Επεκτάσεις και Περαιτέρω Αξιολόγηση Συστήματος Αναγνώρισης Μερών του Λόγου για Ελληνικά Κείμενα

Ιωάννης Χρονάκης

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων Καθηγητής:
Ίων Ανδρουτσόπουλος

Τμήμα Πληροφορικής

Οικονομικό Πανεπιστήμιο Αθηνών

Σεπτέμβριος 2006

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία επεκτείνει μία προηγούμενη μελέτη του προβλήματος της αναγνώρισης των μερών του λόγου (part-of-speech tagging) ελληνικών κειμένων. Πρόκειται για το πρόβλημα της κατάταξης των λέξεων ενός κειμένου ανάλογα με το μέρος του λόγου στο οποίο ανήκουν (ουσιαστικό, ρήμα, κλπ.). Στη γενικότερη μορφή του προβλήματος, η κατάταξη γίνεται ανάλογα και με κλιτικά χαρακτηριστικά των λέξεων, όπως γένος, πτώση, χρόνος κλπ. (π.χ. αρσενικό ουσιαστικό στην ονομαστική ενικού). Η προηγούμενη μελέτη είχε αναπτύξει ένα σύστημα αναγνώρισης μερών του λόγου για ελληνικά κείμενα, το οποίο χρησιμοποιούσε τεχνικές ενεργητικής μάθησης. Στη διάρκεια της παρούσας εργασίας βελτιώθηκε το προϋπάρχον λογισμικό, επαναλήφθηκαν τα πειράματα της προηγούμενης μελέτης με το νέο λογισμικό, ενώ διεξήχθησαν και επιπλέον πειράματα με νέες συλλογές κειμένων.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1:

ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο 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 αυτή. Στη γενικότερη μορφή του προβλήματος, η ετικέτα μπορεί να παριστάνει και επιπλέον κλιτικές πληροφορίες, όπως το γένος, τον αριθμό, την πτώση, το πρόσωπο ή το χρόνο της λέξης. 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 χρησιμοποιούν ήδη μηχανική μάθηση, αλλά οι τεχνικές ενεργητικής μάθησης δεν έχουν ακόμα αξιοποιηθεί επαρκώς στην περιοχή αυτή.

Μία μελέτη πάνω στην αξιοποίηση μεθόδων ενεργητικής μάθησης για τους σκοπούς της αναγνώρισης μερών του λόγου σε ελληνικά κείμενα πραγματοποιήθηκε από τον Πρόδρομο Μαλακασιώτη [Μα05]. Η παρούσα εργασία αποτελεί επέκταση των μεθόδων και αποτελεσμάτων της μελέτης εκείνης. Οι εν λόγω επεκτάσεις μπορούν να χωριστούν σε δύο γενικά επίπεδα:

Σε θεωρητικό και πειραματικό επίπεδο, αναθεωρήθηκαν ελαφρά οι ετικέτες, ώστε να διευκολύνεται η κατάταξη των λέξεων σε κατηγορίες, ενώ διεξήχθησαν και επιπλέον πειράματα με νέες συλλογές κειμένων.

Σε πρακτικό επίπεδο, αναδομήθηκε και βελτιώθηκε το λογισμικό που είχε παραχθεί στη διάρκεια της προηγούμενης μελέτης, προκειμένου να γίνει ταχύτερο και πιο εύχρηστο.

Αξίζει να σημειωθεί επίσης 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

Η εργασία είναι διαρθρωμένη 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 γίνεται μία συνοπτική παρουσίαση του συστήματος από άποψη λογισμικού.
- Το Κεφάλαιο 4 περιγράφει τα κείμενα τα οποία χρησιμοποιήθηκαν κατά την διαδικασία των πειραμάτων, καθώς και τη μέθοδο προεπεξεργασίας την οποία αυτά υπέστησαν. Επίσης περιγράφει την πειραματική διαδικασία και παρουσιάζει τα αποτελέσματά της.
- Τέλος στο Κεφάλαιο 5 συνοψίζονται τα αποτελέσματα 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 πτυχιακής μου εργασίας.

Σίγουρα θα ήθελα επίσης να ευχαριστήσω τον Πρόδρομο Μαλακασιώτη για την διεξοδική διερεύνηση του παρόντος θέματος στη διπλωματική εργασία του, καθώς και για την καίρια καθοδήγηση που μου πρόσφερε, τόσο στον θεωρητικό όσο και στον τεχνικό άξονα.

Τέλος θα ήθελα να ευχαριστήσω το Γιώργο Λουκαρέλλι για το σύστημα διαχωρισμού περιόδων το οποίο ανέπτυξε και μου διέθεσε.

ΚΕΦΑΛΑΙΟ 2:

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Αναγνώριση Μερών του Λόγου

Ένα σύστημα αναγνώρισης μερών του λόγου πρέπει να έχει τη δυνατότητα, δεδομένου ενός συνόλου κατηγοριών, να μπορεί να κατατάξει κάθε λέξη ενός κειμένου στην κατηγορία στην οποία αυτή η λέξη ανήκει.

Στην απλούστερη περίπτωση, ο αριθμός και το είδος των κατηγοριών αντιστοιχούν στη γενική κατηγοριοποίηση των μερών του λόγου για την εκάστοτε γλώσσα (ουσιαστικά, ρήματα, επιρρήματα, κλπ.). Ένα περισσότερο περίπλοκο και δύσκολο πρόβλημα είναι η αναγνώριση επιπλέον πληροφοριών σχετικών με τους τύπους των κλιτών μερών του λόγου, όπως το γένος, ο αριθμός και η πτώση στα ουσιαστικά, ο αριθμός, το πρόσωπο και ο χρόνος στα ρήματα κ.ο.κ. Στην περίπτωση αυτή έχουμε κατηγορίες όπως «αρσενικό ουσιαστικό στην ονομαστική ενικού», «θηλυκό ουσιαστικό στη γενική πληθυντικού», «ενεστωτικός τύπος ρήματος στο α' πρόσωπο ενικού» κλπ., αντί για κατηγορίες που αντιστοιχούν απλά σε μέρη του λόγου, όπως «ουσιαστικό» ή «ρήμα». Οπότε σε αυτή την περίπτωση μπορεί να ειπωθεί ότι η φάση της αναγνώρισης μερών του λόγου υπεισέρχεται και σε επόμενα στάδια της μορφολογικής ανάλυσης ενός κειμένου.

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση (*Machine Learning*) [Mi97] αποτελεί ένα ευρύ ερευνητικό πεδίο της Τεχνητής Νοημοσύνης. Η Μηχανική Μάθηση αφορά στην ανάπτυξη τεχνικών οι οποίες επιτρέπουν στους υπολογιστές να «διδάσκονται» βασισμένοι στην επεξεργασία συνόλων δεδομένων – για παράδειγμα να μαθαίνουν να κάνουν ιατρικές διαγνώσεις, αφού «εκπαιδευθούν» σε ιατρικές διαγνώσεις ανθρώπων-ιατρών του παρελθόντος. Στην περίπτωση της αναγνώρισης μερών του λόγου, η Μηχανική Μάθηση χρησιμοποιείται για να «εκπαιδευτεί» ο υπολογιστής να κατατάσσει κάθε λέξη στη σωστή κατηγορία.

Για τους σκοπούς της εργασίας, θα μας απασχολήσει η Επιβλεπόμενη Μάθηση (Supervised Learning). Στην κατηγορία αυτή κατατάσσονται αλγόριθμοι οι οποίοι δημιουργούν μια συνάρτηση που αντιστοιχεί κάθε μία είσοδο (π.χ. περίπτωση ασθενούς ή στην περίπτωσή μας εμφάνιση λέξεως) σε μία επιθυμητή έξοδο (π.χ. διάγνωση ή στην περίπτωσή μας κατηγορία λέξεως). Η συνάρτηση αυτή δημιουργείται βάσει ενός σώματος δεδομένων εκπαίδευσης. Τα δεδομένα αυτά αποτελούνται από ζεύγη εισόδων και αντίστοιχων επιθυμητών εξόδων. Στην περίπτωση της αναγνώρισης μερών του λόγου, τα δεδομένα εκπαίδευσης είναι συνήθως κείμενα στα οποία έχουν σημειωθεί χειρωνακτικά οι ορθές κατηγορίες όλων των λέξεων. Ο στόχος ενός αλγορίθμου Επιβλεπόμενης Μάθησης είναι να είναι σε θέση, μετά την εκπαίδευσή του, να αποφανθεί για την αντίστοιχη έξοδο οποιασδήποτε δυνατής εισόδου. Γι' αυτό τον σκοπό ο αλγόριθμος καλείται να χρησιμοποιήσει μεθόδους γενίκευσης πάνω στα δεδομένα εκπαίδευσης, προκειμένου να ανακαλύψει μια συνάρτηση που να συσχετίζει κάθε δυνατή είσοδο με την επιθυμητή έξοδο.

Στη γενική περίπτωση, η διαδικασία που ακολουθείται για την επίλυση ενός προβλήματος κατάταξης σε κατηγορίες με τεχνικές επιβλεπόμενης μάθησης είναι η εξής:

1. Ορισμός των κατηγοριών και επιλογή των ιδιοτήτων (*attributes*) που χαρακτηρίζουν κάθε περίπτωση (π.χ. ηλικία, φύλο, εργαστηριακές μετρήσεις κάθε ασθενούς στην περίπτωση των ιατρικών διαγνώσεων, κατάληξη λέξης, κατηγορία προηγούμενης και προ-προηγούμενης λέξης στην περίπτωση της αναγνώρισης μερών του λόγου). Η αποτελεσματικότητα ενός αλγορίθμου μάθησης προϋποθέτει σε μεγάλο βαθμό την επιλογή ενός όσο το δυνατόν πιο πλήρους συνόλου ιδιοτήτων, οι οποίες να παρέχουν χρήσιμες πληροφορίες για το συγκεκριμένο πρόβλημα.
2. Συλλογή του συνόλου παραδειγμάτων εκπαίδευσης. Το σύνολο εκπαίδευσης πρέπει να αποτελεί αντιπροσωπευτικό δείγμα του συνόλου των περιπτώσεων εισόδου που είναι δυνατόν να εμφανιστούν. Κάθε παράδειγμα εισόδου στο σύνολο εκπαίδευσης πρέπει να έχει εκ των προτέρων αντιστοιχηθεί χειρωνακτικά στην έξοδο η οποία θα αναμενόταν ιδανικά από τον αλγόριθμο μετά την εκπαίδευσή του.
3. Εξαγωγή των τιμών των ιδιοτήτων από κάθε παράδειγμα στο σύνολο εκπαίδευσης. Οι τιμές αυτές λέγονται χαρακτηριστικά (*features*) του παραδείγματος. Στο παρόν βήμα, κάθε παράδειγμα συνήθως μετατρέπεται σε ένα διάνυσμα που περιέχει τα χαρακτηριστικά της περίπτωσης (τιμές ιδιοτήτων).

4. Εκτέλεση του αλγορίθμου μάθησης στα παραδείγματα εκπαίδευσης. Συνήθως ο αλγόριθμος επεξεργάζεται τα διανύσματα που παριστάνουν τα χαρακτηριστικά των παραδειγμάτων εκπαίδευσης και παράγει μια συνάρτηση ταξινόμησης, η οποία αντιστοιχεί κάθε δυνατή περίπτωση (στην περίπτωσή μας κάθε μία εμφάνιση λέξεως σε ένα κείμενο) σε μια έξοδο (στην περίπτωσή μας, μία κατηγορία λέξεων).
5. Χρήση της προκύπτουσας συνάρτησης ταξινόμησης σε νέα δεδομένα αγνώστων εξόδων (π.χ. κείμενα στα οποία δεν είναι σημειωμένες οι κατηγορίες των λέξεων). Όπως και κατά την εκπαίδευση του αλγορίθμου, τα νέα δεδομένα μετατρέπονται πρώτα σε διανύσματα χαρακτηριστικών.

Συχνά χρησιμοποιούμενοι αλγόριθμοι επιβλεπόμενης μάθησης είναι ο αλγόριθμος των k κοντινότερων γειτόνων, ο αφελής ταξινομητής Bayes, ο ID3 κλπ. [Mi97].

Για τους σκοπούς της εργασίας χρησιμοποιήθηκε ο αλγόριθμος των k κοντινότερων γειτόνων (*k nearest neighbours – k-NN*), ο οποίος περιγράφεται παρακάτω.

2.3 Αλγόριθμος k Κοντινότερων Γειτόνων

Στον αλγόριθμο των k κοντινότερων γειτόνων, όπως και σε όλους τους αλγορίθμους μάθησης που χρησιμοποιούν διανυσματικές αναπαραστάσεις των περιπτώσεων εισόδου, κάθε ιδιότητα των διανυσμάτων αντιστοιχεί σε μία διάσταση ενός πολυδιάστατου χώρου. Κατά το στάδιο της εκπαίδευσης, ο k-NN απλά αποθηκεύει τα διανύσματα όλων των παραδειγμάτων εκπαίδευσης, μαζί με τις ορθές εξόδους που αντιστοιχούν στο καθένα. Ουσιαστικά, δηλαδή, αποθηκεύει τα σημεία του πολυδιάστατου χώρου που αντιστοιχούν στα παραδείγματα εκπαίδευσης, μαζί με τις κατηγορίες τους.

Κατά την φάση ταξινόμησης, δηλαδή κατά τη χρήση του εκπαιδευμένου k-NN, το σύστημα λαμβάνει εισόδους (περιπτώσεις) για τις οποίες δεν γνωρίζει την έξοδο και υπολογίζει για κάθε μία τη διανυσματική της αναπαράσταση, δηλαδή το αντίστοιχο σημείο στον πολυδιάστατο χώρο. Κατόπιν υπολογίζεται η απόσταση του σημείου της εισόδου από κάθε σημείο που αντιστοιχεί σε αποθηκευμένο παράδειγμα εκπαίδευσης. Αφού υπολογιστούν οι αποστάσεις αυτές, είναι εύκολο να βρεθούν τα k σημεία (περιπτώσεις) εκπαίδευσης με τη μικρότερη απόσταση (βάσει κάποιας μετρικής) από το σημείο της εισόδου. Η είσοδος κατατάσσεται στη συνέχεια στην κατηγορία που είναι πιο συχνή μεταξύ των k κοντινότερων παραδειγμάτων εκπαίδευσης, όπου το k είναι συνήθως ένας περιττός φυσικός αριθμός για να αποφεύγονται οι ισοπαλίες.

Προφανώς, ο αλγόριθμος απαιτεί περισσότερους υπολογισμούς κατά την κατάταξη νέων περιπτώσεων όσο αυξάνει το πλήθος των παραδειγμάτων εκπαίδευσης, αφού υπολογίζεται κάθε φορά η απόσταση της νέας περίπτωσης από όλα τα παραδείγματα εκπαίδευσης. Έχει, επίσης, μεγάλες απαιτήσεις μνήμης, αφού πρέπει να αποθηκεύονται όλα τα παραδείγματα εκπαίδευσης. Από την άλλη πλευρά, όμως, ο αλγόριθμος είναι εξαιρετικά απλός, είναι ταχύτατος κατά την εκπαίδευση (αφού απλά απομνημονεύει τα παραδείγματα εκπαίδευσης) και μπορεί να μάθει υπερ-επιφάνειες διαχωρισμού οποιουδήποτε είδους (σε αντίθεση με γραμμικούς διαχωριστές όπως, για παράδειγμα, το Perceptron).

Όπως και στην προηγούμενη εργασία [Μα05] στην οποία βασίζεται η παρούσα, χρησιμοποιείται μια βελτιωμένη μορφή του k-NN, η οποία: (α) δίνει διαφορετική βαρύτητα σε κάθε ιδιότητα κατά των υπολογισμό των αποστάσεων, ανάλογα με την Αναλογία Πληροφοριακού Κέρδους (*Gain Ratio*) κάθε ιδιότητας και (β) ζυγίζει την ψήφο κάθε ενός από τους k κοντινότερους γείτονες αντιστρόφως ανάλογα με την απόσταση του γείτονα από την προς κατάταξη περίπτωση. Πιο συγκεκριμένα, στο σύστημα της παρούσας εργασίας η απόσταση του προς κατάταξη διανύσματος $\vec{X}$ από ένα παράδειγμα εκπαίδευσης $\vec{Y}$ δίνεται από τον τύπο:

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

όπου $\Delta(\vec{X},\vec{Y})$ είναι η απόσταση μεταξύ των δύο διανυσμάτων. Τα διανύσματα αυτά έχουν n ιδιότητες το καθένα, και $\delta(x_i, y_i)$ είναι η διαφορά των τιμών της i-οστής ιδιότητας μεταξύ των διανυσμάτων $\vec{X}$ και $\vec{Y}$. Ως w_i ορίζεται η Αναλογία Πληροφοριακού Κέρδους της i-στής ιδιότητας.

Η Αναλογία Πληροφοριακού Κέρδους [Qu93] μιας ιδιότητας ορίζεται ως εξής:

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) H(C \mid v)}{-\sum_{v \in V_i} P(v) \log_2 P(v)}$$

όπου $H(C)$ η εντροπία των κατηγοριών, $H(C \mid v)$ η εντροπία των κατηγοριών δεδομένου ότι η ιδιότητα i έχει την τιμή v, και $P(v)$ η πιθανότητα η τιμή της ιδιότητας i να είναι v. Το σύνολο V_i περιέχει όλες τις δυνατές τιμές της ιδιότητας i.

Τέλος, η βαρύτητα της ψήφου κάθε ενός από τους k κοντινότερους γείτονες δίνεται από τον τύπο:

$$d_i = \frac{1}{c + \Delta}$$

όπου Δ η απόσταση του εν λόγω γείτονα, και c μία (θετική) σταθερά ώστε να αποφευχθεί η διαίρεση με το μηδέν.

Στα πειράματα αυτής της εργασίας, το k είχε την τιμή 5.

2.4 Ιδιότητες Λεκτικών Μονάδων

Προκειμένου να χρησιμοποιηθεί ο αλγόριθμος των k Κοντινότερων Γειτόνων στην αναγνώριση μερών του λόγου, είναι αναγκαίο να παρασταθεί κάθε λεκτική μονάδα των κειμένων ως ένα διάνυσμα χαρακτηριστικών (διάνυσμα τιμών ιδιοτήτων).

Βάσει προηγούμενης έρευνας [Μα05], επιλέχθηκαν οι παρακάτω ιδιότητες:

- Η κατάληξη της λεκτικής μονάδας. Ως κατάληξη θεωρούμε τους τρεις τελευταίους χαρακτήρες της λεκτικής μονάδας, ή ολόκληρη την λεκτική μονάδα στην περίπτωση όπου αυτή έχει μήκος μικρότερο από τρεις χαρακτήρες.
- Το μήκος (σε χαρακτήρες) της λεκτικής μονάδας.
- Η ύπαρξη ή όχι αποστρόφου μέσα στη λεκτική μονάδα.
- Η ύπαρξη ή όχι αριθμητικού ψηφίου μέσα στη λεκτική μονάδα.
- Η ύπαρξη ή όχι κόμματος μέσα στη λεκτική μονάδα.
- Η ύπαρξη ή όχι τελείας μέσα στη λεκτική μονάδα.
- Η ύπαρξη ή όχι λατινικού χαρακτήρα μέσα στη λεκτική μονάδα.
- Η ετικέτα αμφισημίας (*ambitag*) της λεκτικής μονάδας. Η έννοια αυτή εξηγείται παρακάτω.
- Η κατάληξη της επόμενης λεκτικής μονάδας.
- Η ετικέτα αμφισημίας της επόμενης λεκτικής μονάδας.
- Η λεκτική μονάδα δύο θέσεις πριν από την τρέχουσα.
- Η κατάληξη της λεκτικής μονάδας δύο θέσεις πριν από την τρέχουσα.
- Η λεκτική μονάδα μία θέση πριν από την τρέχουσα.
- Η κατάληξη της λεκτικής μονάδας μία θέση πριν από την τρέχουσα.
- Η ετικέτα αμφισημίας της λεκτικής μονάδας μία θέση πριν από την τρέχουσα.

Οπότε έχουμε συνολικά δεκαπέντε ιδιότητες. Η Αναλογία Πληροφοριακού Κέρδους κάθε ιδιότητας, και επομένως το βάρος της κατά τον υπολογισμό της απόστασης στον αλγόριθμο των k Κοντινότερων Γειτόνων, υπολογίζεται με δυναμικό τρόπο πάνω στο σώμα εκπαίδευσης κατά την εκτέλεση του αλγορίθμου.

Ως «ετικέτα αμφισημίας» (*ambivalence tag – ambitag*) ορίζουμε μία συμβολοσειρά η κατασκευή της οποίας ακολουθεί τον εξής αλγόριθμο:

```
Για κάθε λεκτική μονάδα t,
        εάν υπάρχουν αντίγραφα της t στο σώμα εκπαίδευσης,
        τότε δημιούργησε το ambitag ενώνοντας γραμμικά τις
        διαφορετικές ετικέτες των αντιγράφων αυτών σε μία
        συμβολοσειρά

        ειδάλλως,
        εάν υπάρχουν λεκτικές μονάδες με ίδια κατάληξη με την t στο
        σώμα εκπαίδευσης,
        τότε δημιούργησε το ambitag ενώνοντας γραμμικά τις
        διαφορετικές ετικέτες των λεκτικών αυτών μονάδων σε μία
        συμβολοσειρά

        ειδάλλως το ambitag είναι η συμβολοσειρά "unknown"
```

Η σύλληψη και ο ορισμός της ετικέτας αμφισημίας έγινε από τους Daelemans κ.ά. [DaZa03]

2.5 Κατηγορίες Λεκτικών Μονάδων

Οι βασικές κατηγορίες στις οποίες κατατάσσει το σύστημα τις λεκτικές μονάδες είναι:

- αντωνυμία
- άρθρο
- αριθμητικό
- επίθετο
- επίρρημα
- σύνδεσμος
- μόριο
- ουσιαστικό
- πρόθεση
- ρήμα
- σημείο στίξης
- άλλο

Ο χρήστης μπορεί να επιλέξει αν θα χρησιμοποιηθεί το παραπάνω σύνολο κατηγοριών (ισοδύναμα, ετικετών, tags) ή ένα εκτενέστερο, το οποίο περιλαμβάνει και τις εξής υποκατηγορίες:

- Για κάθε αντωνυμία, η κατηγορία (ετικέτα) δείχνει επίσης τον τύπο της αντωνυμίας, εάν δηλαδή είναι άκλιτη ή όχι. Στην περίπτωση που είναι κλιτή, δείχνει ακόμη το γένος, τον αριθμό και την πτώση της.
- Για κάθε άρθρο, η κατηγορία δείχνει επίσης τον τύπο του, εάν δηλαδή είναι οριστικό, αόριστο ή εμπρόθετο, καθώς και το γένος, τον αριθμό και την πτώση του.
- Για κάθε επίθετο και ουσιαστικό, η κατηγορία δείχνει επίσης το γένος, τον αριθμό και την πτώση του.
- Για κάθε ρήμα, η κατηγορία δείχνει επίσης το χρόνο και τον αριθμό του. Τα απαρέμφατα και οι ενεργητικές μετοχές κατατάσσονται ως ρήματα. Στην περίπτωση των απαρεμφάτων, σημειώνεται εάν το απαρέμφατο είναι ενεργητικής ή παθητικής φωνής. Οι παθητικές μετοχές δεν σημειώνονται ως ρήματα, αλλά ως επίθετα ή ουσιαστικά, ανάλογα με το συντακτικό τους ρόλο.
- Στην κατηγορία «άλλο» συμπεριλαμβάνονται υπο-κατηγορίες που διαχωρίζουν ακρωνύμια, συντομεύσεις, ξένες λέξεις και διάφορους χαρακτήρες οι οποίοι δεν αποτελούν σημεία στίξης.

Περισσότερες πληροφορίες για τις κατηγορίες και την αναπαράστασή τους δίνονται στο παράρτημα. Οι ετικέτες του εκτενούς συνόλου είναι ελαφρά διαφορετικές από εκείνες της εργασίας [Μα05], προκειμένου να διευκολύνεται η κατάταξη των λεκτικών μονάδων σε κατηγορίες.

Στην περίπτωση περιφραστικών τύπων, το σύστημα επισημειώνει (κατατάσσει) ξεχωριστά κάθε μία λέξη του τύπου. Στους ρηματικούς τύπους του εξακολουθητικού μέλλοντα, που έχουν τη μορφή «θα» + <ενεστωτικός τύπος> (π.χ. «θα παίζω»), το «θα» σημειώνεται ως μόριο και ο ενεστωτικός τύπος ως ρήμα στον ενεστώτα. Όπως εξηγείται στην ενότητα 3.1, είναι δυνατόν σε μια φάση μετα-επεξεργασίας να ομαδοποιούνται οι λέξεις τέτοιων περιφραστικών τύπων και να τους αποδίδεται η κατάλληλη συνολική ετικέτα.

Ομοίως, σε ρηματικούς τύπους του αορίστου υποτακτικής, όπως «να αποκτήσω», το «αποκτήσω» κατατάσσεται ως μέλλοντας. Στην περίπτωση αυτή οι λέξεις του περιφραστικού τύπου δεν ομαδοποιούνται στο στάδιο της μετα-επεξεργασίας, επειδή οι ετικέτες του συστήματος δεν περιέχουν πληροφορίες έγκλισης κι έτσι δεν είναι δυνατόν να σημειωθεί ο περιφραστικός τύπος ως τύπος της υποτακτικής.

2.6 Ενεργητική Μάθηση

Συνήθως είναι εύκολο να παραχθούν ή να βρεθούν μη κατηγοριοποιημένα δεδομένα εκπαίδευσης (στην περίπτωσή μας, κείμενα στα οποία δεν έχουν επισημειωθεί οι κατηγορίες των λέξεων), αλλά η κατάταξή τους σε κατηγορίες θεωρείται ακριβή και επίπονη διαδικασία.

Η Ενεργητική Μάθηση (*Active Learning*) είναι ο τομέας της μηχανικής μάθησης ο οποίος προτείνει μεθόδους που επιτρέπουν στον αλγόριθμο εκπαίδευσης να επιλέγει ο ίδιος τα παραδείγματα (εμφανίσεις λέξεων στην περίπτωσή μας) που πρέπει να επισημειωθούν χειρωνακτικά και να συμπεριληφθούν στο σύνολο εκπαίδευσης. Με τον τρόπο αυτό είναι δυνατόν να επιτύχουμε το ίδιο επίπεδο ορθότητας (accuracy) με λιγότερα δεδομένα εκπαίδευσης, σε σχέση με το πλήθος των δεδομένων που θα απαιτούνταν αν τα παραδείγματα εκπαίδευσης επιλέγονταν τυχαία (ή με τη σειρά, στην περίπτωση των λέξεων ενός κειμένου), κάτι που μειώνει το φόρτο των ανθρώπων που επισημειώνουν τα παραδείγματα εκπαίδευσης. Στην περίπτωση του k-NN η μείωση των παραδειγμάτων εκπαίδευσης οδηγεί επίσης σε αύξηση της ταχύτητας κατάταξης (λιγότεροι υπολογισμοί αποστάσεων) και μείωση της απαιτούμενης μνήμης (αποθήκευση λιγότερων παραδειγμάτων).

Περισσότερες πληροφορίες για την ενεργητική μάθηση παρέχονται στην εργασία [Μα05]. Στο κείμενο εκείνο προτείνεται και ένα μέτρο σημαντικότητας, το οποίο χρησιμοποιείται για την αξιολόγηση των υποψηφίων παραδειγμάτων εκπαίδευσης και την επιλογή των «καλύτερων», που θα επισημειωθούν στη συνέχεια χειρωνακτικά. Το μέτρο αυτό, το οποίο χρησιμοποιείται και στην παρούσα εργασία, ορίζεται ως εξής:

$$W_n = \begin{cases} \dfrac{H_n(x)}{\log\left(\sum_{c\in C} V^c\right)}, H_n(x) \neq 0 \\ -\sum_{c\in C} V^c, H_n(x) = 0 \end{cases}$$

Έστω x ένα τυχαίο υποψήφιο παράδειγμα εκπαίδευσης. Στόχος του μέτρου W_n είναι να μπορεί το μέτρο αυτό να αντιπροσωπεύσει τη χρησιμότητα του παραδείγματος αυτού, βάσει του ήδη υπάρχοντος συνόλου εκπαίδευσης. Υψηλότερες τιμές του μέτρου αντιστοιχούν σε μεγαλύτερη χρησιμότητα του παραδείγματος.

Ως $H_n(x)$ ορίζεται η κανονικοποιημένη τιμή της εντροπίας της κατηγορίας του x, δηλαδή το κατά πόσον είμαστε αβέβαιοι για την κατηγορία του x. Η ποσότητα αυτή ορίζεται σύμφωνα με τους εξής τύπους:

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

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

Ως V^c ορίζεται το άθροισμα των ψήφων των γειτόνων οι οποίοι ανήκουν στην κατηγορία c, οπότε το P(c) αντιστοιχεί στο βαθμό βεβαιότητας του ταξινομητή ότι το x ανήκει στην κατηγορία c. Το σύνολο C περιέχει όλες τις δυνατές κατηγορίες στις οποίες μπορούν να ανήκουν οι γείτονες.

Χαμηλή εντροπία σημαίνει ότι ο αλγόριθμος k-NN βασιζόμενος στο υπάρχον σύνολο εκπαίδευσης μπορεί να αποφανθεί για την κατηγορία στην οποία ανήκει το παράδειγμα με μεγάλη βεβαιότητα. Αντίθετα, εάν το παράδειγμα έχει υψηλή εντροπία, αυτό σημαίνει ότι ο αλγόριθμος k-NN δεν μπορεί να προτείνει κάποια κατηγορία με βεβαιότητα, κάτι που συνήθως είναι ένδειξη ότι πρόκειται για χρήσιμο παράδειγμα.

Η ποσότητα $\sum_{c \in C} V^c$ είναι ίση με το άθροισμα των ψήφων των k γειτόνων. Λόγω της ζύγισης της ψήφου, όπως αναφέρεται στην ενότητα 2.3, μικρή τιμή του $\sum_{c \in C} V^c$ υποδηλώνει μεγάλη απόσταση των k γειτόνων από το υποψήφιο παράδειγμα x. Οπότε μικρή τιμή του $\sum_{c \in C} V^c$ υποδηλώνει ότι είναι επιθυμητό να εισαχθεί το x στα παραδείγματα εκπαίδευσης, αφού πρόκειται για σημείο σε περιοχή του υπερχώρου όπου δεν έχουμε κοντινά παραδείγματα εκπαίδευσης.
Το μέτρο $\sum_{c \in C} V^c$ λογαριθμίζεται προκειμένου οι τιμές που παίρνει να είναι ανάλογου μεγέθους με αυτές που παίρνει και η εντροπία $H_n(x)$.

Τέλος, εάν η εντροπία παίρνει τιμή τέτοια που μηδενίζει το κλάσμα, το μέτρο μετατρέπεται σε $-\sum_{c \in C} V^c$, προκειμένου να αποφευχθεί η εκφυλισμένη μηδενική τιμή. Από τα παραδείγματα που έχουν μηδενική εντροπία, περισσότερο χρήσιμα είναι εκείνα με χαμηλή τιμή $\sum_{c \in C} V^c$, όπως περιγράφηκε ανωτέρω.

ΚΕΦΑΛΑΙΟ 3:

ΤΟ ΣΥΣΤΗΜΑ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στο κεφάλαιο αυτό θα παρουσιαστεί το σύστημα αναγνώρισης μερών του λόγου που αναπτύχθηκε στη διάρκεια της εργασίας.

Το λογισμικό του συστήματος αποτελείται από μία βιβλιοθήκη και δύο κύρια εργαλεία, τα οποία παρουσιάζονται παρακάτω. Ενσωματώνει επίσης το λογισμικό TiMBL [DaZa04], το οποίο παρέχει, μεταξύ άλλων, υλοποίηση της μορφής του k-NN που χρησιμοποιούμε.

3.1 Βιβλιοθήκη POSTagger.dll

Η βιβλιοθήκη αυτή περιέχει τις απαραίτητες βασικές λειτουργίες και συναρτήσεις για την υλοποίηση του αλγορίθμου ενεργητικής μάθησης. Η βιβλιοθήκη ακολουθεί το αντικειμενοστρεφές μοντέλο. Ως γλώσσα υλοποίησης επιλέχθηκε η C++, καθώς οι συναρτήσεις τις οποίες περιέχει η βιβλιοθήκη απαιτούν σημαντική υπολογιστική ισχύ, οπότε μία γλώσσα όπως η C++ θεωρήθηκε φυσική επιλογή λόγω της ταχύτητας επεξεργασίας η οποία την χαρακτηρίζει.

Η βιβλιοθήκη παρέχει προγραμματιστική διεπαφή εφαρμογής (*Application Programming Interface – API*), προκειμένου οι λειτουργίες αυτής να μπορούν σχετικά εύκολα να χρησιμοποιηθούν και από διαφορετικές εφαρμογές με παρόμοιους στόχους. Μάλιστα, η βιβλιοθήκη POSTagger.dll περιέχει και τις απαραίτητες συναρτήσεις–διαμεσολαβητές προκειμένου το API να μπορεί να χρησιμοποιηθεί και από προγράμματα γραμμένα στην γλώσσα Java.

Πιο συγκεκριμένα, οι λειτουργίες τις οποίες υποστηρίζει η βιβλιοθήκη είναι οι εξής:

Αξιολόγηση Λεκτικών Μονάδων Κειμένου κατά την Ενεργητική Μάθηση. Αναλαμβάνει να προσδιορίσει την αξία η οποία χαρακτηρίζει (ως παράδειγμα εκπαίδευσης) κάθε λεκτική μονάδα σε ένα κείμενο. Υποστηρίζονται δύο μέτρα μέσω των οποίων μπορεί να γίνει η αξιολόγηση:

1. $$W_n = \begin{cases} \dfrac{H_n(x)}{\log\left(\sum_{c\in C} V^c\right)}, H_n(x) \neq 0 \\ -\sum_{c\in C} V^c, H_n(x) = 0 \end{cases}$$

   Το μέτρο αυτό έχει περιγραφεί στην ενότητα 2.6.

2. $H_n(x)$

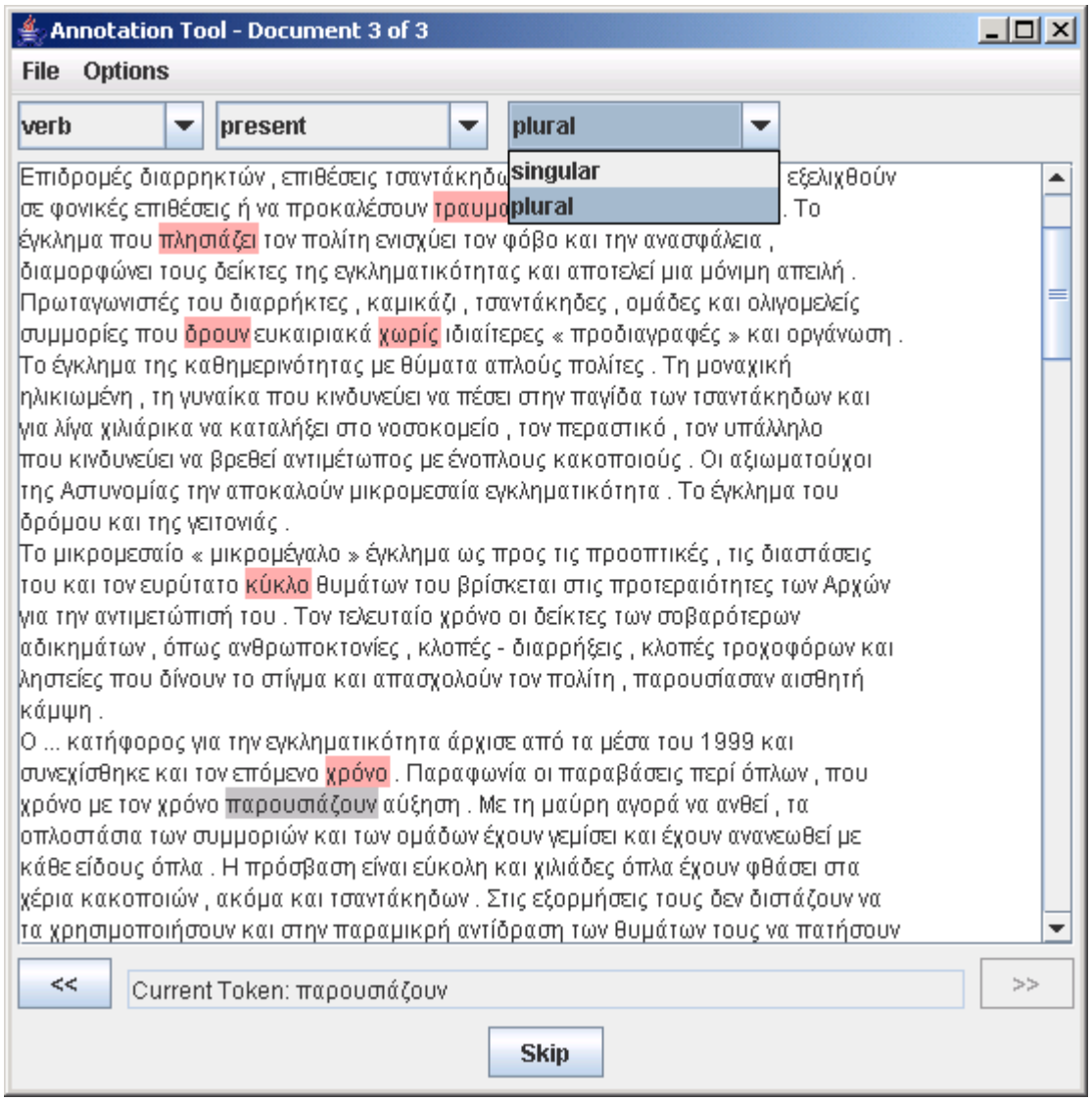
   Το μέτρο αυτό είναι απλούστερο από το πρώτο.

Παρέχονται ξεχωριστές συναρτήσεις για τον υπολογισμό των παραγόντων των παραπάνω μέτρων.

- **Κατάταξη Λεκτικών Μονάδων Κειμένου.** Αναλαμβάνει να κατατάξει σε κατηγορίες τις λεκτικές μονάδες ενός κειμένου βάσει ενός υπάρχοντος σώματος εκπαίδευσης. Για τους σκοπούς της λειτουργίας αυτής, συμπεριλαμβάνεται στο σύστημα και ένα προκαθορισμένο (default) σώμα εκπαίδευσης, το οποίο είναι δυνατόν να αλλάξει.
- **Αξιολόγηση Συστήματος.** Χρησιμοποιώντας την παραπάνω λειτουργία, η βιβλιοθήκη είναι σε θέση να μετρήσει το ποσοστό ορθότητας (accuracy) που επιτυγχάνει το σύστημα σε ένα σώμα ελέγχου (test corpus). Οι κατηγορίες των λεκτικών μονάδων του σώματος ελέγχου πρέπει να έχουν επισημειωθεί χειρωνακτικά.
- **Χωρισμός Κειμένου σε Λεκτικές Μονάδες.** Η συνάρτηση αυτή αναλαμβάνει να διαχωρίσει ένα κείμενο στις συστατικές του λεκτικές μονάδες (tokens).
  Η διαδικασία αυτή περιγράφεται στην ενότητα 3.2.
- **Μετατροπή σε XML.** Η λειτουργία αυτή μετατρέπει ένα (μερικώς ή ολικώς) επισημειωμένο κείμενο σε κείμενο XML, με τέτοιο τρόπο ώστε οι επισημειώσεις να μετατρέπονται σε ετικέτες XML. Οι χρησιμοποιούμενες ετικέτες XML περιγράφονται στο παράρτημα.
  Σε περίπτωση που ο χρήστης το επιθυμεί, η λειτουργία αυτή είναι σε θέση να πραγματοποιήσει και ένα επίπεδο μετα-επεξεργασίας. Η μετα-επεξεργασία αυτή περιλαμβάνει προς το παρόν μόνο τα εξής:
  1. Ανίχνευση μελλοντικών τύπων. Χάριν απλότητας, το σύστημα έχει εκπαιδευθεί να κατατάσσει, για παράδειγμα, πάντα το «παίζει» ως ενεστωτικό ρηματικό τύπο. Αν το «παίζει» συνοδεύεται από το μόριο «θα», η μετα-επεξεργασία σημειώνει το «θα παίζει» συνολικά ως μελλοντικό τύπο. Η μετα-επεξεργασία δεν υποστηρίζει προς το παρόν συντελεσμένους χρόνους (π.χ. «θα έχει παίξει»).
  2. Σημείωση του «για να» ως ενός ενιαίου συνδέσμου.

3.2 Εργαλείο Επισημείωσης Κατηγοριών

Το εργαλείο αυτό επιτρέπει σε ένα χρήστη να σημειώνει τις ορθές κατηγορίες των λέξεων σε ένα κείμενο, όπως φαίνεται παρακάτω, προκειμένου το κείμενο να χρησιμοποιηθεί κατά την εκπαίδευση του συστήματος.



Μέσω του εργαλείου αυτού, ο χρήστης είναι σε θέση να φορτώσει ένα ή περισσότερα αρχεία απλού κειμένου. Τα κείμενα αυτά χωρίζονται αυτόματα σε λεκτικές μονάδες. Κατόπιν, ο χρήστης μπορεί να αρχίσει να σημειώνει τις κατηγορίες των λεκτικών μονάδων (να τους αποδίδει ετικέτες), χρησιμοποιώντας τα αντίστοιχα μενού του εργαλείου.

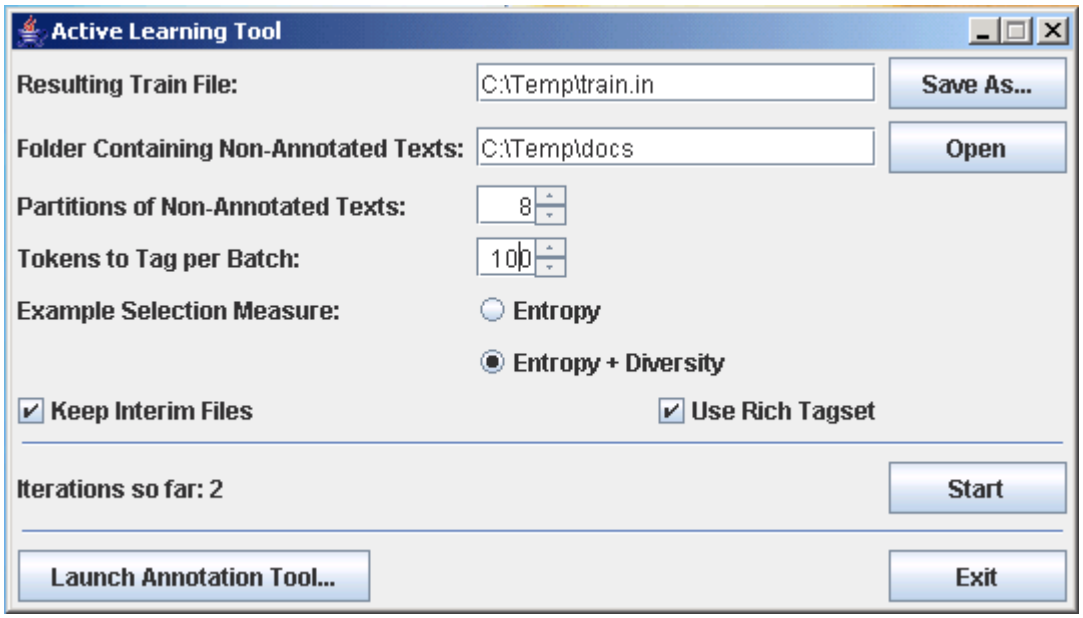
Το σύστημα είναι επίσης σε θέση να αποδώσει αυτόματα ετικέτες στις λεκτικές μονάδες του κειμένου, βάσει ενός ορισμένου από το χρήστη σώματος εκπαίδευσης. Κατόπιν αυτού ο χρήστης μπορεί να επέμβει εκ νέου, διορθώνοντας ετικέτες όπου αυτό είναι σκόπιμο.

Ο χρήστης μπορεί να ελέγξει ο ίδιος το βαθμό πολυπλοκότητας των ετικετών τις οποίες αποδίδει. Στην απλή βαθμίδα πολυπλοκότητας, προσφέρονται μόνο οι βασικές κατηγορίες (ρήμα, ουσιαστικό, επίθετο κλπ.), ενώ στη δεύτερη βαθμίδα οι ετικέτες περιλαμβάνουν και πληροφορίες όπως γένος, αριθμό κλπ. Οι ετικέτες και των δύο βαθμίδων περιγράφονται αναλυτικότερα στην ενότητα 2.5.

Τέλος, ο χρήστης μπορεί να εξαγάγει το τρέχον επισημειωμένο κείμενο σε μορφή XML, όπως αναφέρεται στην ενότητα 3.1.

3.3 Εργαλείο Δημιουργίας Σωμάτων Εκπαίδευσης με Ενεργητική Μάθηση

Το εργαλείο αυτό υποστηρίζει την εκπαίδευση του συστήματος με ενεργητική μάθηση. Βοηθά το χρήστη να κατασκευάσει ένα βαθμιαία μεγαλύτερο και πλουσιότερο σώμα εκπαίδευσης, επιλέγοντας παραδείγματα εκπαίδευσης (εμφανίσεις λέξεων) από μια συλλογή μη επισημειωμένων κειμένων και ζητώντας από το χρήστη να επισημειώσει (κατατάξει) τα επιλεγόμενα παραδείγματα.



Οι περισσότερες παράμετροι μπορούν να οριστούν από τον χρήστη. Συγκεκριμένα, ο χρήστης μπορεί να ορίσει:

- Το όνομα και την τοποθεσία του αρχείου στο οποίο θα εμπεριέχεται το τελικό σώμα εκπαίδευσης.
- Το όνομα και την τοποθεσία του φακέλου που περιέχει τη συλλογή μη επισημειωμένων κειμένων.
- Τον αριθμό των μερών στις οποίες θα χωριστεί το σύνολο των μη επισημειωμένων κειμένων. Επειδή η αξιολόγηση των υποψηφίων παραδειγμάτων εκπαίδευσης είναι χρονοβόρα, η συλλογή των μη επισημειωμένων κειμένων είναι δυνατόν να χωριστεί σε μέρη, και σε κάθε επανάληψη το σύστημα να αξιολογεί και να επιλέγει (κυκλικά) παραδείγματα από μία διαφορετική διαμέριση της συλλογής.
- Τον αριθμό των λεκτικών μονάδων τις οποίες ο χρήστης θα επισημειώνει (κατατάσσει) σε κάθε επανάληψη της ενεργητικής μάθησης. Προκειμένου να μειωθεί ο χρόνος αξιολόγησης και επιλογής παραδειγμάτων εκπαίδευσης, το σύστημα επιλέγει σε κάθε επανάληψη μια δέσμη (batch) παραδειγμάτων εκπαίδευσης, αντί για ένα μεμονωμένο παράδειγμα, και ζητά από το χρήστη να επισημειώσει όλα τα παραδείγματα (λεκτικές μονάδες) της δέσμης, τα οποία προστίθενται στη συνέχεια στο σώμα εκπαίδευσης.
- Το μέτρο σύμφωνα με το οποίο το πρόγραμμα θα αξιολογεί κάθε υποψήφιο παράδειγμα (λεκτική μονάδα των μη επισημειωμένων κειμένων). Υποστηρίζονται τα δύο μέτρα της ενότητας 3.1.

- Εάν ο χρήστης το επιθυμεί, το πρόγραμμα μπορεί να αποθηκεύσει τα ενδιάμεσα αρχεία εκπαίδευσης που παράγονται κατά τη διάρκεια της ενεργητικής μάθησης, και όχι μόνο το τελικό σώμα εκπαίδευσης. Η επιλογή αυτή είναι χρήσιμη για ερευνητικούς σκοπούς.
- Ο χρήστης μπορεί επίσης να επιλέξει το βαθμό πολυπλοκότητας των ετικετών, όπως αναφέρθηκε στην ενότητα 3.2.

Η ενεργητική μάθηση εξελίσσεται ως εξής:

Αφού ο χρήστης δώσει τιμές στις παραπάνω παραμέτρους, το σύστημα χωρίζει τη συλλογή μη επισημειωμένων κειμένων σε ισομεγέθη μέρη, όπως εξηγήθηκε παραπάνω. Ύστερα, και για όσες επαναλήψεις ο χρήστης επιθυμεί, το σύστημα επεξεργάζεται την (κυκλικά επιλεγόμενη σε κάθε επανάληψη) τρέχουσα διαμέριση της συλλογής. Συγκεκριμένα, χρησιμοποιώντας τη βιβλιοθήκη της ενότητας 2.1, αξιολογεί τις λεκτικές μονάδες (υποψήφια παραδείγματα εκπαίδευσης) των κειμένων του τρέχοντος μέρους και ζητά από το χρήστη να επισημειώσει τις σημαντικότερες από αυτές χρησιμοποιώντας το εργαλείο της ενότητας 3.2. Οι επισημειωμένες λεκτικές μονάδες προστίθενται στη συνέχεια στο σύνολο των δεδομένων εκπαίδευσης και η διαδικασία επαναλαμβάνεται.

Καθώς η φάση της αξιολόγησης των υποψηφίων παραδειγμάτων απαιτεί ένα μη κενό σώμα εκπαίδευσης, στην πρώτη επανάληψη οι λεκτικές μονάδες επιλέγονται με αυθαίρετο τρόπο. Συγκεκριμένα, επιλέγονται οι x πρώτες μονάδες του πρώτου κειμένου του πρώτου μέρους της συλλογής εκπαίδευσης, όπου x είναι το πλήθος των λεκτικών μονάδων ανά δέσμη.

ΚΕΦΑΛΑΙΟ 4:

ΠΕΙΡΑΜΑΤΑ

4.1 Σώμα μη επισημειωμένων κειμένων

Για τους σκοπούς της εργασίας, κατασκευάστηκε ένα σώμα μη επισημειωμένων κειμένων που περιέχει άρθρα ελληνικών εφημερίδων. Τα άρθρα συλλέχθηκαν από τους ιστοτόπους δύο εφημερίδων. Από τον ιστότοπο της εφημερίδας "*ΤΑ ΝΕΑ*" συλλέχθηκαν 3033 άρθρα και από τον ιστότοπο της εφημερίδας "*ΤΟ ΒΗΜΑ*" 5489 άρθρα. Τα άρθρα αυτά επιλέχθηκαν με τυχαίο τρόπο από όλες τις ενότητες των εφημερίδων (π.χ. πολιτική, αθλητικά, οικονομικά κλπ.).

Μετά τη συλλογή τους, τα προαναφερθέντα άρθρα υπέστησαν την εξής προεπεξεργασία:

Αρχικά μετατράπηκαν από τη μορφή HTML στην οποία αρχικά βρισκόντουσαν σε μορφή απλού κειμένου.

Στη συνέχεια διαχωρίστηκαν με κενά τα σημεία στίξης και άλλα ειδικά σύμβολα από τους γειτονικούς τους χαρακτήρες, προκειμένου κάθε κείμενο να είναι χωρισμένο σε σαφείς λεκτικές μονάδες (*tokens*). Η διαδικασία αυτή ακολουθήθηκε σε όλες τις περιπτώσεις εμφάνισης σημείων στίξεως και συμβόλων, εκτός από τις εξής περιπτώσεις:

1. Στην περίπτωση των τελειών, ο διαχωρισμός έγινε επιλεκτικά. Στις περιπτώσεις ακρωνυμίων και συντμήσεων, θεωρήθηκε σκόπιμο οι τελείες να θεωρηθούν τμήματά τους. Οπότε διαχωρίζονται από τους γειτονικούς τους χαρακτήρες μόνον οι τελείες οι οποίες σηματοδοτούν λήξη περιόδου. Για τον εντοπισμό των τελειών αυτών χρησιμοποιήθηκε ο διαχωριστής περιόδων (sentence splitter) της εργασίας [Λου05], ο οποίος χρησιμοποιεί μια Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine) και έχει εκπαιδευθεί σε ελληνικά άρθρα εφημερίδων.
2. Στην περίπτωση των αποστρόφων («'», χαρακτήρας 0x27 σύμφωνα με το πρότυπο ASCII), δεν βρέθηκε ικανοποιητικός αλγόριθμος ο οποίος να διακρίνει τη χρήση αποστρόφων ως εισαγωγικών από τη χρήση τους σε περιπτώσεις έκθλιψης ή αφαίρεσης. Οπότε οι απόστροφοι δεν διαχωρίζονται από τους γειτονικούς τους χαρακτήρες.

Προς διευκόλυνση μελλοντικών πειραμάτων, οι προαναφερθείσες μέθοδοι διαχωρισμού συμπεριλήφθηκαν στο εργαλείο επισημείωσης της ενότητας 2.2.

Αξίζει να σημειωθεί ότι, πέρα από την προεπεξεργασία αυτή, τα κείμενα χρησιμοποιήθηκαν ως είχαν. Οι επικεφαλίδες και οι υποκεφαλίδες των άρθρων θεωρήθηκαν ομότιμα τμήματα του κειμένου, και δεν έγινε προσπάθεια εντοπισμού και αφαίρεσης ορθογραφικών και άλλων λαθών.

4.2 Κατασκευή Συνόλων Δεδομένων Εκπαίδευσης

Για τους σκοπούς των πειραμάτων της εργασίας, δημιουργήθηκαν δύο διαφορετικά σύνολα δεδομένων εκπαίδευσης (training sets).

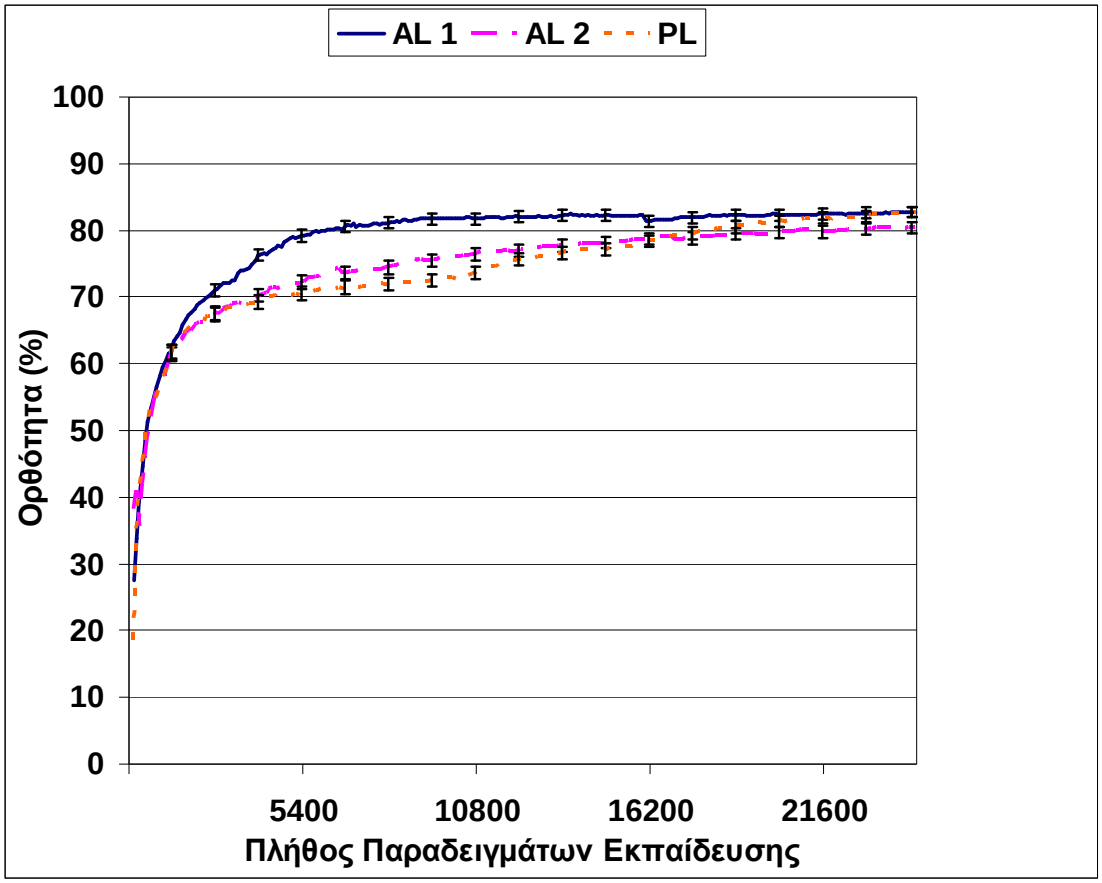
Πρώτα επιλέχθηκαν τυχαία 71 άρθρα πολιτικού, πολιτισμικού και οικονομικού περιεχομένου από το σώμα των μη επισημειωμένων κειμένων της ενότητας 3.1. Τα 3 από αυτά τα άρθρα κρατήθηκαν ολόκληρα, ενώ από τα 68 υπόλοιπα κρατήθηκε μόνο η πρώτη παράγραφος του καθενός. Τα κείμενα που προέκυψαν ελέγχθηκαν διεξοδικά για ορθογραφικά και άλλα λάθη και συνενώθηκαν σε ένα κείμενο. Στη συνέχεια επισημειώθηκαν στο ενιαίο κείμενο οι κατηγορίες όλων των λεκτικών μονάδων, χρησιμοποιώντας το εκτεταμένο σύνολο κατηγοριών της ενότητας 2.5. Για τους σκοπούς των πειραμάτων ενεργητικής μάθησης, δημιουργήθηκε και ένα δεύτερο σύνολο δεδομένων εκπαίδευσης, στο οποίο χρησιμοποιήθηκε πάλι ως αφετηρία το σώμα των μη επισημειωμένων κειμένων της ενότητας 3.1. Για την ακρίβεια, για τους λόγους που εξηγήθηκαν στην ενότητα 3.2 το σώμα των μη επισημειωμένων κειμένων χωρίστηκε σε 700 μέρη και χρησιμοποιήθηκε μέγεθος δέσμης ίσο με 90. Σε κάθε επανάληψη της ενεργητικής μάθησης, η επιλογή των 90 νέων παραδειγμάτων εκπαίδευσης γινόταν από μία (κυκλικά επιλεγόμενη) διαφορετική διαμέριση. Κάθε διαμέριση περιείχε κατά μέσον όρο 11.000 λεκτικές μονάδες και περίπου 12 κείμενα . Συνολικά έγιναν 271 επαναλήψεις του αλγορίθμου της ενεργητικής μάθησης και επισημειώθηκαν 24.390 (271 x 90) λεκτικές μονάδες. Ο ίδιος αριθμός επισημειωμένων λεκτικών μονάδων υπήρχε και στο σύνολο δεδομένων εκπαίδευσης των 71 άρθρων.

4.3 Κατασκευή Συνόλου Δεδομένων Αξιολόγησης

Προκειμένου να αξιολογηθούν οι επιδόσεις του συστήματος, δημιουργήθηκε ένα σύνολο δεδομένων αξιολόγησης (*test set*). Συγκεκριμένα, επιλέχθηκαν 29 άρθρα πολιτικού, πολιτισμικού και οικονομικού περιεχομένου από το σώμα των μη επισημειωμένων κειμένων της ενότητας 3.1, τα οποία είχαν εξαιρεθεί από τη διαδικασία κατασκευής των συνόλων δεδομένων εκπαίδευσης. Τα 2 από αυτά τα άρθρα κρατήθηκαν ολόκληρα, ενώ από τα 27 υπόλοιπα κρατήθηκε μόνο η πρώτη παράγραφος του καθενός. Στη συνέχεια τα κείμενα που προέκυψαν συνενώθηκαν και πάλι σε ένα κείμενο και το ενιαίο κείμενο που προέκυψε ελέγχθηκε για ορθογραφικά και άλλα λάθη και επισημειώθηκε χειρωνακτικά. Προέκυψε έτσι ένα κείμενο αποτελούμενο από συνολικά 8134 χειρωνακτικά επισημειωμένες λεκτικές μονάδες. Τα διανύσματα των λεκτικών μονάδων του κειμένου αυτού χρησιμοποιούνται ως δεδομένα αξιολόγησης

4.4 Πειραματικά Αποτελέσματα

Τα αποτελέσματα που προέκυψαν από τα πειράματα με τα σύνολα δεδομένων που περιγράφηκαν παραπάνω φαίνονται στο εξής γράφημα, όπου χρησιμοποιείται το μεγάλο σύνολο ετικετών (135 ετικέτες) της ενότητας 2.5:



Γράφημα 4.1: Πειράματα με 135 ετικέτες (γενικές κατηγορίες και υποκατηγορίες).

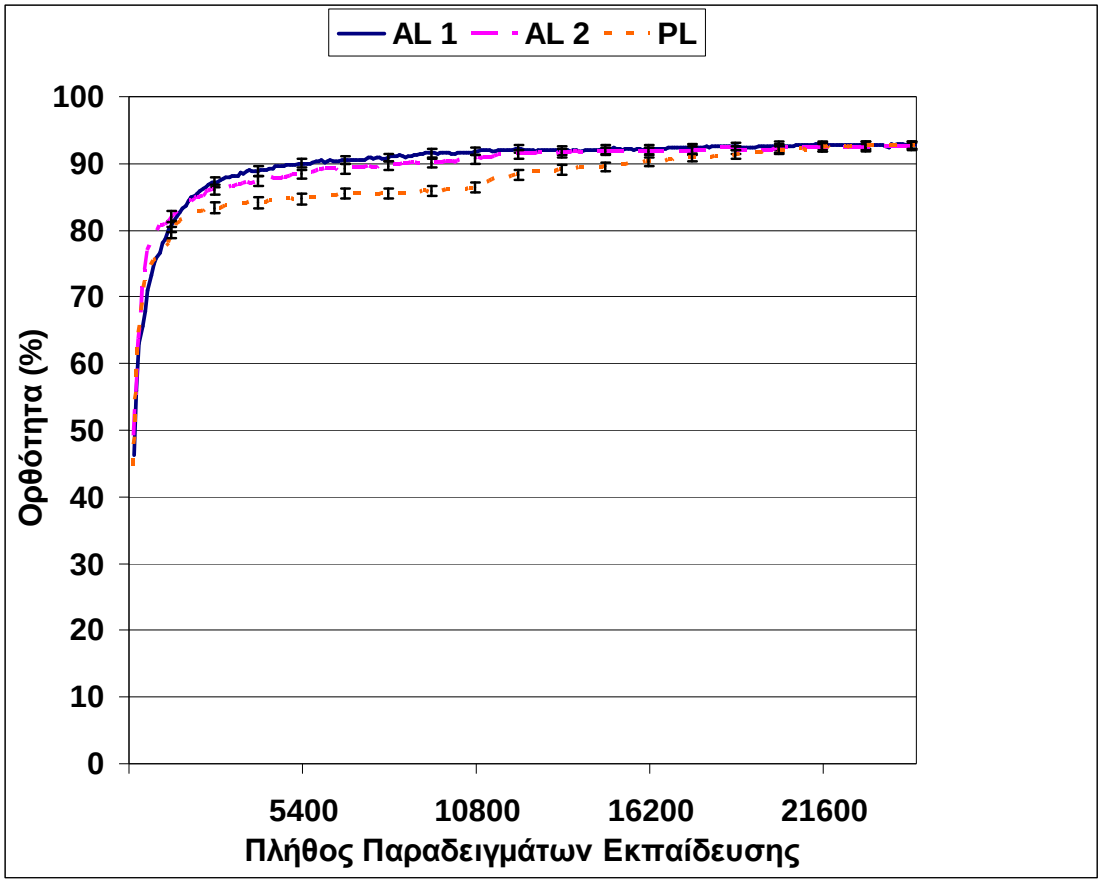
Στο παραπάνω γράφημα ο οριζόντιος άξονας παριστάνει το πλήθος των παραδειγμάτων εκπαίδευσης (χειρωνακτικά επισημειωμένες λεκτικές μονάδες). Ο κατακόρυφος άξονας παριστάνει το ποσοστό ορθότητας (accuracy) που επιτυγχάνεται στο σύνολο δεδομένων αξιολόγησης. Το ποσοστό ορθότητας μετράται ως ο αριθμός των σωστών προβλέψεων (εμφανίσεις λεκτικών μονάδων στις οποίες αποδόθηκαν σωστές ετικέτες) δια του συνόλου των περιπτώσεων (εμφανίσεις λεκτικών μονάδων). Το γράφημα δείχνει και τα διαστήματα εμπιστοσύνης κάθε αποτελέσματος, με βαθμό βεβαιότητας 95%. Διακρίνονται τρεις καμπύλες μάθησης, μία για κάθε είδος πειράματος που διενεργήθηκε:

1. **Καμπύλη PL:** Παθητική μάθηση. Τα παραδείγματα εκπαίδευσης επιλέγονται σειριακά από τη συλλογή εκπαίδευσης των 71 άρθρων της ενότητας 4.2 (συνολικά 24.390 λεκτικές μονάδες). (Μέγιστο διάστημα εμπιστοσύνης ±1,06%.)
2. **Καμπύλη AL1:** Ενεργητική μάθηση, στην οποία τα παραδείγματα εκπαίδευσης επιλέγονται από την ίδια συλλογή της περίπτωσης PL. Στο δεξί άκρο των καμπυλών PL και AL1, το σύστημα εκπαιδεύεται στις ίδιες ακριβώς 24.390 λεκτικές μονάδες. (Μέγιστο διάστημα εμπιστοσύνης ±1,07%.)

3. **Καμπύλη AL2:** Ενεργητική μάθηση, στην οποία τα παραδείγματα εκπαίδευσης επιλέγονται από ολόκληρο το σώμα μη επισημειωμένων κειμένων της ενότητας 3.1. (Μέγιστο διάστημα εμπιστοσύνης ±1,05%.)

Όπως είναι φυσικό, οι καμπύλες PL και AL1 τελικά συγκλίνουν, αφού καταλήγουν να εκπαιδεύονται στα ίδια ακριβώς παραδείγματα. Παρ' όλα αυτά, από τα 2.700 παραδείγματα εκπαίδευσης και πέρα, η καμπύλη AL1 βρίσκεται σημαντικά ψηλότερα από την PL, κάτι που δείχνει ότι η AL1 επιλέγει πιο χρήσιμα παραδείγματα εκπαίδευσης από την PL, που επιλέγει παραδείγματα με τη σειρά που εμφανίζονται στα κείμενα εκπαίδευσης. Αυτό μας έκανε να ελπίζουμε ότι τα αποτελέσματα θα ήταν ακόμα καλύτερα στην περίπτωση της AL2, όπου τα παραδείγματα εκπαίδευσης επιλέγονται από ολόκληρο το σώμα των μη επισημειωμένων κειμένων της ενότητας 3.1, που περιέχει περίπου 3 εκατομμύρια υποψήφια παραδείγματα εκπαίδευσης (εμφανίσεις λεκτικών μονάδων). Τα πειραματικά αποτελέσματα, όμως, δείχνουν ότι η AL2 οδηγεί σε χειρότερη επίδοση από ό,τι η AL1, ενώ το τελικό ποσοστό ορθότητάς της είναι κατώτερο και εκείνου της PL. Το τελικό ποσοστό ορθότητας των PL και AL1 είναι 82,73% (±0,82% για την AL1 και ±0,86% για την PL), ενώ η αντίστοιχη τιμή για την AL2 είναι 80,4373% ±0,82%. Η μη αναμενόμενη αυτή συμπεριφορά διερευνάται περαιτέρω στη συνέχεια.

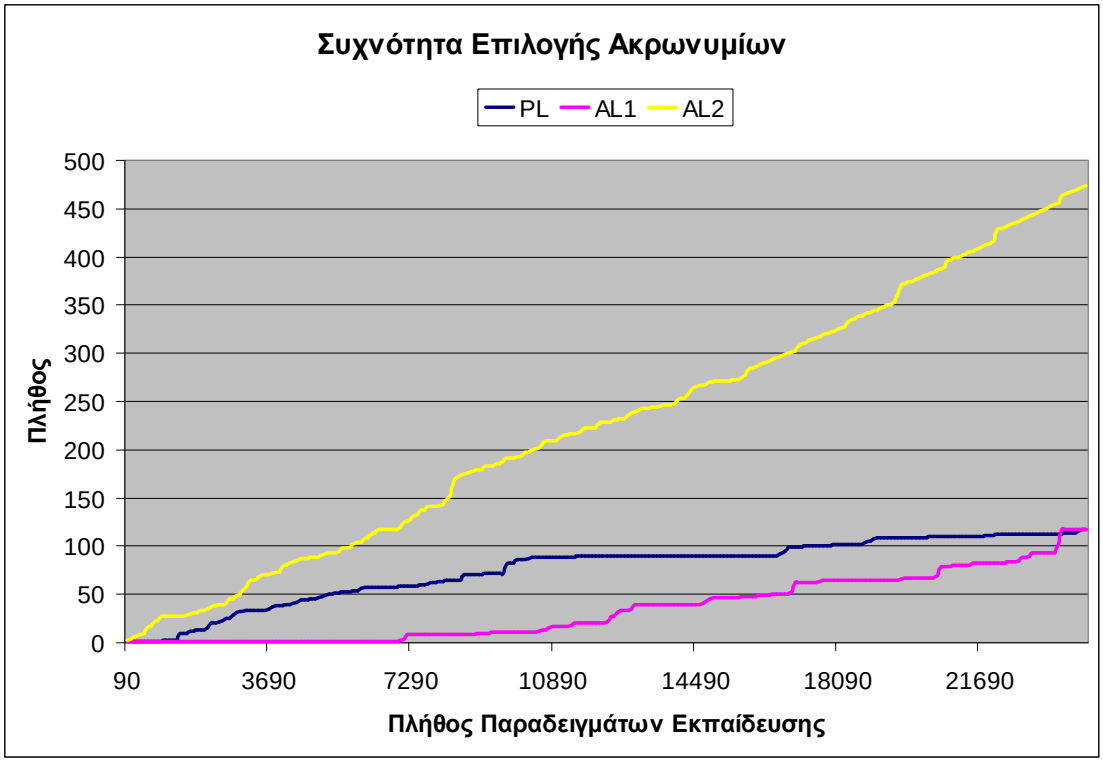
Το γράφημα 4.2 δείχνει τα αντίστοιχα αποτελέσματα όταν χρησιμοποιούνται μόνο οι 12 γενικές κατηγορίες (ετικέτες) της ενότητας 2.5. Προκειμένου να επιταχυνθεί η διεξαγωγή των πειραμάτων, χρησιμοποιήθηκαν ακριβώς τα ίδια παραδείγματα εκπαίδευσης (και αξιολόγησης) που είχαν χρησιμοποιηθεί στα αντίστοιχα πειράματα του μεγάλου συνόλου ετικετών, αλλά οι ετικέτες των παραδειγμάτων εκπαίδευσης (και αξιολόγησης) απλοποιήθηκαν, ώστε να αντιστοιχούν στις 12 γενικές κατηγορίες. Στην περίπτωση αυτή, δηλαδή, τα παραδείγματα εκπαίδευσης της ενεργητικής μάθησης είχαν επιλεγεί και πάλι βάσει των λεπτομερών 135 ετικετών, αντί βάσει των 12 γενικότερων ετικετών.



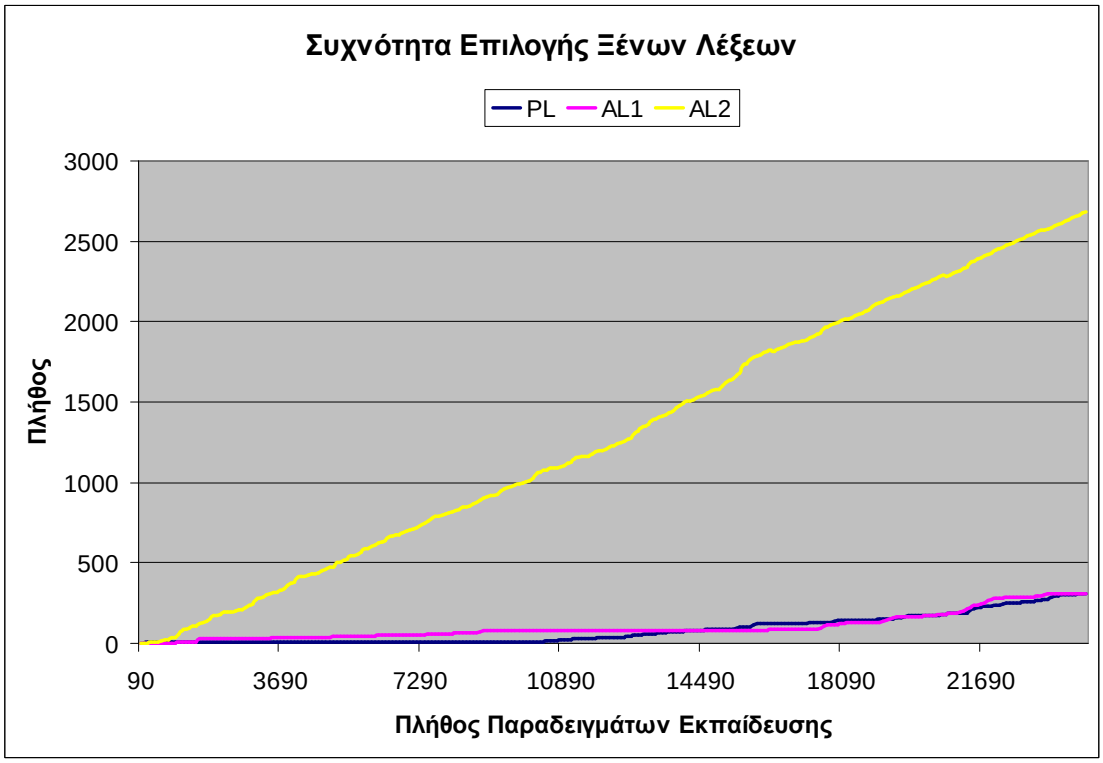
Γράφημα 4.2: Πειράματα με 12 ετικέτες (γενικές κατηγορίες μόνο).

Είναι ορατό και αναμενόμενο ότι οι επιδόσεις του συστήματος σε όλες τις περιπτώσεις (PL, AL1, AL2) είναι σημαντικά καλύτερες συγκρινόμενες με τις αντίστοιχες επιδόσεις του γραφήματος 4.1, όπου χρησιμοποιούνται πολύ περισσότερες κατηγορίες. Η βελτίωση προσεγγίζει ή ακόμη και ξεπερνάει το 10%. Το γεγονός αυτό δείχνει πόσο πιο δύσκολο γίνεται το πρόβλημα όταν επιδιώκει κανείς να αναγνωρίσει όχι μόνο τα μέρη του λόγου αλλά και πληροφορίες όπως γένος, αριθμός, χρόνος κλπ. Βλέπουμε και σε αυτή την περίπτωση ότι η AL2 δεν οδηγεί σε καλύτερα αποτελέσματα από την AL1, αντίθετα από ό,τι αναμέναμε.

Προκειμένου να διερευνηθούν περαιτέρω τα αίτια των μη αναμενόμενων αποτελεσμάτων στην περίπτωση AL2, υπολογίστηκε η συχνότητα με την οποία ο αλγόριθμος ενεργητικής μάθησης επιλέγει ως παραδείγματα εκπαίδευσης ξένες λέξεις και ακρωνύμια. Οι συχνότητες αυτές, συναρτήσει του συνολικού αριθμού των παραδειγμάτων εκπαίδευσης φαίνονται στα παρακάτω γραφήματα:



Γράφημα 4.3

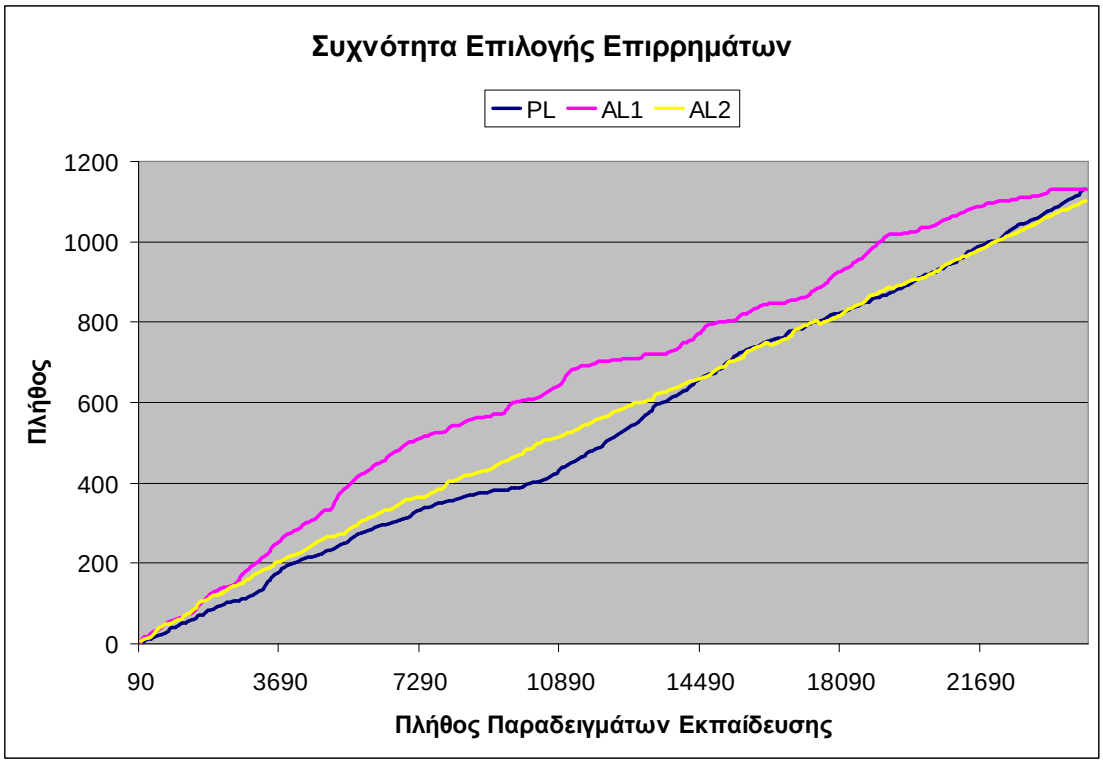


Γράφημα 4.4

Όπως ήταν αναμενόμενο, οι καμπύλες των PL και AL1 τελικά συγκλίνουν, αφού και στις δύο περιπτώσεις το σύστημα εκπαιδεύεται τελικά ακριβώς στα ίδια παραδείγματα. Οι καμπύλες AL2 δείχνουν ότι το μέτρο που χρησιμοποιείται στην ενεργητική μάθηση για την επιλογή παραδειγμάτων ευνοεί ιδιαίτερα τις ξένες λέξεις και τα ακρωνύμια. Αντίθετα από τις περιπτώσεις PL και AL1, στην περίπτωση της

AL2, τα παραδείγματα επιλέγονται από ολόκληρο το σώμα των μη επισημειωμένων κειμένων της ενότητας 3.1, το οποίο περιέχει περίπου 3 εκατομμύρια λεκτικές μονάδες, μεταξύ των οποίων υπάρχει και πολύ μεγάλος αριθμός ξένων λέξεων και ακρωνυμίων. Έτσι στην περίπτωση της AL2, η προτίμηση του μέτρου προς τις ξένες λέξεις και τα ακρωνύμια οδηγεί σε δεδομένα εκπαίδευσης που περιέχουν έναν υπέρμετρο αριθμό λεκτικών μονάδων των κατηγοριών αυτών.

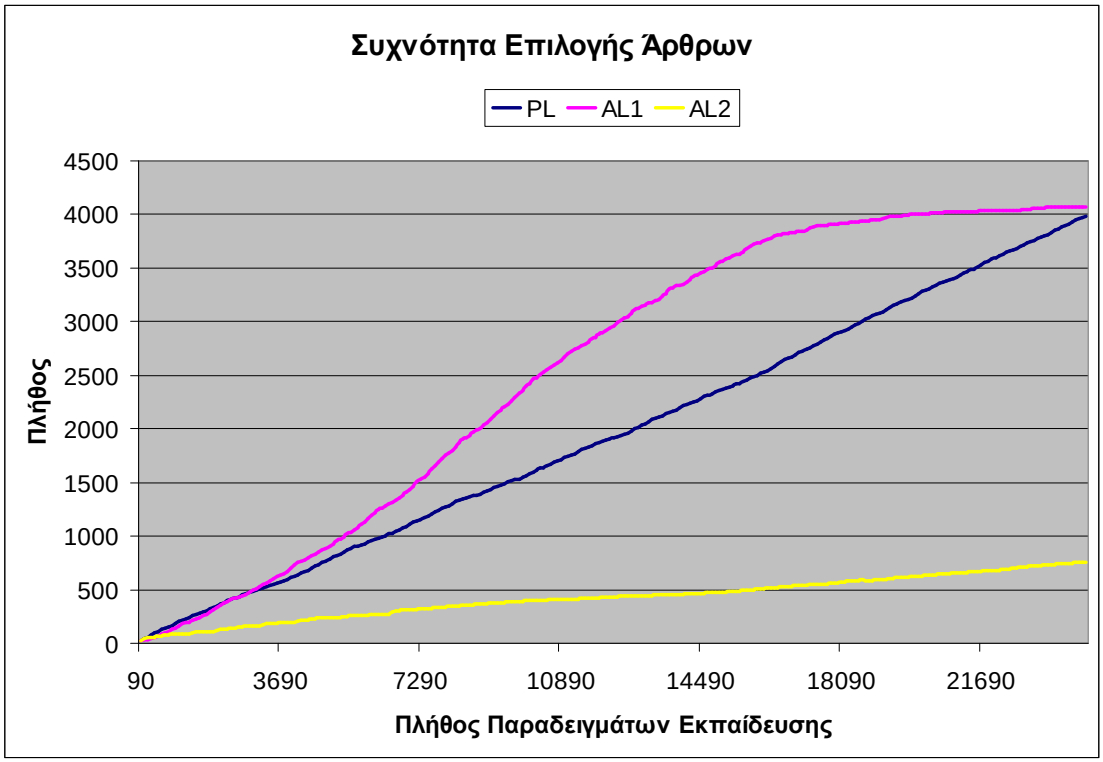
Η προτίμηση των μέτρου της ενεργητικής μάθησης προς τις δύο αυτές κατηγορίες είναι εξηγήσιμη. Η μορφολογία των λεκτικών μονάδων αυτών των κατηγοριών είναι φαινομενικά τυχαία, τουλάχιστον ως προς τα δεδομένα της ελληνικής γλώσσας. Επίσης έχουν τυχαίο γραμματικό ρόλο, καθώς μπορούν να αναπαριστούν ουσιαστικό, επίθετο, κλπ. σε οποιοδήποτε γένος, αριθμό και πτώση. Το αποτέλεσμα είναι ότι ο ταξινομητής είναι ιδιαίτερα αβέβαιος για την ορθή τους κατηγορία (υψηλή εντροπία), ενώ λόγω της «τυχαιότητας» πολλών χαρακτηριστικών τους τα διανύσματά τους συχνά απέχουν πολύ από εκείνα των υπόλοιπων παραδειγμάτων που έχουν επιλεγεί, κάτι που επίσης ευνοεί την επιλογή τους όταν οι ψήφοι των γειτόνων ζυγίζονται βάσει της απόστασής τους (βλ. ενότητα 2.3). Η εισαγωγή τους όμως στο σύνολο εκπαίδευσης δεν βοηθά το σύστημα να επιτύχει μεγαλύτερη ορθότητα, καθώς η τυχαιότητα η οποία τις χαρακτηρίζει μορφολογικά και γραμματικά δεν συντείνει ώστε προηγούμενα παραδείγματα ξένων λέξεων και ακρωνυμίων να είναι κοντινοί γείτονες μελλοντικών λεκτικών μονάδων που ανήκουν στις κατηγορίες αυτές. Εκτός αυτού, η υπέρμετρη εισαγωγή ξένων λέξεων και ακρωνυμίων, οδηγεί στην επιλογή λιγότερων παραδειγμάτων από τις άλλες κατηγορίες, με αποτέλεσμα να αυξάνονται τα λάθη κατάταξης στις υπόλοιπες κατηγορίες. Χαρακτηριστικά παρατίθεται το αντίστοιχο γράφημα συχνότητας για επιρρήματα:



Γράφημα 4.5

Παρατηρούμε ότι η AL1 τείνει να συμπεριλάβει περισσότερα επιρρήματα στο σύνολο εκπαίδευσης απ' ό,τι η παθητική μάθηση. Αντίθετα, η AL2 βρίσκεται πιο κοντά στην παθητική μάθηση παρά στην AL1. Το γεγονός αυτό είναι πιθανό να οφείλεται στην υπερβολική συχνότητα ξένων λέξεων και ακρωνυμίων, τα οποία εκτοπίζουν παραδείγματα επιρρημάτων, τα οποία υπό άλλες συνθήκες θα είχαν προταθεί για επισημείωση.

Αξίζει να σημειωθεί ότι, όπως θα ήθελε κανείς, η AL2 δείχνει πολύ μικρή προτίμηση σε κατηγορίες όπως τα άρθρα, που το σύστημα είναι ιδιαίτερα εύκολο να μάθει να τα κατατάσσει σωστά, επειδή είναι λίγα και ο συνδυασμός μορφολογίας και συντακτικής τοποθέτησης που τα χαρακτηρίζει κάνει εύκολη την αναγνώρισή τους. Αυτό φαίνεται στο επόμενο γράφημα.



Γράφημα 4.6

Ως τώρα χρησιμοποιήθηκαν οι ξένες λέξεις και τα ακρωνύμια ως παραδείγματα κατηγοριών λεκτικών μονάδων στις οποίες η AL2 δείχνει υπέρμετρη προτίμηση. Άλλες τέτοιες κατηγορίες οι οποίες εντοπίστηκαν και αναφέρονται με σύντομο τρόπο είναι οι εξής:

- Ορθογραφικά και τυπογραφικά λάθη, κυρίως εάν τα λάθη αυτά διασπούν μία κατά τα άλλα ενιαία λεκτική μονάδα ή συνενώνουν δύο διαφορετικές.
- Λέξεις οι οποίες δεν μπορούν εύκολα να συγκαταλεχθούν στις κατηγορίες της ενότητας 2.5. Στην ομάδα αυτή περιλαμβάνονται κυρίως κατάλοιπα της δοτικής πτώσης στην νεοελληνική, δηλαδή λέξεις και φράσεις όπως "εν όψει", "λόγω", κ.α.
- Όπως αναφέρθηκε στην ενότητα 4.1, οι επικεφαλίδες και οι υποκεφαλίδες των άρθρων διατηρήθηκαν. Επειδή οι οι φράσεις αυτές συνήθως δεν λήγουν με σημείο στίξης, συχνά εθεωρούντο τμήματα των επομένων περιόδων, με

αποτέλεσμα η τελευταία λέξη της υποκεφαλίδας και η πρώτη λέξη της επόμενης περιόδου να αποτελούν ασυνήθιστες ακολουθίες λέξεων και να επιλέγονται ως παραδείγματα εκπαίδευσης.

Όλες οι προαναφερόμενες κατηγορίες λεκτικών μονάδων παρουσιάζουν χαρακτηριστικά θορυβώδους συμπεριφοράς, η οποία στρεβλώνει την απόδοση του αλγορίθμου ενεργητικής μάθησης, όπως έχει διαπιστωθεί και αλλού [BaBeLa06].

ΚΕΦΑΛΑΙΟ 5:

ΑΝΑΣΚΟΠΗΣΗ

5.1 Συμπεράσματα

Στα πλαίσια της παρούσας εργασίας πραγματοποιήθηκε μία λεπτομερέστερη αξιολόγηση των μεθόδων ενεργητικής μάθησης που είχαν προταθεί σε προηγούμενη μελέτη [Μα05] για την επισημείωση μερών του λόγου ελληνικών κειμένων με τη χρήση ταξινομητή k κοντινότερων γειτόνων. Η αξιολόγηση αυτή έδειξε ότι το μέτρο επιλογής παραδειγμάτων που είχε προταθεί δεν επιφέρει τα αναμενόμενα αποτελέσματα όταν τα παραδείγματα επιλέγονται από πολύ μεγάλες συλλογές μη επισημειωμένων κειμένων, κυρίως λόγω της υπέρμετρης προτίμησης του μέτρου σε κατηγορίες όπως οι ξένες λέξεις και οι συντομογραφίες, που παρουσιάζουν χαρακτηριστικά θορυβώδους συμπεριφοράς. Παράλληλα βελτιώθηκε το λογισμικό της προηγούμενης εργασίας, το οποίο παρέχει τώρα εργαλεία που διευκολύνουν την επανεκπαίδευση και χρήση του συστήματος, ενώ έγιναν πολλές βελτιώσεις στον κώδικα που βελτίωσαν την ταχύτητά του.

5.2 Μελλοντικές Επεκτάσεις

Στην σημερινή του μορφή, ο αλγόριθμος ενεργητικής μάθησης επιλέγει σε υπέρμετρο βαθμό λεκτικές μονάδες των οποίων οι διανυσματικές αναπαραστάσεις οδηγούν σε μεγάλη αβεβαιότητα του ταξινομητή ή τις κάνουν να φαίνονται ασυνήθιστες. Ένας άλλος τρόπος επιλογής θα ήταν να αγνοείται ένα μεγάλο ποσοστό των λεκτικών μονάδων που το υπάρχον μέτρο χαρακτηρίζει ως χαμηλής σημαντικότητας υποψήφια παραδείγματα, και η επιλογή από τις υπόλοιπες να γίνεται με γραμμική ή τυχαία επιλογή. Εναλλακτικά, θα ήταν δυνατόν να τεθεί ένας περιορισμός στον αριθμό παραδειγμάτων κάθε κατηγορίας που επιτρέπεται να επιλεγούν σε κάθε δέσμη (π.χ. ως 5 ξένες λέξεις ανά δέσμη). Θα ήταν, επίσης, χρήσιμο να διερευνηθεί πειραματικά το κατά πόσον είναι πραγματικά χρήσιμοι όλοι οι όροι του μέτρου επιλογής παραδειγμάτων. Θα ήταν, τέλος, ενδιαφέρον να διερευνηθεί η δυνατότητα προσθήκης στο μέτρο επιλογής ενός παράγοντα αντιπροσωπευτικότητας (representativeness), ώστε να επιλέγονται παραδείγματα που να μοιάζουν με πολλά άλλα υποψήφια παραδείγματα εκπαίδευσης.

Μία άλλη μελέτη θα μπορούσε να διερευνήσει την επιρροή που έχει η ύπαρξη εναλλακτικών μορφών της ίδιας λεκτικής μονάδας στην επίδοση του ταξινομητή. Στη σημερινή του μορφή, το σύστημα θεωρεί, για παράδειγμα, τις μονάδες "σήμερα", "σημερα", "Σήμερα", "ΣΗΜΕΡΑ" κλπ. εντελώς διαφορετικές. Δεν έχει μελετηθεί κατά πόσο το γεγονός αυτό συμβάλλει, θετικά ή αρνητικά, στην επίδοση του συστήματος, ή πώς θα μπορούσαν οι διάφορες αυτές μορφές της ίδιας λεκτικής μονάδας να συνενωθούν χωρίς να χαθούν πληροφορίες που είναι χρήσιμες για την κατάταξή τους.

ΠΑΡΑΡΤΗΜΑ:

ΑΝΑΠΑΡΑΣΤΑΣΗ ΤΩΝ ΕΤΙΚΕΤΩΝ ΣΕ XML

Για τους σκοπούς της εργασίας, ορίστηκε μια αναπαράσταση σε XML των κατηγοριών των λεκτικών μονάδων, η οποία δρα ως κανονικοποιημένη μορφή εισόδου και εξόδου. Το σύστημα είναι σε θέση να δέχεται κείμενα εκπαίδευσης στα οποία οι κατηγορίες των λέξεων ακολουθούν αυτή την αναπαράσταση. Το αποτέλεσμα της αυτόματης επισημείωσης (tagging) ενός νέου κειμένου αποθηκεύεται επίσης σε μορφή που ακολουθεί την ίδια αναπαράσταση.

Στην αναπαράσταση αυτή οι κατηγορίες παριστάνονται χρησιμοποιώντας τις ακόλουθες ετικέτες XML:

<article> και </article>

Οι ετικέτες αυτές σηματοδοτούν την έναρξη και λήξη ενός άρθρου εφημερίδας ή άλλου κειμένου μέσα σε ένα αρχείο. Με αυτόν τον τρόπο ένα αρχείο μπορεί να περιέχει πάνω από ένα επισημειωμένα άρθρα.

<sentence> και </sentence>

Οι ετικέτες αυτές σηματοδοτούν την έναρξη και λήξη μίας περιόδου σε ένα άρθρο. Κάθε άρθρο περιέχει τουλάχιστον μία περίοδο. Οι περίοδοι περιέχουν τουλάχιστον μία λεκτική μονάδα.

<token> και </token>

Οι ετικέτες αυτές σηματοδοτούν την έναρξη και λήξη μίας λεκτικής μονάδας (ή μετά την μετα-επεξεργασία, την έναρξη και λήξη ενός περιφραστικού τύπου).

Η ετικέτα <token> δέχεται τις παρακάτω ιδιότητες:

PoS: Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει την κύρια κατηγορία μερών του λόγου στην οποία ανήκει η λεκτική μονάδα. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`adjective, adverb, article, conjunction, noun, numeral, particle, preposition, pronoun, punctuation, verb, other`

case: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει μία από τις τιμές "adjective", "article", "noun", "pronoun", δηλαδή η τρέχουσα λεκτική μονάδα είναι επίθετο, άρθρο, ουσιαστικό ή αντωνυμία, αντίστοιχα. Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει την πτώση της εν λόγω μονάδας. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`nom (`*`nominative`*`), gen (`*`genitive`*`), acc (`*`accusative`*`), voc (`*`vocative`*`)`

gender: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει μία από τις τιμές "adjective", "article", "noun", "pronoun", δηλαδή η τρέχουσα λεκτική μονάδα είναι επίθετο, άρθρο, ουσιαστικό ή αντωνυμία, αντίστοιχα. Η

ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει το γένος της εν λόγω μονάδας. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`masc (`*`masculine`*`), fem (`*`feminine`*`), neut (`*`neuter`*`)`

function: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει τιμή "article", δηλαδή η τρέχουσα λεκτική μονάδα είναι άρθρο. Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει τον τύπο του άρθρου. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`def (`*`definite`*`), indef (`*`indefinite`*`), prep (`*`prepositional`*`) (π.χ. «στον»)`

mode: Η ύπαρξη της ιδιότητας αυτής εξαρτάται από την τιμή που έχει η ιδιότητα PoS. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα mode σε κάθε περίπτωση είναι:

- `inflectionless`, εάν η ιδιότητα PoS έχει τιμή "pronoun", ώστε να σηματοδοτεί άκλιτη αντωνυμία.
- `infinitive`, εάν η ιδιότητα PoS έχει τιμή "verb", ώστε να σηματοδοτεί απαρέμφατο.
- `participle`, εάν η ιδιότητα PoS έχει τιμή "verb", ώστε να σηματοδοτεί μετοχή.

number: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει μία από τις τιμές "adjective", "article", "noun", "pronoun", "verb", δηλαδή η τρέχουσα λεκτική μονάδα είναι επίθετο, άρθρο, ουσιαστικό, αντωνυμία ή ρήμα, αντίστοιχα. Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει τον αριθμό της εν λόγω μονάδας. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`sg (`*`singular`*`), pl (`*`plural`*`)`

tense: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει την τιμή "verb", δηλαδή η τρέχουσα λεκτική μονάδα είναι ρήμα. Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει τον χρόνο της εν λόγω μονάδας. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`present, past, future`

voice: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει την τιμή "verb", δηλαδή η τρέχουσα λεκτική μονάδα είναι ρήμα. Επί πλέον, το εν λόγω ρήμα πρέπει να είναι απαρέμφατο, δηλαδή να έχει επίσης την ιδιότητα mode με τιμή "infinitive". Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει την φωνή του εν λόγω απαρεμφάτου. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`active, passive`

type: Η ύπαρξη της ιδιότητας αυτής προβλέπεται μόνο όταν η ιδιότητα PoS έχει την τιμή "other". Η ιδιότητα αυτή περιέχει μία τιμή η οποία χαρακτηρίζει την επί μέρους κατηγορία της εν λόγω μονάδας. Οι δυνατές τιμές που μπορεί να έχει η ιδιότητα αυτή είναι:
`abbrev (`*`abbreviation`*`), acronym, foreign (`*`foreign word`*`), symbol, undefined`

Η δομή ενός σωστά κατασκευασμένου εγγράφου XML σύμφωνα με τους παραπάνω κανόνες περιγράφεται σε XML Schema στα αρχεία που συνοδεύουν το λογισμικό της εργασίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[BaBeLa06] Balcan, N., Beygelzimer, A. & Langford, J. (2006). "Agnostic Active Learning". *Proceedings of the 23rd International Conference on Machine Learning*, σελ. 65-72.

[Da04] Dasgupta, S. (2004). "Analysis of a greedy active learning strategy". *Advances in Neural Information Processing Systems (NIPS).*

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

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

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

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

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

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