφορούν τη μορφολογία της λέξης θα έχουν την ίδια τιμή (η εξεταζόμενη λέξη είναι η ίδια: «καθαρά») και συνεπώς τα νέα διανύσματα ιδιοτήτων που θα προκύπτουν θα εξακολουθούν να είναι ίδια μεταξύ τους. Όμως στις πέντε πρώτες περιπτώσεις οι λέξεις που βρίσκονται δύο θέσεις πριν την εξεταζόμενη παρέχουν σημαντική πληροφορία για την πτώση της. Πιο συγκεκριμένα στη δεύτερη περίπτωση η λέξη που βρίσκεται δύο θέσεις πριν είναι ρήμα σε παθητική φωνή και η λέξη «καθαρά» είναι επιθετικός προσδιορισμός του υποκειμένου του και συνεπώς έχει την ίδια πτώση με αυτό (ονομαστική). Αντίθετα στις υπόλοιπες τέσσερις περιπτώσεις η λέξη δύο θέσεις πριν είναι ρήμα σε ενεργητική φωνή και η λέξη «καθαρά» είναι επιθετικός προσδιορισμό του αντικειμένου του. Συνεπώς βρίσκεται στην αιτιατική πτώση. Στο συγκεκριμένο παράδειγμα δε μπορούμε να χρησιμοποιήσουμε ως ιδιότητα την ετικέτα της λέξης που βρίσκεται δύο θέσεις πριν, καθώς οι ετικέτες που χρησιμοποιούμε (παράρτημα Ι) δε διακρίνουν μεταξύ παθητικών και

ενεργητικών ρημάτων. Θα μπορούσαμε, όμως, να χρησιμοποιήσουμε ως ιδιότητα την κατάληξη της λέξης που βρίσκεται δύο θέσεις πριν από αυτήν την οποία εξετάζουμε, ελπίζοντας ότι έτσι θα περιλάβουμε έμμεσα στα διανύσματα πληροφορίες για τη φωνή του ρήματος. Παρατηρώντας όμως πιο προσεκτικά το παράδειγμα, διαπιστώνουμε ότι μια τέτοια ιδιότητα δεν παρέχει καμία απολύτως πληροφορία στις περιπτώσεις 6 και 7. Θα πρέπει λοιπόν να βρούμε μία ιδιότητα που από τη μία να παρέχει την πληροφορία που θα διαχωρίζει τα διανύσματα στις περιπτώσεις 1 – 5 αλλά και που δε θα εισάγει θόρυβο στις περιπτώσεις 6 και 7. Μπορούμε να χρησιμοποιήσουμε δύο δυαδικές ιδιότητες εκ των οποίων η μία θα καθορίζει αν η λέξη δύο θέσεις πριν έχει ρηματική κατάληξη ενεργητικής φωνής και η άλλη αν η λέξη δύο θέσεις πριν έχει ρηματική κατάληξη παθητικής φωνής. Με αυτόν τον τρόπο διαχωρίζουμε τις πέντε πρώτες περιπτώσεις αλλά οδηγούμαστε σε λάθος στην περίπτωση 6, καθώς η κατάληξη της λέξης «εκμεταλλεύσεις» είναι ρηματική κατάληξη ενεργητικής φωνής ενώ στην ουσία πρόκειται για ουσιαστικό και οδηγούμαστε εσφαλμένα στο συμπέρασμα ότι η λέξη «καθαρά» είναι σε αιτιατική πτώση. Αν όμως χρησιμοποιήσουμε και μία δυαδική ιδιότητα που να δείχνει αν η λέξη δύο θέσεις πριν από την εξεταζόμενη είναι ρήμα, τότε μπορούμε να πετύχουμε πλήρη διαχωρισμό. Συγκεκριμένα, αν η λέξη δύο θέσεις πριν είναι ρήμα στην παθητική φωνή ή δεν είναι ρήμα, τότε η λέξη «καθαρά» βρίσκεται στην ονομαστική, διαφορετικά βρίσκεται στην αιτιατική. Όπως γίνεται κατανοητό, ακόμα και με την υποβοήθηση από το σύστημα η εύρεση καλών ιδιοτήτων είναι πολύ δύσκολη για ένα πρόβλημα όπως είναι η αναγνώριση μερών του λόγου.

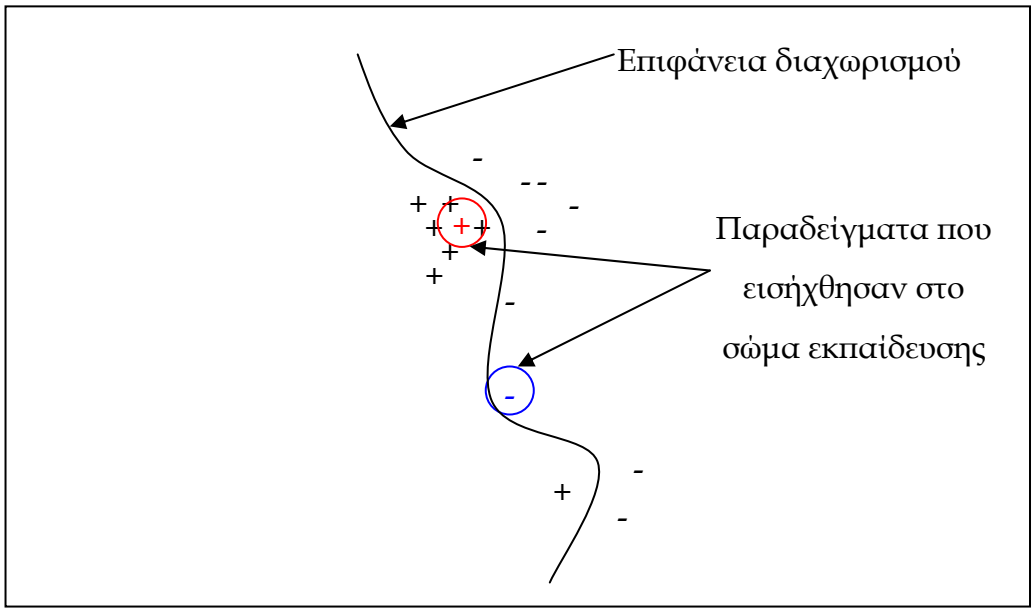
4.2.4 3ο στάδιο ενεργητικής μάθησης (επιλογή των «καλύτερων» παραδειγμάτων)

Το τρίτο στάδιο ενεργητικής μάθησης, είναι το κλασσικό στάδιο ενεργητικής μάθησης, κατά το οποίο το σύστημα επιλέγει προς επισημείωση τα παραδείγματα τα οποία θεωρεί ότι θα βοηθήσουν περισσότερο στην εκπαίδευση του ταξινομητή. Η τεχνική που χρησιμοποιούμε είναι η επιλογή

των παραδειγμάτων που βρίσκονται κοντά σε όσο το δυνατόν περισσότερες υπερ-επιφάνειες διαχωρισμού (ενότητα 4.1.3). Επιπλέον προτιμούμε να εντάξουμε στα δεδομένα εκπαίδευσης (το X της ενότητας 4.1) παραδείγματα που βρίσκονται σε περιοχές για τις οποίες τα υπάρχοντα δεδομένα εκπαίδευσης (τα οποία θα βρίσκονται στη μνήμη του k-NN) δεν περιέχουν ήδη πολλά παραδείγματα. Ο λόγος για την τελευταία προτίμηση είναι ότι αν το υπό εξέταση παράδειγμα βρίσκεται σε μια περιοχή για την οποία τα δεδομένα εκπαίδευσης περιέχουν ήδη πολλά παραδείγματα, η προσθήκη του νέου παραδείγματος δεν θα επηρεάσει σημαντικά την υπερ-επιφάνεια διαχωρισμού. Αντίθετα, αν δεν υπάρχουν ήδη πολλά παραδείγματα στην περιοχή, το νέο παράδειγμα μπορεί να επηρεάσει σημαντικά τη μορφή της υπερ-επιφάνειας διαχωρισμού, ιδιαίτερα αν βρίσκεται κοντά σε αυτή. Ας θεωρήσουμε ένα απλοϊκό παράδειγμα στο οποίο έχουμε μόνο δύο κατηγορίες, θετικά (+) και αρνητικά (-) αντικείμενα. Μία πιθανή κατάσταση των δεδομένων εκπαίδευσης φαίνεται στην εικόνα 4-1. Σύμφωνα με τους παραπάνω συλλογισμούς, το μπλε παράδειγμα είναι πιο χρήσιμο από το κόκκινο, γιατί οδηγεί σε σημαντική τροποποίηση της μορφής της υπερ-επιφάνειας διαχωρισμού σε μια περιοχή όπου δεν υπήρχαν αρκετά παραδείγματα και η μορφή της υπερ-επιφάνειας διαμορφωνόταν από πολύ μακρινά παραδείγματα. Αντίθετα το κόκκινο παράδειγμα δεν επιφέρει καμία ουσιαστική μεταβολή στη μορφή της υπερ-επιφάνειας διαχωρισμού, γιατί η ψήφος του χάνεται μέσα στις ψήφους των πολλών παραδειγμάτων που υπάρχουν ήδη στην περιοχή του (εικόνα 4-2).



Εικόνα 4-1



Εικόνα 4-2

Σε προβλήματα που έχουμε δύο κατηγορίες (+, -) ένα πιθανό μέτρο για να αποδώσουμε τη σημαντικότητα του κάθε υποψήφιου παραδείγματος περιγράφεται από τη σχέση 4.1.

$$W = |V^+ - V^-| \cdot (V^+ + V^-) \qquad (4.1)$$

Στην παραπάνω σχέση με W συμβολίζεται η τιμή του μέτρου, ενώ με V^+ και V^- οι θετικές και οι αρνητικές ψήφοι αντίστοιχα όπως προέκυψαν από την ψηφοφορία των k κοντινότερων γειτόνων. Μικρή τιμή του πρώτου παράγοντα (απόλυτη τιμή της διαφοράς των ψήφων) αντιστοιχεί σε σημείο κοντά στην υπερ-επιφάνεια διαχωρισμού. Επιπλέον, μικρή τιμή του δεύτερου παράγοντα (άθροισμα των ψήφων) υποδεικνύει σημείο για το οποίο δεν υπάρχουν πολλά παρόμοια παραδείγματα στο σώμα εκπαίδευσης. Έτσι, όσο μικρότερη είναι η τιμή του γινομένου, τόσο σημαντικότερο είναι το υποψήφιο παράδειγμα, δηλαδή τόσο πιο επιθυμητό είναι να προστεθεί στο σώμα εκπαίδευσης.

Το μέτρο που μόλις περιγράψαμε έχει σχεδιαστεί για προβλήματα με δύο κατηγορίες. Όμως, όταν έχουμε περισσότερες από δύο κατηγορίες, υπάρχουν πολλές υπερ-επιφάνειες διαχωρισμού και θέλουμε να επιλέγουμε διανύσματα που βρίσκονται κοντά σε όσο το δυνατόν περισσότερες υπερ-επιφάνειες. Έτσι, κρίνεται επιτακτική η ανάγκη ορισμού ενός καλύτερου μέτρου που θα καθορίζει το πότε υπάρχει ισοψηφία. Για το λόγο αυτό επιλέγεται η εντροπία της κατηγορίας του υπό εξέταση παραδείγματος, όπου οι πιθανότητες εκτιμώνται με βάση τις ψήφους των γειτόνων, όπως φαίνεται στη σχέση 4.2.

$$H(x) = -\sum_{c \in C} P(c) \log_2 P(c) \tag{4.2}$$

$$P(c) = \frac{V^c}{\sum_{s \in C} V^s} \tag{4.3}$$

Στη σχέση αυτή με x συμβολίζουμε το υπό εξέταση αντικείμενο ενώ C είναι το σύνολο των κατηγοριών στις οποίες ανήκουν οι k κοντινότεροι γείτονες. Οι πιθανότητα $P(c)$ εκτιμάται από τις ψήφους των k κοντινότερων γειτόνων. Πιο συγκεκριμένα, υπολογίζεται ως το πηλίκο των ψήφων των γειτόνων που ανήκουν στην κατηγορία c (V^c) διά το άθροισμα των ψήφων για όλες τις κατηγορίες στις οποίες ανήκουν οι k κοντινότεροι γείτονες (σχέση

4.3). Όσο πιο μεγάλη είναι η εντροπία ψήφων, τόσο πιο ομοιόμορφη είναι κατανομή των ψήφων στις κατηγορίες και συνεπώς πρόκειται για σημείο που ισαπέχει από πολλές υπερ-επιφάνειες διαχωρισμού. Έτσι, εισάγοντας και τον παράγοντα που μετρά το πλήθος των παραδειγμάτων της περιοχής, το νέο μέτρο σημαντικότητας ενός παραδείγματος εκπαίδευσης είναι το πηλίκο της εντροπίας διά το άθροισμα των ψήφων (σχέση 4.4). Όσο πιο μεγάλη είναι η τιμή του πηλίκου αυτού τόσο πιο σημαντικό είναι το παράδειγμα που εξετάζουμε και τόσο πιο επιθυμητό είναι να ενσωματωθεί στο σώμα εκπαίδευσης.

$$W = \frac{H(x)}{\sum_{c \in C} V^c} \tag{4.4}$$

Όμως και αυτό το μέτρο παρουσιάζει προβλήματα. Συγκεκριμένα, καθώς μεγαλώνει το πλήθος των παραδειγμάτων που υπάρχουν συνολικά στο σώμα εκπαίδευσης (στη μνήμη του k-NN), ο παρονομαστής του πηλίκου παίρνει μεγαλύτερες τιμές, ακόμα και στις περιοχές όπου υπάρχουν λίγα παραδείγματα, λόγω της μικρής (αλλά όχι μηδενικής) συνεισφοράς των μακρινών παραδειγμάτων στο αποτέλεσμα της ψηφοφορίας (τον παρανομαστή). Έτσι, το μέτρο αυτό σταδιακά δίνει μεγαλύτερη βαρύτητα στο κριτήριο του παρανομαστή (να μην υπάρχουν πολλά παραδείγματα στην ίδια περιοχή) από ό,τι στο κριτήριο του αριθμητή (να βρισκόμαστε κοντά σε υπερ-επιφάνειες διαχωρισμού). Επιπλέον, όταν οι κατηγορίες είναι πολλές το κριτήριο του αριθμητή μπορεί να πάρει μεγάλες τιμές και έτσι να αποκτήσει μεγαλύτερη βαρύτητα από ό,τι το κριτήριο του παρονομαστή. Για τους λόγους αυτούς χρησιμοποιούμε την παρακάτω (σχέση 4.5) μορφή του μέτρου, όπου ο λογάριθμος μετριάζει τη βαρύτητα του κριτηρίου του παρανομαστή ενώ στο κριτήριο του αριθμητή η εντροπία κανονικοποιείται διαιρούμενη με το λογάριθμο του συνολικού πλήθους των κατηγοριών (σχέση 4.6).

$$W = \frac{H_n(x)}{\log\left(\sum_{c \in C} V^c\right)} \tag{4.5}$$

$$H_n(x) = -\frac{\sum_{c \in C} P(c) \log_2 P(c)}{\log(|C|)} \tag{4.6}$$

Τέλος, προκειμένου το μέτρο να μην επιστρέφει την ίδια (μηδενική) τιμή σε περιπτώσεις υποψηφίων παραδειγμάτων που έχουν μηδενική εντροπία αλλά να κατατάσσει υψηλότερα τα υποψήφια παραδείγματα που βρίσκονται σε περιοχές όπου δεν υπάρχουν ήδη πολλά παραδείγματα, τροποποιούμε το μέτρο έτσι ώστε στις περιπτώσεις μηδενικής εντροπίας να παίρνει την αντίθετη τιμή του αθροίσματος των ψήφων των k κοντινότερων γειτόνων (σχέση 4.7). Έτσι σε όλα τα υποψήφια παραδείγματα μηδενικής εντροπίας το μέτρο επιστέφει αρνητικές τιμές, αλλά όσα από αυτά βρίσκονται σε περιοχές όπου υπάρχουν ήδη πολλά άλλα παραδείγματα κατατάσσονται χαμηλότερα (πιο αρνητικές τιμές).

$$W = \begin{cases} \dfrac{H_n(x)}{\log\left(\sum_{c \in C} V^c\right)}, & \text{αν } H_n(x) \neq 0 \\ -\sum_{c \in C} V^c, & \text{αν } H_n(x) = 0 \end{cases} \tag{4.5}$$

5 Πειραματική αξιολόγηση και αποτελέσματα

Για την εξαγωγή συμπερασμάτων σχετικά με την αποτελεσματικότητα των παραπάνω μεθόδων, ακολουθήθηκε η παρακάτω πειραματική διαδικασία.

5.1 Σώμα κειμένων

Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκαν ειδησεογραφικά κείμενα από τις εφημερίδες «ΤΑ ΝΕΑ» και «ΒΗΜΑ». Το συνολικό μέγεθος των κειμένων που χρησιμοποιηθήκαν είναι 20374 λέξεις που προέκυψαν από πέντε συνολικά άρθρα. Τα κείμενα υποβλήθηκαν αρχικά στην ακόλουθη προ-επεξεργασία:

1. **Μετατροπή σε μορφότυπο απλού κειμένου (plain text format)**: Τα κείμενα βρίσκονταν αρχικά σε μορφή HTML. Αφαιρέθηκαν οι ετικέτες HTML και τα κείμενα μετατράπηκαν σε μορφότυπο απλού κειμένου.
2. **Ορθογραφικός έλεγχος**: Διορθώθηκαν τα ορθογραφικά λάθη των κειμένων. Επίσης εντοπίστηκαν ελληνικές λέξεις που περιείχαν λατινικούς χαρακτήρες (π.χ. λατινικό «I» αντί για ελληνικό) και οι χαρακτήρες αυτοί αντικαταστάθηκαν από τους αντίστοιχους ελληνικούς. Όλες οι παραπάνω λειτουργίες εκτελέστηκαν χειρωνακτικά.
3. **Διαχωρισμός λέξεων**: Διαχωρίστηκαν χειρωνακτικά οι λέξεις από σημεία στίξης και άλλα ειδικά σύμβολα. Αυτό στο μέλλον θα μπορούσε να γίνεται αυτόματα με ένα διαχωριστή λεκτικών μονάδων (tokenizer) Σε περιπτώσεις όπως οι συντμήσεις και άλλες ειδικές χρήσεις σημείων στίξης και συμβόλων δεν έγινε κανένας διαχωρισμός. Παραδείγματος χάριν, οι τελείες στη σύντμηση «κ.λ.π.» δεν διαχωρίστηκαν αλλά θεωρήθηκε ότι και οι έξι χαρακτήρες αποτελούν μία λέξη. Το ίδιο έγινε και για τις

ημερομηνίες της μορφής «Ημέρα/Μήνας/Έτος» (π.χ. 28/6/2005) όπου δεν έγινε κανένας διαχωρισμός και ολόκληρη η ημερομηνία θεωρήθηκε μία λέξη και επισημειώθηκε ως αριθμητικό.

4. **Επισημείωση**: Επισημειώθηκε χειρωνακτικά κάθε λέξη με την ετικέτα της[1] με τη βοήθεια κατάλληλου εργαλείου που κατασκευάστηκε για το σκοπό αυτό (παράρτημα ΙΙ).
5. **Εύρεση λαθών επισημείωσης**: Χρησιμοποιήθηκε το πρώτο στάδιο ενεργητικής μάθησης για την εύρεση πιθανών λαθών στην επισημείωση. Συνολικά βρέθηκαν 22 λανθασμένα επισημειωμένες λέξεις. Τα παραδείγματα της ενότητας 4.1.2.3 έχουν προέλθει από το στάδιο αυτό. Στη συνέχεια το σώμα κειμένων ελέγχθηκε χειρωνακτικά για την εύρεση τυχόν λαθών στην επισημείωση που δεν προτάθηκαν από το σύστημα χωρίς ωστόσο να βρεθούν άλλα λάθη.

5.2 Το σύνολο των ετικετών

Το σύνολο των ετικετών που χρησιμοποιήσαμε αποτελείται από 112 ετικέτες . Πρόκειται για απλοποιημένη μορφή του συνόλου ετικετών που έχει προταθεί από το Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ) ως αντιστοίχηση του διεθνούς προτύπου PAROLE στην ελληνική γλώσσα. Το απλοποιημένο σύνολο ετικετών που χρησιμοποιήσαμε δημιουργήθηκε αρχικά (με 120 ετικέτες) από το Ινστιτούτο Πληροφορικής του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος». Στη διάρκεια της εργασίας, ορισμένες από τις ετικέτες αυτού του συνόλου αφαιρέθηκαν ή απλοποιήθηκαν (π.χ. τα ειδησεογραφικά κείμενα δεν περιέχουν ουσιαστικά σε κλητική, δεν υπάρχει πλέον ειδική κατηγορία για εμπρόθετα άρθρα κ.λ.π.). Αναλυτικά το σύνολο ετικετών περιγράφεται στο παράρτημα Ι.

[1] Στην πράξη δεν είναι απαραίτητο να επισημειωθεί όλο το σώμα κειμένων αλλά μόνο τα παραδείγματα που επιλέγει η ενεργητική μάθηση. Ο λόγος που εδώ επισημειώθηκε ολόκληρο το σώμα είναι ότι μας χρειάζεται για την εκτέλεση των πειραμάτων.

5.3 Πειράματα με ενεργητική μάθηση

Έχοντας, πλέον, στη διάθεσή μας το σώμα κειμένων εκτελέσαμε τριών ειδών πειράματα. Ένα στο οποίο χρησιμοποιήσαμε μόνο το τρίτο στάδιο ενεργητικής μάθησης, ένα στο οποίο χρησιμοποιήσαμε παθητική μάθηση και ένα στο οποίο χρησιμοποιήσαμε ταυτόχρονα το δεύτερο και τρίτο στάδιο ενεργητική μάθησης. Να υπενθυμίσουμε εδώ, ότι και στις τρεις περιπτώσεις έχει εκτελεστεί το πρώτο στάδιο ενεργητικής μάθησης, κατά τη διαδικασία προ-επεξεργασίας του σώματος κειμένων.

Ο αλγόριθμος εκπαίδευσης που χρησιμοποιήσαμε είναι ο IB1[2] με k=1 (ενότητα 2.2.1.1) με το μέτρο απόστασης MVDM (ενότητα 2.2.1.3), ιδιότητες ζυγισμένες με το μέτρο αναλογία κέρδους (ενότητα 2.2.1.4) και βάρη στις ψήφους των κοντινότερων γειτόνων αντιστρόφως ανάλογα με την απόσταση (ενότητα 2.2.1.5).

Τέλος, το σώμα κειμένων διασπάστηκε σε δύο τμήματα. Το πρώτο έχει μέγεθος 15300 λέξεις και αποτελεί το σώμα εκπαίδευσης (πιο σωστά, αποτελεί τις υποψήφιες προς ένταξη στο σώμα εκπαίδευσης λέξεις), ενώ το δεύτερο έχει μέγεθος 5074 λέξεις και αποτελεί το σώμα ελέγχου.

5.3.1 Πειράματα με το 3ο στάδιο ενεργητικής μάθησης.

Για απλότητα στα παρακάτω θα χρησιμοποιήσουμε τα σύνολα X και U όπως αυτά ορίστηκαν στην ενότητα 4.1.

Στα πειράματα αυτά σκεφτήκαμε και χρησιμοποιήσαμε ένα σύνολο εννέα ιδιοτήτων. Αυτές αφορούν τη μορφολογία τη λέξης, καθώς και τη μορφολογία της προηγούμενης και της επόμενης λέξης (ενότητα 4.2.3.2).

[2] Χρησιμοποιήθηκε η παραλλαγή του IB1 που έχει υλοποιηθεί στο πακέτο TiMBL, που αντί για τους k κοντινότερους γείτονες επιλέγει όλους τους γείτονες που βρίσκονται στις k κοντινότερες αποστάσεις. Έτσι για k=1 μπορεί να υπάρχουν περισσότεροι του ενός γείτονες.

Αρχικά δημιουργούνται τα διανύσματα ιδιοτήτων όλων των λέξεων, τόσο του σώματος ελέγχου όσο και του σώματος εκπαίδευσης. Για να μπορέσει να εκκινήσει η διαδικασία ενεργητικής μάθησης πρέπει να αρχικοποιηθεί το σύνολο X. Για το σκοπό αυτό επιλέγονται τα 51 πρώτα διανύσματα (ισοδύναμα λέξεις) του σώματος εκπαίδευσης, τα οποία και αποτελούν το αρχικό σύνολο X. (Στην πράξη, τα διανύσματα επισημειώνονται κατά την εισαγωγή τους στο X. Στα πειράματά μας όλα τα διανύσματα είναι ήδη επισημειωμένα αλλά οι επισημειώσεις γίνονται ορατές στον αλγόριθμο μάθησης μόνο κατά την εισαγωγή των διανυσμάτων στο X.) Το σύστημα εκπαιδεύεται σε αυτά και επιλέγει από τα εναπομείναντα διανύσματα του U τα καλύτερα 51, τα οποία και εντάσσει στο X. Το σύστημα επανεκπαιδεύεται στο νέο X και η διαδικασία συνεχίζεται μέχρι να τελειώσουν τα υποψήφια για ένταξη στο X διανύσματα του U. Συνολικά δημιουργήθηκαν διαδοχικά 300 σύνολα εκπαίδευσης X, που το καθένα έχει 51 περισσότερα διανύσματα από το προηγούμενό του. Το σύστημα εκπαιδεύτηκε με τη σειρά σε όλα τα σύνολα και ελέγχθηκε, μέσω του σώματος ελέγχου, το ποσοστό ορθότητας (accuracy) που επιτυγχάνεται κάθε φορά.

Τα καλύτερα διανύσματα τα επιλέξαμε χρησιμοποιώντας δύο μέτρα σημαντικότητας. Το πρώτο είναι αυτό που περιγράφεται στην ενότητα 4.2.4 και συνδυάζει την απόσταση από τις υπερ-επιφάνειες διαχωρισμού (εντροπία) με τον αριθμό υπαρχόντων παραδειγμάτων στην περιοχή του υπό εξέταση παραδείγματος, ενώ το δεύτερο είναι μόνο η απόσταση από τις υπερ-επιφάνειες.

5.3.2 Πειράματα με παθητική μάθηση

Στο στάδιο αυτό επιλέξαμε τα νέα διανύσματα εκπαίδευσης, σε ομάδες των 51, με τη σειρά που εμφανίζονται στα κείμενα οι αντίστοιχες λέξεις, δημιουργώντας, όπως και παραπάνω, 300 σύνολα εκπαίδευσης. Χρησιμοποιήθηκαν κι εδώ οι ίδιες 9 ιδιότητες του προηγούμενου πειράματος. Ακολούθως, το σύστημα εκπαιδεύτηκε με τη σειρά σε όλα τα σύνολα και

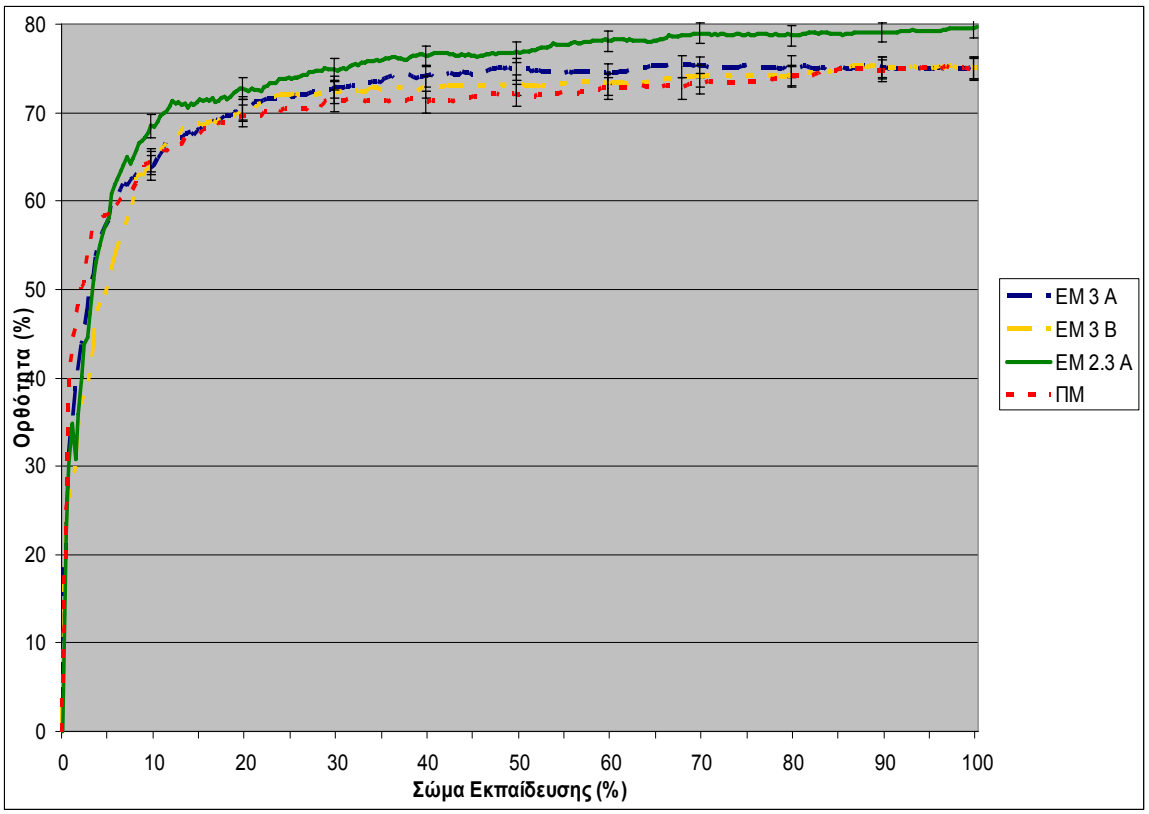
ελέγχθηκε, μέσω του σώματος ελέγχου, το ποσοστό ορθότητας που επιτυγχάνεται κάθε φορά.

5.3.3 Πειράματα με το 2ο και 3ο στάδιο ενεργητικής μάθησης.

Τα πειράματα στην περίπτωση αυτή έγιναν με τον ίδιο περίπου τρόπο που έγιναν και τα πειράματα στα οποία χρησιμοποιήσαμε μόνο το τρίτο στάδιο ενεργητικής μάθησης. Η μόνη διαφορά ήταν ότι μετά την προσθήκη κάθε νέας ομάδας 51 διανυσμάτων στο X εκτελούσαμε το δεύτερο στάδιο ενεργητικής μάθησης με είσοδο τα διανύσματα του X, ώστε να εντοπίσουμε νέες αναγκαίες ιδιότητες και να τις προσθέσουμε στο σύνολο ιδιοτήτων. Μετά την προσθήκη των νέων ιδιοτήτων δημιουργούσαμε εκ νέου τα διανύσματα ιδιοτήτων για τα σώματα εκπαίδευσης και ελέγχου. Στη συγκεκριμένη πειραματική διαδικασία όλες οι ιδιότητες που τελικά χρησιμοποιήθηκαν, καθορίστηκαν στην τρίτη επανάληψη (μετά την προσθήκη της τρίτης 51-άδας διανυσμάτων). Πιο συγκεκριμένα, στην τρίτη επανάληψη το σύστημα μας υπέδειξε περιπτώσεις όπου δε μπορούσε να διαχωρίσει τα αντικείμενα που υπήρχαν στο σύνολο X. Αυτός μας βοήθησε να καθορίσουμε τελικά 185 ιδιότητες οι οποίες αφορούν τη μορφολογία τόσο της υπό εξέταση λέξης όσο και των συμφραζομένων της. Από την επανάληψη αυτή και μετά το σύστημα δε βρήκε όμοια διανύσματα τα οποία κατατάσσονταν σε διαφορετικές κατηγορίες. Το ποσοστό ορθότητας ελέγχθηκε όπως και στις δύο προηγούμενες περιπτώσεις.

5.3.4 Αποτελέσματα

Τα αποτελέσματα που προέκυψαν από τα πειράματα που περιγράφηκαν παραπάνω φαίνονται στο γράφημα 5.1.



Εικόνα 5-1

Στο παραπάνω γράφημα ο οριζόντιος άξονας παριστάνει το ποσοστό του σώματος εκπαίδευσης που χρησιμοποιείται (διανύσματα του X) και ο κατακόρυφος άξονας το ποσοστό ορθότητας που επιτυγχάνεται στο σώμα ελέγχου. Το ποσοστό ορθότητας μετριέται ως ο αριθμός των σωστών προβλέψεων (εμφανίσεις λέξεων στις οποίες αποδόθηκαν σωστές ετικέτες) δια του συνόλου των περιπτώσεων (εμφανίσεις λέξεων). Το γράφημα δείχνει και τα διαστήματα εμπιστοσύνης κάθε αποτελέσματος, με βαθμό βεβαιότητας 95%.

Διακρίνονται τέσσερις καμπύλες εκμάθησης, μία για κάθε είδος πειράματος που διενεργήθηκε:

1. **Καμπύλη ΕΜ 3 Α**: 1ο και 3ο στάδιο ενεργητικής μάθησης με σύνθετο μέτρο (ενότητα 4.2.4) σημαντικότητας παραδειγμάτων εκπαίδευσης (μέγιστο διάστημα εμπιστοσύνης: ±1,22%).
2. **Καμπύλη ΕΜ 3 Β:** 1ο και 3ο στάδιο ενεργητικής μάθησης με μέτρο σημαντικότητας παραδειγμάτων εκπαίδευσης μόνο την

απόσταση από την υπερ-επιφάνεια (μέγιστο διάστημα εμπιστοσύνης: ±1,23%).

3. **Καμπύλη ΕΜ 2.3 Α:** 1ο, 2ο και 3ο στάδιο ενεργητική μάθησης με σύνθετο μέτρο (ενότητα 4.1.4) σημαντικότητας παραδειγμάτων εκπαίδευσης (μέγιστο διάστημα εμπιστοσύνης: ±1,18%).
4. **Καμπύλη ΠΜ**: 1ο στάδιο ενεργητικής μάθησης και παθητική μάθηση (μέγιστο διάστημα εμπιστοσύνης: ±1,25%).

Δυστυχώς από τη μελέτη των τριών καμπυλών ΕΜ 3 Α, ΕΜ 3 Β και ΠΜ δεν είναι δυνατόν να εξαχθούν ασφαλή συμπεράσματα, λόγω της επικάλυψης των διαστημάτων εμπιστοσύνης. Οι καμπύλες, πάντως, προσφέρουν ενδείξεις πως το σύνθετο μέτρο (ενότητα 4.2.4) για τη σημαντικότητα των παραδειγμάτων εκπαίδευσης (καμπύλη ΕΜ 3 Α) είναι καλύτερο από την απλή απόσταση από το επίπεδο διαχωρισμού (καμπύλη ΕΜ 3 Β). Όχι μόνο επιτυγχάνει καλύτερα αποτελέσματα στα περισσότερα σημεία του οριζόντιου άξονα (αριθμός παραδειγμάτων εκπαίδευσης), αλλά και προσεγγίζει το μέγιστο ποσοστό ορθότητας πιο γρήγορα. Συγκεκριμένα, με το σύνθετο μέτρο μπορούμε χρησιμοποιώντας μόνο περίπου το 50% του σώματος εκπαίδευσης να επιτύχουμε τα ίδια ποσοστά ορθότητας με αυτά που μπορούμε να επιτύχουμε χρησιμοποιώντας όλο το σώμα εκπαίδευσης (η καμπύλη ΕΜ 3 Α περίπου οριζοντιώνεται μετά το 50% του σώματος εκπαίδευσης έχοντας φτάσει ήδη το ποσοστό ορθότητας που επιτυγχάνουμε με ολόκληρο το σώμα εκπαίδευσης). Επίσης, τα αποτελέσματα παρέχουν ενδείξεις ότι η χρήση του 3ου στάδιο ενεργητικής μάθησης, ανεξαρτήτως του ποιο από τα δύο μέτρα επιλογής παραδειγμάτων χρησιμοποιείται (καμπύλες ΕΜ 3 Α και ΕΜ 3 Β), υπερτερεί της παθητικής μάθησης (καμπύλη ΠΜ), ως προς το ότι επιτυγχάνει καλύτερα αποτελέσματα με τον ίδιο αριθμό παραδειγμάτων εκπαίδευσης.

Η συμβολή του 2ου σταδίου ενεργητικής μάθησης (καμπύλη ΕΜ 2.3 Α) είναι πολύ καθαρότερη, αφού οδήγησε σε άνοδο του ποσοστού ορθότητας κατά 5% περίπου. Η άνοδος οφείλεται στο μεγάλο αριθμό ιδιοτήτων που προστέθηκαν κατά τη διάρκεια του 2ου σταδίου (μετάβαση από 9 σε 185

ιδιότητες) και στην χρησιμότητα των πληροφοριών που αυτές προφανώς παρείχαν .

6 Συμπεράσματα – μελλοντική έρευνα

6.1 Ανασκόπηση της εργασίας και συμπεράσματα

Στα πλαίσια αυτής της εργασίας προτάθηκε ένας επαναπροσδιορισμός της έννοιας της ενεργητικής μάθησης, υπό την έννοια του διαχωρισμού της σε τρία στάδια, για καθένα από τα οποία παρουσιάστηκε και μία μέθοδος.

Στο πρώτο στάδιο ενεργητικής μάθησης, το σύστημα βρίσκει λάθη στις επισημειώσεις που οφείλονται στον ανθρώπινο παράγοντα. Για το σκοπό αυτό χρησιμοποιήθηκε η μέθοδος των Dickinson και Meurers μέσω της οποίας βρέθηκαν τελικά 22 λανθασμένα επισημειωμένες λέξεις. Παρ' όλο που η μέθοδος δεν είναι καινούρια, δεν είχε ως τώρα αντιμετωπισθεί ως στάδιο της ενεργητικής μάθησης.

Το δεύτερο στάδιο ενεργητικής μάθησης προτείνεται για πρώτη φορά και έχει ως στόχο να βοηθήσει το χρήστη να σκεφτεί και να προσθέσει νέες ιδιότητες που απαιτούνται για το διαχωρισμό παραδειγμάτων που έχουν τις ίδιες τιμές ιδιοτήτων αλλά ανήκουν σε διαφορετικές κατηγορίες. Τα αποτελέσματα ήταν αρκετά ικανοποιητικά, καθώς αυξήθηκε το ποσοστό ορθότητας του συστήματος κατά 4 – 5 %.

Τέλος, στο τρίτο στάδιο ενεργητικής μάθησης το σύστημα επιχειρεί να προτείνει το ίδιο τα καλύτερα παραδείγματα εκπαίδευσης, ώστε αφού επισημειωθούν να ενσωματωθούν στο σύνολο εκπαίδευσης. Οι υπάρχουσες προσεγγίσεις ενεργητικής μάθησης εστιάζονται μόνο σε αυτό το στάδιο. Η καινοτομία της εργασίας σε αυτό το στάδιο έγκειται στο ότι για την αξιολόγηση της σημαντικότητας ενός υποψηφίου παραδείγματος εκπαίδευσης χρησιμοποιείται ένα μέτρο που συνδυάζει την απόσταση από τις υπερ-επιφάνειες διαχωρισμού με το πλήθος των υπαρχόντων παραδειγμάτων εκπαίδευσης που βρίσκονται στην περιοχή του νέου παραδείγματος. Το μέτρο αυτό συγκρίθηκε πειραματικά με το κλασσικό μέτρο, που χρησιμοποιεί μόνο

την απόσταση από τις υπερ-επιφάνειες. Τα αποτελέσματα παρέχουν ενδείξεις ότι το νέο σύνθετο μέτρο λειτουργεί καλύτερα από το κλασσικό, όπως επίσης ότι και τα δύο μέτρα λειτουργούν καλύτερα από την παθητική μάθηση, όπου τα παραδείγματα εκπαίδευσης επιλέγονται τυχαία.

Με βάση τα παραπάνω καταλήγουμε στο συμπέρασμα ότι η εργασία πέτυχε το στόχο της ως προς τη διερεύνηση μεθόδων ενεργητικής μάθησης στην αναγνώριση μερών του λόγου. Αν και δεν πέτυχε ιδιαίτερα υψηλά επίπεδα ορθότητας (το ανώτερο 79,62%), παρήγαγε ενθαρρυντικά αποτελέσματα, που δείχνουν ότι με τη χρήση της ενεργητικής μάθησης το ίδιο σύστημα είναι δυνατόν να έχει καλύτερα αποτελέσματα από αυτά που επιτύγχανε χωρίς τη χρήση ενεργητικής μάθησης. Η εργασία οδήγησε, επίσης, σε μια ευρύτερη θεώρηση της έννοιας της ενεργητικής μάθησης και σε ένα νέο μέτρο αξιολόγησης των υποψηφίων παραδειγμάτων εκπαίδευσης.

Ταυτόχρονα δημιουργήθηκε το πρώτο σύστημα αναγνώρισης μερών του λόγου (τουλάχιστον για την ελληνική γλώσσα) που χρησιμοποιεί ενεργητική μάθηση. Τα ποσοστά ορθότητας δεν είναι το ίδιο υψηλά με αυτά που έχουν ανακοινώσει άλλοι ερευνητές αλλά το σώμα κειμένων αυτής της εργασίας δεν είναι το ίδιο με αυτά των άλλων ερευνητών και άρα τα αποτελέσματα δεν είναι άμεσα συγκρίσιμα. Επίσης, το σύνολο ετικετών αυτής της εργασίας είναι μεγαλύτερο από τα σύνολα ετικετών πολλών άλλων ερευνητών, κάτι που κάνει το πρόβλημα δυσκολότερο.

Περισσότερες λεπτομέρειες σχετικά με την υλοποίηση του συστήματος υπάρχουν στο παράρτημα ΙΙ. Το σύστημα συνοδεύεται από ένα εργαλείο το οποίο διευκολύνει το χρήστη κατά τη χειρωνακτική επισημείωση των κειμένων.

6.2 Μελλοντικές επεκτάσεις

Οι μέθοδοι ενεργητικής μάθησης της εργασίας μπορούν να εφαρμοστούν και σε άλλους τομείς της επεξεργασίας φυσικής γλώσσας. Για

παράδειγμα, και τα τρία στάδια μπορούν να εφαρμοστούν στο πρόβλημα αναγνώρισης ονομάτων οντοτήτων. Το δεύτερο στάδιο μπορεί, επίσης, να αποδειχθεί χρήσιμο σε εφαρμογές διήθησης ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου (spam), όπου λόγω των τεχνασμάτων που χρησιμοποιούν οι αποστολείς ανεπιθύμητων μηνυμάτων για να μπερδεύουν φίλτρα που χρησιμοποιούν μηχανική μάθηση είναι απαραίτητη η περιοδική προσθήκη νέων ιδιοτήτων.

Όσον αφορά το δεύτερο στάδιο, μία καλή βελτίωση θα ήταν να μην εντοπίζει μόνο τα ίδια διανύσματα ιδιοτήτων που κατατάσσονται σε διαφορετικές κατηγορίες, αλλά και διανύσματα τα οποία μοιάζουν πολύ μεταξύ τους αλλά κατατάσσονται σε διαφορετικές κατηγορίες. Η μέτρηση της ομοιότητας θα μπορούσε να γίνεται με το μέτρο απόστασης του k-NN. Για το τρίτο στάδιο, μία βελτίωση θα ήταν το σύστημα να κατασκευάζει δικά του παραδείγματα εκπαίδευσης, αντί απλά να επιλέγει τα καταλληλότερα από μια δεξαμενή υπαρχόντων υποψήφιων παραδειγμάτων. Αυτό θα ήταν χρήσιμο, για παράδειγμα, σε εφαρμογές που δεν είναι εύκολο να συλλεχθούν υποψήφια παραδείγματα.

Τέλος, καθώς οι μηχανές διανυσμάτων υποστήριξης χρησιμοποιούνται όλο και περισσότερο στη μηχανική μάθηση, μία χρήσιμη μελλοντική κατεύθυνση θα ήταν η διερεύνηση του κατά πόσον το σύνθετο μέτρο που προτάθηκε για το στάδιο 3 μπορεί να βελτιώσει την απόδοση μηχανών διανυσμάτων υποστήριξης.

ΠΑΡΑΡΤΗΜΑ Ι: Το σύνολο των ετικετών

Στους πίνακες που ακολουθούν παρατίθεται το σύνολο των ετικετών που χρησιμοποιήθηκε. Δίπλα σε κάθε ετικέτα παρέχεται σύντομη περιγραφή. Οι πίνακες αποτελούν ελαφρά απλουστευμένη μορφή αντιστοίχων πινάκων που διατέθηκαν από το Ινστιτούτο Πληροφορικής του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος».

ΟΡΙΣΤΙΚΑ ΑΡΘΡΑ

AtDfMaSgNm	Οριστικό άρθρο αρσενικό ενικός ονομαστική
AtDfMaSgGe	Οριστικό άρθρο αρσενικό ενικός γενική
AtDfMaSgAc	Οριστικό άρθρο αρσενικό ενικός αιτιατική
AtDfMaPlNm	Οριστικό άρθρο αρσενικό πληθυντικός ονομαστική
AtDfMaPlGe	Οριστικό άρθρο αρσενικό πληθυντικός γενική
AtDfMaPlAc	Οριστικό άρθρο αρσενικό πληθυντικός αιτιατική
AtDfFeSgNm	Οριστικό άρθρο θηλυκό ενικός ονομαστική
AtDfFeSgGe	Οριστικό άρθρο θηλυκό ενικός γενική
AtDfFeSgAc	Οριστικό άρθρο θηλυκό ενικός αιτιατική
AtDfFePlNm	Οριστικό άρθρο θηλυκό πληθυντικός ονομαστική
AtDfFePlGe	Οριστικό άρθρο θηλυκό πληθυντικός γενική
AtDfFePlAc	Οριστικό άρθρο θηλυκό πληθυντικός αιτιατική
AtDfNeSgNm	Οριστικό άρθρο ουδέτερο ενικός ονομαστική
AtDfNeSgGe	Οριστικό άρθρο ουδέτερο ενικός γενική
AtDfNeSgAc	Οριστικό άρθρο ουδέτερο ενικός αιτιατική
AtDfNePlNm	Οριστικό άρθρο ουδέτερο πληθυντικός ονομαστική
AtDfNePlGe	Οριστικό άρθρο ουδέτερο πληθυντικός γενική

AtDfNePlAc	Οριστικό άρθρο ουδέτερο πληθυντικός αιτιατική

ΑΟΡΙΣΤΑ ΑΡΘΡΑ

AtIdMaSgNm	Αόριστο άρθρο αρσενικό ενικός ονομαστική
AtIdMaSgGe	Αόριστο άρθρο αρσενικό ενικός γενική
AtIdMaSgAc	Αόριστο άρθρο αρσενικό ενικός αιτιατική
AtIdMaPlNm	Αόριστο άρθρο αρσενικό πληθυντικός ονομαστική
AtIdMaPlGe	Αόριστο άρθρο αρσενικό πληθυντικός γενική
AtIdMaPlAc	Αόριστο άρθρο αρσενικό πληθυντικός αιτιατική
AtIdFeSgNm	Αόριστο άρθρο θηλυκό ενικός ονομαστική
AtIdFeSgGe	Αόριστο άρθρο θηλυκό ενικός γενική
AtIdFeSgAc	Αόριστο άρθρο θηλυκό ενικός αιτιατική

ΟΥΣΙΑΣΤΙΚΑ

NoMaSgNm	Ουσιαστικό αρσενικό ενικός ονομαστική
NoMaSgGe	Ουσιαστικό αρσενικό ενικός γενική
NoMaSgAc	Ουσιαστικό αρσενικό ενικός αιτιατική
NoMaPlNm	Ουσιαστικό αρσενικό πληθυντικός ονομαστική
NoMaPlGe	Ουσιαστικό αρσενικό πληθυντικός γενική
NoMaPlAc	Ουσιαστικό αρσενικό πληθυντικός αιτιατική
NoFeSgNm	Ουσιαστικό θηλυκό ενικός ονομαστική
NoFeSgGe	Ουσιαστικό θηλυκό ενικός γενική
NoFeSgAc	Ουσιαστικό θηλυκό ενικός αιτιατική
NoFePlNm	Ουσιαστικό θηλυκό πληθυντικός ονομαστική
NoFePlGe	Ουσιαστικό θηλυκό πληθυντικός γενική

NoFePlAc	Ουσιαστικό θηλυκό πληθυντικός αιτιατική
NoNeSgNm	Ουσιαστικό ουδέτερο ενικός ονομαστική
NoNeSgGe	Ουσιαστικό ουδέτερο ενικός γενική
NoNeSgAc	Ουσιαστικό ουδέτερο ενικός αιτιατική
NoNePlNm	Ουσιαστικό ουδέτερο πληθυντικός ονομαστική
NoNePlGe	Ουσιαστικό ουδέτερο πληθυντικός γενική
NoNePlAc	Ουσιαστικό ουδέτερο πληθυντικός αιτιατική

ΕΠΙΘΕΤΑ

AjMaSgNm	Επίθετο αρσενικό ενικός ονομαστική
AjMaSgGe	Επίθετο αρσενικό ενικός γενική
AjMaSgAc	Επίθετο αρσενικό ενικός αιτιατική
AjMaPlNm	Επίθετο αρσενικό πληθυντικός ονομαστική
AjMaPlGe	Επίθετο αρσενικό πληθυντικός γενική
AjMaPlAc	Επίθετο αρσενικό πληθυντικός αιτιατική
AjFeSgNm	Επίθετο θηλυκό ενικός ονομαστική
AjFeSgGe	Επίθετο θηλυκό ενικός γενική
AjFeSgAc	Επίθετο θηλυκό ενικός αιτιατική
AjFePlNm	Επίθετο θηλυκό πληθυντικός ονομαστική
AjFePlGe	Επίθετο θηλυκό πληθυντικός γενική
AjFePlAc	Επίθετο θηλυκό πληθυντικός αιτιατική
AjNeSgNm	Επίθετο ουδέτερο ενικός ονομαστική
AjNeSgGe	Επίθετο ουδέτερο ενικός γενική
AjNeSgAc	Επίθετο ουδέτερο ενικός αιτιατική

AjNePlNm	Επίθετο ουδέτερο πληθυντικός ονομαστική
AjNePlGe	Επίθετο ουδέτερο πληθυντικός γενική
AjNePlAc	Επίθετο ουδέτερο πληθυντικός αιτιατική

ΑΝΤΩΝΥΜΙΕΣ

PnIc	Αντωνυμία άκλιτη π.χ. «που», «κάτι»
PnSgNm	Αντωνυμία ενικός – η αντωνυμία δεν έχει γένος π.χ. «εγώ»- ονομαστική
PnSgGe	Αντωνυμία ενικός – η αντωνυμία δεν έχει γένος π.χ. «σου» - γενική
PnSgAc	Αντωνυμία ενικός – η αντωνυμία δεν έχει γένος π.χ. «εμένα» -αιτιατική
PnPlNm	Αντωνυμία πληθυντικός – η αντωνυμία δεν έχει γένος π.χ. «εμείς»- ονομαστική
PnPlGe	Αντωνυμία πληθυντικός – η αντωνυμία δεν έχει γένος π.χ. «μας»- γενική
PnPlAc	Αντωνυμία πληθυντικός – η αντωνυμία δεν έχει γένος - αιτιατική
PnMaSgNm	Αντωνυμία αρσενικό ενικός ονομαστική
PnMaSgGe	Αντωνυμία αρσενικό ενικός γενική
PnMaSgAc	Αντωνυμία αρσενικό ενικός αιτιατική
PnMaPlNm	Αντωνυμία αρσενικό πληθυντικός ονομαστική
PnMaPlGe	Αντωνυμία αρσενικό πληθυντικός γενική
PnMaPlAc	Αντωνυμία αρσενικό πληθυντικός αιτιατική
PnFeSgNm	Αντωνυμία θηλυκό ενικός ονομαστική
PnFeSgGe	Αντωνυμία θηλυκό ενικός γενική
PnFeSgAc	Αντωνυμία θηλυκό ενικός αιτιατική

PnFePlNm	Αντωνυμία θηλυκό πληθυντικός ονομαστική
PnFePlGe	Αντωνυμία θηλυκό πληθυντικός γενική
PnFePlAc	Αντωνυμία θηλυκό πληθυντικός αιτιατική
PnNeSgNm	Αντωνυμία ουδέτερο ενικός ονομαστική
PnNeSgGe	Αντωνυμία ουδέτερο ενικός γενική
PnNeSgAc	Αντωνυμία ουδέτερο ενικός αιτιατική
PnNePlNm	Αντωνυμία ουδέτερο πληθυντικός ονομαστική
PnNePlGe	Αντωνυμία ουδέτερο πληθυντικός γενική
PnNePlAc	Αντωνυμία ουδέτερο πληθυντικός αιτιατική

ΑΡΙΘΜΗΤΙΚΑ

NmCd	Απόλυτα αριθμητικά και νούμερα

ΡΗΜΑΤΑ - ΜΕΤΟΧΕΣ

VbIs	Απρόσωπο Ρήμα (οποιουδήποτε χρόνου)
VbMnPrSg	Ρήμα προσωπικό παροντικού χρόνου, ενικός
VbMnPaSg	Ρήμα προσωπικό παρελθοντικού χρόνου, πληθυντικός
VbMnXxSg	Ρήμα προσωπικό μελλοντικού χρόνου, ενικός
VbMnPrPl	Ρήμα προσωπικό παροντικού χρόνου, πληθυντικός
VbMnPaPl	Ρήμα προσωπικό παρελθοντικού χρόνου, πληθυντικός
VbMnXxPl	Ρήμα προσωπικό μελλοντικού χρόνου, πληθυντικός
VbMnNfAv	Απαρέμφατο ενεργητικής φωνής
VbMnNfPv	Απαρέμφατο παθητικής φωνής
VbPpPrAv	Μετοχή Παροντικού Χρόνου Ενεργητικής Φωνής π.χ. «κυβερνώντας» (αλλά ο «κυβερνών» σημειώνεται ως ουσιαστικό ή επίθετο)

VbPpPrPvNm	Μετοχή Παροντικού Χρόνου Παθητικής Φωνής Ονομαστική
VbPpPrPvGe	Μετοχή Παροντικού Χρόνου Παθητικής Φωνής Γενική
VbPpPrPvAc	Μετοχή Παροντικού Χρόνου Παθητικής Φωνής Αιτιατική
Οι παρακάτω μετοχές: Παθητική Τετελεσμένη, Παρελθοντικού Χρόνου Ενεργητικής Φωνής και Παρελθοντικού Χρόνου Παθητικής φωνής που χρησιμοποιούνται είτε ως ουσιαστικά (π.χ. αποβιώσας, γράψας) είτε ως επίθετα σημειώνονται ως ουσιαστικά ή ως επίθετα (ανάλογα με τη χρήση τους).	

ΕΠΙΡΡΗΜΑΤΑ

Ad	Επίρρημα

ΠΡΟΘΕΣΕΙΣ

AsPp	Πρόθεση

ΣΥΝΔΕΣΜΟΙ

Cj	Σύνδεσμος

ΜΟΡΙΑ

Pt	Μόριο

ΕΠΙΦΩΝΗΜΑ

Ij	Επιφώνημα

ΣΗΜΕΙΑ ΣΤΙΞΗΣ

Pu	Σημείο στίξης

ΛΟΙΠΕΣ ΚΑΤΗΓΟΡΙΕΣ

RgSy	Σύμβολο
RgAb	Σύντμηση
RgAn	Ακρώνυμο
RgFw	Ξένη λέξη

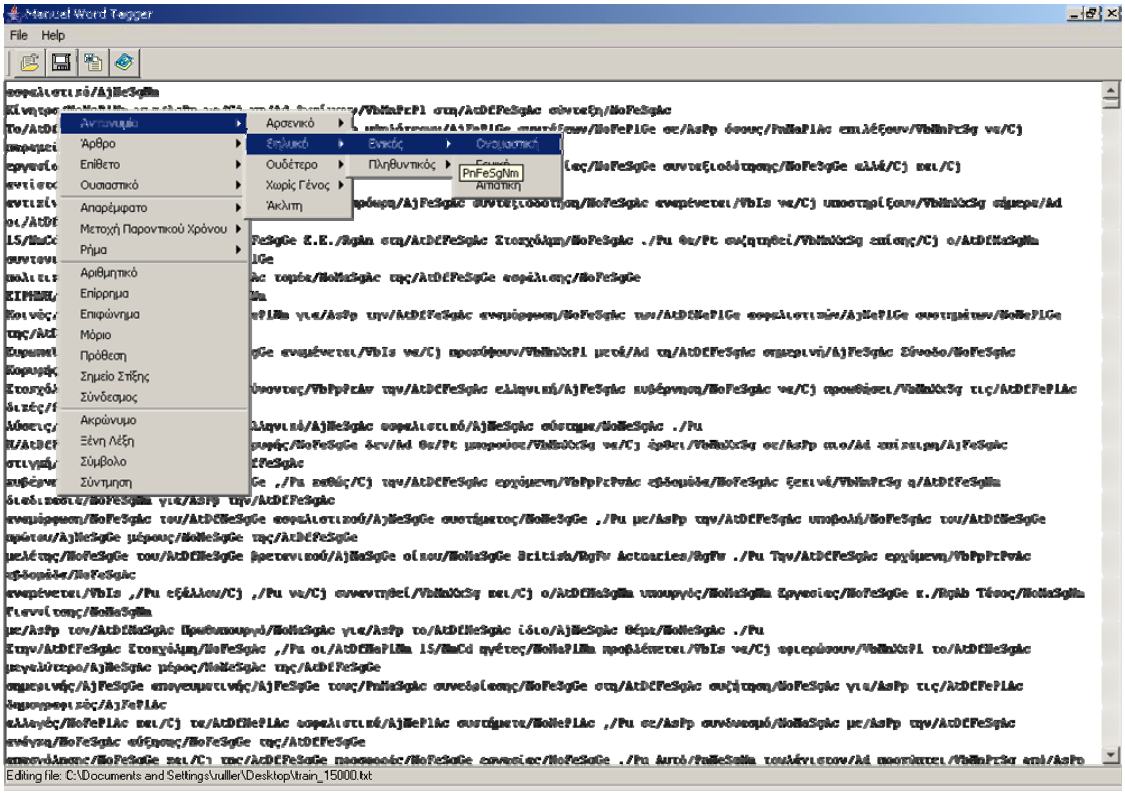
ΠΑΡΑΡΤΗΜΑ ΙΙ: Θέματα υλοποίησης

Για την εκτέλεση των πειραμάτων ήταν απαραίτητη η υλοποίηση των μεθόδων που περιγράψαμε. Για το σκοπό αυτό χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού C++ και Java. Η C++ επιλέχθηκε λόγω του γρήγορου εκτελέσιμου κώδικα που παράγει και χρησιμοποιήθηκε στις υλοποιήσεις των μεθόδων που προτάθηκαν για τα τρία στάδια ενεργητικής μάθησης. Η Java επιλέχθηκε λόγω της ευκολίας που παρέχει στην κατασκευή γραφικών διεπαφών. Χρησιμοποιήθηκε για την κατασκευή ενός φιλικού προς το χρήστη εργαλείου, το οποίο χρησιμοποιήθηκε για την επισημείωση των κειμένων.

Πιο συγκεκριμένα για τα τρία στάδια ενεργητικής μάθησης δημιουργήθηκαν αντίστοιχα οι εφαρμογές ErrorDetector, FindProperties και TrainSelector. Επιπλέον για την κατασκευή των διανυσμάτων δημιουργήθηκε η εφαρμογή VectorCreator. Τέλος για την αναδιοργάνωση των αρχείων μετά από την επιλογή των καλύτερων διανυσμάτων δημιουργήθηκε η εφαρμογή CorpusOrganization. Συγκεκριμένα, η εφαρμογή αυτή διαβάζει από ένα αρχείο τις θέσεις των καλύτερων διανυσμάτων (στο αρχείο που περιέχει όλα τα υποψήφια διανύσματα) τα οποία έχουν επιλεγεί από το σύστημα και στη συνέχεια δημιουργεί δύο αρχεία: ένα με τα καλύτερα διανύσματα και ένα με τα εναπομείναντα. Όλες οι εφαρμογές είναι ανεξάρτητες πλατφόρμας εκτός από την TrainSelector που τρέχει μόνο σε Linux ή Unix καθώς χρησιμοποιεί το API του πακέτου TiMBL το οποίο είναι υλοποιημένο μόνο για Unix και Linux. Για να τρέξει σε Windows απαραίτητη προϋπόθεση είναι να έχει εγκατασταθεί το Cygwin ένα σύστημα που προσομοιώνει μέρος του Unix στα Windows. Κάθε μία από τις παραπάνω εφαρμογές συνοδεύεται από αναλυτικό readme.txt αρχείο στο οποίο εξηγείται αναλυτικά ο τρόπος εκτέλεσης.

Όσον αφορά το εργαλείο χειρονακτικής επισημείωσης η μοναδική απαίτηση είναι το σύστημα να έχει εγκατεστημένη την εικονική μηχανή της Java (Java Virtual Machine). Η εφαρμογή εκτελείται κάνοντας διπλό κλικ στο αρχείο ManualTagger.jar. Η εικόνα I-1. δείχνει το παράθυρο που εμφανίζεται

κατά την εκτέλεση της εφαρμογής. Στη συγκεκριμένη εικόνα έχει φορτωθεί και ένα αρχείο προς επισημείωση. Η επισημείωση γίνεται με δεξί κλικ στο τέλος της λέξης που θέλουμε να επισημειώσουμε οπότε και εμφανίζεται αναδυόμενο μενού που περιέχει όλες τις πιθανές ετικέτες που μπορεί να αποδοθούν σε μία λέξη ομαδοποιημένες σε κατηγορίες. Το αναδυόμενο μενού εμφανίζεται και πατώντας το πλήκτρο Shift. Η εφαρμογή παρέχει επίσης δυνατότητα φορτώματος αρχείου που περιέχει τις θέσεις των λέξεων που αντιστοιχούν στα διανύσματα που μας έχει προτείνει το σύστημα (μέσω της εφαρμογής TrainSelector) για επισημείωση. Αφού φορτωθεί ένα τέτοιο αρχείο ο χρήστης μπορεί να μετακινηθεί από λέξη σε λέξη (με τη σειρά που εμφανίζονται στο κείμενο) με το πλήκτρο F3. Δυνατότητα μετακίνησης προς τα πίσω δεν παρέχεται. Επιπλέον παρέχεται η δυνατότητα μέτρησης των λέξεων που έχουν επισημειωθεί καθώς και των λέξεων που απομένουν.



Εικόνα Ι-1

Τέλος ήταν απαραίτητη μία υλοποίηση του αλγορίθμου εκμάθησης IB1. Ευτυχώς οι Daelemans, Zavrel και λοιποί παρέχουν δωρεάν το πακέτο TIMBL

([DaZa04], [Sl05], http://ilk.uvt.nl), που χρησιμοποιήθηκε στην εργασία, το οποίο περιέχει αρκετούς αλγόριθμους μηχανικής μάθησης, συμπεριλαμβανομένου του IB1 και πολλών παραλλαγών του.

Βιβλιογραφία

[AhKi91] Aha D. W., Kibler D. and Albert M. (1991) Instance-Based Learning Algorithms. *Machine Learning*, 6: 37-66.

[ArDa99] Argamon-Engelson S. and Dagan I. (1999) Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*. 11:335–360.

[BaBr01] Banko M. and Brill E. (2001) Scaling to very large corpora for natural language disambiguation. *Meeting of the Association for Computational Linguistics*. pp. 26-33.

[BaMo04] Banko M. and Moore R. (2004) Part of Speech Tagging in Context. *Proceedings of COLING*.

[Br92] Brill, E. (1992) A Simple Rule-Based Part-of-speech tagger. *Proceedings 3rd Conference on Applied Natural Language Processing, ANLP*, pp. 152-155. ACL.

[Br93] Brill, E. (1993) Automatic Grammar Induction and Parsing Free Text: A Transformation Based Approach. *Proceedings 31st Annual Meeting of the Association for Computational Linguistics*.

[Br95a] Brill E. (1995) Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study on Part of Speech Tagging. *Computational Linguistics* **21** (4): 543-565.

[Br95b] Brill E. (1995) Unsupervised learning of disambiguation rules for part-of-speech tagging. *Proceedings 3rd Workshop on Very Large Corpora*, pp 1-13. Massachusetts. http://www.cs.jhu.edu/~brill/acadpubs.html.

[Bu98] Burges C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. **2**(2):121–167.

[Br03] Brinker K. (2003) Incorporating diversity in active learning with Support Vector Machines. *Proceedings of the 20th International Conference on Machine Learning*, pp. 59-66.

[ClCu03] Clark S., Curran J. R. and Osborne M. (2003) Bootstrapping POS taggers using Unlabelled Data. *Proceedings of the 7th CoNLL conference,* HLT-NAACL. pp. 49-55.

[CoGh95] Cohn D. A., Ghahramani Z. and Jordan M. I. (1995) Active learning with statistical methods. *Advances in Neural Information Processing, volume 7*. pp. 705-712

[CoSa93] Cost S. and Salzberg S. (1993) A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*. 10: 57-78.

[CoTh91] Cover T. M. and Thomas J. (1991) Elements of Information Theory. New York: Wiley.

[CoVa95] Cortes C. and Vapnik V. (1995) Support-vector networks. *Machine Learning*. **20**(3):273–297.

[DaBo92] Daelemans W. and Van den Bosch A. (1992) Generalisation performance of back propagation learning on a syllabification task. *Proceedings of TWLT3: Connectionism and Natural Language Processing*. pp. 27-37.

[DaBo97] Daelemans W.,Van den Bosch A. and Weijters A. (1997) IGTREE: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review* 11: 407-423.

[DaZa96] Daelemans W., Zavrel J., Berck P. and Gillis S. (1996) MBT: A memory-based part-of-speech tagger generator. *Proceedings 4th Workshop on Very Large Corpora*, pp. 14-27. Copenhagen, Denmark.

[DaZa04] Daelemans W., Zavrel J., Van Der Sloot K., Van Den Bosch A. (2004) TiMBL: Tilburg Memory-Based Learner, version 5.1, Reference Guide.

[DeKo95] Dermatas E. and Kokkinakis G. (1995) Automatic Stochastic Tagging of Natural Language Texts. *Computational Linguistics*, *Volume 21, Issue 2*. pp. 137-163.

[DiMe03] Dickinson M. and Meurers W.D. (2003) Detecting Errors in Part-of-Speech Annotation. *Proceedings of EACL*.

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

[He00] Hepple M. (2000) Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

[KiRi03] Kim J. D., Rim H. C. and Tsujii J. (2003) Self-Organizing Markov Models and Their Application to Part-of-Speech Tagging. *ACL.*

[LeGa94] Lewis D. D. and Gale W. A. (1994) A sequential algorithm for training text classifiers. *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*. pp. 3-12.

[LeTs00] Lee S. J., Tsujii J. and Rim H. C. (2000) Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

[Mi97] Mitchell T. M. (1997) Machine Learning. McGraw-Hill.

[NaKu02] Nakagawa T., Kudo T. and Matsumoto Y. (2002) Revision Learning and its Application to Part-of-Speech-Tagging. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

[NgSm04] Nguyen H. T. and Smeulders A. (2004) Active learning using pre-clustering. *21st international conference on Machine learning*.

[OrCh99] Orphanos G.S. and Christodoulakis D.N. (1999) POS Disambiguation and Unknown Words Guessing with Decision Trees. *Proceedings EACL 1999*. pp.134-141

[OrKa99] Orphanos G., Kalles D., Papagelis T. and Christodoulakis D. (1999) Decision Trees and NLP: A Case Study in POS Tagging. *Proceedings of ACAI'99.*

[PaPr00] Papageorgiou H., Prokopidis P., Giouli V. and Piperidis S. (2000) A Unified POS Tagging Architecture and its Application to Greek. *Proceedings of the 2nd Language Resources and Evaluation Conference*. pp.1455-1462.

[PePa99] Petatsis G., Paliouras G., Karkaletsis V., Spyropoulos C.D. and Androutsopoulos I. (1999) Resolving Part-of-Speech ambiguity in the Greek language using learning techniques. *In Fakotakis, N. et al. (Eds.), Machine Learning in Human Language Technology (Proceedings of the ACAI Workshop)*. pp. 29-34.

[PlMo04] Pla F. and Molina A. (2004) Improving part-of-speech tagging using lexicalized HMMs. *Natural Language Engineering* **10** (2): 167-189.

[Qu86] Quinlan J. R. (1986) Induction of Decision Trees. *Machine Learning*. 1: 81-206

[Qu93] Quinlan J. R. (1993) C.4.5. *Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

[Ra96] Ratnaparkhi A. (1996) A maximum entropy part-of-speech tagger. *Proceedings 1st Conference on Empirical Methods in Natural Language Processing, EMNLP*.

[Ro96] Rosenfeld R. (1996) A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language* **10**: 187-228.

[RuNo02] Russel S. and Norvig P. (2002) Artificial Intelligence: A Modern Approach, 2nd edition. Prentice Hall.

[ScCo00] Schohn G. and Cohn D. (2000) Less is More: Active Learning with Support Vector Machines. *Proceedings of the 17th International Conference on Machine Learning*, pp. 839-846.

[SeOp92] Seung H. S., Opper M. and Sompolinsky H. (1992) Query by committee. *Computational learning theory*. pp. 287-294.

[ShZh04] Shen D., Zhang J., Su J., Zhou G. and Tan C. L. (2004) Multi-Criteria-based Active Learning for Named Entity Recognition. *42nd Meeting of the Association of Computational Linguistics*.

[StWa86] Stanfill C. and Waltz D. (1986) Toward Memory-Based Reasoning. *Communications of the ACM*. **29** (12): 1213-1228.

[Sl05] Van Der Sloot K. (2005) TiMBL: Tilburg Memory-Based Learner, version 5.1, API Reference Guide.

[ToKo00] Tong S. and Koller D. (2000) Support vector machine active learning with applications to text classification. *Proceedings of ICML-00, 17th International Conference on Machine Learning*. pp. 996-1006.

[Va98] Vapnik V. N. (1998) *Statistical Learning Theory*. Wiley.

[Vi67] Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Infromation Theory* 260-269, April.

[Vl98] Vlachos A. (2004) Active Learning with Support Vector Machines. http://www.aueb.gr/users/ion/sdep/

[We94] Wettschereck D. (1994) A study of distance-based machine learning algorithms. *Ph.D. thesis, Oregon State University*.

[ΝεΓρ] Νεοελληνική Γραμματική. Αναπροσαρμογή της μικρής νεοελληνικής γραμματικής του Μανόλη Τριανταφυλλίδη. ΟΕΔΒ.