Περιεχόμενα

Περιεχόμενα *1*

1. Εισαγωγή *3*

1.1 Αντικείμενο της εργασίας **3**

1.2 Κυριότερα αποτελέσματα της εργασίας **4**

1.3 Διάρθρωση της εργασίας **5**

1.4 Ευχαριστίες **5**

2. Μηχανική Μάθηση *7*

2.1 Μηχανική Μάθηση **7**

2.2 Αλγόριθμοι **8**

2.2.1 Ο αλγόριθμος k-NN 8

2.2.1.1 Απόδοση βαρών στις ιδιότητες 9

2.2.1.2 Απόδοση βαρών στους γείτονες 10

2.2.2 Δένδρα απόφασης 11

2.2.3 Οι υβριδικοί αλγόριθμοι TRIBL και TRIBL2 13

2.2.4 Μοντέλα Hidden Markov (HMM) 13

2.2.5 TBED 14

3. Προηγούμενες πειραματικές προσπάθειες *15*

3.1 Πειράματα για την αγγλική γλώσσα **16**

3.1.1 Brill – Ο Brill Tagger 16

3.1.2 Daelemans - Πειράματα με τον MBT 17

3.2 Πειράματα για την ελληνική γλώσσα **18**

3.2.1 Πειράματα με συστήματα βασισμένα στα μοντέλα Hidden Markov 19

3.2.2 Πειράματα με συστήματα βασισμένα σε δένδρα απόφασης 20

3.2.2 Προσαρμογή του Brill tagger στα Ελληνικά 21

3.2.2.1 Πειράματα της ερευνητικής ομάδας του «ΕΚΕΦΕ Δημόκριτος» 21

3.2.2.2 Πειράματα της ομάδας του ΙΕΛ 23

4. Το λογισμικό, το σύνολο ετικετών και τα σώματα κειμένων της εργασίας *26*

4.1 Το σύστημα MBT/TiMBL **26**

4.1.1 Ο TiMBL 27

4.1.2 Ο MBT 28

4.2 Το σύστημα επισημείωσης WordTagger **31**

4.2.1 Διεπαφή και λειτουργίες του WordTagger 31

4.2.2 Περιβάλλον κατασκευής συστήματος 32

4.2.3 Χρήση του WordTagger 33

4.3 Το σύστημα ανίχνευσης σφαλμάτων ErrorDetector **35**

4.3.1 Αλγόριθμος 36

4.3.2 Υλοποίηση του συστήματος 38

4.4 Το σύνολο ετικετών **41**

4.5 Τα Σώματα κειμένων **43**

4.5.1 Αρχικό σώμα κειμένων 43

4.5.1.1 Χαρακτηριστικά σώματος 43

4.5.1.2 Προεπεξεργασία 44

4.5.2 Νέο σώμα κειμένων 46

4.5.2.1 Χαρακτηριστικά σώματος κειμένων 46
4.5.2.2 Προεπεξεργασία 47
5. Πειραματική Διαδικασία *49*
5.1 Πειράματα με το αρχικό σώμα κειμένων (σώμα κειμένων "Δημόκριτου") **50**
5.2 Πειράματα στο νέο σώμα κειμένων **56**
5.3 Πειράματα στο σύνολο του μεγέθους των κειμένων **59**
6. Συμπεράσματα & Μελλοντική Έρευνα *61*
6.1 Ανασκόπηση της εργασίας και συμπεράσματα **61**
6.2 Προτάσεις για μελλοντική διερεύνηση **62**
7. Βιβλιογραφικές Αναφορές *64*
Παράρτημα Α *66*
Summary of the thesis in English **66**
Παράρτημα Β *67*
Επεξήγηση του συνόλου ετικετών **67**

1. Εισαγωγή

1.1 Αντικείμενο της εργασίας

Η παρούσα εργασία ασχολείται με το τρόπο με τον οποίο μπορεί η χρήση των τεχνικών της Μηχανικής Μάθησης να βοηθήσει στην κατασκευή συστημάτων τα οποία έχουν την δυνατότητα να πραγματοποιούν αυτόματη αναγνώριση του μέρους του λόγου της κάθε λέξης ενός κειμένου. Και πιο συγκεκριμένα, λέξεων που συναντώνται σε ελληνικά κείμενα. Το πρόβλημα λοιπόν είναι ότι δεδομένου ενός κειμένου και ενός συνόλου ετικετών (tag set) που έχει ορίσει ο ιδιοκτήτης του συστήματος θα πρέπει να γίνει η αντιστοίχηση μιας και μόνο ετικέτας σε κάθε λέξη.

Η ονομασία που δόθηκε στην διαδικασία, οδηγεί τον αναγνώστη στο συμπέρασμα ότι σκοπός της διαδικασίας είναι η απλά η κατηγοριοποίηση των λέξεων στο μέρος του λόγου που αυτές ανήκουν. Αυτό αποτελεί σφάλμα που οφείλεται στην επικράτηση του όρου λόγο της συχνής του χρήσης. Έτσι, πολλές φορές απαιτείται και ο προσδιορισμός πληροφοριών που έχουν να κάνουν με μορφολογικά χαρακτηριστικά της λέξης, όπως γένος, αριθμός, πρόσωπο, πτώση, κ.τ.λ.

Η παραπάνω διαδικασία είναι αρκετά σημαντική για τον τομέα της επεξεργασίας φυσικής γλώσσας. Κι αυτό γιατί οι μορφολογικές πληροφορίες που τελικά αποδίδονται σε κάθε λέξη του κειμένου αποτελούν την βάση για κάθε μορφής επεξεργασία του. Ένα σύστημα που μπορεί να κάνει αναγνώριση του μέρους του λόγου της κάθε λέξης ενός κειμένου μπορεί να αποτελέσει τον πυρήνα για αρκετά άλλα συστήματα. Από απλούς συντακτικούς αναλυτές, διορθωτές κειμένων και μεταφραστές, έως και συστήματα που κάνουν χρήση ή αναγνώριση φωνής.

Το ερώτημα που μπορεί εύκολα να τεθεί είναι το γιατί πρέπει να γίνει χρήση μηχανικής μάθησης και όχι η απευθείας αναγνώριση της λέξης με την χρήση ενός λεξικού της γλώσσας, το οποίο είναι εμπλουτισμένο με μορφολογικούς κανόνες. Η αιτία είναι οι περιορισμοί που υπεισέρχονται σε μια τέτοια λύση. Κατ' αρχάς, η κατασκευή ενός ηλεκτρονικού λεξικού είναι μια διαδικασία με αρκετά υψηλό κόστος,

η οποία έχει συχνά ως αποτέλεσμα την μη ελεύθερη διάθεσή του στο κοινό. Επίσης, το πεπερασμένο σύνολο των λέξεων που περιέχει δεν είναι σε θέση να καλύψει το σύνολο των λέξεων που μπορούν να εμφανιστούν. Τέλος, ένα λεξικό μπορεί να δώσει μόνο όλες τις πιθανές ετικέτες μιας λέξης. Δεν μπορεί όμως, να επιλέξει την σωστή, στην περίπτωση που η ετικέτα εξαρτάται και από τα συμφραζόμενα (π.χ. το «διατάξεις» μπορεί να είναι ρήμα ή ουσιαστικό).

Όπως θα δούμε και στα κεφάλαια 2 και 3, έχουν γίνει αρκετές απόπειρες για την κατασκευή συστημάτων που θα δώσουν λύση στο πρόβλημα. Σε αρκετές περιπτώσεις, και κυρίως σε αυτές που αναφέρονται στα Ελληνικά, τα συστήματα που προέκυψαν δεν είναι διαθέσιμα εκτός του οργανισμού που τα κατασκεύασε, ενώ σε άλλες περιπτώσεις τα συστήματα είναι περιορισμένου εύρους και δεν είναι σε θέση να δώσουν λύση στο σύνολο των αναγκών μας.

Σκοπός της συγκεκριμένης εργασίας είναι η ανάπτυξη ενός συστήματος αυτόματης αναγνώρισης των μερών του λόγου το οποίο θα αφορά ελληνικά κείμενα και το οποίο θα διατίθεται ελεύθερα για ακαδημαϊκή χρήση. Για τον σκοπό αυτό βασιστήκαμε σε ένα υπάρχον σύστημα που διατίθεται επίσης ελεύθερα και το οποίο έχει δοκιμαστεί με επιτυχία σε αρκετές ευρωπαϊκές γλώσσες.

1.2 Κυριότερα αποτελέσματα της εργασίας

Τα κυριότερα αποτελέσματα της εργασίας είναι:

- ✓ Κατασκευή συστήματος χειρωνακτικής επισημείωσης κειμένων, που χρησιμοποιείται για τη δημιουργία σωμάτων εκπαίδευσης και ελέγχου.
- ✓ Κατασκευή εργαλείου αυτόματου εντοπισμού πιθανών λαθών επισημείωσης.
- ✓ Δημιουργία νέου επισημειωμένου σώματος κειμένων αποτελούμενου από κείμενα ειδήσεων, συνολικού μεγέθους 130500 λέξεων, που μπορεί να χρησιμοποιηθεί για την εκπαίδευση και σύγκριση διαφορετικών συστημάτων αναγνώρισης μερών του λόγου.
- ✓ Πειραματική διερεύνηση της ακρίβειας που επιτυγχάνει ένα σύστημα αναγνώρισης μερών του λόγου που βασίζεται σε τεχνικές μάθησης

βασισμένης σε παραδείγματα (instance-based learning) σε δυο σώματα ελληνικών κειμένων (αγγελίες και ειδήσεις).

- ✓ Πειραματική διερεύνηση του κατά πόσον ένα σύστημα εκπαιδευμένο σε ένα σώμα κειμένων ενός είδους (αγγελίες) δίνει ικανοποιητικά αποτελέσματα σε ένα σώμα κειμένων διαφορετικού είδους (ειδήσεις) και αντίστροφα.

1.3 Διάρθρωση της εργασίας

Το υπόλοιπο της εργασίας είναι οργανωμένο ως εξής: Στο κεφάλαιο 2 παρατίθεται το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η εργασία, και αφορά την περιοχή της Μηχανικής Μάθησης. Γίνεται περιγραφή και ανάλυση των αλγορίθμων που χρησιμοποιήθηκαν τόσο στα πλαίσια αυτής της εργασίας, όσο και σε προηγούμενες προσπάθειες επίλυσης του ιδίου προβλήματος από άλλους ερευνητές. Στο κεφάλαιο 3 περιγράφονται τα σχετικά πειράματα που έχουν διεξαχθεί μέχρι σήμερα από άλλους ερευνητές, τόσο για την ελληνική γλώσσα, όσο και για άλλες. Στο κεφάλαιο 4 περιγράφονται οι αρχιτεκτονικές που διερευνήθηκαν στη διάρκεια της εργασίας, το λογισμικό που αναπτύχθηκε, το σύνολο ετικετών που επιλέξαμε να χρησιμοποιήσουμε και τα χαρακτηριστικά των δυο σωμάτων κειμένων στα οποία διεξήχθησαν τα πειράματα. Στο κεφάλαιο 5 γίνεται περιγραφή των πειραμάτων που διεξήχθησαν και των συμπερασμάτων που προκύπτουν από αυτά. Τέλος στο κεφάλαιο 6 ανακεφαλαιώνονται τα ζητήματα που θίχτηκαν και προτείνονται κατευθύνσεις-επεκτάσεις που θα άξιζε να μελετηθούν μελλοντικά.

1.4 Ευχαριστίες

Αρχικά θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου, κ. Ίωνα Ανδρουτσόπουλο για την ουσιαστική και πολύτιμη βοήθειά που μου προσέφερε κατά την διάρκεια της εκπόνησης αυτή της εργασίας, καθώς και στον κ. Θεόδωρο Καλαμπούκη, που αποδέχθηκε τον ρόλο του δεύτερου αξιολογητή. Επίσης θα ήθελα να ευχαριστήσω το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» που μας παραχώρησε ένα, ήδη επισημειωμένο, σώμα κειμένων καθώς και πληροφορίες για τις ετικέτες που χρησιμοποιούνται σε αυτό.

Τέλος θα ήθελα να ευχαριστήσω και το Ινστιτούτο Επεξεργασίας του Λόγου για τις πληροφορίες που μας παρείχε σχετικά με τις ελληνικές ετικέτες του ερευνητικού έργου PAROLE, που χρησιμοποιήθηκαν στην εργασία.

2. Μηχανική Μάθηση

Το κεφάλαιο αυτό παρέχει μια σύντομη εισαγωγή στις έννοιες και τους αλγορίθμους μηχανικής μάθησης που σχετίζονται με το αντικείμενο της εργασίας. Περιλαμβάνονται τόσο αλγόριθμοι που χρησιμοποιήθηκαν σε αυτήν την εργασία όσο και αλγόριθμοι που χρησιμοποιήθηκαν σε αντίστοιχα συστήματα προηγούμενων ερευνητών.

2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Μ.Μ.) αποτελεί έναν από τους παλαιότερους τομείς έρευνας της Τεχνητής Νοημοσύνης. Στόχος της είναι η δημιουργία συστημάτων τα οποία θα μπορούν να βελτιώνουν της απόδοσή τους στην εργασία που επιτελούν εκμεταλλευόμενα αυτόματα προηγούμενη εμπειρία από την εκτέλεση της εργασίας.

Σε γενικές γραμμές, η διαδικασία της μάθησης αποτελείται από τα εξής στάδια:

- Απόκτηση γνώσης και εμπειρίας από την αλληλεπίδραση με το περιβάλλον.
- Επεξεργασία της αποκτηθείσας γνώσης, ώστε να γίνει κατηγοριοποίησή της, και αν αυτό είναι δυνατόν να βρεθούν πιθανές γενικεύσεις ή εξειδικεύσεις.
- Χρησιμοποίηση των αποτελεσμάτων της επεξεργασίας και λήψη ανατροφοδότησης από το περιβάλλον, ώστε να βελτιωθεί περαιτέρω το σύστημα.

Στην εργασία αυτή χρησιμοποιούνται μέθοδοι επιβλεπόμενης επαγωγικής μάθησης. Στο σύστημα παρέχονται παραδείγματα εκπαίδευσης που συνοδεύονται από τις κατηγορίες τους. Στην περίπτωσή μας, παρέχεται ένα σώμα κειμένων, οι λέξεις του οποίου έχουν επισημειωθεί με τις επιθυμητές ετικέτες τους. Ο αλγόριθμος μάθησης, επεξεργάζεται τα παραδείγματα εκπαίδευσης και παράγει επαγωγικά έναν ταξινομητή, ο οποίος χρησιμοποιείται στη συνέχεια για να κατατάσσει σε κατηγορίες νέες περιπτώσεις, των οποίων οι κατηγορίες δεν είναι γνωστές. Στην περίπτωσή μας, ο ταξινομητής αποδίδει ετικέτες στις λέξεις νέων κειμένων, για τις οποίες δεν είναι γνωστές οι σωστές ετικέτες.

2.2 Αλγόριθμοι

2.2.1 Ο αλγόριθμος k-NN

Ο αλγόριθμος k-NN [Mitchell 1997] αποτελεί έναν από τους πιο χαρακτηριστικούς αλγόριθμους για μάθηση βασισμένη σε παραδείγματα. Η πλήρης ονομασία του είναι αλγόριθμος των k κοντινότερων γειτόνων (Nearest Neighbor – NN).

Για τη χρήση και την αξιολόγηση του αλγορίθμου θα πρέπει να υπάρχουν δυο σύνολα από αντικείμενα: το σύνολο εκπαίδευσης TrS και το σύνολο ελέγχου TeS. Τα αντικείμενα των συνόλων αυτών πρέπει να έχουν καταταγεί χειρωνακτικά σε δύο ή περισσότερες κατηγορίες (c_1, c_2, c_3, …, c_k), των οποίων η τομή ανά δύο πρέπει να είναι κενή, δηλαδή δεν μπορεί ένα αντικείμενο να ανήκει σε περισσότερες από μία κατηγορίες. Επίσης θα πρέπει να έχει οριστεί ένα σύνολο ιδιοτήτων $A = \{X_1, X_2, X_3, ..., X_m\}$. Κάθε αντικείμενο των συνόλων TrS και TeS παριστάνεται από ένα διάνυσμα $\vec{x} = \langle x_1, x_2, x_3, ..., x_m \rangle$, κάθε συντεταγμένη του οποίου αποτελεί την τιμή μιας συγκεκριμένης ιδιότητας x_i. Ο αλγόριθμος «εκπαιδεύεται» στα αντικείμενα του συνόλου TrS, ώστε να προβλέπει τη σωστή κατηγορία κάθε αντικειμένου βάση των τιμών του διανύσματός του. Η ακρίβεια του ταξινομητή που προκύπτει από την εκπαίδευση αξιολογείται στο σύνολο TeS, συγκρίνοντας τις αποφάσεις του ταξινομητή με τις σωστές κατηγορίες.

Κατά την εκπαίδευση, ο αλγόριθμος k-NN απλά αποθηκεύει σε μια μνήμη όλα τα διανύσματα των αντικειμένων του συνόλου TrS και τις σωστές κατηγορίες τους. Η κατάταξη νέων αντικειμένων, των οποίων δεν είναι γνωστές οι κατηγορίες, γίνεται ως εξής: Υπολογίζεται η απόσταση του διανύσματος του νέου αντικειμένου από τα διανύσματα όλων των αντικειμένων εκπαίδευσης. Επιλέγονται τα k αντικείμενα εκπαίδευσης με τις μικρότερες αποστάσεις (οι k κοντινότεροι γείτονες) και το νέο αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k αντικειμένων. Ως μέτρο απόστασης μπορεί να χρησιμοποιηθεί, για παράδειγμα, η απόσταση Manhattan, που ορίζεται ως εξής:

$$d(\vec{x}_i, \vec{x}_j) \equiv \sum_{r=1}^{m} \delta(x_{ir}, x_{jr}) \quad \delta(x, y) \equiv \begin{cases} 0, & \text{αν } x = y \\ 1, & \text{διαφορετικά} \end{cases} \tag{2.1}$$

Η απλότητα αλλά και η γενικότητα του παραπάνω αλγορίθμου παρέχει στον χρήστη την ευκολία να τον τροποποιήσει, προσθέτοντας δικά του μέτρα απόστασης, που θα επηρεάσουν την ευαισθησία του αλγορίθμου ως προς του k γείτονες και θα φέρουν τα αποτελέσματα πιο κοντά στις επιθυμητές τιμές.

Μια παραλλαγή του παραπάνω αλγορίθμου, με την ονομασία IB1, εμφανίζεται στο λογισμικό MBT/TiMBL που χρησιμοποιήθηκε στην εργασία. Η παραλλαγή αυτή παρουσιάζει δυο βασικές διαφορές:

1. Η τιμή του k αφορά τις k μικρότερες αποστάσεις και όχι τα k κοντινότερα αντικείμενα. Μπορεί, επομένως, να χρησιμοποιηθούν κατά την κατάταξη περισσότεροι από k γείτονες, αν κάποιοι από αυτούς απέχουν το ίδιο από το αντικείμενο προς κατάταξη.
2. Σε περίπτωση ισοψηφίας μεταξύ των γειτόνων, αν δηλαδή οι γείτονες ισοκατανέμονται στις δύο κατηγορίες, το k αυξάνεται κατά 1 και η απόφαση λαμβάνεται βάσει του νέου k.

2.2.1.1 Απόδοση βαρών στις ιδιότητες

Η χρήση του τύπου (2.1) προϋποθέτει ότι οι ιδιότητες έχουν την ίδια αξία ως προς την λήψη της απόφασης. Υπάρχουν όμως και φορές που επιθυμούμε να δώσουμε μεγαλύτερη αξία στις ιδιότητες που προβλέπουν καλύτερα τη σωστή κατηγορία. Αυτό μπορεί να γίνει υπολογίζοντας το πληροφοριακό κέρδος (Information Gain - IG) που παρέχει κάθε ιδιότητα και χρησιμοποιώντας την ποσότητα αυτή ως βάρος της κάθε ιδιότητας.

Το IG υπολογίζεται μετρώντας τη μέση μείωση της αβεβαιότητας για τη σωστή κατηγορία που προκαλεί η γνώση της τιμής της συγκεκριμένης ιδιότητας. Το πληροφοριακό κέρδος για την ιδιότητα X_i, που χρησιμοποιείται ως βάρος w_i της ιδιότητας, δίνεται από τον παρακάτω τύπο:

$$w_i = H(C) - \sum_{u \in V_i} P(X_i = u) \cdot H(C \mid X_i = u) \quad (2.2)$$

Όπου V_i το σύνολο τιμών της $X_{i,,}$ και H(C) η εντροπία της τυχαίας μεταβλητής C που παριστάνει τη σωστή κατηγορία:

$$H(C) = -\sum_{c_i} P(C = c_i) \log_2 P(C = c_i) \quad (2.3)$$

Ο υπολογισμός του IG με τον τύπο 2.2 τείνει να υπερεκτιμά ιδιότητες με μεγάλα πλήθη τιμών. Γι' αυτό και προτείνεται η χρήση μιας κανονικοποιημένης μορφής του IG που καλείται Gain Ration (GR) και δίνεται από τον τύπο:

$$w_i = \frac{H(C) - \sum_{u \in V_i} P(X_i = u) \cdot H(C \mid X_i = u)}{si(i)} \quad (2.4)$$

όπου

$$si(i) = -\sum_{u \in V_i} P(X_i = u) \log_2 P(X_i = u) \quad (2.5)$$

Έτσι, για τον υπολογισμό της απόστασης χρησιμοποιείται η εξίσωση:

$$d(\vec{x}_i, \vec{x}_j) \equiv \sum_{r=1}^{m} w_i \delta(x_{ir}, x_{jr}) \quad \delta(x, y) \equiv \begin{cases} 0, & \text{αν } x = y \\ 1, & \text{διαφορετικά} \end{cases} \quad (2.6)$$

Στο λογισμικό MBT/TiMBL, η παραλλαγή του αλγορίθμου που χρησιμοποιεί τον τύπο (2.6) ονομάζεται IB1-IG.

2.2.1.2 Απόδοση βαρών στους γείτονες

Εκτός από την απόδοση βαρών στις ιδιότητες, είναι δυνατόν να αποδοθούν βάρη και στους γείτονες, ώστε κοντινότεροι γείτονες να λαμβάνουν μεγαλύτερη αξία. Η βαρύτητα w_j του j-στού γείτονα μπορεί να υπολογιστεί ως εξής:

$$w_j = \begin{cases} \frac{d_k - d_j}{d_k - d_1}, & \text{αν } d_k \neq d_1 \\ 1, & \text{αν } d_k = d_1 \end{cases} \quad (2.8)$$

όπου d_1, d_k και d_j οι αποστάσεις των κοντινότερου (παίρνει βάρος 1), του πιο απομακρυσμένου (παίρνει βάρος 0) και του j-ιοστού γείτονα, αντίστοιχα.

Εκτός όμως από το παραπάνω, έχουν προταθεί και λύσεις όπως η απόδοση βάρους αντιστρόφως ανάλογου από την απόσταση:

$$w_j = \begin{cases} \frac{1}{d_j}, & \text{αν } d_j \neq 0 \end{cases} \quad (2.9)$$

Στην περίπτωση αυτή, αν για κάποιο γείτονα $d_j = 0$, το νέο αντικείμενο κατατάσσεται στην κατηγορία του γείτονα αυτού.

2.2.2 Δένδρα απόφασης

Ο ID3 [Quinlan 1987] είναι ένας αρκετά διαδεδομένος αλγόριθμος επαγωγικής μάθησης, με τον οποίο κατασκευάζονται δένδρα απόφασης. Η λογική του βασίζεται στην χρήση των ιδιοτήτων, μία προς μία, για την διάσπαση του συνόλου των αντικειμένων εκπαίδευσης στους κλάδους ενός δένδρου. Σε γενικές γραμμές, εκτελούνται αναδρομικά τα παρακάτω βήματα:

1. Επιλέγεται η ιδιότητα με το μεγαλύτερο πληροφοριακό κέρδος, εκτιμώντας τις πιθανότητες στο σύνολο εκπαίδευσης.
2. Τοποθετείται στη ρίζα του δένδρου ένας έλεγχος για την ιδιότητα με το μεγαλύτερο πληροφοριακό κέρδος και δημιουργείται ένας κλάδος κάτω από τη ρίζα για κάθε μία δυνατή τιμή της ιδιότητας.
3. Τα αντικείμενα του συνόλου εκπαίδευσης κατανέμονται στους κλάδους, ανάλογα με την τιμή της ιδιότητας που χρησιμοποιήθηκε στη ρίζα.
4. Κάθε κλάδος οδηγεί σε ένα υποδένδρο που κατασκευάζεται αναδρομικά, χρησιμοποιώντας ως σύνολο εκπαίδευσης το υποσύνολο που αντιστοιχεί στον κλάδο και ως σύνολο ιδιοτήτων το αρχικό μείον την ιδιότητα που χρησιμοποιήθηκε στη ρίζα.

Πιο αναλυτικά, ο αλγόριθμος έχει ως εξής:

ID3 (Παραδείγματα, C, Ιδιότητες)

1. Αν όλα τα Παραδείγματα ανήκουν σε μία κατηγορία, επίστρεψε ένα δένδρο ενός κόμβου με απόφαση την κατηγορία αυτή.

2. Αν Ιδιότητες = {}, επίστρεψε ένα δένδρο ενός κόμβου με απόφαση την κατηγορία που είναι πιο συχνή στα Παραδείγματα.
3. Διάλεξε με βάση τα Παραδείγματα την ιδιότητα Χ με το μεγαλύτερο κέρδος πληροφορίας IG(C,X).
4. Φτιάξε ένα δένδρο Δ με ρίζα την ιδιότητα Χ.
5. Για κάθε δυνατή τιμή x_i της Χ:
   I. Πρόσθεσε στη ρίζα του Δ ένα κλαδί για την περίπτωση $X = x_i$.
   II. Έστω Παραδείγματα-i το υποσύνολο των Παραδειγμάτων με $X = x_i$.
   III. Αν Παραδείγματα-i = { } τότε κάνε το κλαδί να οδηγεί σε φύλλο του οποίου η απόφαση να είναι η πιο συχνή τιμή του C στα Παραδείγματα.
   IV. Διαφορετικά, κρέμασε κάτω από το κλαδί το δένδρο ID3(Παραδείγματα-i, C, Ιδιότητες – {X}).
6. Επίστρεψε το δένδρο Δ.

Παραλλαγή αυτού του αλγορίθμου αποτελεί ο αλγόριθμος IGTREE που υποστηρίζεται από το MBT/TiMBL. Η κυριότερη διαφορά είναι πως η αξιολόγηση των ιδιοτήτων γίνεται μόνο μία φορά στην αρχή και όλοι οι κόμβοι του δένδρου που βρίσκονται στο ίδιο βάθος ελέγχουν την ίδια ιδιότητα [Daelemans et al. 1996]. Το δένδρο απόφασης που προκύπτει διαχωρίζει το αντικείμενα εκπαίδευσης σε ομάδες. Κατά το στάδιο της κατάταξης αντικειμένων των οποίων η κατηγορία είναι άγνωστη, εντοπίζεται πρώτα η «κοντινότερη» ομάδα χρησιμοποιώντας το δέντρο απόφασης και το αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί σε εκείνη την ομάδα αντί (όπως στην περίπτωση του k-NN) για την κατηγορία που πλειοψηφεί στο σύνολο των αντικειμένων εκπαίδευσης. Με αυτήν την έννοια ο IGTREE αποτελεί μια προσέγγιση του k-NN. Το κυριότερο πλεονέκτημα του IGTREE είναι ότι η κατάταξη γίνεται πολύ γρηγορότερα, αφού δεν χρειάζεται να εξεταστούν όλα τα παραδείγματα εκπαίδευσης.

Παραλλαγές αλγορίθμων εκμάθησης δένδρων απόφασης χρησιμοποιήθηκαν και από τους Ορφανό κ.ά. [Orphanos,Christodoulakis et al. 1999, Orphanos,Kalles et al. 1999]

2.2.3 Οι υβριδικοί αλγόριθμοι TRIBL και TRIBL2

Οι αλγόριθμοι TRIBL και TRIBL2 αποτελούν συνδυασμούς του IGTREE και του IB1. Σε γενικές γραμμές, και οι δύο χρησιμοποιούν τον IGTREE για να κατασκευάσουν ένα δένδρο απόφασης, που διαχωρίζει τα αντικείμενα εκπαίδευσης σε ομάδες. Κατά την κατάταξη νέων αντικειμένων, επιλέγεται πρώτα η «κοντινότερη» ομάδα χρησιμοποιώντας το δένδρο απόφασης ως κάποιο βάθος. Στη συνέχεια, βρίσκονται οι k κοντινότεροι γείτονες του νέου αντικειμένου μέσα στην ομάδα, και το νέο αντικείμενο κατατάσσεται στην κατηγορία που πλειοψηφεί μεταξύ των k κοντινότερων γειτόνων, όπως στην περίπτωση του IB1. Η διαφορά μεταξύ των TRIBL και TRIBL2 έγκειται στο πότε ακριβώς γίνεται κατά την κατάταξη η μετάβαση από τη χρήση του δένδρου απόφασης στη χρήση του IB1. Ο TRIBL χρησιμοποιεί το δένδρο απόφασης ως κάποιο συγκεκριμένο (το ίδιο πάντα) βάθος, ενώ ο TRIBL2 χρησιμοποιεί το δένδρο απόφασης ως το βάθος όπου δεν βρίσκει κλαδί που να ταιριάζει με τις τιμές των ιδιοτήτων του αντικειμένου προς κατάταξη.

2.2.4 Μοντέλα Hidden Markov (HMM)

Η βέλτιστη σειρά ετικετών $T = <t_1, t_2, t_3, \ldots, t_M>$ των λέξεων $W = <w_1, w_2, w_3, \ldots, w_M>$ ενός κειμένου μπορεί θεωρητικά να υπολογιστεί όπως παρακάτω, χρησιμοποιώντας το θεώρημα του Bayes [Dermatas & Kokkinakis 1995]:

$$T_0 = \text{argmax}_T\, P(T \mid W) = \text{argmax}_T \frac{P(W \mid T) * P(T)}{P(W)} = \text{argmax}_T (P(W \mid T) * P(T) \qquad (2.10)$$

Η διαδικασία εύρεσης των ετικετών μπορεί να αναπαρασταθεί με ένα HMM, θεωρώντας ότι η εμφάνιση μιας ετικέτας είναι ένα τυχαίο κρυφό γεγονός που εξαρτάται μόνο από τις προηγούμενες N ετικέτες και ότι η (κρυφή) εμφάνιση μιας ετικέτας παράγει μια λέξη στο κείμενο. Στην περίπτωση αυτή, η βέλτιστη σειρά ετικετών μπορεί να παρασταθεί ως [Dermatas & Kokkinakis 1995]:

$$T_0^{(HMM-TS)} = \arg_{t_1,\ldots,t_M} \max P(t_1) \prod_{i=2}^{N} P(t_i \mid t_{i-1},\ldots,t_1) \prod_{i=N+1}^{M} P(t_i \mid t_{i-1},\ldots,t_{i-N}) \prod_{i=1}^{M} P(w_i \mid t_i) \qquad (2.11)$$

Η βέλτιστη σειρά ετικετών μπορεί τότε να υπολογιστεί με τον αλγόριθμο Viterbi. Εναλλακτικά στατιστικά μοντέλα και πειραματική σύγκρισή τους περιγράφονται από τους Δερματά και Κοκκινάκη [Dermatas & Kokkinakis 1995].

2.2.5 TBED

Ένας άλλος αλγόριθμος μάθησης που χρησιμοποιείται συχνά στο πρόβλημα της εύρεσης των μερών του λόγου είναι ο Transformation-Based Error-Driven Learning (TBED) [Brill 1995]. Ο αλγόριθμος κατασκευάζει μια ακολουθία κανόνων, κάθε ένας από τους οποίους επιχειρεί να διορθώσει τα λάθη που προέκυψαν ή απέμειναν από την εφαρμογή των προηγούμενων. Σε γενικές γραμμές, η λειτουργία του είναι η εξής. Αρχικά, σε κάθε λέξη του σώματος εκπαίδευσης αποδίδεται μια προεπιλεγμένη ετικέτα (π.χ. η πιο συχνή ετικέτα). Από ένα σύνολο υποψήφιων κανόνων (που προσδιορίζονται π.χ. από ένα πρότυπο κανόνων και τις διαθέσιμες ιδιότητες), επιλέγεται εκείνος ο κανόνας που οδηγεί στη μεγαλύτερη μείωση των λανθασμένων ετικετών στο σώμα εκπαίδευσης. Ο κανόνας εφαρμόζεται στο σώμα εκπαίδευσης, και επιλέγεται στη συνέχεια ο κανόνας που οδηγεί στη μεγαλύτερη περαιτέρω μείωση των λαθών κ.ο.κ. Μετά την εκπαίδευση, κατά τη χρήση του συστήματος, οι κανόνες που προέκυψαν εφαρμόζονται σε νέα κείμενα, με την ίδια σειρά, ξεκινώντας από τις ίδιες αρχικές ετικέτες όπως και στην εκπαίδευση.

3. Προηγούμενες πειραματικές προσπάθειες

Στο κεφάλαιο αυτό παρουσιάζονται οι σημαντικότερες πειραματικές προσπάθειες που έχουν γίνει στο παρελθόν στον τομέα της χρήσης Μηχανικής Μάθησης για τον αυτόματο εντοπισμό των μερών του λόγου των λέξεων ενός κειμένου. Οι προσπάθειες που περιγράφονται παρακάτω αφορούν κατά κύριο λόγο την αγγλική και την ελληνική γλώσσα. Υπάρχει όμως και αναφορά σε προσπάθειες που πραγματοποιήθηκαν για άλλες ευρωπαϊκές γλώσσες, ως επεκτάσεις αντιστοίχων μελετών για τα Αγγλικά ή τα Ελληνικά.

Ένα σημείο στο οποίο θα πρέπει να δοθεί ιδιαίτερη προσοχή είναι το πόσες πληροφορίες περιλαμβάνει η ετικέτα της κάθε λέξης του κειμένου. Κάθε κατηγορία μερών του λόγου (π.χ. ουσιαστικά, ρήματα) μπορεί να εμπεριέχει ή όχι υποκατηγορίες (π.χ. αρσενικά ουσιαστικά, θηλυκά ουσιαστικά). Το βάθος σε αυτό το «δένδρο των κατηγοριών» που χρησιμοποιεί ο κάθε ερευνητής μπορεί να προσδώσει στις ετικέτες περισσότερες ή λιγότερες πληροφορίες. Αυτό επηρεάζει τον πληθάριθμο του συνόλου ετικετών και, επομένως, τη δυσκολία του προβλήματος και τα αποτελέσματα που λαμβάνουμε από το εκάστοτε χρησιμοποιούμενο σύστημα. Λόγω αυτής της διαφοροποίησης στο χρησιμοποιούμενο σύνολο ετικετών, δεν είναι δυνατή μια αντικειμενική σύγκριση μεταξύ των προσπαθειών προηγούμενων ερευνητών.

Κάτι αντίστοιχο συμβαίνει με τα ελληνικά σώματα κειμένων. Η έλλειψη ενός επισημειωμένου σώματος ελληνικών κειμένων που να τυγχάνει γενικής αποδοχής έχει ως αποτέλεσμα κάθε ερευνητική ομάδα να χρησιμοποιεί διαφορετικό σώμα κειμένων, το οποίο συχνά επισημειώνει η ίδια. Έτσι, υπάρχει πάντα και η πιθανότητα το επιλεγμένο σύνολο ετικετών να «ταιριάζει» καλύτερα με τα κείμενα του συγκεκριμένου σώματος της έρευνας και τα αποτελέσματα να είναι καλύτερα από ό,τι σε άλλα σώματα κειμένων.

Σε κάποιες από τις ερευνητικές προσπάθειες άλλων ομάδων που αναφέρονται παρακάτω και στα πειράματα αυτής της εργασίας χρησιμοποιήθηκε η τεχνική της δεκαπλής διασταυρωμένης επικύρωσης (10-fold cross validation). Σε αυτήν, το επισημειωμένο σώμα κειμένων χωρίζεται σε 10 τμήματα με όσο το δυνατόν

μικρότερη διαφορά μεγέθους μεταξύ τους. Κάθε πείραμα επαναλαμβάνεται 10 φορές. Σε κάθε επανάληψη, ένα διαφορετικό τμήμα χρησιμοποιείται για έλεγχο και τα υπόλοιπα 9 για εκπαίδευση. Τα τελικά αποτελέσματα είναι οι μέσοι όροι των 10 επαναλήψεων.

3.1 Πειράματα για την αγγλική γλώσσα

Στο τμήμα αυτό παρατίθενται οι προσπάθειες που έγιναν ως απόπειρα επίλυσης του προβλήματος για την αγγλική γλώσσα.

3.1.1 Brill – O Brill Tagger

Για την διεξαγωγή των συγκεκριμένων πειραμάτων, που έγιναν στην αγγλική γλώσσα, χρησιμοποιήθηκε το σώμα κειμένων Penn Treebank Tagged Wall Street Journal Corpus και ο αλγόριθμος μάθησης TBED [Brill 1995].

Αρχικά διεξήχθησαν πειράματα με γνωστές λέξεις (με λέξεις που στο σύνολό τους υπήρχαν στο σώμα εκπαίδευσης). Για σύγκριση χρησιμοποιήθηκε και ένα σύστημα βασισμένο στα μοντέλα Hidden Markov [Dermatas and Kokkinakis 1995]. Η ακρίβεια του τελευταίου ήταν 96.3% όταν το σώμα εκπαίδευσης ήταν μεγέθους 64K λέξεων, και 96.7% όταν το σώμα εκπαίδευσης ήταν 1000K λέξεων. Σε αντίθεση, ο TBED πέτυχε το ίδιο αποτέλεσμα με μόλις 64K λέξεις, για να το ανεβάσει στο 97.2% όταν εκπαιδεύτηκε σε 600K λέξεις.

Ακολούθησε έρευνα για την ικανότητα του συστήματος να αποδίδει ετικέτες σε άγνωστες λέξεις (δηλαδή λέξεις που δεν εμφανίζονταν στα σώματα εκπαίδευσης), με ένα σώμα ελέγχου αποτελούμενο από 150K λέξεις. Στην φάση της εκπαίδευσης, για την δημιουργία των κανόνων επισημείωσης άγνωστων λέξεων χρησιμοποιήθηκαν 350K λέξεις και για την εξαγωγή των κανόνων περιεχομένου 600K λέξεις (βλ. [Brill 1995] για λεπτομέρειες). Η ακρίβεια που επιτεύχθηκε έφτασε το 82.2% για τις άγνωστες λέξεις και 96.6% για το σύνολο των λέξεων.

3.1.2 Daelemans - Πειράματα με τον MBT

Ιδιαίτερα σημαντικά είναι τα αποτελέσματα που προέκυψαν από την έρευνα της ομάδας των Daelemans-Zavrel, των κατασκευαστών του συστήματος MBT που χρησιμοποιήθηκε στην παρούσα εργασία. Κι αυτό γιατί η έρευνά τους επεκτάθηκε και στην χρήση άλλων τεσσάρων γλωσσών, πλην της αγγλικής. Η έρευνα ξεκίνησε με την χρήση ενός σώματος κειμένων της Αγγλικής γλώσσας [Daelemans et al. 1996]. Αργότερα, η έρευνα επεκτάθηκε και σε ένα δεύτερο σώμα Αγγλικών κειμένων, καθώς και σε κείμενα γραμμένα στα Δανέζικα, τα Τσέχικα, τα Ισπανικά και τα Σουηδικά [Zavrel et al. 1999]. Τα σύνολα ετικετών που χρησιμοποιήθηκαν δεν ήταν τα ίδια σε όλες τις γλώσσες.

Όπως προαναφέραμε, τα πρώτα πειράματα [Daelemans et al. 1996] πραγματοποιήθηκαν σε Αγγλικά κείμενα, και συγκεκριμένα σε ένα σώμα κειμένων της Wall Street Journal, μεγέθους 2200K λέξεων. Το σύνολο των ετικετών που επελέγη περιελάμβανε 44 στοιχεία. Κατ' αρχάς, χρησιμοποιήθηκε ένα τμήμα του συνολικού σώματος, μεγέθους 110K λέξεων (100K για εκπαίδευση και 10K για έλεγχο) για την σύγκριση των αλγορίθμων IB1, IB1-IG και IGTREE του συστήματος. Η χρήση του IG βελτιώνει την απόδοση του αλγορίθμου κατά 3.5% (φτάνει στο 96%), χωρίς παράλληλα να απαιτεί σημαντικό ποσό επιπλέον χρόνου. Η χρήση, όμως, του IGTREE πετυχαίνει την ίδια ακρίβεια (96%) μειώνοντας τον απαιτούμενο χρόνο πάνω από 100 φορές (περίπου 46 λεπτά έναντι 29 δευτερολέπτων). Αυτό συνέτεινε στην χρήση του στο σύνολο των πειραμάτων που πραγματοποιήθηκαν και που αναφέρονται παρακάτω. Έπειτα πραγματοποιήθηκε έρευνα στο σύνολο του κειμένου, με εκπαίδευση στα 2000K λέξεις και έλεγχο στις 200K λέξεις. Οι ρυθμίσεις που χρησιμοποιήθηκαν ήταν οι *ddfa*[1] για τις γνωστές και *pdFasss* για τις άγνωστες λέξεις. Στα πειράματα έγινε χρήση της δεκαπλής διασταυρωμένης επικύρωσης, ενώ δημιουργήθηκε και η καμπύλη μάθησης, με βήμα 100K λέξεων. Τα αποτελέσματα ανέδειξαν συνολική ακρίβεια του συστήματος 96.4%, ενώ παράλληλα υπέδειξαν το σημείο των 700K λέξεων ως ελάχιστο μέγεθος του συνόλου εκπαίδευσης για την λήψη ικανοποιητικής ακρίβειας.

[1] Λεπτομέρειες για τις ρυθμίσεις των παραμέτρων του MBT/TiMBL αναφέρονται στην ενότητα 4.1

Στο δεύτερο σύνολο πειραμάτων [Zavrel et al. 1999], χρησιμοποιήθηκε και ένα δεύτερο σύνολο από κείμενα Αγγλικών, το LOB [Johansson 1986], μεγέθους 1046K λέξεων (931K λέξεις για εκπαίδευση και 115K για έλεγχο). Στην διαδικασία χρησιμοποιήθηκαν 170 ετικέτες, ενώ οι ρυθμίσεις του MBT ήταν ίδιες με αυτές που αναφέρθηκαν παραπάνω. Στην προκειμένη περίπτωση η ακρίβεια έφτασε στο 97%. Ακολούθως, πραγματοποιήθηκε έρευνα για την ακρίβεια που επιτυγχάνει το σύστημα σε κείμενα άλλων γλωσσών. Τα μεγέθη των κειμένων, καθώς και των συνόλων ετικετών που χρησιμοποιήθηκαν, κυμαίνεται ανάλογα με την γλώσσα. Το ίδιο ισχύει και με τις παραμέτρους που θα πρέπει να δοθούν στον MBT.

Γλώσσα	**Ρυθμίσεις**	**Μέγεθος συνόλου ετικετών**	**Μέγεθος κειμένων (λέξεις x1000)**	**% Ακρίβεια**
Δανέζικα	ddfa / pdFasss	13	711	95.7
Τσέχικα	ddfa / pdFasss	42	595	93.6
Ισπανικά	ddfwaa / chndFasss	484	800	97.8
Σουηδικά	ddwfwaa / chndFasss	23	1167	95.6

Πίνακας 3.1.2.1

Τέλος, χρησιμοποιώντας το δεύτερο σώμα αγγλικών κειμένων, επιχειρήθηκε μια προσπάθεια σύγκρισης των αποτελεσμάτων με αυτά άλλων συστημάτων, του TDEB, ενός που χρησιμοποιεί [Steetskamp 1995] trigrams και ενός που χρησιμοποιεί την μέθοδο της Μέγιστης Εντροπίας [Ratnaparkhi 1996]. Τα αποτελέσματα έδειξαν μια σαφή υπεροχή του MBT κατά 0.5% και 0.9% έναντι της ακρίβειας που επιτυγχάνουν οι δυο πρώτοι, όμως υπολείπεται από το σύστημα της Μέγιστης Εντροπίας κατά 0.4% (97.4%). Τέλος θα πρέπει να αναφερθεί ότι επιτεύχθηκε βέλτιστο αποτέλεσμα 97.9% με συνδυασμό των αποτελεσμάτων και των τεσσάρων μεθόδων.

3.2 Πειράματα για την ελληνική γλώσσα

Στη συνέχεια του κεφαλαίου, θα παρουσιαστούν οι σημαντικότερες προσπάθειες διερεύνησης της χρήσης αλγορίθμων Μ.Μ. για την εύρεση των μερών του λόγου λέξεων που βρίσκονται σε κείμενα γραμμένα στην ελληνική γλώσσα.

3.2.1 Πειράματα με συστήματα βασισμένα στα μοντέλα Hidden Markov

Στο τμήμα αυτό περιγράφονται οι προσπάθειες που έγιναν από την ομάδα των Δερματά-Κοκκινάκη [Dermatas and Kokkinakis 1995], και οι οποίες βασίζονται στην χρήση στοχαστικών συστημάτων που χρησιμοποιούν μοντέλα Hidden Markov (HMM). Στην προσπάθεια αυτή γίνεται χρήση πέντε, συνολικά, παραλλαγών μοντέλων HMM. Επίσης η προσπάθεια αυτή επεκτείνεται και στην έρευνα της απόδοσης της μεθόδου και σε άλλες γλώσσες, συνολικά σε επτά, που είναι: Ελληνικά, Αγγλικά όπου και χρησιμοποιούνται δυο σύνολα κειμένων (ένα με άρθρα εφημερίδων και ένα με νομικά κείμενα), Γερμανικά, Δανέζικα, Γαλλικά, Ιταλικά και Ισπανικά. Επίσης, για την επισημείωση των λέξεων χρησιμοποιούνται δύο σύνολα από ετικέτες, ένα μικρό και ένα εκτεταμένο. Φυσικά το μέγεθός τους δεν είναι το ίδιο για όλες τις γλώσσες.

Στον πίνακα που ακολουθεί παρατίθενται τα μεγέθη των συνόλων ετικετών και των κειμένων που χρησιμοποιήθηκαν στις δυο σειρές πειραμάτων.

Γλώσσα		Μέγεθος		
		Μικρό σύνολο ετικετών	Μεγάλο σύνολο ετικετών	Αρχείου εκπαίδευσης
Δανέζικα		9	50	110K λέξεις
Αγγλικά	Ειδήσεις	10	43	180K λέξεις
	Νομικά	10	36	110K λέξεις
Γαλλικά		10	14	100K λέξεις
Γερμανικά		11	116	100K λέξεις
Ελληνικά		11	443	120K λέξεις
Ιταλικά		10	121	160K λέξεις
Ισπανικά		10	121	60K λέξεις

Η πειραματική διαδικασία γίνεται ξεκινώντας με μέγεθος εκπαίδευσης τις 10K λέξεις και αυξάνοντάς το επίσης κατά 10K κάθε φορά. Με την χρήση του μικρού συνόλου ετικετών, το ποσοστό ακρίβειας φτάνει σε αρκετά υψηλά επίπεδα. Ξεκινώντας με τις 10000 λέξεις, είναι ιδιαίτερα μικρό στην αρχή και αυξάνεται κατακόρυφα μέχρι το πρώτο 25% του συνόλου εκπαίδευσης. Στο υπόλοιπο τμήμα αυξάνεται και πάλι αλλά

με μικρότερους ρυθμούς, για να φτάσει στο τέλος 93-96% για τις περισσότερες γλώσσες εκτός των ισπανικών που φτάνει στο 88%.

Από την άλλη, η αύξηση του μεγέθους του συνόλου ετικετών προκαλεί πολύ περισσότερα προβλήματα, δηλαδή σημαντική αύξηση του ποσοστού σφαλμάτων. Η γενική μορφή των καμπυλών μάθησης είναι παρόμοια. Το τελικό αποτέλεσμα όμως διαφέρει σημαντικά από πριν. Έτσι, σε όλες σχεδόν τις περιπτώσεις (εκτός από το 2^{ο} σώμα κειμένων των αγγλικών όπου το τελικό ποσοστό ακρίβειας παραμένει σταθερό περίπου στο 96%) η επιτευχθείσα ακρίβεια μειώνεται σε επίπεδα που βρίσκονται ανάμεσα στο 83 και 88%.

3.2.2 Πειράματα με συστήματα βασισμένα σε δένδρα απόφασης

Συστήματα που βασίζονται στην εκμάθηση δένδρων απόφασης χρησιμοποιήθηκαν στις προσπάθειες που έγιναν από την ομάδα των Ορφανού-Χριστοδουλάκη [Orphanos,Christodoulakis et al. 1999, Orphanos,Kalles et al. 1999]. Επίσης, χρησιμοποιήθηκε και ο αλγόριθμος IGTREE του συστήματος TiMBL, με σκοπό την σύγκριση της απόδοσής του σε σχέση με τις άλλες παραλλαγές αλγορίθμων εκμάθησης δένδρων απόφασης που χρησιμοποιήθηκαν.

Για την πραγματοποίηση των πειραμάτων χρησιμοποιήθηκε ένα σύνολο κειμένων με μέγεθος 137,765 λέξεων. Η προέλευση των τμημάτων που το συνθέτουν ποικίλλει και αποτελείται από γραπτά φοιτητών, τμήματα λογοτεχνικών κειμένων, άρθρα από εφημερίδες, τεχνικά, οικονομικά και αθλητικά περιοδικά. Έγινε διαχωρισμός των λέξεων και αφέθηκε σε ένα λεξικό η αυτόματη, πρώτη απόδοση των ετικετών. Το μέγεθος των ετικετών δεν καθορίστηκε αυστηρά, και αφέθηκε στην αρμοδιότητα ενός προγράμματος που συνεργάζονταν με ένα λεξικό να πραγματοποιήσει την αρχική επισημείωση και να επιλέξει το μέγεθος της ετικέτας που θα αποδοθεί στις λέξεις που ανήκουν σε κάθε μέρος του λόγου. Ακολούθησε χειρωνακτική διόρθωση των κειμένων.

Στα πειράματα χρησιμοποιήθηκε δεκαπλή διασταυρωμένη επικύρωση. Η σύγκριση των αποτελεσμάτων υποδεικνύει μια σαφή υπεροχή της τάξης του 2% του IGTREE

έναντι στους άλλους αλγορίθμους, τόσο στις γνωστές όσο και στις άγνωστες λέξεις (για τις γνωστές είναι ~5.5% έναντι των ~7.4% και ~6.5% ενώ για τις άγνωστες ~16% σε αντίθεση με τα ~18% και ~16%). Αντίθετα, η χρήση της τρίτης παραλλαγής, βελτιώνει κατά πολύ την ακρίβεια της διαδικασίας, ρίχνοντας τα ποσοστά έως το ~4.8% για τις γνωστές και ~12.3% για τις άγνωστες λέξεις).

3.2.2 Προσαρμογή του Brill tagger στα Ελληνικά

Σε αυτή την παράγραφο παρατίθενται δυο προσπάθειες που έγιναν για την ελληνική γλώσσα, βασισμένες στο ίδιο σύστημα, τον Brill tagger. Τα αποτελέσματα δεν μπορούν να τύχουν σύγκρισης μεταξύ τους, όμως έχει ιδιαίτερο ενδιαφέρον η παρατήρησή τους σε σχέση και με τις αλλαγές που έγιναν στο σύστημα.

3.2.2.1 Πειράματα της ερευνητικής ομάδας του «ΕΚΕΦΕ Δημόκριτος»

Η πρώτη από τις προσπάθειες που βασίστηκε στον αλγόριθμο TBED, ήταν αυτή της ερευνητικής ομάδας του «Δημόκριτου» [Petatsis et al. 1999]. Η επιλογή του παραπάνω συστήματος έγινε βάσει του γεγονότος ότι είναι ένα προϊόν που είναι ελεύθερα διαθέσιμο στο κοινό, και το οποίο έχει αποδείξει την υψηλή απόδοση που μπορεί να επιτύχει. Το τελευταίο επιβεβαιώνεται από τα αποτελέσματα των πειραμάτων που έχουν γίνει τόσο σε κείμενα της αγγλικής γλώσσας, όσο και σε κείμενα γραμμένα στα γερμανικά, τα γαλλικά, τα ιταλικά και τα εσθονικά. Τέλος, θα πρέπει να τονιστεί ότι ιδιαίτερη σημασία δόθηκε και το γεγονός πως βάσει των αποτελεσμάτων των πειραμάτων που είχαν αναφερθεί από άλλους ερευνητές, συστήματα που βασίζονται σε κανόνες (rule-based taggers) επιτυγχάνουν, κατά κανόνα, μεγαλύτερο βαθμό ακρίβειας από άλλα συστήματα που εκτελούν την ίδια εργασία.

Το σύνολο ετικετών που χρησιμοποιήθηκε για την διερεύνηση ήταν αρκετά περιορισμένο, αποτελούνταν μόλις από 58 στοιχεία, αν σκεφτεί κανείς το πλήθος των χαρακτηριστικών που μπορούν να αποδοθούν σε μια ελληνική λέξη. Ο λόγος όμως που έγινε μια τέτοια επιλογή ήταν η επίτευξη μεγαλύτερης συνέπειας. Επίσης, θα

πρέπει να προστεθεί, ότι η φύση της γλώσσας και των κειμένων, απαίτησε την προσθήκη ενός κανόνα στον κώδικα του συστήματος, με σκοπό την αναγνώριση και κατηγοριοποίηση και των ξένων λέξεων (ας μην ξεχνάμε ότι το αρχικό σύστημα σχεδιάστηκε και χρησιμοποιήθηκε σε γλώσσες με λατινογενή αλφάβητα).

Για την διεξαγωγή των πειραμάτων χρησιμοποιήθηκαν δυο σύνολα κειμένων. Το πρώτο αποτελούνταν από άρθρα μιας ελληνικής εφημερίδας που είχαν θέμα τις μετακινήσεις στελεχών επιχειρήσεων. Το σώμα περιείχε 65K λέξεις. Στην φάση της προεπεξεργασίας, ένα τμήμα του, αποτελούμενο από περίπου 36K λέξεις, έτυχε χειρωνακτικής επισημείωσης, με σκοπό την χρήση του στα πειράματα. Το δεύτερο σύνολο αποτελούσε μια συλλογή κειμένων ποικίλου ενδιαφέροντος. Αποτελούνταν από 125K λέξεις, και είχε, επίσης, τύχει χειρωνακτικής επισημείωσης, στο σύνολό του.

Με σκοπό την εξαγωγή πιο αμερόληπτων συμπερασμάτων, κατά την φάση των πειραμάτων χρησιμοποιήθηκε δεκαπλή διασταυρωμένη επικύρωση, για κάθε σύνολο κειμένων ξεχωριστά. Και στις δυο περιπτώσεις, η ακρίβεια του συστήματος, φαινόταν να αυξάνει με την μεγέθυνση του σώματος εκπαίδευσης, καταλήγοντας σε ισορροπία στα επίπεδα του 95%. Η ισορροπία, μάλιστα, αυτή ξεκινά να παρουσιάζεται από το μέγεθος των 18K λέξεων, σημείο που αποτελεί και το προτεινόμενο μέγεθος για το σώμα εκπαίδευσης. Επίσης θα πρέπει να τονιστεί ότι η παρουσία παρόμοιων αποτελεσμάτων αποδεικνύει ότι η απόδοση του συστήματος δεν εξαρτάται από τα χαρακτηριστικά του κειμένου που καλείται να επισημειώσει.

Παράλληλα, έγινε και έρευνα για το μέγεθος του συνόλου των κανόνων (λεκτικών και θεματικών) σε σχέση με το μέγεθος του σώματος εκπαίδευσης. Τα αποτελέσματα έδειξαν γραμμική αύξηση του πλήθους των κανόνων. Επίσης απέδειξαν ότι οι λεκτικοί κανόνες αυξάνουν με πιο γρήγορο ρυθμό από τους θεματικούς. Το παραπάνω αποδόθηκε στην αδυναμία (από κατασκευής) του tagger να χειρίζεται τις μορφολογικές ιδιαιτερότητες που παρουσιάζονται στην ελληνική γλώσσα. Αυτό, όμως, μπορεί να αποτελέσει και «πηγή προβλήματος» για το σύστημα, κυρίως σε περιπτώσεις που θα πρέπει να χρησιμοποιηθεί εκτεταμένου μεγέθους σώμα εκπαίδευσης και η διαδικασία θα πρέπει να βασιστεί στη χρήση των λεκτικών κανόνων.

3.2.2.2 Πειράματα της ομάδας του ΙΕΛ

Η πιο σημαντική, ίσως, από τις ερευνητικές απόπειρες στον τομέα της αυτόματης αναγνώρισης των μερών του λόγου λέξεων ελληνικών κειμένων ήταν αυτή που διεξήχθη από την ερευνητική ομάδα του Ινστιτούτου Επεξεργασίας Λόγου (ΙΕΛ). Η προσπάθεια αυτή, όπως και η προηγούμενη (παράγραφος 3.2.2.1) βασίστηκε στο σύστημα που κατασκεύασε ο E.Brill. Στην προκειμένη περίπτωση έγινε χρήση του συγκεκριμένου συστήματος σε δυο μορφές: η πρώτη ήταν η αρχική του μορφή, όπως δηλαδή προσφέρεται ελεύθερα από τον E.Brill. Η δεύτερη αποτελεί μια μετατροπή του παραπάνω συστήματος, που χρησιμοποιεί χαρακτηριστικά του κειμένου για την αναγνώριση των λέξεων. Εκτός αυτού, από την ερευνητική ομάδα κατασκευάστηκε και ένας σημαντικός αριθμός από ενισχυτικά προγράμματα τα οποία έχουν σκοπό να βελτιώσουν την συνολική διαδικασία. Στα προγράμματα αυτά περιλαμβάνεται ένα σύστημα διαχωρισμού τον λέξεων (tokenizer), ένα σύστημα επισημείωσης λέξεων με γραφική διεπαφή, ένα σύστημα που μπορεί να οπτικοποιεί τα στατιστικά αποτελέσματα που προκύπτουν από τις μετρήσεις και ένα σύστημα που μπορεί να διαχειρίζεται τους πόρους του συστήματος.

Το σημαντικό στοιχείο που δίνει εξέχουσα σημασία στην συγκεκριμένη προσπάθεια σε σχέση με όλες τις άλλες που αφορούν την ελληνική γλώσσα είναι ότι το σύνολο ετικετών που χρησιμοποιήθηκε είναι ιδιαίτερα εκτεταμένο (φυσικά δεν μπορεί να γίνει σύγκριση με άλλες γλώσσες όπως τα αγγλικά, όπου το σύνολο ετικετών είναι υπερβολικά περιορισμένο). Για την ακρίβεια αποτελείται από 584 σύμβολα, περιλαμβάνοντας έτσι το σύνολο των ετικετών του έργου PAROLE, όπως αυτό προσαρμόστηκε για την ελληνική γλώσσα. Το συγκεκριμένο σύνολο ετικετών είναι σε θέση να καλύψει το 100% των περιπτώσεων γραμματικής κατάταξης μιας λέξης, παρέχοντας παράλληλα στον χρήστη το μέγιστο δυνατό ποσό πληροφορίας που μπορεί να λάβει για μια λέξη.

Το σύνολο κειμένων που χρησιμοποιήθηκε για τα πειράματα αποτελείται από 210 αρχεία συνολικού μεγέθους, περίπου, 447K λέξεων. Ιδιαίτερη προσοχή δόθηκε ώστε το περιεχόμενό τους να καλύπτει όσο το δυνατόν περισσότερα και διαφορετικά

αντικείμενα, χωρίς όμως κάποιο από αυτά να υπερβαίνει, σε μέγεθος, κάποιο από τα άλλα. Έτσι συλλέχθηκαν από συνολικά 17 διαδικτυακές πηγές κείμενα που αφορούν από οικονομικές έως πολιτικές ειδήσεις, και από συνεντεύξεις έως και αποτελέσματα ελέγχου υλικού ηλεκτρονικών υπολογιστών.

Τα κείμενα πέρασαν από αρκετά στάδια επεξεργασίας, τόσο αυτοματοποιημένης όσο και ανθρώπινης, φτάσουν στην τελική τους μορφή. Πρώτα διαχωρίστηκαν οι λέξεις με την χρήση του προγράμματος διάσπασης των λέξεων, ενώ παράλληλα έγινε και μια πρώτη επισημείωση ορισμένων σημείων όπως οι αριθμοί, τα σύμβολα, οι ημερομηνίες, κ.τ.λ. Έπειτα, με την χρήση ενός μικρού, ήδη επισημειωμένου, σώματος κειμένων με το οποίο εκπαιδεύτηκε ο TBED, έγινε μια πρώτη επισημείωση του σώματος κειμένων. Τέλος, εισήχθη ο ανθρώπινος παράγοντας. Δυο γλωσσολόγοι ανάλαβαν, εργαζόμενοι επί ένα τρίμηνο, να διορθώσουν τα εξαγόμενα της παραπάνω διαδικασίας. Παράλληλα ανάλαβαν να «επισημοποιήσουν» και να κατηγοριοποιήσουν τους κανόνες που χρησιμοποιούσαν. Για την δημιουργία των τελικών κειμένων, έγινε διόρθωση ώστε κάθε γραμμή να περιλαμβάνει μια περίοδο. Έπειτα χωρίστηκαν σε 3 τμήματα, που περιείχαν μόνο ολόκληρες περιόδους: ένα που αντιστοιχεί στο 20% του αρχικού (με περίπου 90000 λέξεις) και δυο άλλα από 40% το καθένα (με περίπου 178K λέξεις). Το πρώτο ήταν το σύνολο ελέγχου ενώ τα άλλα δυο τα σύνολα εκπαίδευσης. Από τα τελευταία το πρώτο χρησιμοποιήθηκε για την εξαγωγή των λεκτικών και το δεύτερο για την εξαγωγή των «περιβαλλοντικών» κανόνων. Από το πρώτο δημιουργήθηκε και ένα λεξικό, με όλες τις διαφορετικές λέξεις (περίπου 25000) ακολουθούμενες από όλες τις πιθανές τους ετικέτες.

Ακολούθως έγιναν πειράματα τα οποία όπως αναφέραμε και παραπάνω αφορούσαν δυο μορφές του συστήματος. Για την πρώτη μορφή (TDEB), πραγματοποιήθηκε εκπαίδευση του συστήματος για την παραγωγή των κανόνων που απαιτούνταν. Για την δεύτερη (FBT), και προκειμένου να μειωθεί ο χρόνος χρησιμοποιήθηκε ένα λεξικό προθέματος-ετικέτας που είχε δημιουργηθεί από το σώμα εκπαίδευσης, για την απόδοση των αρχικών τιμών στις ετικέτες των λέξεων. Επίσης για την εύρεση των κανόνων του περιεχομένου, διεξήχθη εκπαίδευση σε τέσσερα στάδια. Έτσι πραγματοποιήθηκε κλιμακωτή δημιουργία της τελικής ετικέτας, η οποία σε κάθε στάδιο εμπλουτίζονταν και με περισσότερα στοιχεία που αφορούσαν τα χαρακτηριστικά της λέξης.

Τα αποτελέσματα των πειραμάτων έδειξαν μια μικρή υπεροχή του FBT έναντι του TBED. Η υπεροχή αυτή υπάρχει τόσο όταν χρησιμοποιούνται τα βασικά χαρακτηριστικά, όσο και όταν γίνεται χρήση του συνόλου αυτών. Μια δεύτερη σειρά πειραμάτων πραγματοποιήθηκε μετά από τον διαχωρισμό των κειμένων στις κατηγορίες τις οποίες ανήκουν, για να μελετηθεί η απόδοση του FBT σε σχέση με το περιεχόμενο και την δομή του κειμένου. Οι κατηγορίες ήταν πέντε: γενικά, τεχνικά, οικονομικά κείμενα, ενημερώσεις τύπου και διάλογοι. Η απόδοση του FBT φαίνεται να είναι αρκετά υψηλή στις τέσσερις πρώτες κατηγορίες (περίπου κατά Μ.Ο. 91%) με βέλτιστη αυτή των οικονομικών κειμένων που βρίσκεται περίπου στο 92%, για το σύνολο των χαρακτηριστικών. Εξαίρεση αποτελεί η κατηγορία των διαλόγων, που λόγω της απότομης μεταπήδησης, δημιουργεί προβλήματα ροή και ρίχνει την συνολική απόδοση στο 86%.

4. Το λογισμικό, το σύνολο ετικετών και τα σώματα κειμένων της εργασίας

Το κεφάλαιο αυτό αποτελεί ουσιαστικά μια αναλυτική περιγραφή της φάσης που προηγήθηκε της πειραματικής διαδικασίας. Στις παραγράφους που ακολουθούν γίνεται εκτενής αναφορά σε ό,τι σχετίζεται με το προπαρασκευαστικό στάδιο των πειραμάτων. Αρχικά γίνεται μια λεπτομερής αναφορά στο MBT/TiMBL, το σύστημα που χρησιμοποιήθηκες για την πραγματοποίηση των πειραμάτων. Έπειτα ακολουθεί η περιγραφή των συστημάτων WordTagger και ErrorDetector, δυο υποστηρικτικών συστημάτων απαραίτητων για την διαδικασία, τα οποία αναπτύχθηκαν στα πλαίσια της εργασίας. Τέλος, γίνεται αναφορά στο σύνολο ετικετών που επελέγη να χρησιμοποιηθεί αλλά και στην προεπεξεργασία που υπέστησαν τα δυο σώματα κειμένων που χρησιμοποιήθηκαν για τα πειράματα που περιγράφουμε στο κεφάλαιο 5.

4.1 Το σύστημα MBT/TiMBL

Το σύστημα MBT/TiMBL, αποτελεί ένα προϊόν λογισμικού, που κατασκευάστηκε από την ερευνητική ομάδα των Daelemans, Zavrel, van der Sloot και van den Bosch του πανεπιστημίου του Tilbourg σε συνεργασία με την αντίστοιχη ομάδα από το πανεπιστήμιο του Antwerp. Είναι ένα σύστημα το οποίο εμπεριέχει υλοποίηση αρκετών αλγορίθμων μηχανικής μάθησης, προσαρμοσμένων έτσι ώστε να πετυχαίνουν όσο το δυνατόν καλύτερο αποτέλεσμα στο πρόβλημα που καλούνται να επιλύσουν. Το σύστημα διατίθεται ελεύθερα σε μορφή open source μέσο του διαδικτύου[2]. Αποτελείται από δυο τμήματα: το πρώτο είναι ο TiMBL (Tilburg Memory-Based Learner) που είναι το σύστημα μάθησης, αυτό που περιέχει την υλοποίηση των αλγορίθμων. Το δεύτερο είναι ο MBT (Memory-Based Tagger) που αποτελεί τον επισημειωτή.

[2] Η διεύθυνση στην οποία μπορεί να βρεθεί είναι: http://ilk.kub.nl/software.html

Κατά την διαδικασία της εκπαίδευσης, το σύστημα πραγματοποιεί ανάγνωση του κειμένου-παραδείγματος που του παρασχέθηκε, και δημιουργεί τα αρχεία-λεξικά (περιέχουν τις ετικέτες των λέξεων στις οποίες εκπαιδεύτηκε) που του είναι απαραίτητα για την διαδικασία της αναγνώρισης. Επίσης αποθηκεύει και πληροφορία που αφορά τις ιδιότητες της λέξης που θα πρέπει να βασιστεί και τους αλγορίθμους που θα πρέπει να χρησιμοποιήσει. Όλα αυτά, όπως φαίνονται και παρακάτω, μπορούν να του δοθούν ως παράμετροι κατά την κλήση του.

4.1.1 Ο TiMBL

Το όνομα TiMBL προκύπτει από τα αρχικά των Tilbourg Memory-Based Learner, και αποτελεί το τμήμα του συστήματος που υλοποιεί την διαδικασία της μάθησης. Με σκοπό την επίτευξη όσο το δυνατόν καλύτερων αποτελεσμάτων, οι δημιουργοί του υλοποίησαν ένα σημαντικό αριθμό από διαφορετικούς αλγορίθμους, εξοπλίζοντας παράλληλα τον καθένα με ένα μεγάλο πλήθος παραμέτρων βελτιστοποίησης.

Οι σημαντικότερες παράμετροι που μπορεί να ορίσει ο χρήστης είναι:

- Επιλογή του αλγορίθμου με αναζήτησης. Μετά το σημείο **–a** τοποθετούμε:
  - **0**, για τον IB1, μια παραλλαγή του k-NN
  - **1**, για τον IGTREE
  - **2**, για τον TRIBL, έναν υβριδικό αλγόριθμο, που παράγεται από συνδυασμό των δυο προηγούμενων
  - **4**, για τον TRIBL2
- Επιλογή τρόπου απόδοσης βάρους στις ιδιότητες. Μετά το σημείο **–w** τοποθετούμε:
  - **0**, για να έχουμε ίδιο βάρος σε όλα τα χαρακτηριστικά (w=1)
  - **1**, για την χρήση του Gain Ratio
  - **2**., για την χρήση του Information Gain
- Επιλογή του αριθμού των κοντινότερων γειτόνων, στην περίπτωση που έχει επιλεγεί ένας εκ των αλγορίθμων IB1, IB2, TRIBL, TRIBL2. Την παράμετρο **–k** ακολουθεί ο αριθμός που δηλώνει τον αριθμό των γειτόνων. Η προεπιλεγμένη τιμή είναι το 1.

4.1.2 Ο MBT

Ο MBT αποτελεί το μέρος του συστήματος που επιτελεί την λειτουργία της επισημείωσης. Είναι αυτό που παρέχει στον χρήστη και την διεπαφή βάσει της οποίας θα ορίσει τις παραμέτρους που αυτός επιθυμεί. Αποτελείται από δυο επιμέρους προγράμματα: τον Mbtg και τον Mbt.

Ο Mbtg (Memory-Based Tagger Generator) είναι η εφαρμογή που χρησιμοποιείται κατά την φάση της εκπαίδευσης του συστήματος. Κατά την κλήση του, δίνεται το όνομα του χειρωνακτικά επισημειωμένου αρχείου και οι παράμετροι που αφορούν τη διαδικασία και παράγονται τα αρχεία λεξικού που χρειάζεται η εφαρμογή για να πραγματοποιήσει την επισημείωση ενός «καθαρού» κειμένου.

Η κλήση του Mbtg γίνεται με την ακόλουθη εντολή:
Mbtg –T <filename> -O "<TiMBL options>" –p <options> -P <options>

Με <filename> συμβολίζουμε το όνομα του επισημειωμένου αρχείου. Αυτό θα πρέπει να είναι μορφοποιημένο με τέτοιο τρόπο, ώστε κάθε γραμμή να περιέχει και από μια λέξη, η οποία θα ακολουθείται από το tag της. Ο διαχωρισμός των δυο θα γίνεται με την χρήση ενός tab. Εκτός αυτού, στο τέλος της κάθε περιόδου θα πρέπει να εισάγεται μια γραμμή με το σύμβολο <*utt*>, έτσι ώστε κατά την ανάγνωση του κειμένου να μπορεί να γίνει αντιληπτή η αλλαγή της.

Στη θέση <options> μετά τα –p και -P, τοποθετούνται από μια σειρά χαρακτήρων που συμβολίζουν τα στοιχεία της κάθε λέξης που θα πρέπει να λάβει υπόψη του το πρόγραμμα, τόσο για την κατασκευή του λεξικού, όσο και για την φάση της επισημείωσης. Στην πρώτη περίπτωση τοποθετούνται οι ρυθμίσεις που αφορούν τις γνωστές, ενώ στην δεύτερη αυτές που αφορούν τις άγνωστες λέξεις. Οι χαρακτήρες που επιτρέπεται να χρησιμοποιηθούν είναι οι εξής:

1) Για γνωστές και άγνωστες λέξεις
   a) **d**: Αφορά την ετικέτα λέξεων που προηγούνται και για τις οποίες έχει αποσαφηνιστεί η ετικέτα τους.
   b) **a**: Αφορά την ετικέτα των λέξεων που ακολουθούν και για τις οποίες δεν έχει αποσαφηνιστεί η ετικέτα που πρέπει να τους αποδοθεί.

c) **w**: Αφορά τις ίδιες λέξεις που προηγούνται ή ακολουθούν (εξαρτάται από την θέση στην οποία έχει τοποθετηθεί το σύμβολο).

2) Μόνο για τις γνωστές λέξεις
   a) **f**: Αναφέρεται στην λέξη της οποίας το tag προσπαθούμε να προσδιορίσουμε. Στην συγκεκριμένη ιδιότητα προσδίδονται όλες οι πιθανές ετικέτες που μπορεί να αποδοθούν στη λέξη και από τις οποίες θα επιλεγεί η τελική της ετικέτας.
3) Μόνο για τις άγνωστες λέξεις
   a) **F**: Αναφέρεται στην λέξη της οποίας την ετικέτα προσπαθούμε να προσδιορίσουμε.
   b) **c**: Αναγνώριση ύπαρξης κεφαλαίων γραμμάτων στην διερευνούμενη λέξη.
   c) **h**: Αναγνώριση ύπαρξης αποστρόφου στην διερευνούμενη λέξη.
   d) **n**: Αναγνώριση ύπαρξης αριθμητικών συμβόλων στην διερευνούμενη λέξη.
   e) **p**: Χρήση ενός χαρακτήρα που ανήκει στην αρχή της λέξης.
   f) **s**: Χρήση ενός χαρακτήρα που ανήκει στο τέλος της λέξης.

Στο παράδειγμα του πίνακα 4.1 που ακολουθεί φαίνεται πως χρησιμοποιούνται τα παραπάνω χαρακτηριστικά για τον προσδιορισμό της ετικέτας των λέξεων της πρότασης “Pierre Vinken, 21 years old”. Αριστερά παρουσιάζεται το ενδεχόμενο όλες οι λέξεις να θεωρούνται γνωστές, ενώ δεξιά το ενδεχόμενο η κάθε λέξη να είναι άγνωστη.

Λέξεις	**Γνωστές**					**Άγνωστες**						
	d	**d**	**f**	**a**	**Αποτέλεσμα**	**p**	**d**	**a**	**s**	**s**	**s**	**Αποτέλεσμα**
Pierre	-	-	np	np	np	P	-	np	r	r	e	np
Vinken	-	np	np	,	np	V	np	,	k	e	n	np
,	np	np	,	cd	,	-	-	-	-	-	-	-
61	np	,	cd	nns	cd	6	,	nns	-	6	1	cd
years	,	cd	nns	jj-np	nns	y	cd	jj-np	a	r	s	nns
old	cd	nns	jj-np	,	jj	o	nns	,	o	l	d	jj

Πίνακας 4.1

Σε περίπτωση που η λέξη είναι γνωστή (έχει συναντηθεί στο κείμενο εκπαίδευσης), λαμβάνονται υπόψη οι ετικέτες των δυο προηγούμενων λέξεων (όπως αυτές

αναγνωρίστηκαν από το σύστημα, η ετικέτα που πιθανόν να ανήκει στην επόμενη λέξη (εφόσον αυτή θεωρείται γνωστή), καθώς και όλες οι πιθανές ετικέτες που μπορούν να αποδοθούν στην διερευνούμενη λέξη, βάσει της συχνότητας εμφάνισής τους στο κείμενο εκπαίδευσης. Αντίθετα, αν η λέξη είναι άγνωστη, προσπαθεί να προσδιοριστεί η ετικέτα της βάσει του πρώτου γράμματός της, των τριών γραμμάτων της κατάληξης, της ετικέτας της προηγούμενης λέξης και της πιθανής ετικέτας της επόμενης.

Το τμήμα **-O "<TiMBL options>"** και χρησιμοποιείται μόνο στις περιπτώσεις που υπάρχει η επιθυμία χρήσης ενός άλλου αλγορίθμου. Οι παράμετροι για τους νέους αλγόριθμους τοποθετούνται στην θέση του <TiMBL options>.

Ο Mbt (Memory-Based Tagger) αποτελεί την δεύτερη εφαρμογή του συστήματος και είναι αυτή που πραγματοποιεί την επισημείωση ενός κειμένου. Η κλήση του γίνεται ως εξής:

Mbt –s <file > -t/-T <filename>

Η παράμετρος <file> αποτελεί το όνομα ενός αρχείου με κατάληξη *.settings*. Το αρχείο αυτό παράγεται από τον Mbtg κατά την φάση της εκπαίδευσης και περιέχει πληροφορίες σχετικά με τα αρχεία που βρίσκονται αποθηκευμένες οι πληροφορίες εκπαίδευσης (λεξικά), αλλά και στοιχεία που αφορούν τις παραμέτρους που χρησιμοποιήθηκαν στην φάση της εκπαίδευσης και οι οποίες θα χρησιμοποιηθούν και τώρα, στην φάση της επισημείωσης.

Η παράμετρος <filename>, που έπεται του συμβόλου –t ή –T, αποτελεί το όνομα του αρχείου που προορίζεται για την επισημείωση. Η χρησιμοποίηση του –t γίνεται στην περίπτωση που θέλουμε να επισημειώσουμε ένα καθαρό αρχείο. Το κείμενο βρίσκεται σε κανονική μορφή, διαβάζεται από το σύστημα και εμφανίζεται στην έξοδο επισημειωμένο. Σε αντίθεση, η χρήση του –T γίνεται για την περίπτωση που θέλουμε να ελέγξουμε την απόδοση ορισμένων ρυθμίσεων. Το κείμενο του οποίου το όνομα δίνουμε είναι ήδη επισημειωμένο και μορφοποιημένο σύμφωνα με το μορφότυπο εκπαίδευσης του Mbtg, που περιγράφαμε σε προηγούμενη παράγραφο. Το σύστημα διαβάζει το κείμενο, κάνει επισημείωση και συγκρίνει τα tags που απέδωσε το ίδιο με τα σωστά. Στο τέλος εμφανίζει το κείμενο με κάθε λέξη να

διαθέτει και τις δυο ετικέτες, ενώ ακολούθως εμφανίζει και τα αποτελέσματα των μετρήσεων που διεξήγαγε σχετικά με τα ποσοστά επιτυχίας.

4.2 Το σύστημα επισημείωσης WordTagger

Όπως παραδέχεται το σύνολο των ερευνητών που ασχολούνται με το πρόβλημα της επισημείωσης κειμένων, η ύπαρξη ενός σώματος κειμένων, επισημειωμένων με συνέπεια και ορθότητα (ενός Golden Corpus, όπως χαρακτηριστικά ονομάζεται), είναι ζωτικής σημασίας για την ομαλή πορεία της έρευνας, αλλά και για την εξαγωγή ορθών συμπερασμάτων. Επίσης, είναι γενικά αποδεκτό ότι ένα σύνολο κειμένων με αυτά τα χαρακτηριστικά είναι αδύνατο να δημιουργηθεί αν δεν υπεισέλθει και ο ανθρώπινος παράγοντας μέσα στην διαδικασία, στον ρόλο του ελεγκτή. Για το σκοπό αυτό, κρίθηκε απαραίτητο, στα πλαίσια της συγκεκριμένης εργασίας, να δημιουργηθεί ένα σύστημα χειρονακτικής επισημείωσης ελληνικών κειμένων, ένα προϊόν λογισμικού που θα βοηθάει τον χρήστη να αποδώσει, χειροκίνητα, από μια ετικέτα σε κάθε λέξη του κειμένου. Το σύστημα αυτό ονομάστηκε WordTagger, και κατασκευάστηκε με γνώμονα την φιλικότητα στον χρήστη, αλλά και την μεταφερσιμότητα.

4.2.1 Διεπαφή και λειτουργίες του WordTagger

Όπως αναφέρθηκε παραπάνω, πρωταρχικός στόχος ήταν η δημιουργία ενός εύχρηστου προϊόντος λογισμικού που θα βοηθήσει τον χρήστη να επισημειώσει ένα ελληνικό κείμενο, σε σχετικά μικρό χρονικό διάστημα. Επίσης κρίθηκε απαραίτητο το προϊόν αυτό να μπορεί να λειτουργήσει σε διάφορες πλατφόρμες, με όσο το δυνατόν λιγότερες αλλαγές στον κώδικά του.

Η πρώτη απαίτηση, αυτή δηλαδή της ευχρηστίας, ανέδειξε ως πρώτη ανάγκη την ύπαρξη γραφικής διεπαφής. Δεν υπάρχει καμία αμφιβολία, ότι ένα περιβάλλον τύπου γραμμής εντολών δεν είναι δυνατόν να βοηθήσει έναν χρήστη να εκτελέσει γρήγορα μια εργασία σαν την παραπάνω, που βασίζεται στην γρήγορη και σωστή επιλογή ενός σημείου ανάμεσα σε δεκάδες άλλα. Περαιτέρω έρευνα της διαδικασίας, φανέρωσε

πως η παραπάνω εφαρμογή θα πρέπει και να έχει την δυνατότητα να βοηθάει τον χρήστη να διορθώνει ένα κείμενο που περιέχει σφάλματα. Τέλος, και επειδή για την συγκεκριμένη εργασία χρησιμοποιήθηκε το σύστημα MBT, κρίθηκε απαραίτητο ο WordTagger να εμπλουτιστεί και με τη δυνατότητα αναγνώρισης του μορφότυπου (format) κειμένου που αναγνωρίζει ο MBT.

Από πλευράς λειτουργικότητας, προκύπτει ότι οι σημαντικότερες λειτουργίες, είναι αυτές της πλοήγησης μέσα στο κείμενο, της απόδοσης μιας ετικέτας σε μια λέξη, αλλά και της εμφάνισης της πληροφορίας που έχει αποδοθεί σε μια λέξη. Αυτό σημαίνει ότι οι λειτουργίες που αφορούν το άνοιγμα και το κλείσιμο των αρχείων, αλλά και της εφαρμογής είναι δευτερεύουσας σημασίας, και γι' αυτό και τοποθετούνται σε ένα menubar. Αντίθετα, οι υπόλοιπες τοποθετούνται στην κεντρική επιφάνεια της εφαρμογής. Όπως φαίνεται και στην Εικόνα 4.2.1.1, στο επάνω τμήμα εμφανίζεται το κείμενο, έτσι ώστε ο χρήστης να έχει μια καλή άποψη του περιβάλλοντος γύρω από την λέξη που τον ενδιαφέρει. Ακολουθούν τα πλήκτρα πλοήγησης και τα πεδία που εμφανίζουν την πληροφορία της εκάστοτε λέξης. Τέλος, εμφανίζονται σε μέσα σε combo boxes οι ετικέτες, χωρισμένες σε κατηγορίες. Επειδή, μάλιστα, ο χρόνος εύρεσης μιας ετικέτας μέσα σε ένα combo box αποτελεί σημαντικό παράγοντα καθυστέρησης, δίπλα σε καθένα από αυτά έχει τοποθετηθεί ένα πλήκτρο που αποδίδει στην επιλεγμένη λέξη την τελευταία επιλεγμένη ετικέτα από αυτό το combo box.

4.2.2 Περιβάλλον κατασκευής συστήματος

Η ανάγκη της μεταφερσιμότητας, σε συνδυασμό με τα όσα αναφέρθηκαν προηγουμένως, ήταν αυτά που οδήγησαν και στην επιλογή του περιβάλλοντος κατασκευής. Είναι σίγουρο ότι είναι αρκετές οι γλώσσες των οποίων ο κώδικας μπορεί να μεταφραστεί και σε άλλα συστήματα χωρίς αλλαγές, δίνοντας έτσι ένα πρόγραμμα που να τρέχει σωστά και στο νέο σύστημα. Η ανάγκη όμως για γραφική διεπαφή δυσχεραίνει τα πράγματα και μειώνει τις επιλογές. Η γλώσσα Java της SUN κρίθηκε τελικά ως η πιο κατάλληλη για την συγκεκριμένη εργασία. Κι αυτό γιατί λόγω του τρόπου λειτουργίας της, τα αρχεία java-bytecode που προκύπτουν από την φάση της μεταγλώττισης είναι ικανά να λειτουργήσουν, χωρίς καμιά άλλη ενέργεια

σε οποιοδήποτε σύστημα διαθέτει το περιβάλλον εκτέλεσης Java, JRE (Java Runtime Environment).

Έτσι, ο WordTagger, δημιουργήθηκε σε περιβάλλον MS Windows XP Pro., με την χρήση του Borland JBuilder 5, που χρησιμοποιεί το Sun JDK 1.3. Είναι όμως δυνατόν να λειτουργήσει κανονικά, χωρίς να πρέπει να γίνει καμιά άλλη επέμβαση στον κώδικα και χωρίς να υπάρχει η ανάγκη αναμεταγλώττισης σε οποιοδήποτε περιβάλλον που διαθέτει εγκατεστημένη την έκδοση 1.3 (ή και νεότερη) του Sun JRE.

4.2.3 Χρήση του WordTagger

Το πρόγραμμα, όπως προαναφέρθηκε, είναι γραμμένο σε γλώσσα Java. Αυτό σημαίνει ότι δεν διατίθεται εκτελέσιμο αρχείο. Επίσης θα πρέπει μέσα στον υπολογιστή να υπάρχει εγκατεστημένο το περιβάλλον εκτέλεσης προγραμμάτων Java (JRE), τουλάχιστον στην έκδοση 1.3.

Η έναρξη του προγράμματος γίνεται με την εκτέλεση της εντολής:
javaw -classpath "classpath" wordtagger.MainApp
αφού πρώτα τροποποιηθεί το classpath της εντολής που περιέχει, ώστε να αναφέρει το full path του φακέλου που περιέχει τον φάκελο “wordtagger” με τα αρχεία “.class” (για το περιβάλλον των MS Windows μπορεί να γίνει εκτέλεση του αρχείου **run.bat** που συνοδεύει την εφαρμογή, εφόσον μέσα σε αυτό γίνουν οι αλλαγές που αναφέρθηκαν παραπάνω).

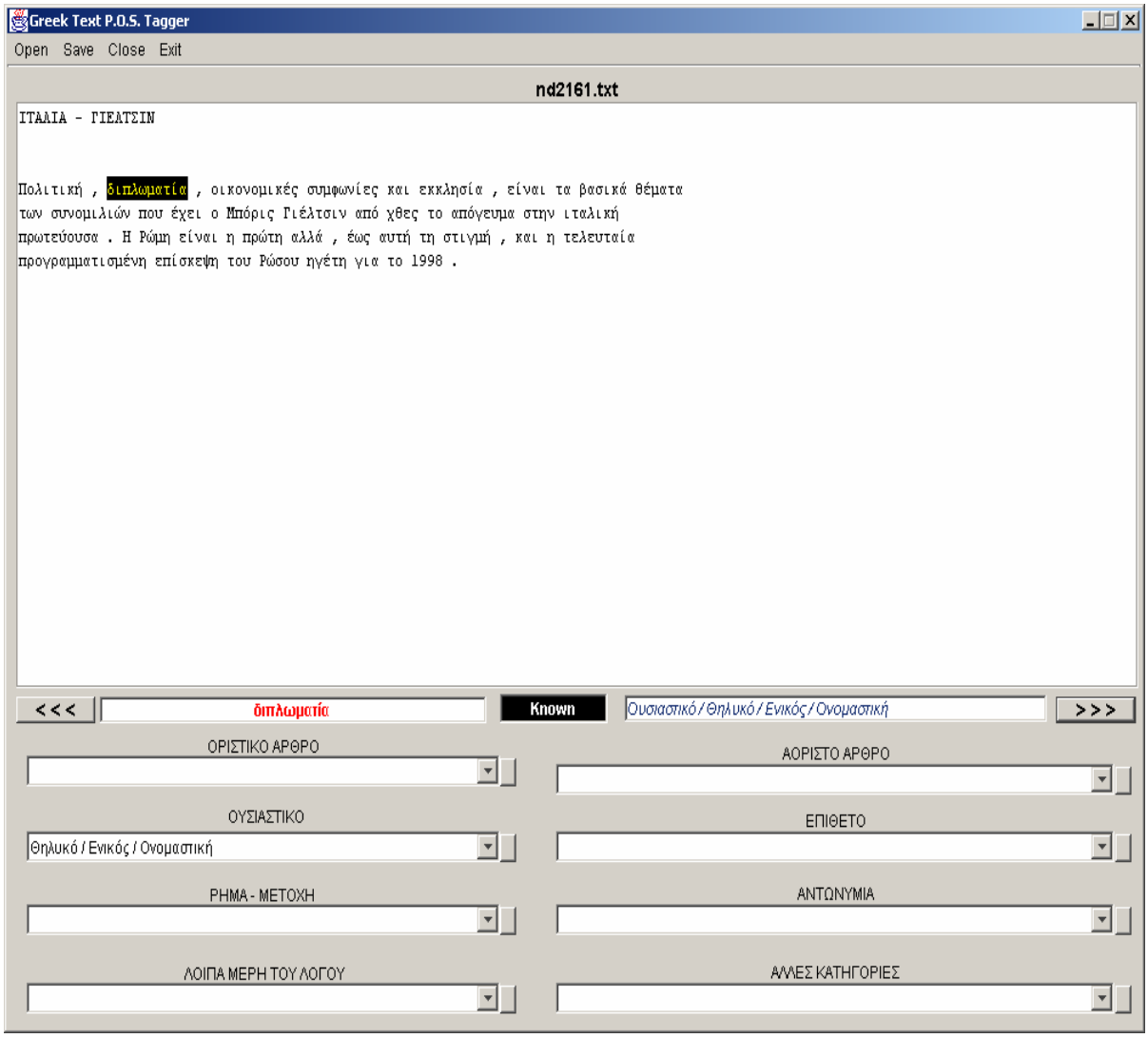
Στο παράθυρο της εφαρμογής που εμφανίζεται επιλέγεται από το menu Open το είδος του αρχείου που θέλουμε να επεξεργαστούμε (“Open Untagged File” για μη επισημειωμένα κείμενα και “Open Tagged File” για επισημειωμένα κείμενα) και μετά επιλέγουμε το όνομα του αρχείου με το κείμενο, το οποίο και θα πρέπει να είναι σε μορφή text-file.

Πατώντας ένα από τα κουμπιά “<<<” ή “>>>” πλοηγούμαστε στο κείμενο προχωρώντας στην επόμενη ή στην προηγούμενη λέξη. Η λέξη την οποία

επεξεργαζόμαστε εμφανίζεται με μαύρο φόντο μέσα στο κείμενο. Επίσης εμφανίζεται και στην θέση 1, ενώ αν έχει ήδη αποδοθεί κάποιο tag, αυτό εμφανίζεται στην θέση 3. Στην 2 εμφανίζεται το αν η λέξη είναι «Γνωστή» ή «Άγνωστη» (το παραπάνω λειτουργεί στην περίπτωση που το κείμενο έχει υποστεί επισημείωση από το MBT/TiMBL). Στην περίπτωση που επιθυμούμε να αποδώσουμε κάποια άλλη ετικέτα στην λέξη, επιλέγουμε αυτή που επιθυμούμε από το αντίστοιχο combo box. Στην περίπτωση, δε, που η ετικέτα έχει ήδη επιλεγεί σε προηγούμενη επιλογή, αρκεί να πιέσουμε το μικρό κουμπί που βρίσκεται δεξιά του combo box (το παραπάνω προστέθηκε με σκοπό την επιτάχυνση της διαδικασίας, αφού η αναζήτηση της ετικέτας μέσα στα στοιχεία της λίστας του combo box αποτελεί σημαντική καθυστέρηση και αυξάνει αρκετά τον χρόνο επισημείωσης).

Τέλος, στην περίπτωση που θέλουμε να αποθηκεύσουμε το κείμενο που επεξεργαστήκαμε, επιλέγουμε μια από τις λειτουργίες του menu “Save”. Αυτές είναι οι: “Save Tagged Files”, “Save Train File” και “Save MBT Tagged File”. Από αυτές, η πρώτη αποθηκεύει το κείμενο στην μορφή στην οποία εξάγει το αποτέλεσμα ο MBT (λέξη/ετικέτα) και η δεύτερη αποθηκεύει το κείμενο στο format εκπαίδευσης του MBT (μια λέξη ανά γραμμή , χωριζόμενη από την ετικέτα της με ένα tab). Τέλος η τρίτη, που λειτουργεί μόνο για την περίπτωση διόρθωσης επισημειωμένου κειμένου, αποθηκεύει το κείμενο παρέχοντας και την πληροφορία που είχε αποδοθεί από τον MBT, βοηθώντας τον χρήστη να κάνει συγκρίσεις και να εξάγει συμπεράσματα.

Η επιλογή Close αδειάζει τον WordTagger από τα στοιχεία του κειμένου. Είναι λειτουργία που θα πρέπει να εφαρμόζεται πάντα πριν από την επεξεργασία ενός αρχείου.



Εικόνα 4.2.1.1

4.3 Το σύστημα ανίχνευσης σφαλμάτων ErrorDetector

Στην αρχή της προηγούμενης παραγράφου τονίστηκε η σημασία ύπαρξης ενός σωστά επισημειωμένου συνόλου κειμένων. Επίσης, δόθηκε ιδιαίτερη έμφαση στην σημασία της ανθρώπινης παρέμβασης κατά την διάρκεια της όλης διαδικασίας. Η παρέμβαση αυτή, όμως, δεν εξασφαλίζει την τελειότητα του αποτελέσματος. Αντίθετα, θα λέγαμε ότι αποτελεί λόγο βεβαιότητας για την ύπαρξη ενός, αν και ελάχιστου, ποσοστού σφαλμάτων. Η εμφάνιση των σφαλμάτων μπορεί να οφείλεται σε κάποια από τις παρακάτω αιτίες:

- ✓ Παράλειψη διόρθωσης υπάρχοντος σφάλματος.
- ✓ Λανθασμένη επιλογή ετικέτας λόγω απροσεξίας.

- ✓ Λανθασμένη επιλογή ετικέτας από εσφαλμένη αναγνώριση της περίπτωσης (αυτό γίνεται όταν η ετικέτα της λέξης εξαρτάται από το είδος και την δομή της πρότασης).
- ✓ Λανθασμένη «διόρθωση» της ετικέτας μιας σωστά επισημειωμένης λέξης.
- ✓ Εσκεμμένη εισαγωγή σφαλμάτων.

Για τον λόγο αυτό Dickinson και Meurers [Dickinson and Meurers 2003] προτείνουν μια μέθοδο για την κατασκευή ενός συστήματος που θα εντοπίζει και θα αναφέρει πιθανές περιπτώσεις λανθασμένης επισημείωσης λέξεων μέσα σε ένα κείμενο. Το υπόλοιπο κομμάτι της παραγράφου 4.3 αναφέρεται σε αυτή την ιδέα. Αρχικά γίνεται αναλυτική περιγραφή του αλγορίθμου που οι δυο ερευνητές προτείνουν ως λύση στο συγκεκριμένο πρόβλημα. Έπειτα γίνεται περιγραφή του συστήματος που κρίθηκε απαραίτητο να υλοποιηθεί στα πλαίσια αυτής της εργασίας και το οποίο βασίστηκε στην ιδέα των Dickinson-Meurers.

4.3.1 Αλγόριθμος

Το υποκείμενο που αναλαμβάνει την διαδικασία της επισημείωσης (υπολογιστικό σύστημα ή ανθρώπινος παράγοντας) καλείται να επιλέξει από ένα σύνολο συμβόλων αυτό που ταιριάζει καλύτερα σε κάθε λέξη του κειμένου και να το αναθέσει σε αυτήν. Στην λήψη της σωστής απόφασης δεν συμβάλουν μόνο τα χαρακτηριστικά της λέξης, αλλά και αυτά των λέξεων που βρίσκονται γύρω από αυτήν, δηλαδή της γειτονιάς της. Αυτό το τελευταίο είναι και η κεντρική ιδέα της πρότασης. Έτσι, είναι αρκετά φυσικό μια λέξη να εμφανίζεται μέσα σε ένα κείμενο περισσότερες από μια φορές. Επίσης, δεν είναι και υποχρεωτικό η λέξη αυτή να έχει πάντοτε το ίδιο tag (π.χ. η λέξη *διατάξεις* μπορεί να είναι Ουσιαστικό-Θηλυκό-Πληθυντικός-Ονομαστική (*οι διατάξεις*), Ουσιαστικό-Θηλυκό-Πληθυντικός-Αιτιατική (*τις* διατάξεις) ή Ρήμα-Μέλλοντας-Υποτακτική (*να διατάξεις*). Αυτό το οποίο είναι όμως περίεργο και σηματοδοτεί, με μεγάλη πιθανότητα, την ύπαρξη λάθους είναι η εμφάνιση της λέξης σε δυο ίδιες ή παρόμοιες γειτονιές, με την ένδειξη όμως διαφορετικού συμβόλου, κάθε φορά. Η "γειτονιά" που αναφέραμε παραπάνω, αποτελεί μια συνεχή «αλυσίδα» λέξεων που περιέχει την προς διερεύνηση λέξη και έχει μήκος n στοιχείων. Η

αλυσίδα αυτή ονομάζεται n-gram και αποτελεί το σημαντικότερο στοιχείο της μεθόδου.

Ο αλγόριθμος που προτείνεται, έχει να κάνει με την δημιουργία και επεξεργασία μεγάλων n-grams. Αρχικά γίνεται εξαγωγή όλων των n-grams του κειμένου χρησιμοποιώντας την παρακάτω διαδικασία:

1. Υπολογίζουμε όλα τα 1-grams που περιέχει το κείμενο και τα αποθηκεύουμε, μαζί με την πληροφορία που αφορά την τοποθεσία τους μέσα στο κείμενο.
2. Ξεκινώντας από τα παραπάνω και βασισμένοι στην πρόταση ότι "ένα n-gram μπορεί να δημιουργηθεί από ένα (n-1)-gram, αν το επεκτείνουμε κατά μια λέξη δεξιά ή αριστερά, αν αυτό είναι δυνατό", δημιουργούμε όλα να 2-grams, τα οποία και αποθηκεύουμε.
3. Συνεχίζουμε αναδρομικά την διαδικασία 2, μέχρι να φτάσουμε σε σημείο που δεν μπορούμε να επεκταθούμε άλλο.

Έπειτα, χρησιμοποιούνται ευριστικές μέθοδοι για την ανακάλυψη των πιθανών σφαλμάτων. Η πρώτη από αυτές ασχολείται με το περιεχόμενο του n-gram. Όπως προείπαμε, το περιεχόμενο του n-gram είναι καθοριστικό για την ετικέτα της λέξης. Συγκρίνοντας τα περιεχόμενα δυο n-grams που περιέχουν την ίδια λέξη με διαφορετικές ετικέτες, χρησιμοποιώντας μια μέθοδο όπως αυτή του αλγορίθμου k-NN, μπορεί να βρεθεί η απόσταση ανάμεσα στα δυο διανύσματα. Όσο, δε, πιο μεγάλο είναι το n και όσο πιο όμοια βρεθούν τα δυο n-grams, τόσο αυξάνεται η πιθανότητα να υπάρχει σφάλμα στην επισημείωση. Εκτός όμως από αυτό, σημασία έχει και η θέση της λέξης στο n-gram. Το γεγονός αυτό είναι που χρησιμοποιεί η δεύτερη μέθοδός. Έτσι, λοιπόν το αν η λέξη βρίσκεται στην αρχή ή στο τέλος του κειμένου ή το αν χρησιμοποιείται μαζί με κάποια συγκεκριμένη λέξη (π.χ. το *φάει*, που είναι ρήμα έχει διαφορετικό χρόνο αν έπεται της λέξης *έχει* (Παρακείμενος)*, θα*(Μέλλον) ή *να*(Ενεστώτας)) σε μεγάλο ποσοστό των σημείων, είναι ένα εξίσου ενδιαφέρον στοιχείο που μπορεί να αποτελέσει ένδειξη για εμφάνιση σφάλματος.

4.3.2 Υλοποίηση του συστήματος

Αρχικά θα πρέπει να αναφέρουμε ότι το σύστημα ErrorDetector αναπτύχθηκε σε περιβάλλον MS Windows με την χρήση της γλώσσας ANSI C++. Στην επιλογή της παραπάνω γλώσσας συνέβαλλαν αρκετοί λόγοι. Σημαντικότερος από όλους ήταν η ταχύτητα που προσφέρει σε σχέση με άλλες γλώσσες, στοιχείο απαραίτητο αν αναλογιστεί κανείς τον τεράστιο όγκο στοιχείων και πράξεων που προκύπτουν στην επεξεργασία των n-grams. Επίσης, η μη απαίτηση ύπαρξης γραφικού περιβάλλοντος και η ανάγκη, όπως και ο WordTagger, το σύστημα να μπορεί να μεταφερθεί σε άλλη πλατφόρμα χωρίς πολλές αλλαγές συνέβαλαν σημαντικά στην τελική επιλογή.

Επίσης θα πρέπει να τονίσουμε ότι με σκοπό την βελτίωση της ταχύτητας του συστήματος, αλλά και την επίτευξη καλύτερου τελικού αποτελέσματος, πραγματοποιήθηκαν κάποιες, μικρής έκτασης, αλλαγές σε ορισμένα σημεία των αλγορίθμων που προτάθηκαν παραπάνω, οι οποίες όμως σε καμία περίπτωση δεν αλλοιώνουν την ουσία τους. Η πρώτη από αυτές αφορά τα n-grams. Ο προτεινόμενος αλγόριθμος συνιστά την εύρεση όλων των n-grams του κειμένου, στοιχείο που προσθέτει μεγάλο φόρτο στο σύστημα. Αντίθετα, εδώ επιλέγουμε να δημιουργήσουμε μόνο τα n-grams που τελικά χρειάζονται (και τα οποία είναι αυτά που περιέχουν λέξεις για τις οποίες υπάρχει υποψία εμφάνισης σφάλματος). Επίσης στο σημείο της αξιολόγησης, επιλέγεται η σύγκριση των λέξεων του κάθε n-gram. Με σκοπό την εύρεση μεγαλύτερου πλήθους σφαλμάτων, διαφοροποιήσαμε τον αλγόριθμο και σε αυτό το τμήμα. Έτσι, εδώ, ελέγχουμε την γειτονιά της λέξης όχι μόνο ως προς την εμφάνιση (ως προς αυτές καθ' αυτές τις λέξεις) αλλά και ως προς την φύση της (ως προς τις ετικέτες των λέξεων). Κι αυτό, γιατί, ενώ είναι σχετικά δύσκολο να βρεθεί μια λέξη μέσα σε ακριβώς ίδιο n-gram 2 ή περισσότερες φορές, είναι πολύ πιο πιθανό να βρεθεί μια λέξη σε περισσότερες από 2 γειτονιές ίδιας φύσεως (π.χ. αν μελετάμε την λέξη ***θα*** που έχει διαφορετικά tags μέσα στο ίδιο κείμενο, δύσκολα θα δούμε να επαναλαμβάνεται το «... ***θα*** *φύγω αύριο* ...», ενώ είναι πολύ πιο πιθανό να υπάρχουν οι φράσεις «... ***θα*** *φύγω αύριο* ...» και «... ***θα*** *πάω μετά* ...», που είναι ίδιας φύσεως και μπορούν να προδώσουν την ύπαρξη λάθους). Φυσικά, η δεύτερη περίπτωση έχει και αυτή τους κινδύνους της, αφού πιθανή ύπαρξη λανθασμένης ετικέτας σε μια λέξη της γειτονιάς μπορεί να αποκρύψει κάποιο σφάλμα (Error1) ή να εμφανίσει κάποια σωστή επισημείωση για λανθασμένη (Error2). Επειδή όμως ο ErrorDetector είναι

σύστημα που παράγει αναφορές (και δεν επεμβαίνει στο κείμενο το ίδιο), και επειδή με την τεχνική αυτή έχουμε μεγαλύτερη αποκάλυψη λαθών, η εμφάνιση των Error1 και Error2 είναι αποδεκτή.

Παρακάτω παρατίθεται ο τρόπος λειτουργίας του προγράμματος, με την μορφή ενός γενικού παραδείγματος, για την ευκολότερη κατανόησή του (στο διάγραμμα 4.3.1 απεικονίζεται διαγραμματικά η λειτουργία του προγράμματος). Αρχικά, εκτελείται το πρόγραμμα με την χρήση της ακόλουθης εντολής:

ErrorDetector <filename> [n]

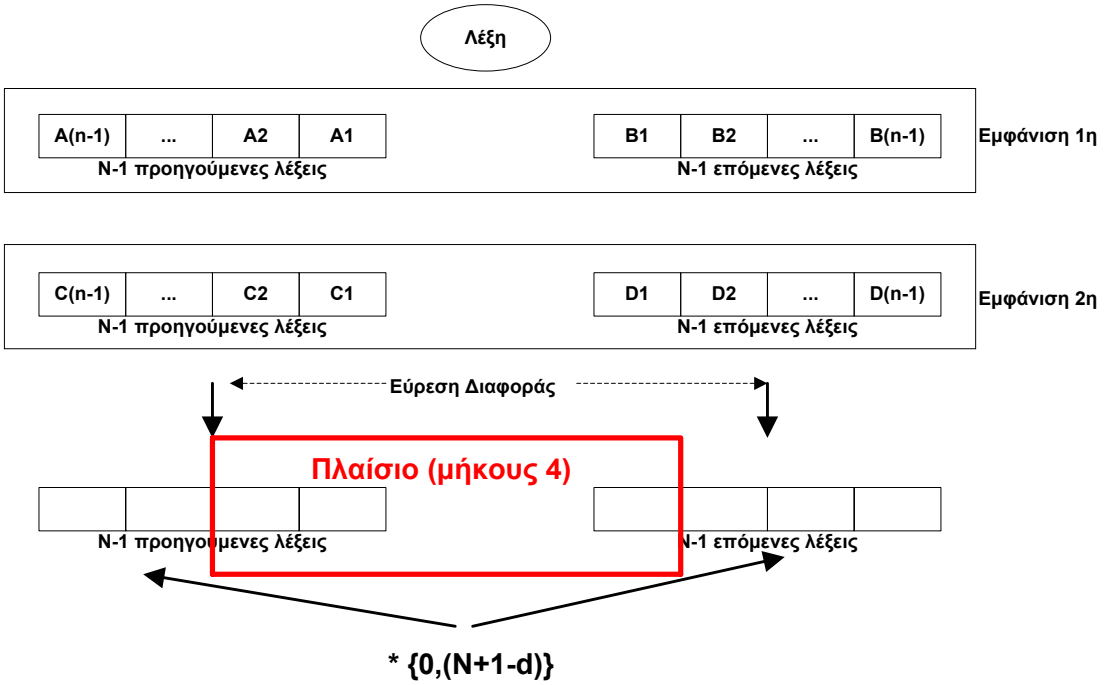
Στην θέση του *filename* τοποθετούμε το όνομα του αρχείου που θέλουμε να ελεγχθεί. Το αρχείο πρέπει να είναι επισημειωμένο και να βρίσκεται στο «κλασσικό» μορφότυπο επισημειωμένου αρχείου (δηλαδή, <λέξη>/<ετικέτα>). Το *n* αντιπροσωπεύει το μέγιστο μήκος που επιθυμούμε να έχουν τα n-grams που θα συγκριθούν και σε περίπτωση που δεν δοθεί χρησιμοποιείται η προεπιλεγμένη τιμή που είναι το 10 (η τιμή αυτή προκύπτει από το γεγονός ότι λέξεις που βρίσκονται σε απόσταση μεγαλύτερη του 5 από την διερευνούμενη λέξη δεν είναι σε θέση να επηρεάζουν σημαντικά την ετικέτα που πρέπει να της αποδοθεί).

Γίνεται ανάγνωση των λέξεων και των ετικετών τους. Από αυτές διατηρούνται μόνο οι λέξεις που εμφανίζονται τουλάχιστον δυο φορές. Έπειτα απορρίπτονται αυτές για τις οποίες δεν έχουμε εμφάνιση διαφορετικής ετικέτας. Για καθεμιά από τις υπόλοιπες διατηρούμε (σε μορφή διανύσματος) τις n-1 προηγούμενες και n-1 επόμενες λέξεις αλλά και ετικέτες. Για κάθε δυο εμφανίσεις της λέξης που υπάρχει διαφορετική ετικέτα, γίνεται σύγκριση ανάμεσα στα διανύσματά τους. Η σύγκριση αυτή γίνεται με την χρήση του κανόνα:

$$Difference = \begin{cases} 0\ ,\ \lambda\acute{\epsilon}\xi\eta\ \ 1\ \neq\ \ \lambda\acute{\epsilon}\xi\eta\ \ 2 \\ n\ +\ 1\ -\ dist\ \ ,\ \lambda\acute{\epsilon}\xi\eta\ \ 1\ =\ \ \lambda\acute{\epsilon}\xi\eta\ \ 2 \end{cases}$$

Έτσι, αποδίδεται ταυτόχρονα και ένα βάρος στην λέξη. Αυτό είναι αρκετά μεροληπτικό ως προς τις λέξεις που βρίσκονται πλησίον της διερευνούμενης, πράγμα αρκετά λογικό αν σκεφτεί κανείς ότι μόνον αυτές είναι σε θέση να επηρεάσουν την ετικέτα. Ακολούθως, θα πρέπει να ελέγξουμε όλα τα n-grams που αφορούν την λέξη και που έχουν μήκος από 2 έως n. Έτσι ξεκινώντας με m=2, δημιουργούμε ένα «πλαίσιο» μήκους m το οποίο κινείται πάνω στα δυο διανύσματα διαφοράς και ξεκινάει πάντα περιλαμβάνοντας τα m-1 δεξιότερα στοιχεία του διανύσματος προς

αριστερά, και μετακινείται κάθε φορά 1 θέση προς τα δεξιά μέχρι να φτάσει να περιλαμβάνει τα m-1 αριστερότερα στοιχεία του διανύσματος προς τα δεξιά. Έτσι σε κάθε φάση, μέσα στο πλαίσιο υπάρχει και από ένα διαφορετικό m-gram του κειμένου που περιέχει την λέξη. Προσθέτοντας την αξία των σημείων που περιλαμβάνει το πλαίσιο, την συγκρίνουμε με το μισό αυτής που θα είχε αν υπήρχε πλήρης συμφωνία ανάμεσα στα δυο διανύσματα που συγκρίνουμε. Αν την ξεπερνάει, σημειώνουμε το γεγονός. Αυτό γίνεται για όλα τα m-grams της λέξης έως ότου m=n. Τέλος απαριθμούμε τα αποδεκτά διανύσματα και τα συγκρίνουμε με το συνολικό τους πλήθος. Αν ξεπερνάνε το 1/3 από αυτά, σημειώνουμε τις δυο λέξεις ως υποψήφιες για διαφωνία. Η διαδικασία αυτή εκτελείται τόσο με τις λέξεις αυτές καθ' αυτές, όσο και με τις ετικέτες τους.



Διάγραμμα 4.3.1

Το αποτέλεσμα της όλης διαδικασίας είναι ένα αρχείο με την ονομασία *<filename>.ErrorReport.txt*, το οποίο εμφανίζει μια αναφορά με τις λέξεις που πιθανόν να έχουν επισημειωθεί λάθος και τα σημεία που αυτές εμφανίζονται μέσα στο κείμενο. Το μορφότυπο του αρχείου εμφανίζεται παρακάτω:

```
***** Possible-Error Report *****

Word: <word 1>
Lines: <line 1>, <line 2>, …, <line N>

Word: < word 2>
Lines: <line 1>, <line 2>, …, <line N>

…

**************************
```

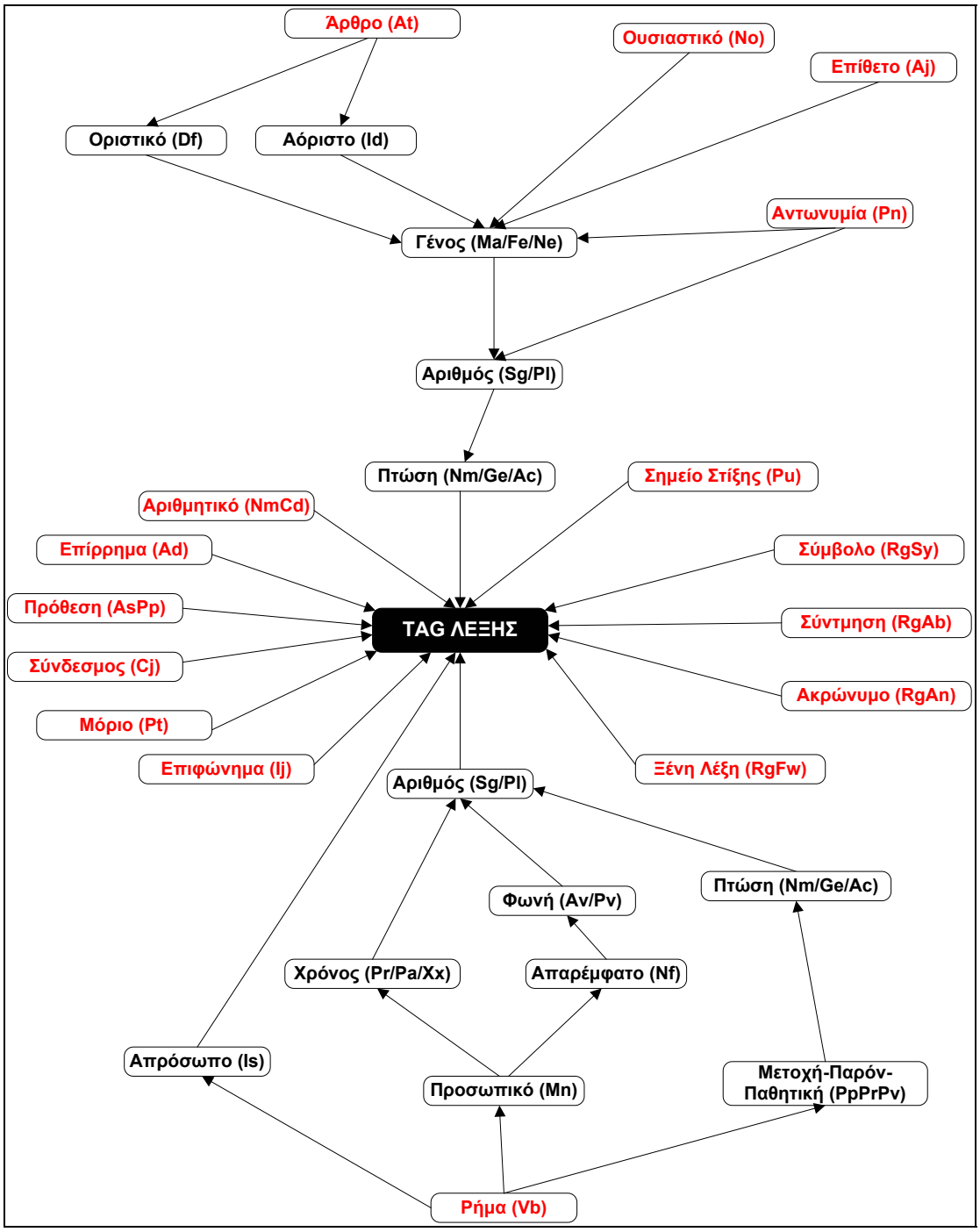
4.4 Το σύνολο ετικετών

Στο τμήμα αυτό θα παρουσιαστεί το σύνολο των ετικετών (tag set) που χρησιμοποιήθηκε για την επισημείωση των λέξεων των κειμένων με τα οποία θα εργαστούμε.

Αναλυτικά, το χρησιμοποιούμενο σύνολο, μαζί με την επεξήγηση του κάθε συμβόλου παρατίθεται στους πίνακες του παραρτήματος Β. Το συγκεκριμένο σύνολο ετικετών βασίζεται στο πρότυπο που προτείνεται από το ΙΕΛ ως αντιστοίχηση του διεθνούς προτύπου PAROLE για την ελληνική γλώσσα. Από αυτό (αποτελείται από συνολικά 584 σύμβολα), επιλέξαμε να χρησιμοποιήσουμε ένα αρκετά εκτεταμένο υποσύνολο αποτελούμενο από 120 ετικέτες, έτσι ώστε να καλύπτονται όλες οι κατηγορίες λέξεων, αλλά και να αποδίδουμε ένα αρκετά μεγάλο ποσό πληροφορίας στην κάθε λέξη που επισημειώνουμε. Έτσι, το κάθε tag μπορεί να αποτελείται από 1 (για τα σημεία στίξης) έως και 5 (για τα άρθρα) τμήματα. Στο Διάγραμμα 4.4.1 που ακολουθεί φαίνεται η διαδικασία με την οποία συντίθεται η κάθε ετικέτα από τα επιμέρους χαρακτηριστικά που θέλουμε να προσδώσουμε στην λέξη.

Όπως είπαμε, το μεγάλου εύρους χρησιμοποιούμενο σύνολο ετικετών μας βοηθάει να καλύψουμε όλες τις κατηγορίες της ελληνικής γραμματικής. Παρόλ' αυτά, όμως, η σύνθεση λέξεων, φαινόμενο πολύ συνηθισμένο στην ελληνική γλώσσα, μπορεί να δημιουργήσει λέξεις που να ανήκουν σε κάποια ιδιάζουσα κατηγορία, η οποία να μην προβλέπεται από το επιλεγμένο σύνολο ετικετών. Σε τέτοιες περιπτώσεις θα πρέπει

να γίνουν ορισμένες παραδοχές. Η μοναδική τέτοια περίπτωση που παρατηρήθηκε κατά την διεξαγωγή των πειραμάτων : η σύνθεση της πρόθεσης «*σε*» με ένα οριστικό άρθρο, π.χ. «*το*» δημιουργεί την λέξη «*στο*» το οποίο ανήκει στην υποκατηγορία των σύνθετων άρθρων, που υπάγονται στην κατηγορία των προθέσεων. Για τους σκοπούς της συγκεκριμένης εργασίας, όμως, θα το κατατάξουμε σε αυτήν των άρθρων.



Διάγραμμα 4.4.1

4.5 Τα Σώματα κειμένων

Στα πειράματα που διεξήχθησαν, κατά την διάρκεια της εργασίας, για την εξαγωγή των μετρήσεων χρησιμοποιήθηκαν δυο σύνολα από κείμενα. Περισσότερες λεπτομέρειες για αυτά υπάρχουν στις παραγράφους που ακολουθούν. Αυτό που θα πρέπει να αναφερθεί εδώ είναι ότι και στα δυο εφαρμόστηκε επισημείωση βάσει του συνόλου ετικετών που περιγράφηκε στο τμήμα 4.4. Επίσης, για την συγκεκριμένη διαδικασία χρησιμοποιήθηκαν ως βοήθεια πληροφορίες από το βιβλίο της «Νεοελληνικής Γραμματικής» του Χ.Τσολάκη (σχολικό βιβλίο Γυμνασίου) και το «Λεξικό της Νέας Ελληνικής Γλώσσας» του Γ.Μπαμπινιώτη. Τα χαρακτηριστικά των δυο σωμάτων κειμένων, καθώς και η επεξεργασία την οποία τύχανε ώστε να καταστούν έτοιμα για χρήση, περιγράφονται παρακάτω.

4.5.1 Αρχικό σώμα κειμένων

Το πρώτο σώμα κειμένων, που χρησιμοποιήθηκε για τα πειράματα, μας παρασχέθηκε από το ΕΚΕΦΕ «Δημόκριτος». Τα κείμενα ήταν ήδη επισημειωμένα και μάλιστα βάσει του συνόλου ετικετών που χρησιμοποιήθηκε και σε αυτήν την εργασία. Περισσότερες λεπτομέρειες παρατίθενται στις ενότητες που ακολουθούν.

4.5.1.1 Χαρακτηριστικά σώματος

Το σώμα αυτό αποτελεί μια συλλογή από 86 κείμενα, συνολικού μεγέθους 31943 λέξεων και αποτελείται στο σύνολό του από αγγελίες που αφορούν την αναζήτηση ατόμων για την κάλυψη θέσεων εργασίας. Το τελευταίο, αποτελεί σημαντικό γεγονός, αν σκεφτεί κανείς την δομή την οποία έχει μια αγγελία. Οι αγγελίες είναι, συνήθως, κείμενα μικρού μεγέθους. Αποτελούνται, κατά κύριο λόγο, από ουσιαστικά, επίθετα και μετοχές, διαχωριζόμενα από πληθώρα σημείων στίξεως, ενώ διαθέτουν ελάχιστο αριθμό ρημάτων. Επίσης, παρόλο που τηρούν τους κανόνες του συντακτικού, η σύνταξη που διαθέτουν είναι στοιχειώδης. Οι προτάσεις που τις απαρτίζουν είναι μικρού μήκους και κοφτές, ενώ πολύ σπάνια παρατηρείται σε κάποια από αυτές κάποια πιο «επιτηδευμένη» μορφή σύνταξης. Γενικά, θα λέγαμε

ότι, συντακτικά τουλάχιστον, μια αγγελία δεν μπορεί σε καμία περίπτωση να μοιάζει με ένα «μέσο κείμενο». Ακολούθως παρατίθεται, ως υπόδειγμα, ένα από τα κείμενα της συλλογής:

Byte
Σύμβουλος Πωλήσεων Απαραίτητα προσόντα :
Απόφοιτοι ΑΕΙ - ΤΕΙ Πληροφορικής
Ικανότητα συνεργασίας και επικοινωνίας
Εκπληρωμένες στρατιωτικές υποχρεώσεις
Ηλικία μέχρι 35 ετών
Εμπειρία τουλάχιστον 3 ετών σε αντίστοιχη θέση
Καλή γνώση Αγγλικής
Προσφέρονται :
Εξέλιξη σε δυναμικό περιβάλλον
Πρόσθετη ασφάλιση
Πρόσθετη ιατροφαρμακευτική περίθαλψη

4.5.1.2 Προεπεξεργασία

Παρόλο που, όπως προαναφέραμε, τα κείμενα ήταν ήδη επισημειωμένα όταν μας παραδόθηκαν, η μορφή στην οποία βρίσκονταν δεν ήταν τέτοια που να επιτρέπει την χρήση τους άμεσα. Έτσι πριν από την έναρξη της πειραματικής διαδικασίας, προηγήθηκε η φάση της προεπεξεργασίας. Αυτή περιελάμβανε τα εξής στάδια:

1) *Διαχωρισμός του κειμένου στα επιμέρους κείμενα που το αποτελεί.*

   Το σώμα κειμένων μας παραδόθηκε ως ένα αρχείο, το οποίο περιελάμβανε όλα τα κείμενα του σώματος, με το όνομα του καθενός να προηγείται του περιεχομένου του. Με την βοήθεια ενός μικρού προγράμματος, που δημιουργήθηκε γι' αυτό το σκοπό, έγινε ανάγνωση του κειμένου, αναγνωρίστηκε το όνομα του κάθε αρχείου και το περιεχόμενό του. Πραγματοποιήθηκε διάσπαση και τελικά προέκυψαν τα 89 αρχεία.

2) *Έλεγχος συνέπειας διαχωρισμού των λέξεων*

   Σε αρκετά σημεία του κειμένου παρατηρήθηκε το φαινόμενο του μη συνεπούς διαχωρισμού των λέξεων. Αυτό παρατηρήθηκε κυρίως στα σημεία που υπήρχαν λέξεις ενωμένες με αριθμούς ή λέξεις που περιείχαν απόστροφο. Έτσι βρέθηκαν σημεία όπου γράμματα και αριθμοί (ή η απόστροφος):

   - βρίσκονταν στην μορφή της μιας λέξης με μια ετικέτα

- ήταν διασπασμένα με το καθένα να διαθέτει τη δική του ετικέτα
- ή ήταν διασπασμένα ώστε το καθένα να είναι επισημειωμένο με τη σωστή ετικέτα και η συνολική λέξη επισημειωμένη ως άγνωστη.

Η παραπάνω ασυμφωνία παρατηρήθηκε ακόμα και σε πολλαπλές εμφανίσεις της ίδιας λέξης, με την κάθε εμφάνιση να είναι επισημειωμένη και με διαφορετικό από τους τρεις προαναφερθέντες τρόπους. Ύστερα από έλεγχο όλου του κειμένου εντοπίστηκαν και διορθώθηκαν όλες οι παραπάνω προβληματικές περιπτώσεις, με σωστή διάσπαση των λέξεων. Επίσης διασπάστηκαν και όλες οι λέξεις που περιείχαν στοιχεία από περισσότερες από μια από τις παρακάτω κατηγορίες: γράμματα, αριθμούς, σημεία στίξης, σύμβολα. Τέλος, διασπάστηκαν και εμφανίσεις συνεχόμενων συμβόλων ή σημείων στίξης.

3) *Έλεγχος συνέπειας της επισημείωσης.*

Ορισμένες από τις λέξεις του κειμένου, όπως οι ξένες λέξεις και ορισμένα σύμβολα, μπορούν να επισημειωθούν με μια και μόνο ετικέτα. Παραβίαση του παραπάνω κανόνα, σημαίνει την ύπαρξη σφάλματος, το οποίο όμως, λόγω της φύσης των ιδίων των λέξεων, μεταφράζεται και σε έλλειψη συνέπειας. Με έλεγχο του κειμένου εντοπίστηκαν τέτοιες περιπτώσεις και διορθώθηκαν.

4) *Έλεγχος ύπαρξης σφαλμάτων.*

Μια επέκταση του παραπάνω, αλλά για το σύνολο των λέξεων, αποτελεί ο έλεγχος για την ύπαρξη σφαλμάτων. Ο παραπάνω έλεγχος, έγινε με σκοπό την επιβεβαίωση ότι δεν υπάρχουν σφάλματα επισημείωσης στις υπόλοιπες λέξεις του κειμένου. Η διαδικασία αυτή πραγματοποιήθηκε σε δυο στάδια. Αρχικά έγινε με την παρατήρηση του κειμένου. Έπειτα ακολούθησε έλεγχος με την χρήση του ErrorDetector, ο οποίος και ανακάλυψε ένα ποσοστό σφαλμάτων της τάξης του 1.1%, που εντοπίζεται κατά κύριο λόγο σε λάθος επισημείωση των σημείων στίξεως.

5) *Μετατροπή του μορφότυπου.*

Όπως προαναφέρθηκε, το σύστημα MBT αναγνωρίζει ένα συγκεκριμένο μορφότυπο επισημειωμένου κειμένου για εκπαίδευση αλλά και για έλεγχο. Έτσι, δημιουργήθηκε και χρησιμοποιήθηκε και εδώ ένα μικρό πρόγραμμα, το οποίο βοήθησε στο να μετατραπούν τα κείμενα στην σωστή μορφή μέσα σε μικρό χρονικό διάστημα.

4.5.2 Νέο σώμα κειμένων

4.5.2.1 Χαρακτηριστικά σώματος κειμένων

Στηριζόμενοι στο γεγονός ότι, για τα ελληνικά, δεν είχαμε στη διάθεσή μας ένα σώμα κειμένων, σωστά επισημειωμένο, το οποίο να έχει ικανοποιητικό μέγεθος, κρίθηκε αναγκαίο να δημιουργηθεί ένα. Η νέα συλλογή αποτελείται από 640 κείμενα και έχει συνολικό μέγεθος 130500 λέξεις. Τα κείμενα που την απαρτίζουν αποτελούνται στο σύνολό τους από άρθρα που αφορούν την καθημερινή ειδησιογραφία, σε επίπεδο κοινωνικό, πολιτικό και οικονομικό. Η συγκεκριμένη συλλογή μπορεί να θεωρηθεί πιο «αντιπροσωπευτική» από την προηγούμενη, αφού πλέον τα κείμενα διαθέτουν πλησιάζουν πιο πολύ στο «μέσο κείμενο». Τα άρθρα διαθέτουν λέξεις από όλο το σύνολο της γραμματικής, ενώ παράλληλα η σύνταξή τους είναι τέτοια ώστε να παρατηρούνται περιπτώσεις από το μεγαλύτερο εύρος του συντακτικού. Ο λόγος τους είναι πιο «στρωτός» και πλησιάζει πολύ περισσότερο αυτόν με τον οποίο ο καθημερινός άνθρωπος συντάσσει ένα κείμενο. Στο πλαίσιο που ακολουθεί παρατίθεται, ως παράδειγμα, ένα από τα κείμενα του σώματος:

Εμμένει στην πολιτική των προκλήσεων η Άγκυρα

ΟΣΟ η κυβέρνηση εμμένει στην πρόταση για παραπομπή των όποιων ελληνοτουρκικών διαφορών στο Διεθνές Δικαστήριο της Χάγης, προκαλώντας τις έντονες επικρίσεις της αξιωματικής αντιπολίτευσης, τόσο η Άγκυρα φροντίζει να απαντά με την πάγια θέση της για απευθείας διάλογο μεταξύ των δύο χωρών, ενώ παράλληλα συνεχίζει τις επικίνδυνες και προκλητικές της ενέργειες στο Αιγαίο. Προχθές, δύο τουρκικές τορπιλοπυραυλάκατοι παραβίασαν τους κανόνες αβλαβούς διέλευσης, πράξη την οποία χθες το τουρκικό υπουργείο Εξωτερικών διέψευσε την ώρα που ο Έλληνας πρεσβευτής στην Άγκυρα διαβίβαζε διάβημα διαμαρτυρίας της ελληνικής κυβέρνησης.

4.5.2.2 Προεπεξεργασία

Όπως και με το αρχικό σώμα κειμένων, έτσι και εδώ το κάθε κείμενο της συλλογής θα πρέπει να υποστεί κάποιας μορφής προεπεξεργασία πριν γίνει έτοιμο για χρήση. Η διαδικασία εδώ περιλαμβάνει τα εξής στάδια:

1) *Ορθογραφικός έλεγχος.*

   Αρχικά, γίνεται ένας έλεγχος των κειμένων, προκειμένου να βρεθούν τυχόν ορθογραφικά λάθη που υπάρχουν σε αυτό. Επίσης γίνεται έλεγχος και διόρθωση των χαρακτήρων που περιέχει η κάθε λέξη. Κι αυτό γιατί υπάρχουν χαρακτήρες των ελληνικών που είναι ίδιοι με τους ελληνικούς. Η παρεμβολή ενός τέτοιου μέσα σε μια λέξη μπορεί να δημιουργήσει πρόβλημα, τόσο στην διάσπαση της λέξης, όσο και στην προσπάθεια αναγνώρισής της.

2) *Έλεγχος διαχωρισμού λέξεων.*

   Σε αυτή την φάση ελέγχονται οι λέξεις για το πόσο σωστά είναι διαχωρισμένες, Διαχωρίζονται οι αριθμητικοί χαρακτήρες, τα σύμβολα και τα σημεία στίξης από τις λέξεις που περιέχουν αλφαβητικούς χαρακτήρες, ελληνικούς ή λατινικούς, αλλά και μεταξύ τους. Έτσι στο τελικό κείμενο δεν θα πρέπει να υπάρχει αριθμός ενωμένος με χαρακτήρα, σημείο στίξης ή σύμβολο. Επίσης δεν θα πρέπει να παρατηρείται εμφάνιση συνεχόμενων συμβόλων ή σημείων στίξης μέσα στην ίδια λέξη.

3) *Επισημείωση κειμένου.*

   Η φάση αυτή είναι η σημαντικότερη, αφού όπως αναφέραμε στην αρχή το κείμενο αυτό ήταν σε καθαρή μορφή. Προκειμένου να εξοικονομηθεί χρόνος, η διαδικασία αυτή πραγματοποιήθηκε χρησιμοποιώντας και τις δυο μεθόδους επισημείωσης (αυτόματη και χειροκίνητη). Προηγήθηκε επισημείωση του σώματος με την χρήση του MBT, με σώμα εκπαίδευσης το αρχικό σώμα κειμένων, και ακολούθησε διόρθωση των σφαλμάτων με την χρήση του WordTagger.

4) *Έλεγχος συνέπειας επισημείωσης και εμφάνισης σφαλμάτων.*

   Όπως και με το αρχικό, έτσι και με αυτό το σύνολο κειμένων, έγινε έλεγχος για την συνέπεια του τρόπου επισημείωσης. Επίσης χρησιμοποιήθηκε αποκλειστικά το σύστημα ErrorDetector για την ανακάλυψη τυχόν σφαλμάτων. Το γεγονός ότι το σύνολο της διαδικασίας επισημείωσης πραγματοποιήθηκε από το ίδιο πρόσωπο, συνέβαλε στην ελαχιστοποίηση του

ποσοστού εμφάνισης σφαλμάτων, με αποτέλεσμα αυτό να βρίσκεται κοντά στο μηδέν.

5. Πειραματική Διαδικασία

Στις παραγράφους που ακολουθούν περιγράφονται αναλυτικά τα βήματα που γίνανε στην διάρκεια της διεξαγωγής των πειραμάτων. Επίσης γίνεται παράθεση των αποτελεσμάτων και γίνεται προσπάθεια ποιοτικής ερμηνείας τους, ώστε να μπορέσουμε να εξάγουμε τα επόμενα βήματα έρευνας που θα πρέπει να ακολουθήσουμε.

Για την διεξαγωγή των πειραμάτων χρησιμοποιούμε το σύστημα MBT/TiMBL (στο κεφάλαιο 4.1 έγινε αναλυτική περιγραφή του). Αυτή η επιλογή έγινε γιατί ένα τέτοιο σύστημα μπορεί να δώσει αρκετά καλά αποτελέσματα. Η παραπάνω παραδοχή, όμως, στέκεται αυθαίρετη, αφού πουθενά δεν εμφανίζεται ένα μέτρο σύγκρισης, ένα κατώτερο όριο απόδοσης (baseline), πάνω από το οποίο μπορούμε να πούμε ότι τα αποτελέσματα είναι αποδεκτά. Έτσι πριν ξεκινήσουμε θα πρέπει να ορίσουμε αυτό το όριο.

Όπως, συχνά, γίνεται σε αυτές τις περιπτώσεις, το κατώτερο όριο ορίζεται από τα αποτελέσματα που εμφανίζει μια στοιχειώδης διεργασία που πραγματοποιεί την συγκεκριμένη λειτουργία χωρίς να «σκέφτεται», παίρνοντας σε κάθε σημείο που απαιτείται μια αυθαίρετη, τυχαία απόφαση. Έτσι, και εδώ θα ορίσουμε τον ελάχιστο επίπεδο αποδεκτής απόδοσης ως την απόδοση που παρέχει ένας «χαζός» επισημειωτής κειμένου. Ένας τέτοιος επισημειωτής είναι ένα σύστημα το οποίο διαβάζει τις λέξεις του κειμένου εκπαίδευσης, ανακαλύπτει την πιο συχνά χρησιμοποιούμενη ετικέτα και την αποδίδει σε όλες τις λέξεις του κειμένου που καλείται να επισημειώσει. Από την εφαρμογή ενός τέτοιους συστήματος σε καθένα από τα σύνολα κειμένων προέκυψε ότι το ποσοστό επιτυχίας στο πρώτο είναι κοντά στο 7% ενώ στο δεύτερο πλησιάζει 4%. Ευνόητο και αναμενόμενο είναι λοιπόν πως όποια και να είναι τα εξαγόμενα του MBT, θα είναι σαφώς καλύτερα από τα προηγούμενα.

5.1 Πειράματα με το αρχικό σώμα κειμένων (σώμα κειμένων "Δημόκριτου")

Όπως αναφέρθηκε και στην ενότητα 4, το αρχικό σώμα κειμένων αποτελείται από ένα σύνολο 68 κειμένων, συνολικού μεγέθους περίπου 32000 λέξεων (στο κείμενο εμπεριέχονται 15000 διαφορετικές λέξεις).

Η πρώτη φάση των πειραμάτων περιλαμβάνει δεκαπλή διασταυρωμένη επικύρωση. Σκοπός της είναι η εύρεση των ρυθμίσεων εκείνων του MBT, οι οποίες θα δώσουν τα καλύτερα δυνατά αποτελέσματα στην διαδικασία της επισημείωσης. Το παραπάνω σύνολο κειμένων χωρίστηκε σε 10 τμήματα. Σε κάθε φάση της διαδικασίας επιλέγουμε και συγχωνεύουμε τα 9 τμήματα ώστε να τα χρησιμοποιήσουμε ως σώμα εκπαίδευσης, ενώ το εναπομείναν τμήμα χρησιμοποιείται ως σώμα ελέγχου.

Ξεκινώντας τις δοκιμές, ασχολούμαστε μόνο με το configuration του MBT, γεγονός που σημαίνει ότι το σύστημα χρησιμοποιεί τις προεπιλεγμένες ρυθμίσεις του TiMBL, οι οποίες είναι:
i) Για τις γνωστές λέξεις ο αλγόριθμος IGTREE.
ii) Για τις άγνωστες λέξεις ο αλγόριθμος k-NN όπου k=1 και την χρήση Gain Ratio για την απόδοση «βάρους» στα χαρακτηριστικά.

Επιλέγοντας τις παραμέτρους έναρξης, λαμβάνουμε υπόψη μας μόνο την επίδραση του περιβάλλοντος της λέξης (γειτονιά λέξεων) στην επιλογή της ετικέτας της. Έτσι, δίνουμε στον tagger τις ελάχιστες δυνατές ρυθμίσεις, που είναι το να λαμβάνει υπόψη του μια λέξη πριν ή μετά από την λέξη που πρέπει να σημειωθεί (wf/wF, fw/Fw, wfw/wFw). Τα αποτελέσματα δεν είναι και τόσο ικανοποιητικά, αφού ενώ η επίλυση των ασαφειών στις γνωστές λέξεις κυμαίνεται, κατά μέσο όρο στο 92%, η δυνατότητά του να μαντεύει άγνωστες λέξεις βρίσκεται στο απογοητευτικό 36%. Από τις παραπάνω δυνατότητες, προτιμάται η τρίτη (wfw/wFw), αφού είναι αυτή που έδωσε τα καλύτερα ποσοστά, και μάλιστα με τις μικρότερες διακυμάνσεις. Παρατηρώντας τα σφάλματα επισημείωσης, σημειώνουμε ότι το σύνολο των λέξεων του κειμένου που αποτελούνται από αριθμητικούς χαρακτήρες έχει επισημειωθεί με λάθος τρόπο (το ποσοστό αυτό φτάνει το 97-99%). Βάσει αυτού του γεγονότος,

εμπλουτίζουμε τις ρυθμίσεις του tagger για τις άγνωστες λέξεις με την δυνατότητα να αναγνωρίζει αριθμητικά σύμβολα (wfw/wnFw). Τα πειράματα που ακολουθούν επαληθεύουν θετικά την αλλαγή αυτή, και η επιτυχία στις άγνωστες λέξεις ανέρχεται σε ποσοστό άνω του 40% (περίπου στο 40,7%).

Εν συνεχεία, ρυθμίζουμε τον tagger ώστε να χρησιμοποιεί περισσότερες λέξεις στα δεξιά ή αριστερά. Η ελάχιστη θετική μεταβολή που εμφανίζεται στα ποσοστά κάποιων κειμένων (3-4/10), δεν είναι σε θέση να αντισταθμιστεί από την σημαντικότατη πτώση σε αυτά των υπολοίπων, και γι' αυτό η παραπάνω αλλαγή απορρίπτεται.

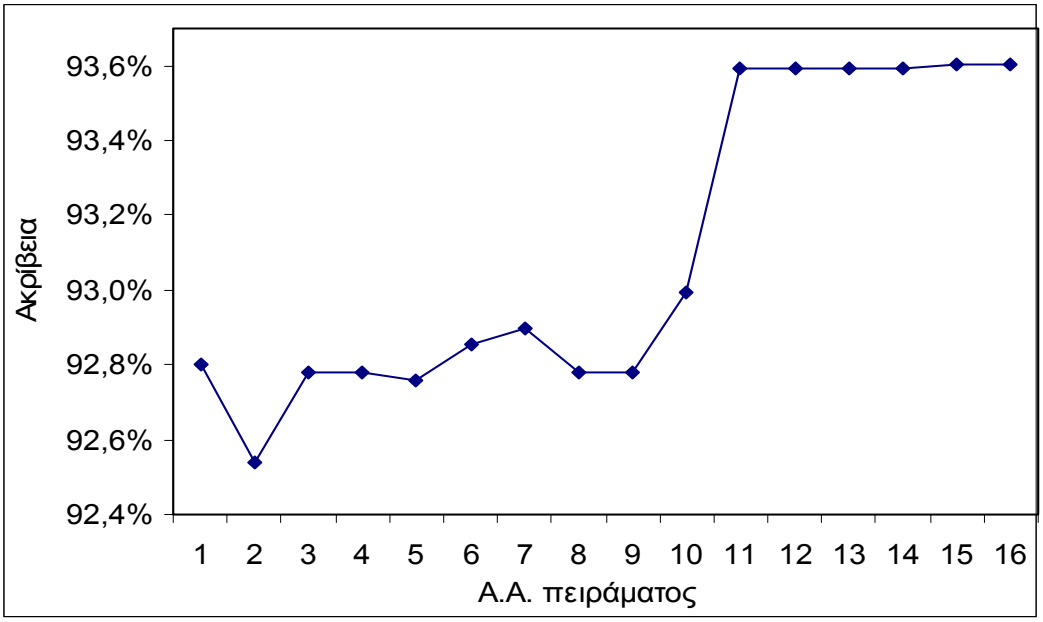
Ακολούθως, εισάγουμε στον tagger ρυθμίσεις που του επιτρέπουν να χρησιμοποιήσει και στοιχεία που αφορούν την ίδια την λέξη. Από μελέτη της ελληνικής γραμματικής, προκύπτει ότι μια επιλογή του ενός ή των δυο τελευταίων γραμμάτων μιας λέξης ως κριτήριο προσδιορισμού της αποτελεί λανθασμένη κίνηση, μιας και το μεγαλύτερο ποσοστό των καταλήξεων αποτελούνται από τουλάχιστον 3 γράμματα. Η προσαύξηση των παραπάνω στις ρυθμίσεις των άγνωστων λέξεων, εκτοξεύει τα ποσοστά επιτυχίας σε αυτές στο 68% και ωθεί το συνολικό ποσοστό λίγο πάνω από το 88%. Επειδή όμως, όπως προαναφέραμε, οι ελληνικές καταλήξεις αποτελούνται από τουλάχιστον 3 γράμματα, μεταβάλλουμε και πάλι τις ρυθμίσεις, λαμβάνοντας υπόψη ένα ακόμα (wfw/sssswnFw). Τα αποτελέσματα αυξάνονται κατά ακόμα 1% για τις άγνωστες λέξεις και κατά 0,2% για το σύνολο. Η περαιτέρω αύξηση της γραμμάτων σε πέντε, βάσει του γεγονότος ότι οι καταλήξεις των ρημάτων και των μετοχών της παθητικής φωνής μπορεί να αποτελούνται από 5 ή και 6 γράμματα, δεν επαληθεύει τις προβλέψεις και μειώνει σημαντικά τα τελικά ποσοστά. Επίσης, η χρήση της δυνατότητας του tagger να χρησιμοποιεί και κάποια από τα αρχικά γράμματα της λέξης για την αναγνώρισή της, προκάλεσε σύγχυση στο σύστημα και ελάττωσε τα συνολικά ποσοστά επιτυχίας σε σημαντικό βαθμό (1-5%). Στο ίδιο αποτέλεσμα οδήγησε και η χρησιμοποίηση της δυνατότητας να αναγνωρίζει κεφαλαία γράμματα μέσα στις λέξεις, ρύθμιση που επίσης απορρίφθηκε.

Τέλος, το μοναδικό κομμάτι που απέμεινε ήταν η δυνατότητα χρήσης ετικετών λέξεων. Αυτές μπορεί να είναι ετικέτες που έχουν αποσαφηνιστεί (disambiguated tags) και αφορούν τις λέξεις που βρίσκονται αριστερά ή ετικέτες που δεν έχουν

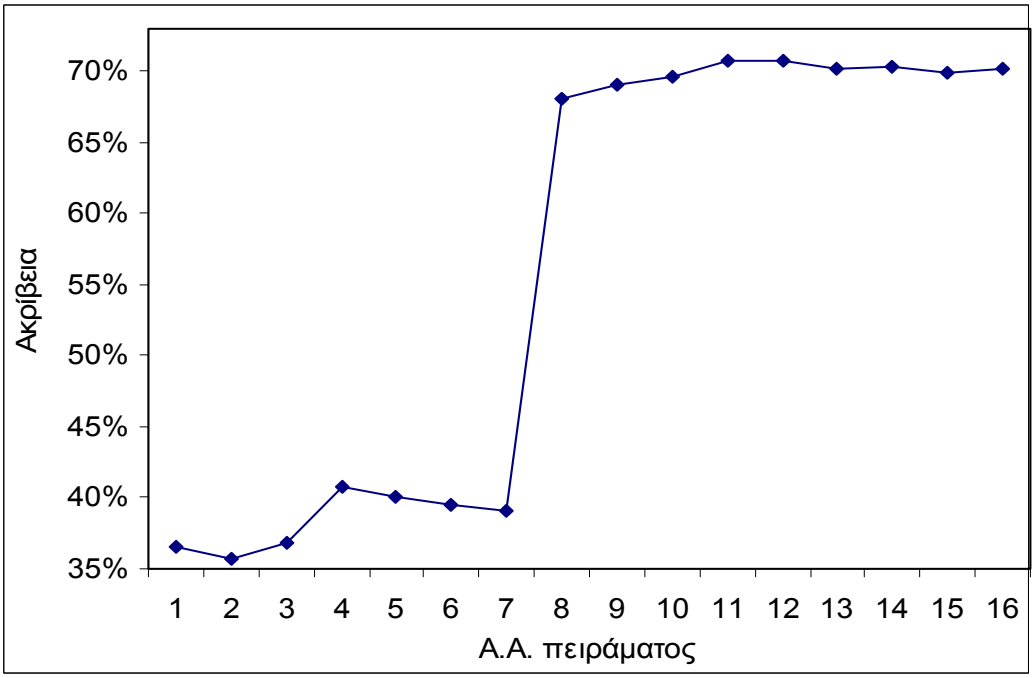
αποσαφηνιστεί (abmbitags) και αφορούν τις λέξεις που έπονται για τις οποίες δεν έχει γίνει η διαδικασία της ανάθεσης. Ξεκινώντας και πάλι με στοιχειώδεις ρυθμίσεις (dwfw/sssdwnFw, wfaw/sssswnFaw, dwfaw/sssdwnFaw), παρατηρούμε μια αύξηση κατά Μ.Ο. 2% τόσο στις γνωστές όσο και στις άγνωστες λέξεις, και άνοδο του συνολικού ποσοστού επιτυχίας στο 90%. Μελετώντας τα σφάλματα που προέκυψαν, επιχειρούμε μια αύξηση των χρησιμοποιούμενων ετικετών. Και εδώ όμως τα αποτελέσματα αποδεικνύουν ότι η κίνηση ήταν λανθασμένη. Έτσι λοιπόν καταλήγουμε ότι οι ρυθμίσεις που παρέχουν το καλύτερο αποτέλεσμα είναι οι: ***dwfaw*** για τις γνωστές λέξεις και ***ssssdwnFaw*** για τις άγνωστες. Στον πίνακα 5.1 που ακολουθεί παρατίθενται τα αποτελέσματα των μέσων όρων των παραπάνω πειραμάτων, ενώ στα διαγράμματα 5.1, 5.2 και 5.3 παρουσιάζονται γραφικά οι παραπάνω μεταβολές.

Παράμετροι			**Ποσοστά Επιτυχίας**		
Α.Α.	**Γνωστές**	**Άγνωστες**	**Γνωστές**	**Άγνωστες**	**Σύνολο**
1	wf	wF	92,8021%	36,6018%	83,9297%
2	fw	Fw	92,5377%	35,6524%	83,5343%
3	wfw	wFw	92,7791%	36,8608%	83,9450%
4	wfw	wnFw	92,7791%	40,7590%	84,5348%
5	wwfw	wwnFw	92,7597%	40,0931%	84,4241%
6	wfww	wnFww	92,8573%	39,5607%	84,4240%
7	wwfww	wwnFww	92,8970%	39,0166%	84,3866%
8	wfw	ssswnFw	92,7791%	68,0708%	88,8304%
9	wfw	sssswnFw	92,7791%	69,1025%	88,9885%
10	dwfw	ssssdwnFw	92,9942%	69,6510%	89,2525%
11	**dwfaw**	**ssssdwnFaw**	**93,5936%**	**70,6926%**	**89,9179%**
12	ddwfaw	sssddwnfaw	93,5953%	70,6817%	89,9200%
13	dwfaaw	ssssdwnFaaw	93,5935%	70,2513%	89,8491%
14	dwfaw	pssssdwnFaw	93,5936%	70,3685%	89,8738%
15	dwfaw	psssssdwnFaw	93,6046%	69,9486%	89,8184%
16	dwfaw	psssssdwncFaw	93,6046%	70,2331%	89,8641%

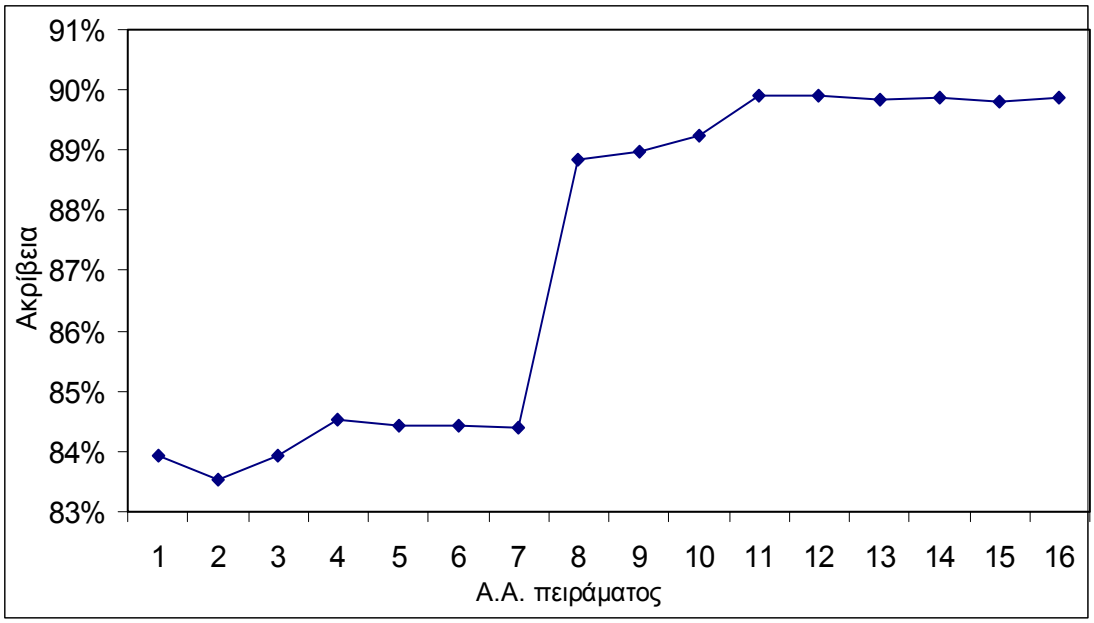
Πίνακας 5.1.1 Ακρίβεια πειραμάτων



Διάγραμμα 5.1.1 Ακρίβεια στην εύρεση των γνωστών λέξεων



Διάγραμμα 5.1.2 Ακρίβεια στην εύρεση των άγνωστων λέξεων



Διάγραμμα 5.1.3 Ακρίβεια στο σύνολο των λέξεων

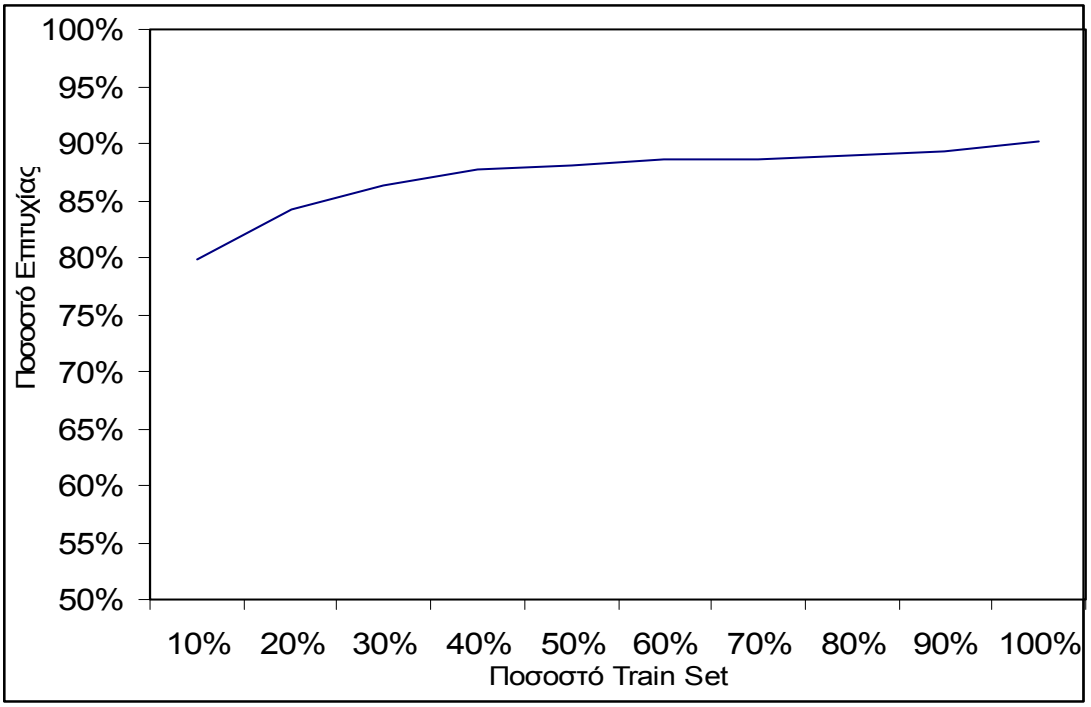
Όπως προαναφέραμε, στα παραπάνω πειράματα δεν έγινε καμιά τροποποίηση στις ρυθμίσεις που χρησιμοποιούνται από τον TiMBL. Σύμφωνα όμως με τον Daelemans, η επιλογή του κατάλληλου αλγορίθμου του TiMBL για τις διαδικασίες αναγνώρισης των γνωστών και των άγνωστων λέξεων μπορεί να επηρεάσει σημαντικά το τελικό αποτέλεσμα. Αποδεχόμενοι το παραπάνω γεγονός, και κρατώντας σταθερές τις ρυθμίσεις που προέκυψαν από την προηγούμενη διαδικασία, διενεργήσαμε εκ νέου πειράματα, μεταβάλλοντας πλέον τον αλγόριθμο της διαδικασίας. Τα αποτελέσματα απέδειξαν ότι η χρήση του IGTREE για την αποσαφήνιση των γνωστών λέξεων αυξάνει την επιτυχία κατά το στοιχειώδες 0,1%, σε σχέση με την προηγούμενοι περίπτωση, σε αντίθεση με τους υπόλοιπους αλγορίθμους, οι οποίοι αποκλειστικά και μόνο μειώνουν το αποτέλεσμα. Στην περίπτωση των αγνώστων λέξεων, βέλτιστος αποδείχθηκε ο αλγόριθμος k-NN. Αποκλειστικά γι' αυτόν πραγματοποιήθηκε πειραματική διαδικασία για την εύρεση του βέλτιστου πλήθους γειτόνων. Ξεκινώντας από k=1 και επεκτεινόμενοι έως k=12, αποδείχθηκε ότι η βέλτιστη επιλογή ήταν αυτή των 2, η οποία πέτυχε να βελτιώσει το ποσοστό έως και 3%,ανεβάζοντας τον συνολικό Μ.Ο. έως το 91%. Αξιοσημείωτο είναι ότι τα αποτελέσματα σημείωναν ανοδική πορεία έως το k=2, ενώ από εκεί και πέρα παρουσίασαν σημαντική πτώση, διαψεύδοντας τις εκτιμήσεις που ανάμεναν τις βέλτιστες επιδώσεις για k=4 ή k=5.

Στον πίνακα 5.2 παρατίθενται τα αποτελέσματα της διαδικασίας με την χρήση των βέλτιστων αλγορίθμων.

Ρυθμίσεις	Ποσοστά Επιτυχίας		
	Γνωστές	Άγνωστες	Σύνολο
Με τις προεπιλεγμένες ρυθμίσεις αλγορίθμων	93,5936%	70,6926%	89,9179%
Με τους βέλτιστους αλγόριθμους	93,6047%	72,7740%	90,2571%

Πίνακας 5.1.2

Στη συνέχεια ακολουθήθηκε μια παραλλαγή της παραπάνω διαδικασίας, η οποία είχε ως σκοπό να αξιολογήσει της επάρκεια του μεγέθους του σώματος εκπαίδευσης. Σε αυτή την περίπτωση, χρησιμοποιείται ένα ποσοστό 10x % (όπου x=1..9) του σώματος εκπαίδευσης για την επισημείωση του 100% του σώματος ελέγχου. Η μελέτη της συμπεριφοράς της καμπύλης επιτυχίας σε σχέση με το μέγεθος του σώματος εκπαίδευσης μπορεί να δώσει πληροφορίες σχετικά με το αν το μέγεθος του σώματος που έχουμε επιλέξει είναι επαρκές. Στο διάγραμμα που ακολουθεί φαίνεται ότι η συμπεριφορά των κειμένων. Ξεκινώντας με ένα αρκετά μικρό ποσοστό για το 10%, η απόδοση αυξάνεται σημαντικά για την χρήση του 20 και 30%. Από εκεί και πέρα ακολουθεί μια επίσης ανοδική πορεία, με ελάχιστή όμως κλίση. Το συμπέρασμα που προκύπτει από τα παραπάνω είναι ότι το σώμα κειμένων που έχουμε χρησιμοποιήσει είναι επαρκές. Έτσι μια αύξηση του μεγέθους του με κείμενα της ίδιας ποιότητας, όσο σημαντική κι αν είναι αυτή, δεν θα είναι σε θέση να αυξήσει το τελικό αποτέλεσμα σε τέτοιο βαθμό που να μπορεί να αιτιολογήσει αυτή την ενέργεια.



Διάγραμμα 5.1.4 Καμπύλη μάθησης

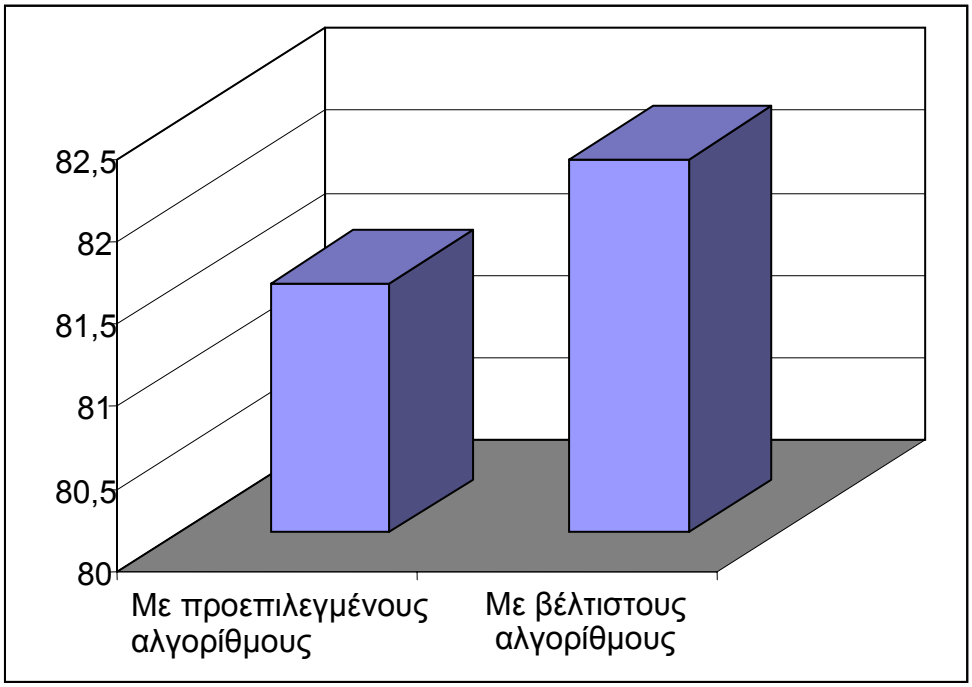
5.2 Πειράματα στο νέο σώμα κειμένων

Το συγκεκριμένο σώμα, αποτελεί το δεύτερο στο οποίο πραγματοποιήθηκαν πειράματα. Πρόκειται για μια συλλογή από 640 κείμενα, συνολικού μεγέθους 130500 λέξεων, αποτελούμενη από 50000 διαφορετικές λέξεις.

Η διαδικασία, εδώ, διαφέρει από αυτήν που περιγράφηκε στην προηγούμενη παράγραφο. Κι αυτό, γιατί ο σημαντικότερος στόχος είναι η μελέτη της συμπεριφοράς των ρυθμίσεων που προέκυψαν από την προηγούμενη διαδικασία, σε αυτό το corpus. Ξεκινώντας λοιπόν την έρευνα, διεξάγουμε πειράματα με ρυθμίσεις *dwfaw* για τις γνωστές λέξεις και *ssssdwnFaw* για τις άγνωστες. Σε πρώτη φάση γίνεται χρήση των default ρυθμίσεων του TiMBL και έπειτα με την χρήση των βέλτιστων αλγορίθμων που προκύψανε από την προηγούμενη διαδικασία. Τα αποτελέσματα που προκύπτουν, παρουσιάζουν μια σημαντικότατη πτώση σε σχέση με αυτά που λάβαμε κατά την μελέτη του 1^{ου} σώματος κειμένων. Στην προκειμένη περίπτωση θα πρέπει να τονιστεί η αδυναμία του συστήματος να μαντέψει το είδος μια άγνωστης λέξης – το 60% είναι ποσοστό που δεν μπορεί να δημιουργεί μεγάλη

αισιοδοξία -, σε σχέση με την σχεδόν άριστη ικανότητά του να ξεχωρίζει την ετικέτα μιας γνωστής λέξης (τα ποσά φτάνουν κατά Μ.Ο. το 98%).

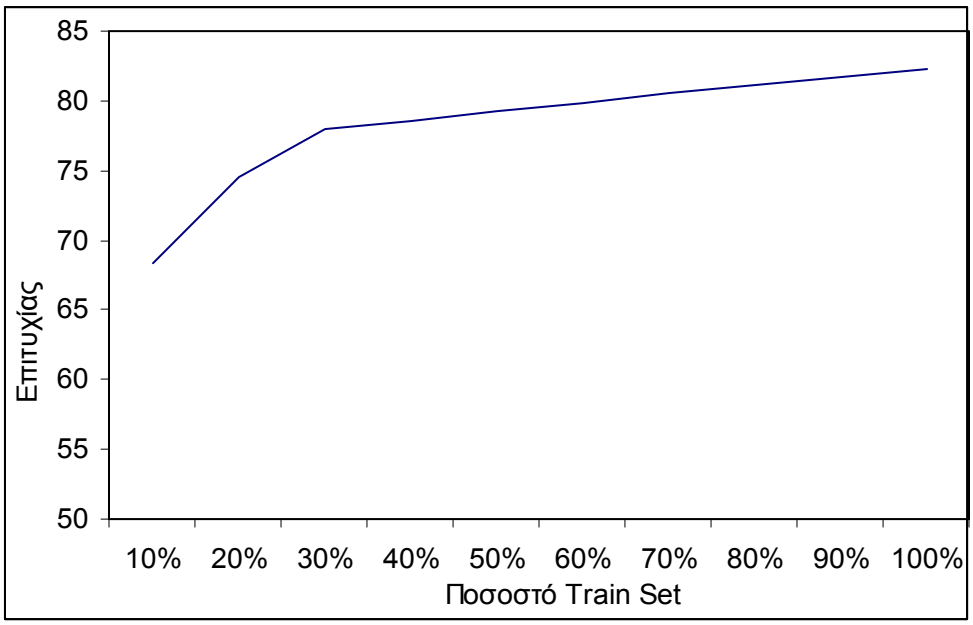
Τα προαναφερθέντα χαμηλά ποσοστά (με την χρήση των βέλτιστων αλγορίθμων φτάσαμε στο 82% - Διάγραμμα 5.5), σε συνδυασμό με την πληροφορία που διαθέτουμε για την δομή των κειμένων (στα συγκεκριμένα κείμενα υπάρχει μια πολύ μεγαλύτερη ομοιογένεια στον τρόπο δόμησης, καθώς και πολύ πιο οργανωμένη σύνταξη), δημιουργεί μια πρώτη έκπληξη για τα χαμηλά επίπεδα επιτυχίας. Μια πιο προσεκτική έρευνα όμως αποδεικνύει ότι τελικά αυτός είναι ο λόγος που τα προκαλεί όλα. Όπως προαναφέραμε, το σύστημα χρησιμοποιεί ρυθμίσεις που αφορούν την λέξη που ερευνούμε, καθώς και τις γειτονικές αυτής. Η εισαγωγή όμως μιας δευτερεύουσας πρότασης ανάμεσα στο υποκείμενο και το ρήμα ή η αντιστροφή των θέσεων επιθέτων και ουσιαστικών (οι παραπάνω δυο μορφές εμφανίζονται πολύ συχνά στα κείμενα) είναι παράγοντες που «καταστρέφουν» την ομαλή ροή του κειμένου. Αποτέλεσμα αυτού είναι η δημιουργία μεγάλου πλήθους αλληλουχιών λέξεων, οι οποίες είναι δύσκολο να καλυφθούν στο σύνολό τους από την φάση της εκπαίδευσης, και οι οποίες έχουν ως αποτέλεσμα την εμφάνιση σφαλμάτων.



Διάγραμμα 5.2.1 Ακρίβεια με το νέο σώμα κειμένων

Η προσπάθεια να μεταβληθούν οι παραπάνω ρυθμίσεις, κυρίως προς την κατεύθυνση την επέκτασης της γειτονιάς των λέξεων που λαμβάνεται υπόψη, ιδίως στην περίπτωση των άγνωστων λέξεων, όχι μόνο απέβη άκαρπη, αλλά και «καταπόντισε» τα αποτελέσματα έως και 6%. Στο ίδιο σημείο οδήγησε και η προσπάθεια μεταβολής των αλγορίθμων ή η προσπάθεια μεταβολής των παραμέτρων αυτών (όπως στην απόπειρα αύξησης του k για τον αλγόριθμο k-NN). Τα παραπάνω αποδεικνύουν ότι τελικά αυτές είναι και οι ρυθμίσεις που δίνουν βέλτιστα αποτελέσματα και σε αυτό το σώμα κειμένων.

Τέλος η μελέτη της καμπύλης μάθησης οδήγησε στα ίδια συμπεράσματα με την προηγούμενη περίπτωση. Η απόδοση ξεκινάει από μια τιμή αρκετά χαμηλή (68%) και ανέρχεται με αρκετά γοργούς ρυθμούς για τα επόμενα 3x10%. Από εκεί και πέρα η άνοδος γίνεται με πολύ χαμηλό ρυθμό, παρουσιάζοντας μια τάση σταθεροποίησης. Το γεγονός αυτό δηλώνει και την επάρκεια του μεγέθους του σώματος εκπαίδευσης για την διαδικασία της επισημείωσης.

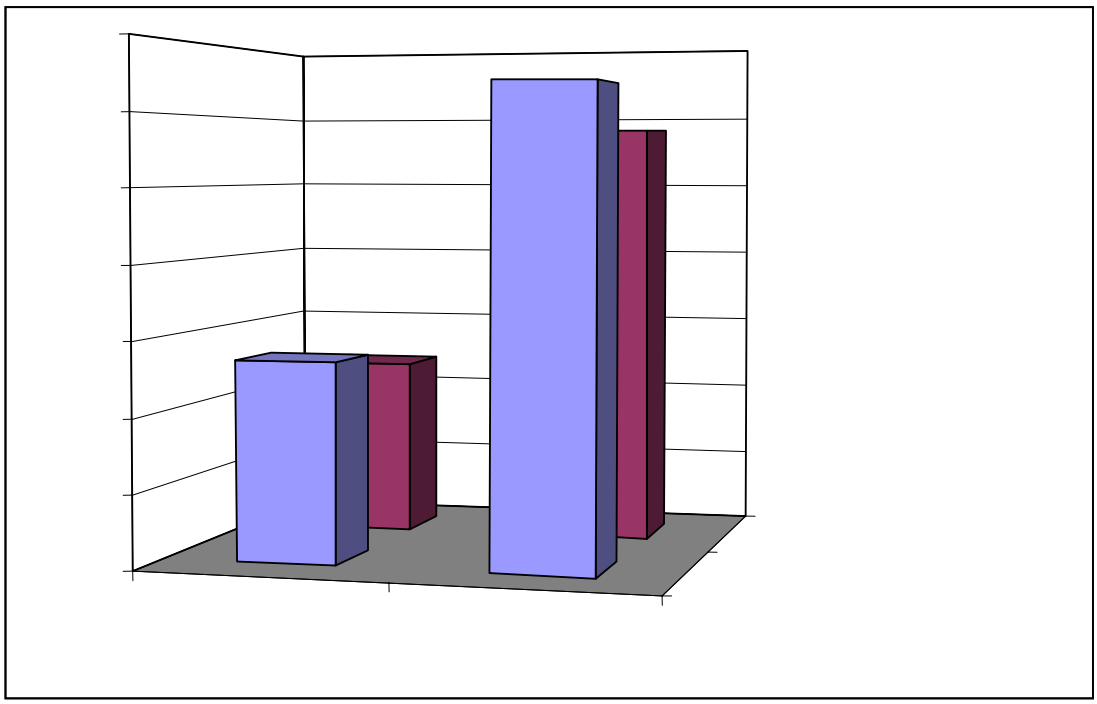


Διάγραμμα 5.2.2 Καμπύλη μάθησης

5.3 Πειράματα στο σύνολο του μεγέθους των κειμένων

Στο τμήμα αυτό αναφέρονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν με την χρήση του συνόλου των κειμένων, τόσο για εκπαίδευση όσο και για έλεγχο. Αρχικά πραγματοποιήσαμε εκπαίδευση του συστήματος τόσο στο πρώτο όσο και στο δεύτερο κείμενο, και ακολούθως πραγματοποιήσαμε έλεγχο για την ακρίβεια που επιτυγχάνεται όταν καθένα από αυτά καλείται να επισημειώσει τον εαυτό του. Το αποτέλεσμα που λάβαμε και στις δυο περιπτώσεις ήταν επιτυχία 100% . Αυτό μπορεί εύκολα να εξηγηθεί ως εξής: κατ' αρχάς, το σύστημα καλείται να λειτουργήσει πάνω σε ένα κείμενο για το οποίο ήδη γνωρίζει όλες τις λέξεις, άρα εργάζεται μόνο με λέξεις που ήδη γνωρίζει τις πιθανές τους ετικέτες. Εκτός αυτού, οι περιπτώσεις γειτόνων που συναντάει και που καλείται να χρησιμοποιήσει, είναι επίσης γνωστές, αφού έχει ήδη εκπαιδευτεί πάνω σε αυτές. Άρα έχει όλη την πληροφορία που χρειάζεται για να δώσει σε όλες τις περιπτώσεις το σωστό αποτέλεσμα.

Έτσι λοιπόν, πλέον το ενδιαφέρον στρέφεται στην διερεύνηση του τρόπου που συμπεριφέρεται το κάθε κείμενο, όταν καλείται να αναγνωρίσει λέξεις κειμένου που ανήκει και σε διαφορετική κατηγορία (αναφερόμαστε τόσο σε μορφολογικά χαρακτηριστικά, όσο και σε χαρακτηριστικά που αφορούν το περιεχόμενο του κειμένου). Για το συγκεκριμένο πείραμα πραγματοποιήθηκε εκπαίδευση στο σύνολο του πρώτου σώματος και ακολούθησε απόπειρα επισημείωσης του δεύτερου σώματος, χρησιμοποιώντας τις ρυθμίσεις που προέκυψαν από την $1^{η}$ φάση. Η λειτουργία αυτή πραγματοποιήθηκε χωρίς, αλλά και με την χρήση των διαφοροποιημένων παραμέτρων για τους αλγόριθμους του TiMBL. Έπειτα, ακολούθησε η ίδια διαδικασία με αντιστροφή, όμως, των ρόλων των σωμάτων κειμένων. Τα αποτελέσματα που παρουσιάζονται παρακάτω και δείχνουν παρόμοια ποσοστά επιτυχίας και στις δυο κατευθύνσεις. Το γεγονός αυτό δείχνει ότι η διαδικασία είναι ανεξάρτητη του ύφους του σώματος κειμένων που χρησιμοποιούνται για την εκπαίδευση ή την επισημείωση.



Διάγραμμα 5.3.1 Σύγκριση ακρίβειας

6. Συμπεράσματα & Μελλοντική Έρευνα

6.1 Ανασκόπηση της εργασίας και συμπεράσματα

Αντικείμενο της παρούσας εργασίας αποτέλεσε η μελέτη για την εφαρμογή τεχνικών Μηχανικής Μάθησης στην διαδικασία της επισημείωσης κειμένων της ελληνικής γλώσσας. Σε πρώτη φάση έγινε μελέτη του θεωρητικού υπόβαθρου της περιοχής. Η σχετική βιβλιογραφία, έδειξε ότι στο παρελθόν έχουν γίνει αρκετές αξιόλογες προσπάθειες ώστε να δοθεί κάποια λύση στο πρόβλημα αυτό. Οι προσπάθειες ξεκίνησαν με την χρήση γνωστών συστημάτων που βασίζονταν στην εύρεση του tag υπολογίζοντας πιθανότητες βάσει καθορισμένων χαρακτηριστικών. Αργότερα έγινε προσπάθεια να δημιουργηθούν κι άλλα συστήματα Μηχανικής Μάθησης για την παροχή καλύτερων λύσεων. Τα πειράματα που πραγματοποιήθηκαν, παρόλο που δεν μπορούν να συγκριθούν μεταξύ τους, λόγο του διαφορετικού συνόλου ετικετών και των διαφορετικών κειμένων που επελέγησαν για το καθένα, έδειξαν ικανοποιητικά αποτελέσματα, με ποσοστό επιτυχίας, συνήθως, άνω του 90%. Σε παρόμοια επίπεδα κινήθηκαν και οι προσπάθειες που αφορούσαν τα ελληνικά.

Στα πλαίσια της συγκεκριμένης εργασίας τα πειράματα διεξήχθησαν με την βοήθεια του συστήματος MBT/TiMBL. Πριν όμως από αυτά, προηγήθηκε η διαδικασία της προετοιμασίας. Αυτή περιελάμβανε, την κατασκευή του WordTagger, ενός προγράμματος που δίνει στον χρήστη του την δυνατότητα να επισημειώσει χειροκίνητα ένα κείμενο, με σχετικά γρήγορα ρυθμό. Παράλληλα δημιουργήθηκε και το σύστημα ErrorDetector, ένα προϊόν το οποίο βασίζεται στις οδηγίες των Dickinson-Meurers, και έχει ως στόχο την ανακάλυψη σφαλμάτων επισημείωσης μέσα σε ένα κείμενο. Τέλος, ακολούθησε η δημιουργία ενός νέου, αρκετά εκτεταμένου σώματος επισημειωμένων κειμένων.

Στα πειράματα που ακολούθησαν, έγινε έρευνα σε δυο σώματα κειμένων. Η διαδικασία ξεκίνησε με την χρήση δεκαπλή διασταυρωμένη επικύρωση για την ανεύρεση των βέλτιστων ρυθμίσεων του MBT. Αυτές ήταν *dwfaw* για τις γνωστές λέξεις και *ssssdwnFaw* για τις άγνωστες. Επίσης, προέκυψε ότι η χρήση των

αλγορίθμων IGTREE για τις πρώτες και k-NN με k=2 για τις δεύτερες, έδινε, σε κάθε περίπτωση, καλύτερα αποτελέσματα έως και 2%. Συγκρίνοντας μεταξύ τους τα αποτελέσματα που έδωσαν τα δυο σύνολα κειμένων, παρατηρούμε σημαντική διαφορά, 90% έναντι 82%, γεγονός που οφείλεται στον τρόπο σύνταξης του δεύτερου, και πιο συγκεκριμένα στην διάσπαση της ροής του λόγου που προκαλεί η συνεχής εισαγωγή δευτερευουσών προτάσεων.

Επίσης διεξήχθησαν πειράματα με μεταβολή του μεγέθους του σώματος εκπαίδευσης, τα οποία παρουσίασαν παρόμοια συμπεριφορά και στις δυο περιπτώσεις κειμένων, η οποία συνίστατο από απότομη αύξηση της απόδοσης στα πρώτα στάδια και πιο αργή αργότερα, για να συγκλίνει σε ένα όριο κοντά στα ποσοστά που αναφέραμε παραπάνω. Αυτό αποδεικνύει ότι το μέγεθος των κειμένων είναι αποδεκτό και περαιτέρω αύξησή τους δεν είναι σε θέση να αντισταθμιστεί από ανάλογη αύξηση των αποτελεσμάτων. Επίσης, η διεξαγωγή πειραμάτων από το ένα κείμενο στο άλλο, έδειξε παρόμοιο βαθμό απόδοσης, γεγονός που αποδεικνύει ότι, το περιεχόμενο δεν επηρεάζει την διαδικασία.

6.2 Προτάσεις για μελλοντική διερεύνηση

Ξεκινώντας τις προτάσεις για την μελλοντική έρευνα θα πρέπει να αναφερθούμε σε μια ημιτελή απόπειρα που ξεκίνησε ως τμήμα της παρούσης εργασίας. Αυτή είναι η προσπάθεια διερεύνησης με χωριζόμενες ετικέτες. Όπως περιγράφηκε παραπάνω, στην φάση της επισημείωσης αποδίδεται μονομιάς σε μια λέξη ένα συγκεκριμένο tag, το οποίο περιέχει το σύνολο της πληροφορίας. Το πρόβλημα είναι ότι σφάλμα σε ένα από τα τμήματα της ετικέτας σημαίνει και σφάλμα στο σύνολό του. Η πρόταση που προκύπτει ως λύση του παραπάνω είναι η σταδιακή απόδοση των χαρακτηριστικών της ετικέτας στην κάθε λέξη. Έτσι γίνεται κλιμακωτή αναγνώριση της κατηγορίας που ανήκει η κάθε λέξη, έως ότου συμπληρωθεί το σύνολο της ετικέτας (π.χ. πρώτα αναγνωρίζεται αν η λέξη είναι ρήμα, ουσιαστικό, επίθετο, ... Έπειτα ο χρόνος. Ακολούθως η φωνή, κ.ο.κ.).

Για τον σκοπό αυτό έγινε απόπειρα τροποποίησης του MBT. Ο λόγος που δεν επελέγη η κατασκευή ενός νέου συστήματος ήταν κυρίως το γεγονός ότι το υπάρχον σύστημα παρέχει ένα σημαντικό αριθμό αλγορίθμων που μπορούν να δοκιμαστούν στην παραπάνω διαδικασία. Από την μελέτη που πραγματοποιήθηκε, προέκυψε ότι η τροποποίηση της διαδικασίας δεν ήταν υποχρεωτικό να περιλαμβάνει και την φάση της εκπαίδευσης. Έτσι, χρησιμοποιώντας τα αρχεία λεξικών που είχαν προκύψει από την εκπαίδευση με τις ενιαίες ετικέτες, θα μπορούσαν να χρησιμοποιηθούν και στην προκειμένη περίπτωση, αρκεί να γινόταν σε κάθε στάδιο η επιλογή του κατάλληλου τμήματος της ετικέτας.

Η δεύτερη πρόταση προκύπτει από τα αποτελέσματα των πειραμάτων πάνω στο δεύτερο σώμα κειμένων. Όπως είδαμε, η παρεμβολή δευτερευουσών προτάσεων μέσα σε μια άλλη δημιουργεί προβλήματα στην διαδικασία, αφού δημιουργεί σημεία «ασυνέχειας» που μπερδεύουν το σύστημα. Έτσι, θα είχε μεγάλο ενδιαφέρον αν μπορούσε ο MBT να συνδυαστεί με ένα σύστημα που έχει την δυνατότητα να αποκόπτει τέτοιες προτάσεις, επιτρέποντας του να εργάζεται με μια μόνο πρόταση την κάθε φορά.

Η τελευταία πρόταση αφορά την επέκταση της μεθόδου και ο συνδυασμός της με τεχνικές Ενεργού Μηχανικής Μάθησης. Η χρήση της Ενεργούς Μηχανικής Μάθησης, έγκειται στην δημιουργία ενός συστήματος το οποίο χρησιμοποιώντας μικρότερου μεγέθους σώματα κειμένων θα είναι σε θέση να ανακαλύπτει ένα το σωστό tag της κάθε λέξης, χρησιμοποιώντας την δυνατότητα αλληλεπίδρασης με τον χρήστη. Το σύστημα θα μπορεί να ανακαλύπτει περιπτώσεις που το «προβληματίζουν» και αφού λάβει την ανάλογη ανατροφοδότηση από τον χρήστη θα μπορεί να συνεχίσει την διαδικασία και να βελτιώνει την ακρίβειά του.

7. Βιβλιογραφικές Αναφορές

Βιβλία

Βλαχάβας, Κεφάλας, Βασιλειάδης, Ρεφάνης, Κόκορας, Σακελαρίου "Τεχνητή Νοημοσύνη", εκδόσεις Γαρταγάνη, 2002

Mitchell, T.M. "Machine Learning", McGraw-Hill International Editions, 1997

Quinlan R. "Applications of Expert Systems", Addison-Wesley, 1987

Άρθρα

Brill E. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study on Part of Speech Tagging", Computational Linguistics, 1995/12

Daelemans W., Zavrel J., Van Den Bosch A., Van Der Sloot K. "MBT: Memory-Based Tagger, version 1.0, Reference Guide"

Daelemans W., Zavrel J. "MBT: A Memory-Based Part of Speech Tagger-Generator", Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark, pp.14-27, 1996

Daelemans W., Zavrel J., Van Der Sloot K., Van Den Bosch A. "TiMBL: Tilburg Memory-Based Learner, version 4.3, Reference Guide"

Dermatas E., Kokkinakis G. "Automatic Stochastic Tagging of Natural Language Texts", Computational Linguistics**,** Volume 21, Issue 2, pp. 137-163, 1995/06

Dickinson M., Meurers W.D. "Detecting Errors in Part-of-Speech Annotation", Proceedings of EACL, 2003

Johansson S. "The tagged LOB Corpus: User's Manual", Bergen, Norway: Norwegian Computing Centre for Humanities, 1986

Orphanos G.S., Christodoulakis D.N. "POS Disambiguation and Unknown Words Guessing with Decision Trees", In proceedings EACL 1999, pp.134-141, 1999

Orphanos G., Kalles D., Papagelis T., Christodoulakis D. "Decision Trees and NLP: A Case Study in POS Tagging", In proceedings of ACAI'99, 1999

Papageorgiou H., Prokopidis P., Giouli V., Piperidis S. "A Unified POS Tagging Architecture and its Application to Greek", In Proceedings of the 2nd Language Resources and Evaluation Conference, pp.1455-1462, Athens, 2000

Petatsis G., Paliouras G., Karkaletsis V., Spyropoulos C.D., Androutsopoulos I. "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, Chania, Greece, 1999

Ratnaparkhi A. "A maximum entropy part-of-speech tagger", In Proc. Of the Conference on Empirical Methods in Natural Language Processing, May 17-19 1996, University of Pennsylvania, 1996

Steetskamp R. "An implementation of a probabilistic tagger", MSc Thesis, TOSCA Research Group, University of Nijmegen, Nijmegen, The Netherlands, 1995

Sakkis G., Androutsopoulos I., Paliouras G., Karkaletsis V., Spyropoulos C.D., Stamatopoulos P. "A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists", Kluwer Academic Publishers, 2003

Van Der Sloot K. "TiMBL: Tilburg Memory-Based Learner, version 4.3, API Reference Guide"

Zavrel J., Daelemans W. "Recent Advances in Memory-Based Part-of-Speech Tagging", VI Simposio Internacional de Comunicacion Social, Santiago de Cuba, pp.590-597, 1999. ILK-9903, 1999

Παράρτημα Α

Summary of the thesis in English

This thesis explores the use of machine learning techniques in part of speech (POS) tagging of Greek texts. Having studied several previous learning-based approaches to POS tagging for English and Greek texts, we decided to use the MBT/TiMBL instance-based learning system, which has been used in several European languages. We adopted the Greek tag-set of the PAROLE project (120 tags), and conducted experiments with several configurations of MBT/TiMBL. The experiments were performed with two collections of manually tagged Greek texts, one consisting of advertisements (32K words), and another one consisting of newspaper articles (130K words). The latter was tagged during the work of this thesis. Two supporting pieces of software were also developed, a tool that helps a user tag manually collections of Greek texts with PAROLE tags, and a tool that detects possible tagging errors. The best of the MBT/TiMBL Greek configurations that we studied achieved an accuracy score of approximately 90% on the collection of advertisements, and 82% on the collection of news articles. Training on one of the collections and testing on the other one led in both cases to an accuracy score of approximately 73%.

Παράρτημα Β

Επεξήγηση του συνόλου ετικετών

Στους πίνακες που ακολουθούν παρατίθεται το σύνολο των ετικετών που χρησιμοποιήθηκε κατά την διαδικασία της επισημείωσης. Δίπλα σε κάθε ετικέτα αναφέρεται και το πλήρες μονοπάτι της κατηγορίας που αυτό χαρακτηρίζει.

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
AtIdFePlNm	Αόριστο άρθρο θηλυκό πληθυντικός ονομαστική

AtIdFePlGe	Αόριστο άρθρο θηλυκό πληθυντικός γενική
AtIdFePlAc	Αόριστο άρθρο θηλυκό πληθυντικός αιτιατική
AtIdNeSgNm	Αόριστο άρθρο ουδέτερο ενικός ονομαστική
AtIdNeSgGe	Αόριστο άρθρο ουδέτερο ενικός γενική
AtIdNeSgAc	Αόριστο άρθρο ουδέτερο ενικός αιτιατική
AtIdNePlNm	Αόριστο άρθρο ουδέτερο πληθυντικός ονομαστική
AtIdNePlGe	Αόριστο άρθρο ουδέτερο πληθυντικός γενική
AtIdNePlAc	Αόριστο άρθρο ουδέτερο πληθυντικός 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

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