ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή εργασία

Ένας νέος ελληνικός επισημειωτής μερών του λόγου, βασισμένος σε ταξινομητή μεγίστης εντροπίας

Ευαγγελία Κολέλη

Επιβλέπων: Ίων Ανδρουτσόπουλος

Βοηθός επίβλεψης: Πρόδρομος Μαλακασιώτης

Μάρτιος 2011

Περιεχόμενα

Κεφάλαιο 1:

Εισαγωγή

1.1 Αντικείμενο της εργασίας

Η εργασία αυτή έχει ως σκοπό τη δημιουργία ενός συστήματος επισημείωσης μερών του λόγου (part-of-speech tagger) για την ελληνική γλώσσα. Υπάρχουν τρεις προηγούμενες εργασίες που ασχολήθηκαν με το ίδιο πρόβλημα (Μαλακασιώτης, 2005) (Χρονάκης, 2006) (Παππάς, 2008). Τόσο τα συστήματα των προηγούμενων, όσο και αυτής της εργασίας κατατάσσουν κάθε λέξη ενός δοθέντος κειμένου σε κατηγορίες. Οι κατηγορίες μπορεί να αντιστοιχούν στα μέρη του λόγου (π.χ. ρήμα, επίθετο, άρθρο) ή να είναι πιο λεπτομερείς και να παρέχουν επιπλέον πληροφορίες (π.χ. επίθετο αρσενικού γένους στην ονομαστική ενικού).

Στις προηγούμενες εργασίες χρησιμοποιήθηκε ο αλγόριθμος μάθησης των κ κοντινότερων γειτόνων για την εκπαίδευση ενός ταξινομητή που αποφασίζει σε ποια κατηγορία ανήκει κάθε λέξη. Σε αυτή την εργασία, δοκιμάσαμε ένα διαφορετικό αλγόριθμο μάθησης, και συγκεκριμένα τον ταξινομητή Μέγιστης Εντροπίας. Επίσης, ενώ οι παλιότερες εργασίες ασχολήθηκαν και με την ενεργητική μάθηση, στην οποία το ίδιο το σύστημα προτείνει κατά το στάδιο της εκπαίδευσης λέξεις προς επισημείωση, στη συγκεκριμένη εργασία ασχοληθήκαμε μόνο με την παθητική μάθηση, χρησιμοποιώντας ένα προϋπάρχον επισημειωμένο (με τις σωστές κατηγορίες) σώμα κειμένων εκπαίδευσης. Τέλος, αντίθετα από τα συστήματα των προηγούμενων εργασιών, το σύστημα της παρούσας εργασίας παρέχει Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface ή API), μέσω της οποίας είναι ευκολότερο να ενσωματωθεί σε μεγαλύτερα συστήματα.

1.2 Διάρθρωση της εργασίας

Η εργασία αποτελείται από τα εξής κεφάλαια:

- Στο κεφάλαιο 2 περιγράφεται αναλυτικότερα το πρόβλημα που προσπαθούμε να λύσουμε, καθώς και το σύστημα που αναπτύξαμε.
- Στο κεφάλαιο 3, περιγράφουμε το σύνολο των δεδομένων που χρησιμοποιήσαμε για την εκπαίδευση και την αξιολόγηση του συστήματος. Επίσης, περιγράφουμε τα πειράματα που κάναμε μέχρι να καταλήξουμε στην τελική έκδοση του συστήματος.
- Τέλος, στο κεφάλαιο 4 συνοψίζονται τα προηγούμενα κεφάλαια και τα συμπεράσματα της εργασίας, ενώ προτείνονται και μελλοντικές επεκτάσεις της.

1.3 Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Ίωνα Ανδρουτσόπουλο, ο οποίος με καθοδηγούσε και με συμβούλευε σε όλη τη διάρκεια της πτυχιακής μου εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον Πρόδρομο Μαλακασιώτη για την ουσιαστική βοήθειά του, με τον οποίο συνεργάστηκα για τη λύση προβλημάτων όχι μόνο σε θεωρητικό επίπεδο, αλλά επίσης σε θέματα υλοποίησης. Επιπλέον, ευχαριστώ το Γεράσιμο Λάμπουρα για τις πολύτιμες συμβουλές του, καθ' όλη τη διάρκεια της πτυχιακής μου εργασίας.

Τέλος, ευχαριστώ όλα τα μέλη της Ομάδας Επεξεργασίας Φυσικής Γλώσσας του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών για την ενθάρρυνση και τη βοήθεια που μου παρείχαν.

Κεφάλαιο 2:

Περιγραφή του νέου συστήματος

2.1 Αναγνώριση μερών του λόγου για τα Ελληνικά

Το πρόβλημα που μελετάμε είναι η αυτόματη κατάταξη λέξεων ελληνικών κειμένων σε μέρη του λόγου ή λεπτομερέστερες κατηγορίες. Κάποιος ίσως σκεφτεί ότι το πρόβλημα αυτό θα μπορούσε να λυθεί με τη χρήση ενός λεξικού. Η προσέγγιση, όμως, αυτή δεν είναι επαρκής· για παράδειγμα, η λέξη «δημοσιεύσεις» μπορεί να είναι είτε ρήμα είτε ουσιαστικό, ανάλογα με τα συμφραζόμενα.

Τα συστήματα αναγνώρισης μερών του λόγου για ελληνικά κείμενα που προαναφέραμε (Μαλακασιώτης, 2005) (Χρονάκης, 2006) (Παππάς, 2008) χρησιμοποιούσαν ταξινομητή k κοντινότερων γειτόνων (k-NN) και μεθόδους ενεργητικής μάθησης. Το σύστημα αυτής της εργασίας χρησιμοποιεί την υλοποίηση του ταξινομητή Μέγιστης Εντροπίας (Maximum Entropy classifier) του Πανεπιστημίου Stanford.[1]

2.2 Κατηγορίες

Το σύστημα αυτής της εργασίας, όπως και των τριών προηγούμενων, υποστηρίζει δύο σύνολα κατηγοριών. Το πρώτο είναι το βασικό σύνολο κατηγοριών, που περιλαμβάνει 12 κατηγορίες:

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

Το δεύτερο σύνολο κατηγοριών είναι πιο εκτεταμένο. Περιλαμβάνει 170 κατηγορίες, οι οποίες παρέχουν περισσότερες πληροφορίες, όπως αριθμός, γένος, πτώση και άλλα. Το σύνολο αυτό περιγράφεται αναλυτικά σε προηγούμενη εργασία (Παππάς, 2008).

[1] Βλ. http://nlp.stanford.edu/software/index.shtml.

2.3 Ιδιότητες

Κάθε μία λέξη (για την ακρίβεια εμφάνιση λέξης) σε ένα κείμενο αναπαριστάται ως ένα διάνυσμα ιδιοτήτων που παρέχει πληροφορίες για τη λέξη και τα συμφραζόμενά της. Ως συμφραζόμενα σε αυτή την εργασία θεωρούμε τις λέξεις (τη «γειτονιά») που βρίσκονται πριν και μετά από την υπό κατάταξη λέξη (πριν και μετά τη λέξη που παριστάνεται με το συγκεκριμένο διάνυσμα). Το σύστημα αυτής της εργασίας χρησιμοποιεί διαφορετικό σύνολο ιδιοτήτων από τα συστήματα των προηγούμενων τριών εργασιών. Επίσης, ενώ στο πιο πρόσφατο από τα τρία συστήματα (Παππάς, 2008) οι ιδιότητες ήταν στην πλειοψηφία τους συμβολοσειρές, στο νέο σύστημα χρησιμοποιούμε κυρίως ιδιότητες με πραγματικές τιμές. Μετά από πειράματα, καταλήξαμε να χρησιμοποιούμε διαφορετικές ιδιότητες για τα δύο σύνολα κατηγοριών (βασικό και εκτεταμένο). Στο επόμενο κεφάλαιο, φαίνεται αναλυτικά η πορεία των πειραμάτων και πώς τελικά καταλήξαμε να χρησιμοποιούμε τις συγκεκριμένες ιδιότητες.

Όταν χρησιμοποιείται το βασικό σύνολο κατηγοριών, κάθε λέξη (ακριβέστερα, κάθε λεκτική μονάδα, token) παριστάνεται ως ένα διάνυσμα με τις τιμές των παρακάτω ιδιοτήτων. Η έννοια της «ιδιότητας αμφισημίας» (ambitag) εξηγείται παρακάτω.

1. Ιδιότητες αμφισημίας της λεκτικής μονάδας.
2. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας.
3. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας.
4. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας.
5. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων).
6. Ύπαρξη αποστρόφου στη λεκτική μονάδα.
7. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα.
8. Ύπαρξη τελείας στη λεκτική μονάδα.
9. Ύπαρξη κόμματος στη λεκτική μονάδα.
10. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα.
11. Ιδιότητες αμφισημίας της επόμενης λεκτικής μονάδας.
12. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της επόμενης λεκτικής μονάδας.
13. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της επόμενης λεκτικής μονάδας.
14. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της επόμενης λεκτικής μονάδας.

Όταν χρησιμοποιείται το εκτεταμένο σύνολο κατηγοριών, τα διανύσματα περιέχουν τις τιμές των παραπάνω ιδιοτήτων και επιπλέον τις τιμές των παρακάτω ιδιοτήτων:

15. Ιδιότητες αμφισημίας της προηγούμενης λεκτικής μονάδας.
16. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προηγούμενης λεκτικής μονάδας.
17. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προηγούμενης λεκτικής μονάδας.
18. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προηγούμενης λεκτικής μονάδας.
19. Ιδιότητες αμφισημίας της προ-προηγούμενης λεκτικής μονάδας.
20. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προ-προηγούμενης λεκτικής μονάδας.
21. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας.

22. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας.

Οι ιδιότητες αμφισημίας μίας λεκτικής μονάδας είναι τόσες, όσες είναι και οι κατηγορίες. Οι τιμές των ιδιοτήτων αυτών είναι πραγματικοί αριθμοί. Κάθε μία ιδιότητα αμφισημίας δείχνει πόσο συχνά (ποσοστό) εμφανίζεται η λεκτική μονάδα (την οποία παριστάνει το διάνυσμα) στα παραδείγματα εκπαίδευσης με κάποια συγκεκριμένη κατηγορία. Για παράδειγμα, αν η λέξη «πολύ» εμφανίζεται 10 φορές στα παραδείγματα εκπαίδευσης, τις 3 φορές με την κατηγορία *επίρρημα*, ενώ τις 7 με την κατηγορία *επίθετο*, οι ιδιότητες αμφισημίας του διανύσματος που παριστάνει μια εμφάνιση αυτής της λέξης θα έχουν τις παρακάτω τιμές, υποθέτοντας ότι χρησιμοποιείται το μικρό σύνολο κατηγοριών:

isAdverb =0.3, isAdjective = 0.7, όπου isAdverb και isAdjective οι ιδιότητες αμφισημίας για τις κατηγορίες *επίρρημα* και *επίθετο* αντίστοιχα (βλ. παράρτημα Α). Οι ιδιότητες αμφισημίας για τις υπόλοιπες κατηγορίες θα έχουν τιμή 0.

Αν μία λεκτική μονάδα δεν εμφανίζεται στα δεδομένα εκπαίδευσης, τότε όλες οι ιδιότητες αμφισημίας της έχουν τιμή 0.

Ομοίως υπολογίζονται οι ιδιότητες αμφισημίας των καταλήξεων. Για παράδειγμα, αν μια κατάληξη (π.χ. «-εις») εμφανίζεται ως κατάληξη ρήματος με συχνότητα 0.7 στα δεδομένα εκπαίδευσης και ως κατάληξη ουσιαστικού με συχνότητα 0.3, τότε οι ιδιότητες αμφισημίας της θα έχουν τις τιμές:

isVerb =0.7, isNoun = 0.3, όπου isVerb και isNoun οι ιδιότητες αμφισημίας για τις κατηγορίες *ρήμα* και *ουσιαστικό* αντίστοιχα (βλ. παράρτημα Α). Οι ιδιότητες αμφισημίας για τις υπόλοιπες κατηγορίες θα έχουν τιμή 0.

Χρησιμοποιούμε ιδιότητες αμφισημίας για ολόκληρες λεκτικές μονάδες, καθώς και για τις καταλήξεις μήκους ενός, δύο και τριών χαρακτήρων των λεκτικών μονάδων. Σε περίπτωση που μια λεκτική μονάδα έχει λιγότερα από τρία γράμματα, τότε:

- Αν η λεκτική μονάδα έχει ακριβώς δύο χαρακτήρες (γράμματα), τότε οι ιδιότητες αμφισημίας για την κατάληξη τριών χαρακτήρων είναι ίδιες με τις ιδιότητες αμφισημίας για την κατάληξη δύο χαρακτήρων και ίδιες με τις ιδιότητες αμφισημίας για ολόκληρη τη λεκτική μονάδα.
- Αν η λεκτική μονάδα έχει ακριβώς ένα χαρακτήρα, τότε οι ιδιότητες αμφισημίας για την κατάληξη δύο χαρακτήρων, τριών χαρακτήρων καθώς για ολόκληρη τη λεκτική μονάδα είναι ίδιες με τις ιδιότητες αμφισημίας για την κατάληξη ενός χαρακτήρα.

2.4 Λεξικό και κανόνες διόρθωσης λαθών

Κάποιες λέξεις ανήκουν αποκλειστικά σε μία κατηγορία. Για παράδειγμα, η λέξη «και» είναι σύνδεσμος και δεν υπάρχει περίπτωση να τη συναντήσουμε ως άλλο μέρος του λόγου. Το σύστημα της παρούσας εργασίας, όπως και το αμέσως προηγούμενο, δημιουργεί στη διάρκεια της εκπαίδευσης ένα λεξικό. Στο λεξικό αυτό αποθηκεύονται οι λεκτικές μονάδες των παραδειγμάτων εκπαίδευσης που δεν είναι δυνατόν να ανήκουν σε παραπάνω από μία κατηγορίες. Πιο συγκεκριμένα, όταν χρησιμοποιείται το βασικό σύνολο κατηγοριών, το λεξικό αποθηκεύει τις

λεκτικές μονάδες των παραδειγμάτων εκπαίδευσης που ανήκουν (σύμφωνα με τις χειρωνακτικά επισημειωμένες ορθές κατηγορίες των παραδειγμάτων εκπαίδευσης) στις κατηγορίες:

- σημείο στίξης,
- μόριο,
- σύνδεσμος.

Όταν χρησιμοποιείται το εκτεταμένο σύνολο κατηγοριών, το λεξικό αποθηκεύει τις λεκτικές μονάδες των παραδειγμάτων εκπαίδευσης που ανήκουν στις κατηγορίες:

- οριστικό άρθρο στην ονομαστική ενικού, γένους αρσενικού και θηλυκού,
- εμπρόθετο άρθρο στην αιτιατική ενικού και πληθυντικού, γένους θηλυκού,
- εμπρόθετο άρθρο στην αιτιατική πληθυντικού, γένους αρσενικού,
- άκλιτη αντωνυμία,
- συντομογραφία,
- ξένη λέξη,
- σύμβολο,
- άλλο,
- σημείο στίξης,
- μόριο,
- σύνδεσμος.

Όταν θέλουμε να κατατάξουμε μία λεκτική μονάδα, ελέγχουμε αρχικά αν περιλαμβάνεται στο λεξικό. Αν ναι, κατατάσσουμε τη λεκτική μονάδα την κατηγορία που υποδεικνύει το λεξικό, ενώ σε αντίθετη περίπτωση, δίνουμε τη λεκτική μονάδα στον ταξινομητή για να την κατατάξει.

Κατά τη διερεύνηση των λαθών τα οποία έκανε το σύστημα στα δεδομένα αξιολόγησης με το παραπάνω σύνολο ιδιοτήτων, παρατηρήσαμε λάθη στην κατάταξη των αντωνυμιών και των άρθρων. Τις περισσότερες φορές, ισχύει ότι αμέσως μετά από τα άρθρα ακολουθεί είτε επίθετο είτε ουσιαστικό. Επίσης, ισχύει ότι συνήθως οι αντωνυμίες προηγούνται των ρημάτων. Για παράδειγμα, η λέξη «του» μπορεί να είναι είτε αντωνυμία είτε άρθρο. Στην φράση «του μεγάλου μυστικού» είναι άρθρο, ενώ στη φράση «του είπα» είναι αντωνυμία. Με βάση την παραπάνω παρατήρηση, για να πετύχουμε μεγαλύτερη ορθότητα, αποφασίσαμε να κάνουμε τον παρακάτω έλεγχο. Αφού γίνει η κατάταξη των λεκτικών μονάδων ενός κειμένου, εξετάζουμε τις κατηγορίες που έχουν δοθεί από το σύστημά μας στις λεκτικές μονάδες. Αν μια λεκτική μονάδα έχει καταταγεί ως άρθρο, ενώ η επόμενη έχει καταταγεί ως ρήμα, τότε αλλάζουμε την κατηγορία της πρώτης σε αντωνυμία. Αντίθετα, αν μια λεκτική μονάδα έχει καταταγεί ως αντωνυμία, ενώ η επόμενη ως επίθετο ή ουσιαστικό, τότε αλλάζουμε την κατηγορία της πρώτης σε άρθρο.

Κεφάλαιο 3:

Πειραματικά αποτελέσματα

3.1 Σύνολο δεδομένων

Το σύστημα που αναπτύξαμε χρησιμοποιεί τα δεδομένα εκπαίδευσης και αξιολόγησης που χρησιμοποιήθηκαν και στην αμέσως προηγούμενη εργασία (Παππάς, 2008). Πιο συγκεκριμένα, το σύνολο δεδομένων αποτελείται από 31.553 λεκτικές μονάδες που έχουν αντληθεί από ειδησεογραφικά κείμενα των εφημερίδων «ΤΑ ΝΕΑ» και «ΒΗΜΑ». Από αυτές οι 7.878 χρησιμοποιούνται για αξιολόγηση, ενώ οι υπόλοιπες 23.675 εκπαίδευση.

3.2 Πειράματα με το βασικό σύνολο κατηγοριών

Πριν καταλήξουμε στο συγκεκριμένο σύνολο ιδιοτήτων για το βασικό (το μικρό) σύνολο κατηγοριών που περιγράψαμε στο προηγούμενο κεφάλαιο, δοκιμάσαμε πολλές παραλλαγές του.

Στο κεφάλαιο 2, αναφέραμε ότι οι τιμές που παίρνουν οι ιδιότητες αμφισημίας είναι πραγματικοί αριθμοί και δείχνουν τις συχνότητες εμφάνισης (ποσοστά) των λεκτικών μονάδων ή καταλήξεων στα παραδείγματα εκπαίδευσης με κάποια συγκεκριμένη κατηγορία. Εντούτοις, στην πρώτη έκδοση του συστήματος οι ιδιότητες αμφισημίας είχαν τιμές Boolean, οι οποίες δήλωναν αν η λεκτική μονάδα ή κατάληξη εμφανίζεται στα παραδείγματα εκπαίδευσης με τη συγκεκριμένη κατηγορία ή όχι. Για παράδειγμα, οι ιδιότητες αμφισημίας για τη λέξη «του», που την έχουμε συναντήσει στα παραδείγματα εκπαίδευσης και ως άρθρο και ως αντωνυμία, θα ήταν οι παρακάτω:

> isVerb=0, isAdverb=0, isArticle=1, isAdjective=0, isNoun=0, isConjunction=0, isNumber=0, isPunctuation=0, isParticle=0, isPronoun=1, isOther=0, isPreposition=0

Στην πρώτη έκδοση, χρησιμοποιούσαμε επίσης ιδιότητες αμφισημίας όχι μόνο για την τρέχουσα και την επόμενη λεκτική μονάδα (και τις καταλήξεις τους), αλλά και για την προηγούμενη και την προ-προηγούμενη λεκτική μονάδα (και τις καταλήξεις τους). Επιπλέον, το σύστημα χρησιμοποιούσε λεξικό (ενότητα 2.4), καθώς και τις παρακάτω ιδιότητες:

1. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων)
2. Ύπαρξη αποστρόφου στη λεκτική μονάδα
3. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα
4. Ύπαρξη τελείας στη λεκτική μονάδα
5. Ύπαρξη κόμματος στη λεκτική μονάδα
6. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα

Το ποσοστό ορθότητας (accuracy, στα παραδείγματα αξιολόγησης) που επιτυγχανόταν με αυτό το σύστημα ήταν 89,44%. Δοκιμάσαμε, επίσης, να μη χρησιμοποιήσουμε την ιδιότητα του μήκους της

λεκτικής μονάδας ή να κανονικοποιήσουμε τις τιμές της ιδιότητας αυτής στο διάστημα [0, 1], αλλά τα αποτελέσματα ήταν χειρότερα (88,58% και 88,60% αντίστοιχα, πάλι στα δεδομένα αξιολόγησης).

Στη συνέχεια, αποφασίσαμε να αλλάξουμε τις ιδιότητες αμφισημίας, ώστε αντί για τιμές Boolean να έχουν τιμές που να δείχνουν τη συχνότητα εμφάνισης (στα παραδείγματα εκπαίδευσης) μιας λεκτικής μονάδας ή κατάληξης με μια συγκεκριμένη κατηγορία. Με αυτή την αλλαγή, η ορθότητα βελτιώθηκε (έφτασε στο 89,87**%** στα δεδομένα αξιολόγησης), κάτι που ήταν αναμενόμενο, αφού πλέον οι ιδιότητες αμφισημίας παρέχουν περισσότερη πληροφορία· η βελτίωση, όμως, ήταν μικρή. Δοκιμάσαμε και πάλι να μη χρησιμοποιούμε την ιδιότητα του μήκους ή να κανονικοποιούμε τις τιμές της στο διάστημα [0,1], αλλά και στις δύο περιπτώσεις η ορθότητα μειώθηκε (σε 88,19% και 88,22% αντίστοιχα στα δεδομένα αξιολόγησης).

Στη συνέχεια, προκειμένου να μη χρησιμοποιείται το σύνολο των δεδομένων αξιολόγησης στα πειράματα επιλογής ιδιοτήτων, αρχίσαμε να χρησιμοποιούμε (κατά την επιλογή ιδιοτήτων) δεκαπλή διασταυρωμένη επικύρωση (10-fold cross validation) αποκλειστικά στα δεδομένα εκπαίδευσης. Το σύνολο των παραδειγμάτων εκπαίδευσης, δηλαδή, χωρίστηκε σε δέκα (σχεδόν) ίσα τμήματα και γίνονταν δέκα επαναλήψεις κάθε πειράματος επιλογής ιδιοτήτων. Σε κάθε επανάληψη, χρησιμοποιούσαμε ένα διαφορετικό τμήμα των δεδομένων εκπαίδευσης για την αξιολόγηση του συστήματος και τα υπόλοιπα εννέα για την εκπαίδευσή του. Τα ποσοστά ορθότητας που αναφέρουμε στις περιπτώσεις δεκαπλής διασταυρωμένης επικύρωσης είναι οι μέσοι όροι των δέκα επαναλήψεων. Η καλύτερη από τις παραπάνω μορφές του συστήματός μας επιτύγχανε ορθότητα 89,87% στα δεδομένα εκπαίδευσης με δεκαπλή διασταυρωμένη επικύρωση.

Έπειτα, για να βελτιώσουμε τα αποτελέσματα, χρησιμοποιήσαμε λεξικό καθώς και τις παρακάτω ιδιότητες, όπου κάθε ιδιότητα αμφισημίας είχε ως τιμή την αντίστοιχη συχνότητα (ποσοστό):

1. Ιδιότητες αμφισημίας της λεκτικής μονάδας
2. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας
3. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας
4. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας
5. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων), όχι κανονικοποιημένο
6. Ύπαρξη αποστρόφου στη λεκτική μονάδα
7. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα
8. Ύπαρξη τελείας στη λεκτική μονάδα
9. Ύπαρξη κόμματος στη λεκτική μονάδα
10. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα
11. Ιδιότητες αμφισημίας της επόμενης λεκτικής μονάδας
12. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της επόμενης λεκτικής μονάδας
13. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της επόμενης λεκτικής μονάδας
14. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της επόμενης λεκτικής μονάδας
15. Ιδιότητες αμφισημίας της προηγούμενης λεκτικής μονάδας
16. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προηγούμενης λεκτικής μονάδας
17. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προηγούμενης λεκτικής μονάδας
18. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προηγούμενης λεκτικής μονάδας
19. Ιδιότητες αμφισημίας της προ-προηγούμενης λεκτικής μονάδας

20. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προ-προηγούμενης λεκτικής μονάδας
21. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας
22. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας

Το σύστημα αυτό πέτυχε ορθότητα 89,46% (μετρώντας στο σύνολο των δεδομένων αξιολόγησης), που είναι λίγο χαμηλότερο από το ποσοστό ορθότητας του προηγούμενου συστήματος. Διευρύναμε, κατόπιν, τη γειτονιά της υπό κατάταξη λεκτικής μονάδας (3 λεκτικές μονάδες πριν και 3 μετά), ενώ αφήσαμε όλες τις υπόλοιπες παραμέτρους όπως είχαν (ενεργοποιημένο λεξικό). Οι ιδιότητες του νέου συστήματος ήταν:

1. Ιδιότητες αμφισημίας της τρέχουσας λεκτικής μονάδας
2. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας
3. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας
4. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας
5. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων), όχι κανονικοποιημένο
6. Ύπαρξη αποστρόφου στη λεκτική μονάδα
7. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα
8. Ύπαρξη τελείας στη λεκτική μονάδα
9. Ύπαρξη κόμματος στη λεκτική μονάδα
10. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα
11. Ιδιότητες αμφισημίας της επόμενης λεκτικής μονάδας
12. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της επόμενης λεκτικής μονάδας
13. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της επόμενης λεκτικής μονάδας
14. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της επόμενης λεκτικής μονάδας
15. Ιδιότητες αμφισημίας της μεθεπόμενης λεκτικής μονάδας
16. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της μεθεπόμενης λεκτικής μονάδας
17. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της μεθεπόμενης λεκτικής μονάδας
18. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της μεθεπόμενης λεκτικής μονάδας
19. Ιδιότητες αμφισημίας της λεκτικής μονάδας μετά τη μεθεπόμενη
20. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας μετά τη μεθεπόμενη
21. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας μετά τη μεθεπόμενη
22. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας μετά τη μεθεπόμενη
23. Ιδιότητες αμφισημίας της προηγούμενης λεκτικής μονάδας
24. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προηγούμενης λεκτικής μονάδας
25. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προηγούμενης λεκτικής μονάδας
26. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προηγούμενης λεκτικής μονάδας
27. Ιδιότητες αμφισημίας της προ-προηγούμενης λεκτικής μονάδας
28. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προ-προηγούμενης λεκτικής μονάδας

29. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας
30. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας
31. Ιδιότητες αμφισημίας της λεκτικής μονάδας πριν την προ-προηγούμενη
32. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας πριν την προ-προηγούμενη
33. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας πριν την προ-προηγούμενη
34. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας πριν την προ-προηγούμενη

Οι επιπλέον ιδιότητες πιστεύαμε ότι θα βοηθούσαν τον ταξινομητή να επιτύχει καλύτερες επιδόσεις. Όμως, το σύστημα που προέκυψε είχε χαμηλότερο ποσοστό ορθότητας (89,14%) σε σχέση με την ακριβώς προηγούμενη έκδοση (γειτονιά με δυο λεκτικές μονάδες πριν και μία μετά).

Έπειτα, δοκιμάσαμε όλους τους πιθανούς συνδυασμούς από συμμετρικές γειτονιές και τα αποτελέσματα που πήραμε φαίνονται στον παρακάτω πίνακα.

Γειτονιά	**Ορθότητα (Cross Validation)**
1 λέξη πριν και 1 λέξη μετά	90,24%
2 λέξεις πριν και 2 λέξεις μετά	89,39%
3 λέξεις πριν και 3 λέξεις μετά	89,14%

Από τα αποτελέσματα του παραπάνω πίνακα φαίνεται ότι όσο πιο μικρή είναι η γειτονιά τόσο μεγαλύτερη ορθότητα πετυχαίνουμε. Έτσι, αποφασίσαμε να δοκιμάσουμε σε μία νέα έκδοση του συστήματος να κρατήσουμε ως γειτονιά μόνο την προηγούμενη λεκτική μονάδα και σε μία άλλη έκδοση να κρατήσουμε ως γειτονιά μόνο την επόμενη λεκτική μονάδα. Τα αποτελέσματα αυτών των συστημάτων φαίνονται στον παρακάτω πίνακα:

Γειτονιά	**Ορθότητα (Cross Validation)**
Προηγούμενη λέξη	87,64%
Επόμενη λέξη	90,56%

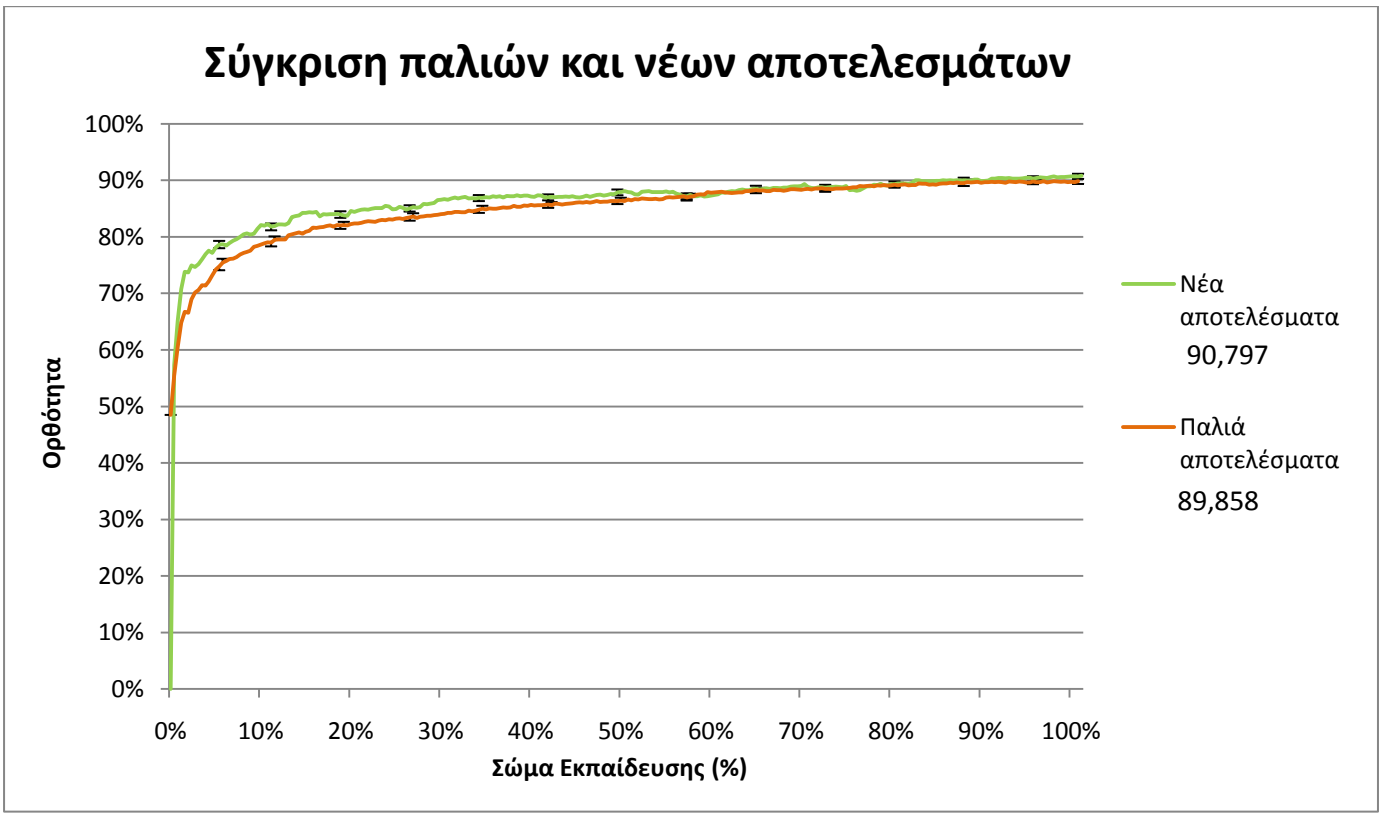
Το σύστημα στο οποίο εξετάζουμε μόνο την τρέχουσα λέξη και την επόμενή της επιτυγχάνει ορθότητα στα δεδομένα αξιολόγησης 90,75% και ξεπερνά το σύστημα της προηγούμενης εργασίας (Παππάς, 2008) που είχε ορθότητα 89,86% με τα ίδια δεδομένα εκπαίδευσης και αξιολόγησης.

Άρα, καταλήξαμε να χρησιμοποιούμε λεξικό και τις παρακάτω ιδιότητες:

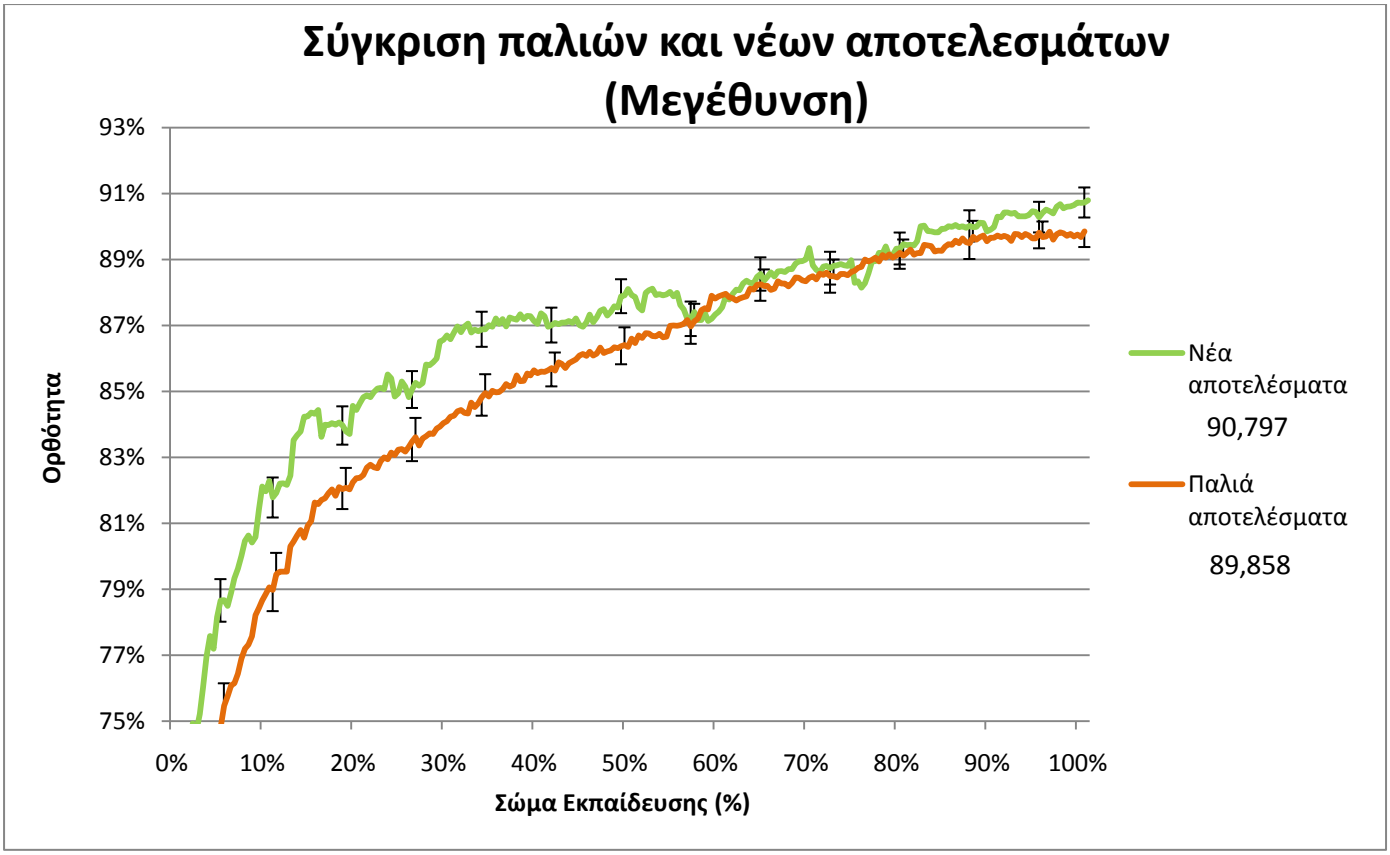
1. Ιδιότητες αμφισημίας της τρέχουσας λεκτικής μονάδας
2. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας
3. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας
4. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας
5. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων), όχι κανονικοποιημένο

6. Ύπαρξη αποστρόφου στη λεκτική μονάδα
7. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα
8. Ύπαρξη τελείας στη λεκτική μονάδα
9. Ύπαρξη κόμματος στη λεκτική μονάδα
10. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα
11. Ιδιότητες αμφισημίας της επόμενης λεκτικής μονάδας
12. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της επόμενης λεκτικής μονάδας
13. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της επόμενης λεκτικής μονάδας
14. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της επόμενης λεκτικής μονάδας

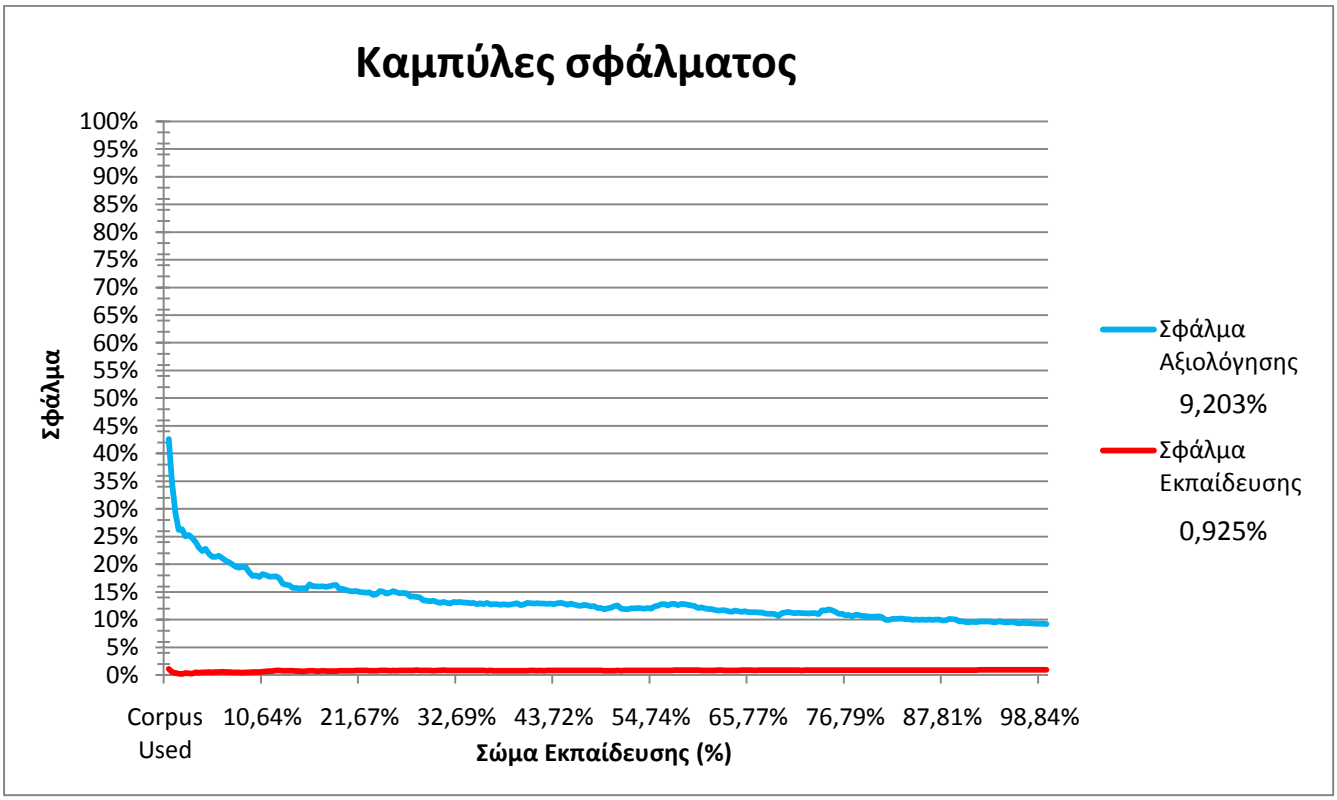
Στα σχήματα 3.2.1 και 3.2.2 εμφανίζονται οι καμπύλες μάθησης (με ποσοστό ορθότητας, accuracy, στον κατακόρυφο άξονα) του συστήματος της παρούσας εργασίας (με το τελικό σύνολο ιδιοτήτων) και εκείνου της προηγούμενης εργασίας (Παππάς, 2008), με διαστήματα εμπιστοσύνης 95%. Στην περίπτωση αυτή, τα συστήματα εκπαιδεύονται σε αυξανόμενου μεγέθους μέρος των δεδομένων εκπαίδευσης και αξιολογούνται στα (ξεχωριστά) δεδομένα αξιολόγησης. Επιπλέον τα σχήματα 3.2.3 και 3.2.4 δείχνουν τις καμπύλες σφάλματος (error rate) του νέου συστήματος στα δεδομένα εκπαίδευσης και αξιολόγησης. Όπως ήταν αναμενόμενο, το σφάλμα στα δεδομένα εκπαίδευσης (τα οποία έχει συναντήσει το σύστημα κατά την εκπαίδευσή του) είναι χαμηλότερο από ό,τι στα δεδομένα αξιολόγησης (νέες περιπτώσεις, που το σύστημα δεν είχε συναντήσει κατά την εκπαίδευσή του). Το σφάλμα στα δεδομένα εκπαίδευσης μπορεί να θεωρηθεί ένα κάτω φράγμα του σφάλματος στα δεδομένα αξιολόγησης.



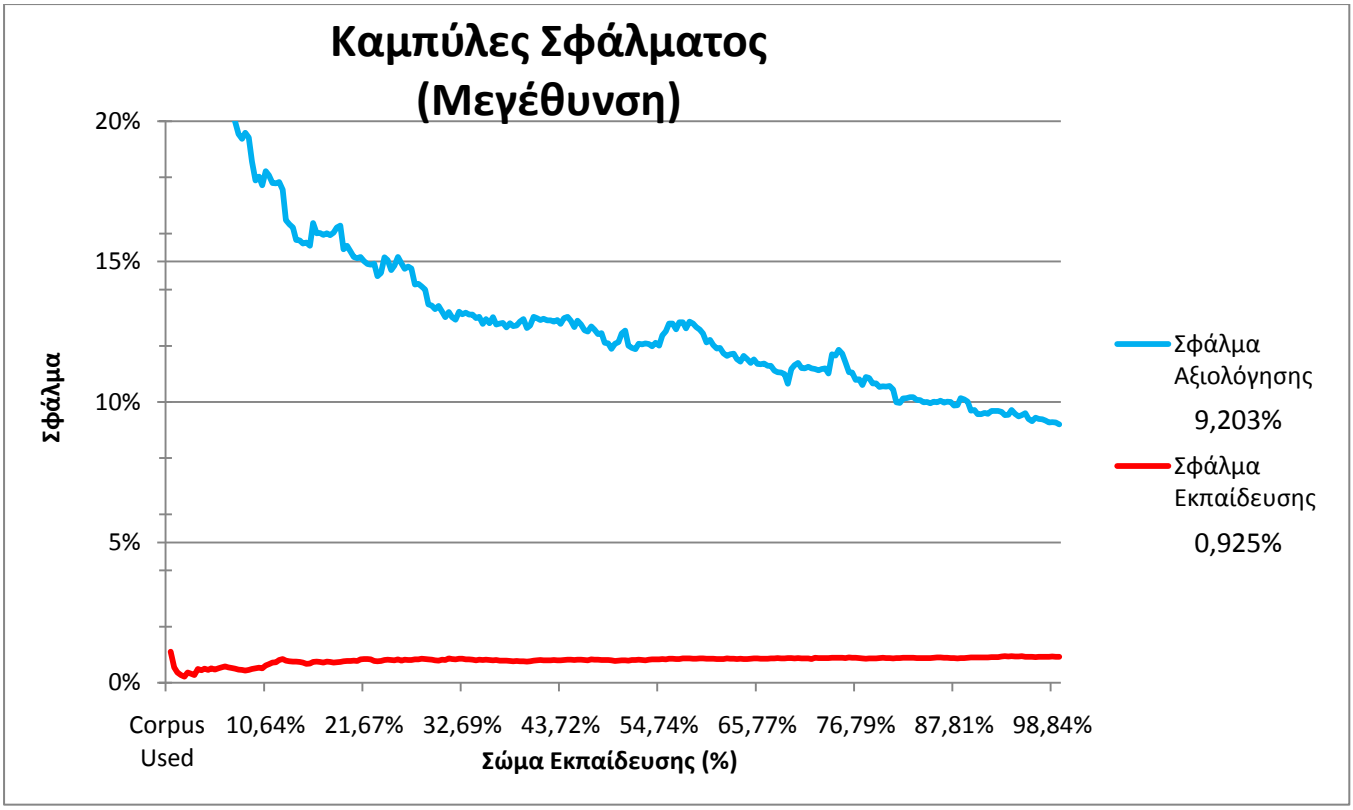
Σχήμα 3.2.1 Συγκριτικές καμπύλες μάθησης παλαιού και νέου συστήματος, με το βασικό σύνολο κατηγοριών.



Σχήμα 3.2.2 Μεγέθυνση καμπυλών μάθησης παλαιού και νέου συστήματος, με το βασικό σύνολο κατηγοριών.



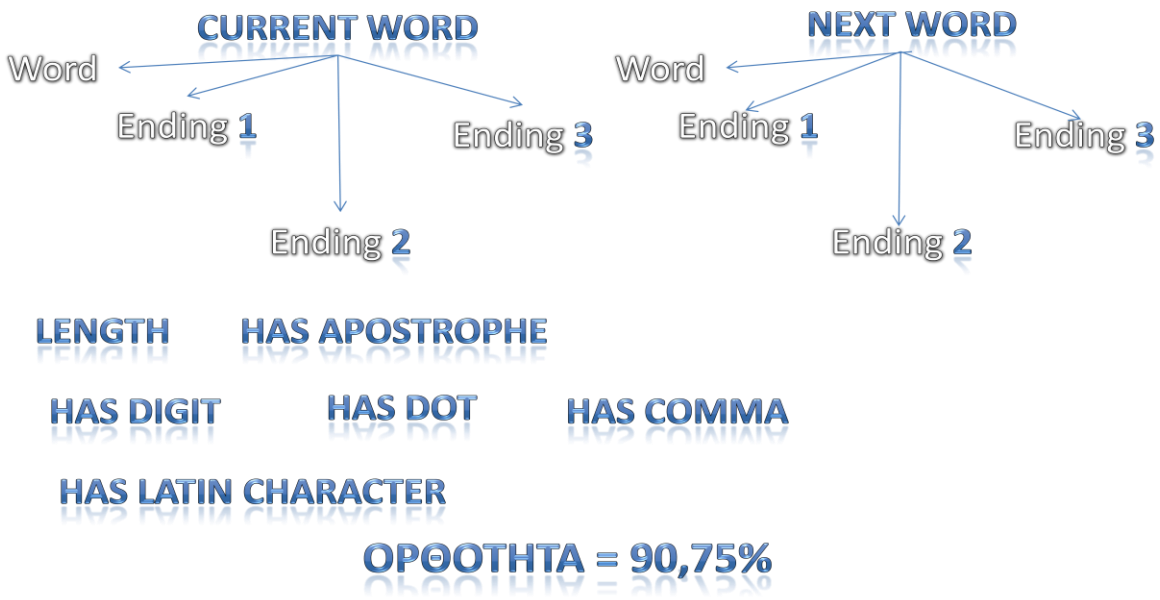
Σχήμα 3.2.3 Καμπύλες σφάλματος του νέου συστήματος, με το βασικό σύνολο κατηγοριών.



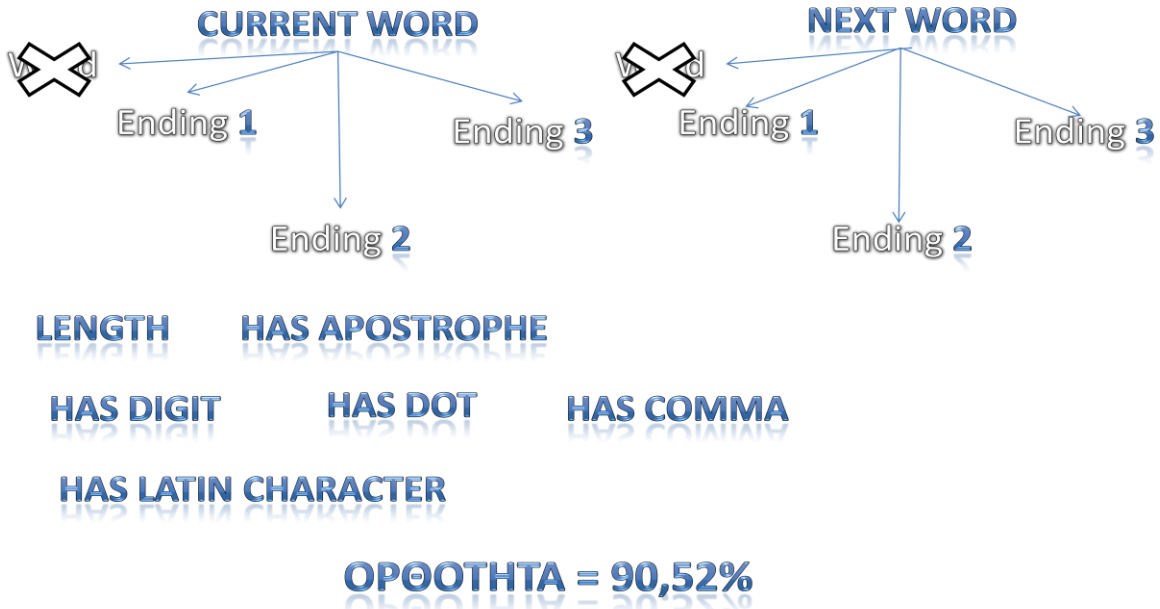
Σχήμα 3.2.4 Μεγέθυνση καμπυλών σφάλματος του νέου συστήματος, με το μικρό σύνολο κατηγοριών.

Από τις καμπύλες των σχημάτων 3.2.1 και 3.2.2 φαίνεται ότι πετύχαμε μεγαλύτερη ορθότητα από εκείνη της προηγούμενης εργασίας (Παππάς, 2008). Επιπλέον, το σφάλμα στα παραδείγματα αξιολόγησης φαίνεται ότι μειώνεται σταδιακά όσο προσθέτουμε παραδείγματα εκπαίδευσης και υπάρχει αρκετό περιθώριο μέχρι την καμπύλη (κάτω φράγμα) του σφάλματος στα δεδομένα εκπαίδευσης. Επιπλέον το σύστημα φαίνεται να τα πάει πολύ καλά στα παραδείγματα εκπαίδευσης. Άρα, φαίνεται να έχουμε πρόβλημα high-variance (χοντρικά, πρόβλημα υπερ-εφαρμογής). Μία πιθανή βελτίωση, επομένως, είναι η προσθήκη περισσότερων παραδειγμάτων εκπαίδευσης. Μία άλλη πιθανή βελτίωση είναι η ακόμα καλύτερη επιλογή ιδιοτήτων ή η εξαγωγή πιο σύνθετων ιδιοτήτων (π.χ. με Principal Components Analysis).

Οι ιδιότητες του συστήματος φαίνονται στο παρακάτω σχήμα. Τα ποσοστά ορθότητας τώρα μετριούνται στο σύνολο των δεδομένων αξιολόγησης.

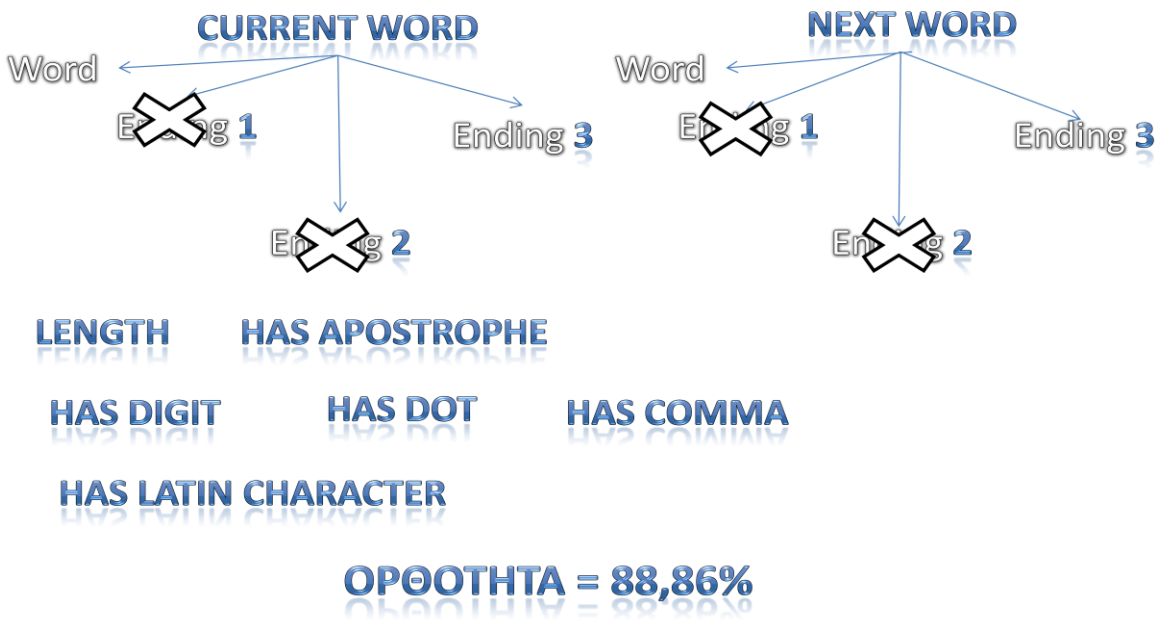


Αποφασίσαμε να κάνουμε κάποια πειράματα για να δούμε ποιες από αυτές παρέχουν τελικά την περισσότερη πληροφορία. Αρχικά, κάναμε πειράματα χωρίς να κρατάμε πληροφορία για ολόκληρες τις λεκτικές μονάδες, αλλά μόνο για τις 3 καταλήξεις, όπως φαίνεται στο παρακάτω σχήμα:



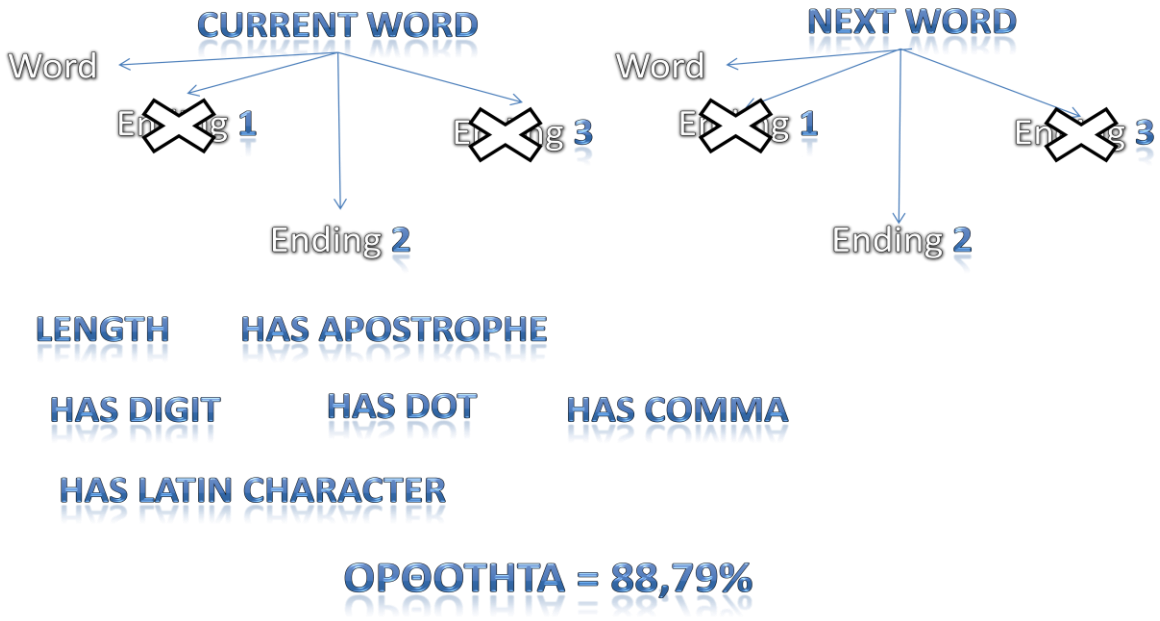
Παρατηρούμε ότι η ορθότητα μειώθηκε πολύ λίγο, κατά 0,23%, άρα οι ιδιότητες αμφισημίας για ολόκληρες τις λέξεις δεν μας δίνουν πολλή πληροφορία.

Έπειτα, κάναμε πειράματα κρατώντας μόνο τις ιδιότητες αμφισημίας ολόκληρων των λέξεων και των καταλήξεων τριών χαρακτήρων, όπως φαίνεται στο παρακάτω σχήμα:



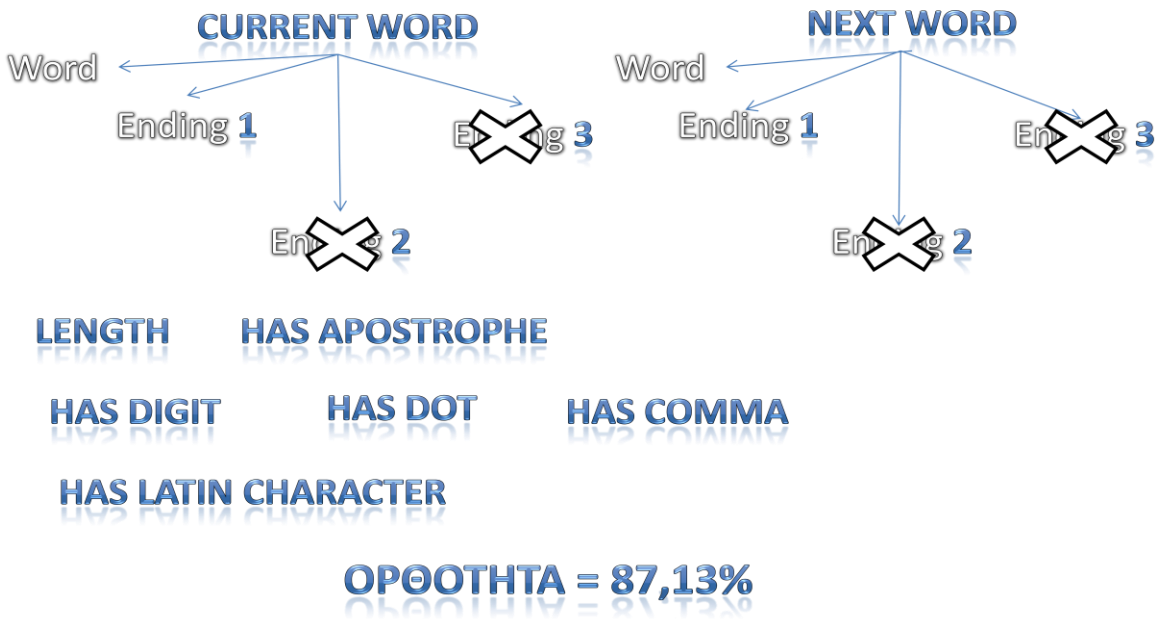
Βλέπουμε ότι η ορθότητα έχει μειωθεί κατά 1,89%, άρα οι καταλήξεις ενός και δύο χαρακτήρων μας προσφέρουν πληροφορία.

Στη συνέχεια, κάναμε πειράματα κρατώντας μόνο τις ιδιότητες αμφισημίας ολόκληρων των λέξεων και των καταλήξεων δύο χαρακτήρων, όπως φαίνεται στο παρακάτω σχήμα:



Βλέπουμε ότι η ορθότητα έχει μειωθεί κατά 1,96%, άρα οι καταλήξεις ενός και τριών χαρακτήρων μας προσφέρουν πληροφορία.

Τέλος, κάναμε πειράματα κρατώντας μόνο τις ιδιότητες αμφισημίας ολόκληρων των λέξεων και των καταλήξεων ενός χαρακτήρα, όπως φαίνεται στο παρακάτω σχήμα:



ΟΡΘΟΤΗΤΑ = 87,13%

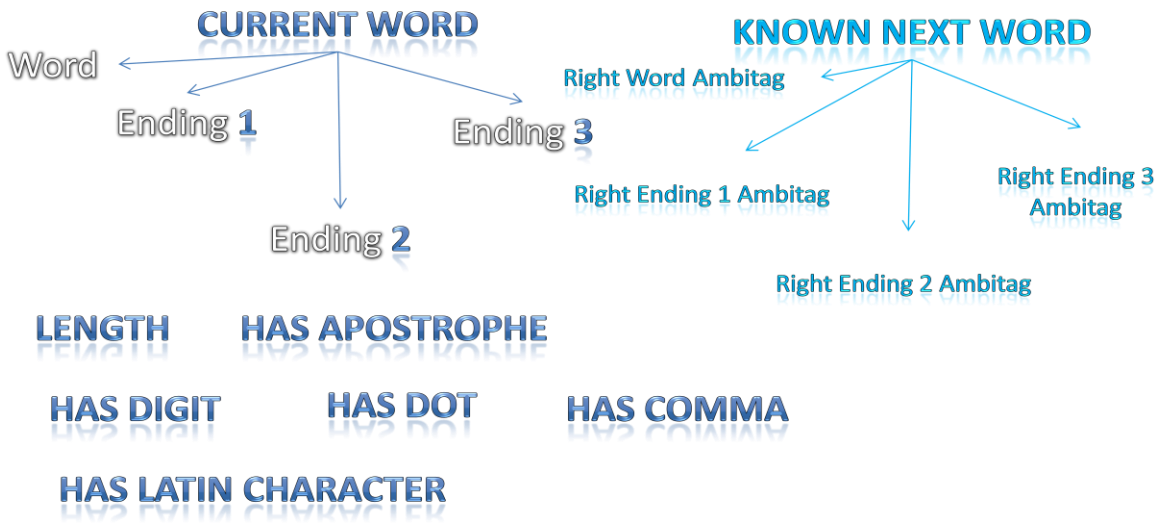
Βλέπουμε ότι η ορθότητα έχει μειωθεί κατά 3,62%, άρα οι καταλήξεις δυο και τριών χαρακτήρων μας προσφέρουν αρκετή πληροφορία.

Από όλα τα παραπάνω πειράματα, καταλήξαμε στο συμπέρασμα ότι τις πιο χρήσιμες πληροφορίες τις προσφέρουν οι ιδιότητες αμφισημίας που αφορούν τις τρεις καταλήξεις κάθε λέξης και της επόμενής της.

Στη συνέχεια, σκεφτήκαμε ότι θα ήταν καλό να καταλάβουμε πόσο σημαντική είναι η γνώση της επόμενης λέξης. Κάναμε, λοιπόν, πειράματα θεωρώντας ότι είναι γνωστή η (σωστή) κατηγορία στην οποία ανήκει η επόμενη λέξη. Στη θέση των ιδιοτήτων αμφισημίας, έχουμε πια την ιδιότητα της σωστής κατηγορίας με τιμή 1 και τις υπόλοιπες ιδιότητες με τιμή 0. Για παράδειγμα, όταν η επόμενη λέξη ανήκει στην κατηγορία «ουσιαστικό», οι ιδιότητες αμφισημίας για ολόκληρη την επόμενη λέξη και για τις τρεις καταλήξεις της είναι:

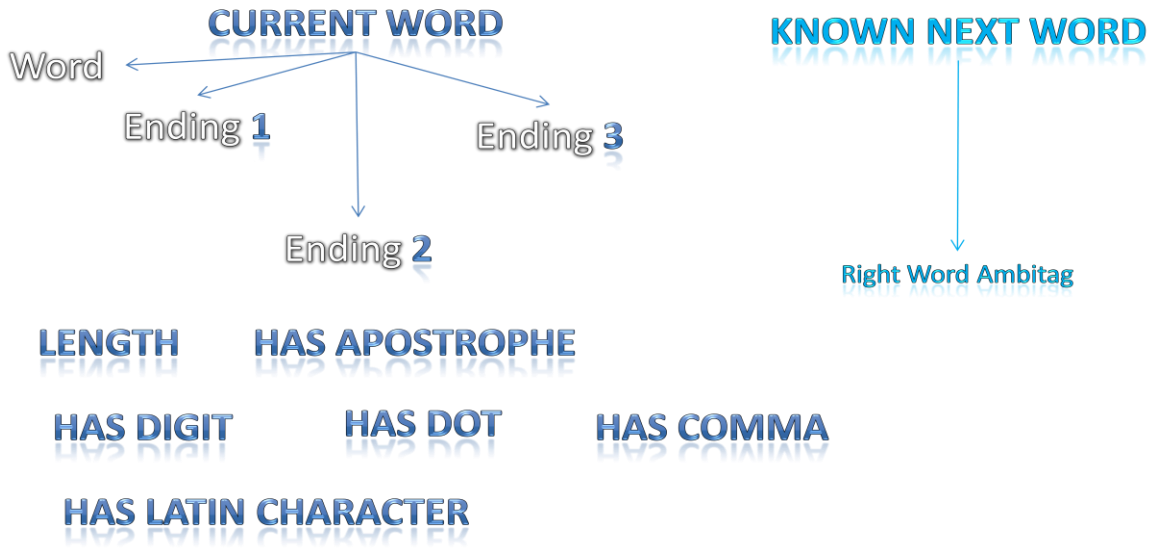
> isVerb=0, isAdverb=0, isArticle=0, isAdjective=0, isNoun=1, isConjunction=0, isNumber=0, isPunctuation=0, isParticle=0, isPronoun=0, isOther=0, isPreposition=0

Αρχικά, απλά αντικαταστήσαμε τις ιδιότητες αμφισημίας για ολόκληρη την επόμενη λέξη και τις τρεις καταλήξεις της με τις σωστές τιμές, όπως φαίνεται στο παρακάτω σχήμα:



ΟΡΘΟΤΗΤΑ = 90,91%

Η ορθότητα αυξήθηκε κατά πολύ μικρό ποσοστό (0,16%). Εν τούτοις παρατηρήσαμε ότι σ' αυτή την περίπτωση έχουμε ουσιαστικά επανάληψη πληροφορίας, κάτι που δεν είναι επιθυμητό. Έτσι, καταργήσαμε τις ιδιότητες που αφορούν τις καταλήξεις και κρατήσαμε μόνο εκείνες που αφορούν ολόκληρη την επόμενη γνωστή λέξη.



ΟΡΘΟΤΗΤΑ = 91,07 %

Μετά από τις παραπάνω αλλαγές, παρατηρείται αύξηση της ορθότητα κατά 0,32%. Αυτό μας δείχνει ότι το να είμαστε σίγουροι για την κατηγορία της επόμενης λέξης μάς προσφέρει κάποια πληροφορία, χωρίς όμως να αλλάζει δραματικά τις επιδόσεις του συστήματος.

Μέχρι εδώ, λοιπόν, έχουμε καταλήξει σε ένα σύστημα που εξετάζει πληροφορίες μόνο για την τρέχουσα λεκτική μονάδα και την επόμενή της, καθώς επίσης και ένα λεξικό με λέξεις που μπορούν

να ανήκουν μόνο σε μία κατηγορία (ενότητα 2.4). Για να βελτιώσουμε ακόμα περισσότερο τα αποτελέσματά μας, αποφασίσαμε να δούμε τα παραδείγματα αξιολόγησης εκείνα, τα οποία το σύστημα κατατάσσει λανθασμένα. Εκεί, παρατηρήσαμε ότι πολλά παραδείγματα εκπαίδευσης που ανήκαν στην κατηγορία «άρθρο» είχαν καταταγεί στην κατηγορία «αντωνυμία» και το αντίθετο. Αποφασίσαμε να εκμεταλλευτούμε το γεγονός ότι συνήθως το άρθρο ακολουθείται από επίθετα ή ουσιαστικά, ενώ οι αντωνυμίες ακολουθούνται από ρήματα. Έτσι, προσθέσαμε τους κανόνες διόρθωσης λαθών της ενότητας 2.4. Επίσης, κάναμε τις παρακάτω παραδοχές που αφορούν την κατηγορία «αριθμητικό»:

- Οι λέξεις που αρχίζουν από αριθμό, ανήκουν πάντα στην κατηγορία «αριθμητικό».

  Για παράδειγμα, οι λέξεις «$3^{η}$», «$2^{α}$», «$12^{ος}$» ανήκουν στην κατηγορία «αριθμητικό»

- Όλα τα αόριστα άρθρα επίσης κατατάσσονται στην κατηγορία «αριθμητικό».
  Για παράδειγμα, οι λέξεις «ένας», «μία», «μιας», «έναν» και άλλες, ανήκουν πλέον στην κατηγορία αριθμητικό.

Με τις παραπάνω αλλαγές, πετύχαμε ορθότητα 90,82% στο σύνολο αξιολόγησης.

Επίσης, διερευνώντας τα λάθη του συστήματος, παρατηρήσαμε ότι μεγάλο ποσοστό λέξεων που έχουν καταταγεί λανθασμένα βρίσκονταν στην κατηγορία «άλλο». Επειδή η κατηγορία αυτή περιέχει ξένες λέξεις, σύμβολα και ακρωνύμια, είναι δύσκολο να αναγνωριστούν από τον ταξινομητή και κάναμε την παραδοχή ότι δεν υπολογίζονται στην συνολική ορθότητα, ώστε να αξιολογήσουμε το σύστημα πιο δίκαια. Αγνοήσαμε, δηλαδή, κατά τον υπολογισμό της ορθότητας τις λέξεις που κατατάξαμε σωστά στην κατηγορία «άλλο» καθώς και αυτές που πραγματικά ανήκουν στην κατηγορία αυτή. Τελικά, η ορθότητα που πετυχαίνουμε (στο σύνολο αξιολόγησης) μετά από την παραπάνω παραδοχή φτάνει το **92%**.

3.3 Πειράματα με το εκτεταμένο σύνολο κατηγοριών

Αντίστοιχα πειράματα έγιναν με το εκτεταμένο σύνολο κατηγοριών, με το οποίο φτάσαμε τελικά να έχουμε ορθότητα **81,56%** στο σύνολο αξιολόγησης. Αντί για 12, χρησιμοποιούμε τώρα 170 κατηγορίες που μας δίνουν πολύ περισσότερες πληροφορίες για κάθε λεκτική μονάδα. Οι πολλές κατηγορίες, όμως, οδήγησαν εύλογα σε χαμηλότερο ποσοστό ορθότητας, συγκρινόμενο με εκείνο του βασικού συνόλου κατηγοριών· οδήγησαν, επίσης, σε αύξηση του χρόνου εκτέλεσης και της μνήμης που χρειάζεται για να εκπαιδευτεί το σύστημα.

Ξεκινήσαμε από ένα σύστημα με λεξικό (ενότητα 2.4) και τις παρακάτω ιδιότητες:

1. Ιδιότητες αμφισημίας της λεκτικής μονάδας
2. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της λεκτικής μονάδας
3. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της λεκτικής μονάδας
4. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της λεκτικής μονάδας
5. Μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων)
6. Ύπαρξη αποστρόφου στη λεκτική μονάδα
7. Ύπαρξη αριθμητικού χαρακτήρα στη λεκτική μονάδα

8. Ύπαρξη τελείας στη λεκτική μονάδα
9. Ύπαρξη κόμματος στη λεκτική μονάδα
10. Ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα
11. Ιδιότητες αμφισημίας της επόμενης λεκτικής μονάδας
12. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της επόμενης λεκτικής μονάδας
13. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της επόμενης λεκτικής μονάδας
14. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της επόμενης λεκτικής μονάδας
15. Ιδιότητες αμφισημίας της προηγούμενης λεκτικής μονάδας
16. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προηγούμενης λεκτικής μονάδας
17. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προηγούμενης λεκτικής μονάδας
18. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προηγούμενης λεκτικής μονάδας
19. Ιδιότητες αμφισημίας της προ-προηγούμενης λεκτικής μονάδας
20. Ιδιότητες αμφισημίας της κατάληξης ενός χαρακτήρα της προ-προηγούμενης λεκτικής μονάδας
21. Ιδιότητες αμφισημίας της κατάληξης δύο χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας
22. Ιδιότητες αμφισημίας της κατάληξης τριών χαρακτήρων της προ-προηγούμενης λεκτικής μονάδας

Με αυτό το σύστημα, που χρησιμοποιούσε ιδιότητες αμφισημίας για 2 λέξεις πριν την τρέχουσα και μια λέξη μετά, πετύχαμε ορθότητα 81,53%. Ένα σημαντικό πρόβλημα, όμως, ήταν ότι ο χρόνος που χρειαζόταν για την εκπαίδευση και την αξιολόγηση του συστήματος ήταν πολύ μεγάλος, με αποτέλεσμα να μην μπορούμε να δημιουργήσουμε καμπύλες μάθησης. Οι ιδιότητες αμφισημίας είναι τόσες όσες και οι πιθανές κατηγορίες, 170 συνολικά. Έτσι, χρησιμοποιώντας ιδιότητες αμφισημίας για την τρέχουσα, την επόμενη, την προηγούμενη και την προ-προηγούμενη λεκτική μονάδα, καθώς και τις τρεις πιθανές καταλήξεις τους, χρειαζόμαστε (μαζί με τις υπόλοιπες) 2.720 ιδιότητες, που συνδυαζόμενες με τον αριθμό των παραδειγμάτων εκπαίδευσης (23.675) οδηγούν εύλογα σε μικρή ταχύτητα εκτέλεσης και μεγάλες απαιτήσεις μνήμης. Πιο συγκεκριμένα, οι απαιτούμενοι χρόνοι ήταν:

- Εκπαίδευση (σε 23.675 παραδείγματα εκπαίδευσης): περίπου 21.345 δευτερόλεπτα.
- Ταξινόμηση (7.878 παραδείγματα αξιολόγησης): 560,32 δευτερόλεπτα.
- Αξιολόγηση (υπολογισμός της ορθότητας): 0,36 δευτερόλεπτα.

Προκειμένου να μειώσουμε το χρόνο εκπαίδευσης και ταξινόμησης, αποφασίσαμε να κάνουμε επιλογή ιδιοτήτων. Χρησιμοποιήσαμε το Weka[2] για να υπολογίσουμε το λόγο κέρδους πληροφορίας (Information Gain Ratio) της κάθε ιδιότητας και κρατήσαμε μόνο εκείνες (1.326 από τις 2.720) που είχαν κέρδος πληροφορίας μεγαλύτερο του μηδενός. Παρ' όλα αυτά, ο χρόνος δεν βελτιώθηκε αρκετά, οπότε η ιδέα εγκαταλείφθηκε.

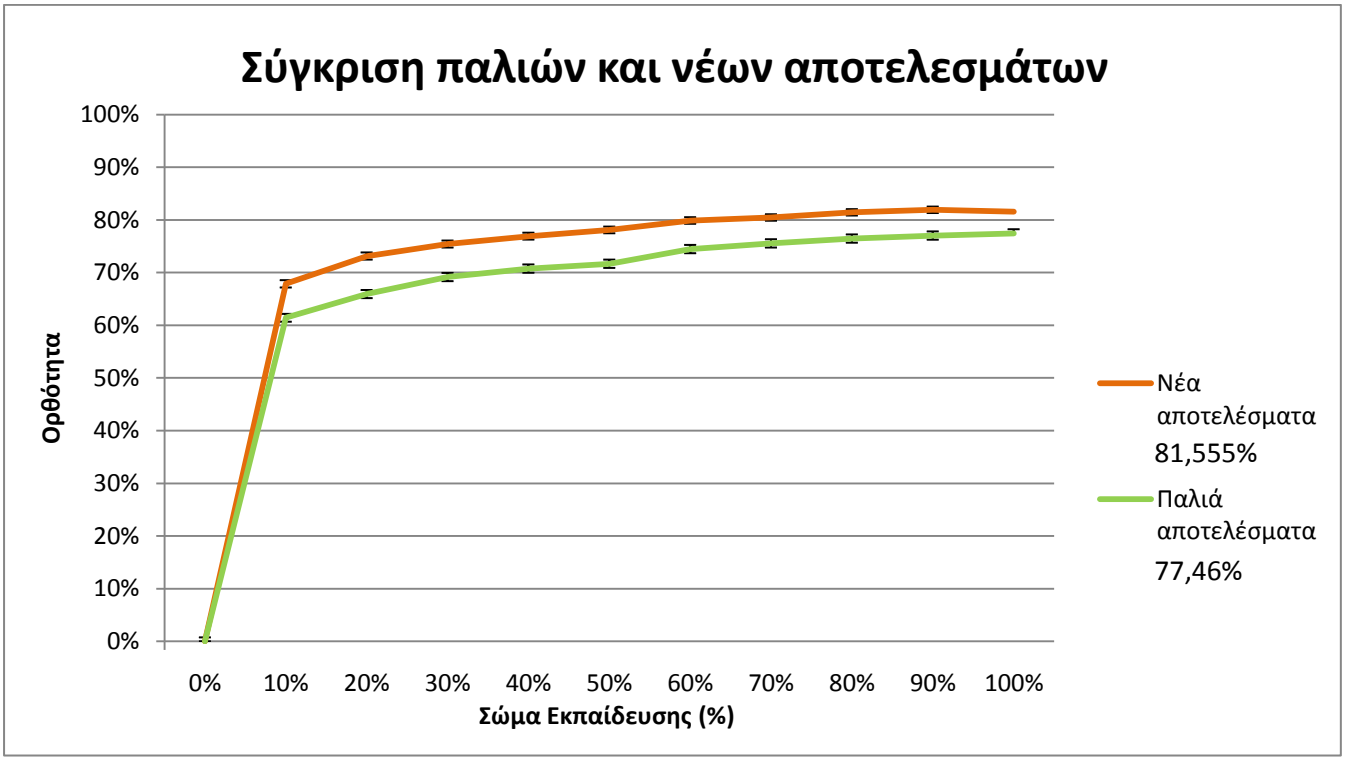
Οι αλλαγές που κάναμε στο σύστημα που χρησιμοποιεί το βασικό σύνολο κατηγοριών κατά τη διάρκεια των πειραμάτων και έδωσαν καλύτερα αποτελέσματα, έγιναν και στο σύστημα που

[2] Βλ. http://www.cs.waikato.ac.nz/ml/weka/.

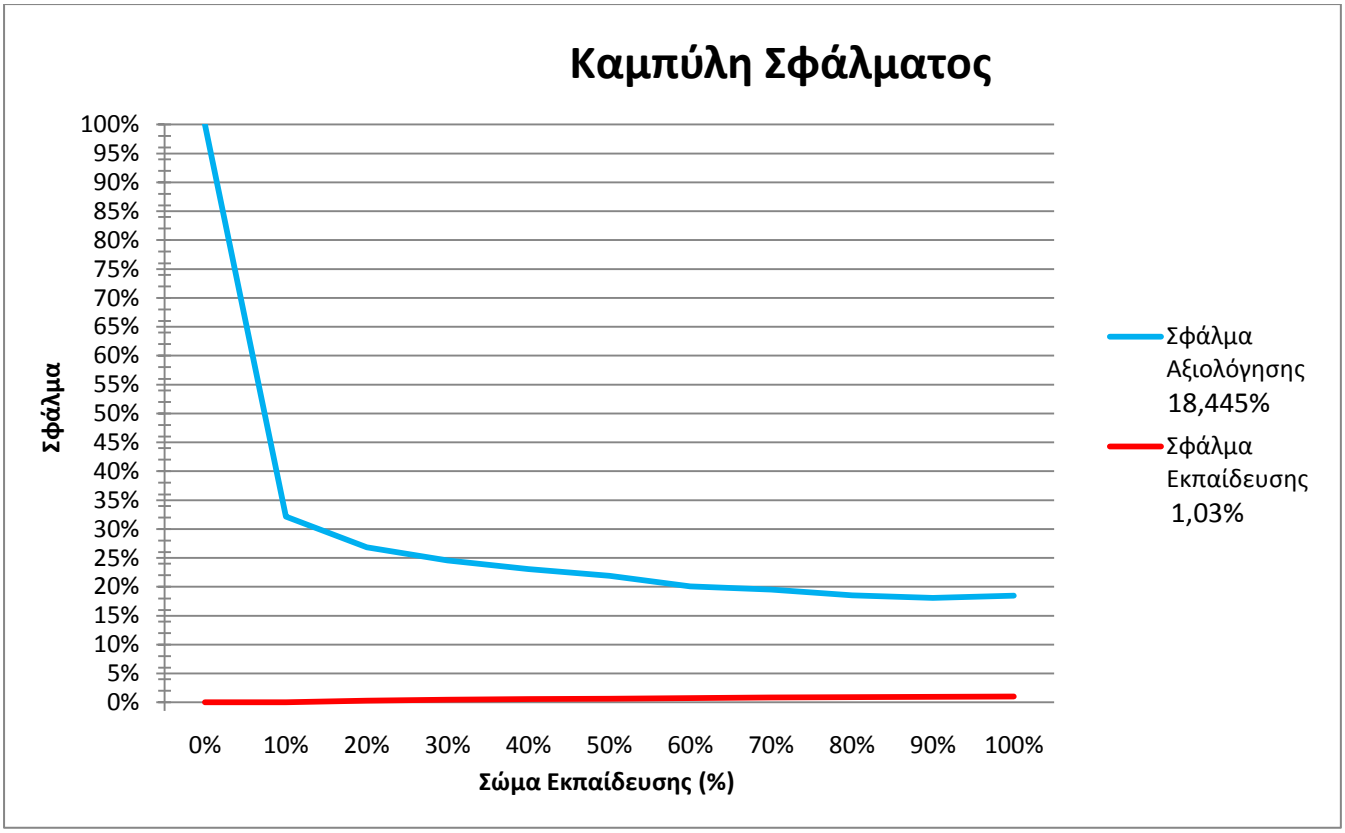
χρησιμοποιεί το εκτεταμένο σύνολο. Έτσι, χρησιμοποιήσαμε και εδώ τις παρακάτω παραδοχές που αφορούν την κατηγορία «αριθμητικό»:

- Οι λέξεις που αρχίζουν με αριθμητικό ψηφίο, ανήκουν πάντα στην κατηγορία «αριθμητικό».
- Όλα τα αόριστα άρθρα επίσης κατατάσσονται στην κατηγορία «αριθμητικό».

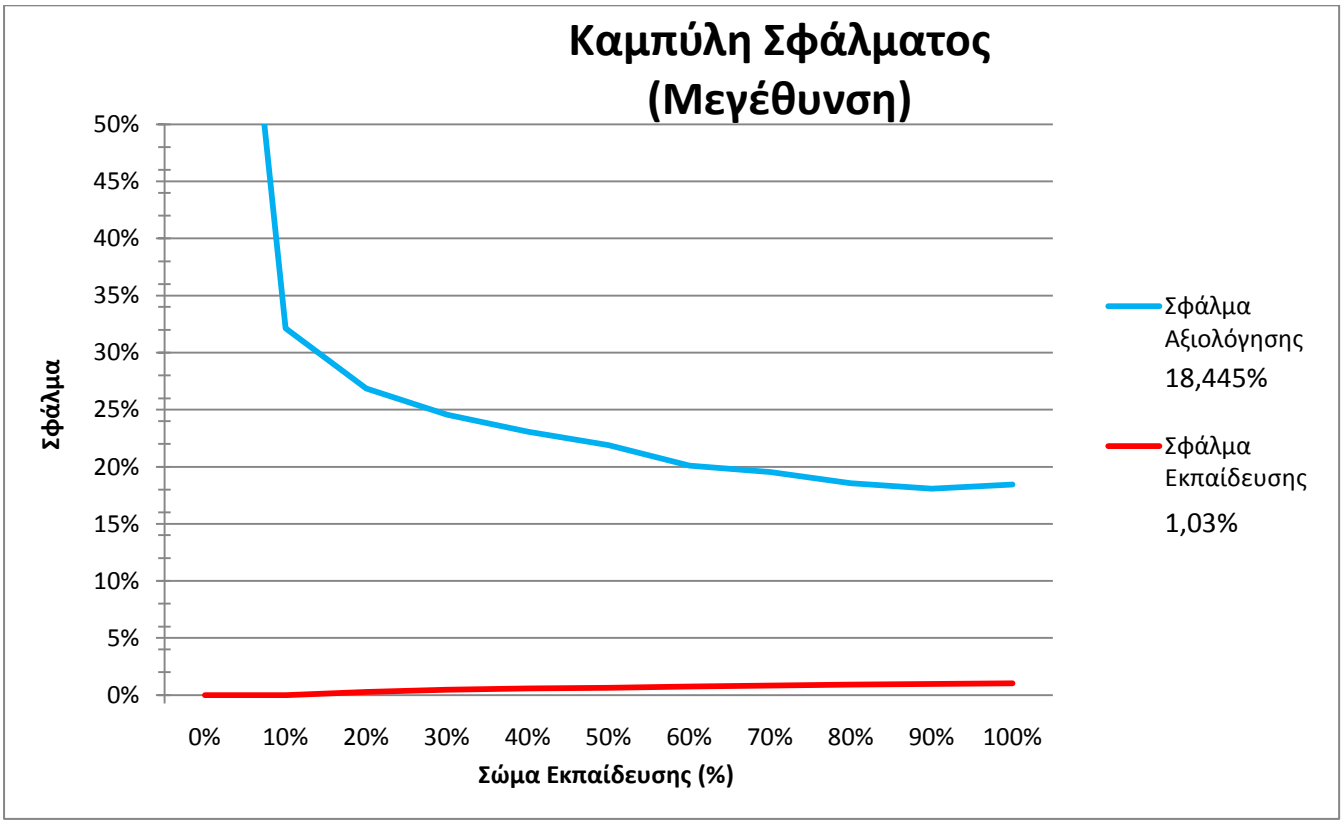
Με τις παραδοχές αυτές πετύχαμε ορθότητα 81,56%, ίδια σχεδόν με το σύστημα χωρίς τις παραπάνω παραδοχές. Για το σύστημα αυτό, υπολογίσαμε τις καμπύλες μάθησης σε σχέση με το σύστημα της προηγούμενης εργασίας (Παππάς, 2008) και τις καμπύλες σφάλματος. Στις καμπύλες μάθησης δείχνουμε και τα αντίστοιχα διαστήματα εμπιστοσύνης 95%.



Σχήμα 3.3.1 Συγκριτικές καμπύλες μάθησης παλαιού και νέου συστήματος με το εκτεταμένο σύνολο κατηγοριών.



Σχήμα 3.3.2 Καμπύλες σφάλματος συστήματος εκτεταμένου συνόλου κατηγοριών.



Σχήμα 3.3.3 Μεγέθυνση καμπύλης σφάλματος συστήματος εκτεταμένου συνόλου κατηγοριών

Οι καμπύλες μάθησης δείχνουν ότι το σύστημά μας ξεπερνάει το προηγούμενο σύστημα (Παππάς, 2008) πετυχαίνοντας ορθότητα υψηλότερη κατά 4,01% και η διαφορά αυτή είναι στατιστικά σημαντική όπως φαίνεται από τα διαστήματα εμπιστοσύνης.

Όσον αφορά τις καμπύλες σφάλματος, παρατηρούμε ότι και οι δύο έχουν σχεδόν γίνει οριζόντιες, ενώ η απόσταση μεταξύ τους είναι αρκετά μεγάλη. Αυτό δείχνει ότι η προσθήκη περισσότερων παραδειγμάτων εκπαίδευσης μάλλον δεν θα βοηθούσε. Ενδεχομένως θα βοηθούσε η καλύτερη επιλογή ή εξαγωγή ιδιοτήτων (π.χ. με Principal Components Analysis), όμως δεν υπήρχε χρόνος για να γίνει αυτό στη διάρκεια της εργασίας.

Κεφάλαιο 4:

Επίλογος

4.1 Ανασκόπηση

Η εργασία αυτή είχε ως στόχο τη δημιουργία ενός βελτιωμένου συστήματος επισημείωσης μερών του λόγου για την ελληνική γλώσσα. Όπως φάνηκε και στα προηγούμενα κεφάλαια, ο στόχος αυτός επιτεύχθηκε και το καινούριο σύστημα πετυχαίνει αρκετά καλύτερα αποτελέσματα τόσο με το βασικό σύνολο κατηγοριών όσο και με το εκτεταμένο, συγκρινόμενο με το προηγούμενο σύστημα που είχε επίσης αναπτυχθεί από την Ομάδα Επεξεργασίας Φυσικής Γλώσσας του ΟΠΑ (Παππάς, 2008). Το νέο σύστημα χρησιμοποιεί διαφορετικό αλγόριθμο μάθησης (ταξινομητή Μέγιστης Εντροπίας) από εκείνον του προηγούμενου συστήματος (k-NN) και διαφορετικό σύνολο ιδιοτήτων. Τέλος, το νέο σύστημα παρέχει προγραμματιστική διεπαφή (API), που διευκολύνει τη χρήση του σε μεγαλύτερα συστήματα.

4.2 Μελλοντικές επεκτάσεις

Σε μια επόμενη εργασία θα ήταν σκόπιμο να προστεθεί στο νέο σύστημα γραφική διεπαφή χρήστη, παρόμοια με εκείνη που υπήρχε στο προηγούμενο σύστημα (Παππάς, 2008). Επίσης, θα ήταν μάλλον σκόπιμο να κατασκευαστούν περισσότερα δεδομένα εκπαίδευσης για την περίπτωση όπου χρησιμοποιείται το βασικό σύνολο κατηγοριών, ενδεχομένως χρησιμοποιώντας μεθόδους ενεργητικής μάθησης, παρόμοιες με εκείνες του προηγούμενου συστήματος (Παππάς, 2008). Ακόμη, θα ήταν σκόπιμο να δοκιμαστούν μέθοδοι εξαγωγής σύνθετων (και λιγότερων) ιδιοτήτων (π.χ. Principal Components Analysis), ιδιαίτερα όταν χρησιμοποιείται το εκτεταμένο σύνολο κατηγοριών.

Παράρτημα Α:

Ιδιότητες αμφισημίας για το βασικό σύνολο κατηγοριών

- **isVerb**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «ρήμα» στα παραδείγματα εκπαίδευσης.
- **isAdverb**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «επίρρημα» στα παραδείγματα εκπαίδευσης.
- **isArticle**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «άρθρο» στα παραδείγματα εκπαίδευσης.
- **isAdjective**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «επίθετο» στα παραδείγματα εκπαίδευσης.
- **isNoun**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «ουσιαστικό» στα παραδείγματα εκπαίδευσης.
- **isConjunction**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «σύνδεσμος» στα παραδείγματα εκπαίδευσης.
- **isNumber**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «αριθμητικό» στα παραδείγματα εκπαίδευσης.
- **isPunctuation**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «σημείο στίξης» στα παραδείγματα εκπαίδευσης.
- **isParticle**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «μόριο» στα παραδείγματα εκπαίδευσης.
- **isPronoun**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «άρθρο» στα παραδείγματα εκπαίδευσης.
- **isOther**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «άλλο» στα παραδείγματα εκπαίδευσης.
- **isPreposition**: το ποσοστό εμφανίσεων της λεκτικής μονάδας που ανήκουν στην κατηγορία «πρόθεση» στα παραδείγματα εκπαίδευσης.

Παράρτημα Β:

Συγκεντρωτικός πίνακας αποτελεσμάτων με το βασικό σύνολο κατηγοριών

Στα παρακάτω συστήματα, οι ιδιότητες αμφισημίας αποτελούνται από ιδιότητες για ολόκληρη τη λέξη και τις τρεις καταλήξεις της, εκτός αν αναφέρεται κάτι διαφορετικό. Επίσης, σε όλα τα συστήματα συμπεριλαμβάνονται οι ιδιότητες «μήκος της λεκτικής μονάδας (αριθμός χαρακτήρων)», «ύπαρξη αποστρόφου στη λεκτική μονάδα», «ύπαρξη αριθμητικού στη λεκτική μονάδα», «ύπαρξη τελείας στη λεκτική μονάδα», «ύπαρξη κόμματος στη λεκτική μονάδα» και «ύπαρξη λατινικού χαρακτήρα στη λεκτική μονάδα», εκτός αν αναφέρεται κάτι διαφορετικό.

Σύστημα	Ορθότητα στα παραδείγματα αξιολόγησης	Ορθότητα σε cross validation
Σύστημα με Boolean ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα	89,43%	--
Σύστημα με Boolean ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα και το μήκος της τρέχουσας λέξης κανονικοποιημένο	88,59%	--
Σύστημα με Boolean ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα, χωρίς το μήκος της τρέχουσας λέξης	88,57%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα	89,77%	89,87%
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα και	88,19%	--

το μήκος της τρέχουσας λέξης κανονικοποιημένο		
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, χωρίς λίστα, χωρίς το μήκος της τρέχουσας λέξης	88,21%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την προηγούμενη και την προ προηγούμενη, με λίστα	89,46%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την μεθεπόμενη, την λέξη μετά την μεθεπόμενη, την προηγούμενη, την προ προηγούμενη και τη λέξη πριν την προ προηγούμενη, με λίστα	--	89,11%
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη και την προηγούμενη με λίστα	--	90,23%
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, την επόμενη, την μεθεπόμενη, την προηγούμενη και την προ προηγούμενη, με λίστα	--	89,39%
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη και την προηγούμενη, με λίστα	--	87,63%
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη και την επόμενη, με λίστα	90,75%	90,56%
Σύστημα με ιδιότητες αμφισημίας(χωρίς τις ιδιότητες που αφορούν ολόκληρη τη λέξη) για την τρέχουσα λέξη και την επόμενη, με λίστα	90,52%	--
Σύστημα με ιδιότητες αμφισημίας(χωρίς τις ιδιότητες που αφορούν την κατάληξη ενός και δύο χαρακτήρων της λέξης) για την τρέχουσα λέξη και την επόμενη, με λίστα	88,86%	--
Σύστημα με ιδιότητες	88,79%	--

αμφισημίας(χωρίς τις ιδιότητες που αφορούν την κατάληξη ενός και τριών χαρακτήρων της λέξης) για την τρέχουσα λέξη και την επόμενη, με λίστα		
Σύστημα με ιδιότητες αμφισημίας(χωρίς τις ιδιότητες που αφορούν την κατάληξη δύο και τριών χαρακτήρων της λέξης) για την τρέχουσα λέξη και την επόμενη, με λίστα	87,13%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, με γνωστή την επόμενη λέξη, με λίστα	90,91%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη, με γνωστή την επόμενη λέξη(χωρίς τις ιδιότητες που αφορούν τις καταλήξεις), με λίστα	91,07%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη και την επόμενη, με λίστα και έλεγχο των άρθρων και των αντωνυμιών μετά την αξιολόγηση	90,82%	--
Σύστημα με ιδιότητες αμφισημίας για την τρέχουσα λέξη και την επόμενη, με λίστα, έλεγχο των άρθρων και των αντωνυμιών μετά την αξιολόγηση και η κατηγορία «άλλο» δεν λαμβάνεται υπ' όψιν στη συνολική ορθότητα	92,71%	--

Βιβλιογραφία[3]

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

Παππάς Κωνσταντίνος Επανυλοποίηση, βελτίωση, αξιολόγηση και τεκμηρίωση ελληνικού επισημειωτή μερών του λόγου που χρησιμοποιεί μηχανική μάθηση, Πτυχιακή Εργασία.
Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, Ιούνιος 2008.

Χρονάκης Ιωάννης Επεκτάσεις και περαιτέρω αξιολόγηση συστήματος αναγνώρισης μερών του λόγου για ελληνικά κείμενα, Πτυχιακή Εργασία.
Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.

[3] Βλ. και τη βιβλιογραφία των τριών προηγούμενων εργασιών.