ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Βελτιώσεις και περαιτέρω αξιολόγηση μεθόδου χειρισμού ερωτήσεων ορισμού για συστήματα ερωταποκρίσεων φυσικής γλώσσας

Γιακουμής Ευστάθιος
Α.Μ. 3000011

Επιβλέπων Καθηγητής : Ίων Ανδρουτσόπουλος

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ 2005

Περιεχόμενα

Περίληψη

Τα συστήματα ερωταποκρίσεων είναι το επόμενο βήμα για τις μηχανές αναζήτησης, μιας και κάνουν την επικοινωνία μεταξύ μηχανής και χρήστη πιο άμεση και προπαντός πιο γρήγορη, δίνοντας στον χρήστη η δυνατότητα να κάνει ερωτήσεις στην φυσική γλώσσα. Η εργασία εστιάζεται στην βελτίωση ενός προηγούμενου τέτοιου συστήματος, ερευνώντας για μια καλύτερη μέθοδο κατασκευής αυτόματων δεδομένων εκπαίδευσης και περαιτέρω αξιολόγηση του με δεδομένα από το διαδύκτιο.

1. Εισαγωγή

Τα τελευταία χρόνια, η εξάπλωση του διαδικτύου είναι ραγδαία και ο όγκος των πληροφοριών στις οποίες ο χρήστης έχει πρόσβαση είναι τεράστιος. Η χρησιμοποίηση αυτού του τεράστιου όγκου πληροφοριών, ο οποίος αφορά μία μεγάλη ποικιλία θεμάτων, είναι πολύ σημαντική υπόθεση για ένα μεγάλο αριθμό ανθρώπων. Επιστήμονες, ερευνητές ,μαθητές, φοιτητές και πολλοί εργαζόμενοι χρησιμοποιούν το διαδίκτυο για την άντληση πληροφοριών που είναι χρήσιμες για την εργασία τους. Όμως, ο εντοπισμός των απαιτούμενων πληροφοριών δεν είναι μια εύκολη υπόθεση, αφού απαιτείται η γνώση των συγκεκριμένων ιστοσελίδων οι οποίες τις φιλοξενούν. Για να είναι δυνατή η αναζήτηση και εύρεση των κατάλληλων ιστοσελίδων έχουν δημιουργηθεί οι μηχανές αναζήτησης, οι οποίες επιτρέπουν στο χρήστη να αναζητήσει ιστοσελίδες που περιέχουν συγκεκριμένους όρους. Για να είναι ακόμα πιο εύκολος ο εντοπισμός των επιθυμητών ιστοσελίδων οι μηχανές αναζήτησης δεν παρουσιάζουν με τυχαία σειρά τις σελίδες που εντοπίζουν, αλλά τις κατατάσσουν με κάποια κριτήρια προσπαθώντας να παρουσιάσουν στις πρώτες θέσεις αυτές οι οποίες είναι οι πιο χρήσιμες στο χρήστη.

Παρά τις πολλές υπηρεσίες που προσφέρουν οι σημερινές μηχανές αναζήτησης είναι επιθυμητό να εξελιχθούν κατάλληλα έτσι ώστε να γίνουν πιο αποδοτικές. Συγκεκριμένα, σε μια επόμενη γενιά μηχανών αναζήτησης είναι επιθυμητό να γίνονται ερωτήσεις σε φυσική γλώσσα και να μην επιστρέφεται στο χρήστη μία λίστα ιστοσελίδων, αλλά μία σύντομη απάντηση και συγχρόνως ένας σύνδεσμος προς την ιστοσελίδα στην οποία περιέχεται η απάντηση . Ένα σύστημα αυτού του είδους ονομάζεται σύστημα ερωταποκρίσεων (Question Answering System). Πολλά ελπιδοφόρα αποτελέσματα έχουν αναφερθεί για τα συστήματα ερωταποκρίσεων τα τελευταία χρόνια και κυρίως μετά το 1999 στα πλαίσια του Question Answering Track του TREC [1](Text Retrieval Conference). Τα καλύτερα συστήματα μπορούν και απαντούν σωστά πάνω από τα 2/3 των ερωτήσεων. Τα αποτελέσματα αυτά, καθώς και η απαίτηση των χρηστών για την δημιουργία τέτοιων συστημάτων με καλές επιδόσεις, έχουν προκαλέσει μεγάλο ενδιαφέρον και πολλή ερευνητική δραστηριότητα σε αυτόν τον τομέα.

Η συγκεκριμένη εργασία στοχεύει στη διερεύνηση θεμάτων που αφορούν στην κατασκευή ενός συστήματος ερωταποκρίσεων, το οποίο χρησιμοποιεί μηχανική μάθηση και εξάγει κατάλληλες απαντήσεις για τις αντίστοιχες ερωτήσεις από τις ιστοσελίδες που επιστρέφει μία μηχανή αναζήτησης. Η εργασία εστιάζεται σε μια συγκεκριμένη κατηγορία ερωτήσεων, τις ερωτήσεις ορισμού (π.χ. «Τι είναι η θαλασσαιμία;», «Ποιος ήταν ο Μπιζέ;»). Πιο συγκεκριμένα, βασίζεται σε μια μέθοδο εντοπισμού σύντομων απαντήσεων σε ερωτήσεις ορισμού, που προτάθηκε σε προηγουμένη εργασία (Μηλιαράκη 2003). Η μέθοδος αυτή χρησιμοποιεί μια Μηχανή Διανυσμάτων Υποστήριξης (ΜΔΥ Support Vector Machine), έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης, για να κατατάσσει τις υποψήφιες απαντήσεις σε αποδεκτούς και μη ορισμούς. Για την εκπαίδευση, όμως, της ΜΔΥ απαιτούνται παραδείγματα εκπαίδευσης, η χειρωνακτική δημιουργία των οποίων είναι επίπονη και χρονοβόρα. Η παρούσα εργασία διερευνά τρόπους αυτόματης παραγωγής παραδειγμάτων εκπαίδευσης για τη ΜΔΥ, σε μια προσπάθεια βελτίωσης μιας προηγούμενης εργασίας (Γαλάνης 2004). Με τον τρόπο αυτό, η μέθοδος χειρισμού των ερωτήσεων ορισμού μετατρέπεται ουσιαστικά σε μη επιβλεπόμενη, αφού είναι δυνατόν να παραχθούν αυτόματα όσα παραδείγματα εκπαίδευσης της ΜΔΥ επιθυμούμε.

[1] http://trec.nist.gov/

Στα πλαίσια αυτής της εργασίας δημιουργήθηκε μια καλύτερη μέθοδος αυτόματης δημιουργίας παραδειγμάτων εκπαίδευσης, ενώ παράλληλα δοκιμάστηκε με επιτυχία μια νέα ΜΔΥ τα αποτελέσματα της οποίας κρίνονται ικανοποιητικά. Κάνοντας πειράματα καταλήξαμε σε ένα βέλτιστο αριθμό ιδιοτήτων για τη ΜΔΥ και βέλτιστο αριθμό όρων στο σώμα εκπαίδευσης της.

2. Συστήματα ερωταποκρίσεων

2.1 Κατηγορίες ερωτήσεων

Οι ερωτήσεις προς ένα σύστημα ερωταποκρίσεων απαιτούν συνήθως μία σύντομη απάντηση, η οποία βρίσκεται αυτούσια σε κάποιο κείμενο της συλλογής εγγράφων που ερευνά η μηχανή αναζήτησης (factual questions). Οι ερωτήσεις αυτές μπορούν να χωριστούν σε κατηγορίες, χρησιμοποιώντας ως κριτήριο διαχωρισμού τον τύπο της απαιτούμενης απάντησης:

Ερωτήσεις προσώπου π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;».
Ερωτήσεις τοποθεσίας π.χ. «Πού βρίσκεται το Στάδιο Ειρήνης και Φιλίας;».
Ερωτήσεις ορισμού π.χ. «Τι είναι η καφεΐνη;».
Ερωτήσεις ποσότητας π.χ. «Πόσες μέρες διήρκεσαν οι Ολυμπιακοί Αγώνες του 2004;».
Ερωτήσεις χρόνου π.χ. «Πότε έγινε η Άλωση της Κωνσταντινούπολης;»

Η εργασία εστιάζεται σε αγγλικές ερωτήσεις ορισμού, η γενική μορφή των οποίων είναι: «What/Who is/are/was <*όρος-στόχος*> ?». Για παράδειγμα:

Who was Galileo?
What is poliomyelitis?
What are sunspots?
What is bipolar disorder?
What is Teflon?

Ως «όρο-στόχο» ονομάζουμε στο εξής τον όρο (π.χ. «Galileo», «poliomyelitis», «bipolar disorder») του οποίου ζητείται ο ορισμός. Ο όρος-στόχος, όπως είναι φανερό, μπορεί να αποτελείται από μία ή παραπάνω λέξεις.

2.2 Αρχιτεκτονική συστημάτων ερωταποκρίσεων

Τα περισσότερα σύγχρονα συστήματα ερωταποκρίσεων υιοθετούν σε γενικές γραμμές την αρχιτεκτονική του σχήματος 1, του οποίου οι μονάδες επεξηγούνται παρακάτω:

1. Επεξεργασία ερωτήσεων

Η ερώτηση αναλύεται, π.χ. μορφολογικά ή συντακτικά, και από την ανάλυση προκύπτουν πληροφορίες όπως η κατηγορία της ερώτησης, οι όροι της, συνώνυμά τους, το συντακτικό της δέντρο κλπ. Στην εργασία αυτή, θεωρούμε πως υπάρχει διαθέσιμη μια μονάδα ανάλυσης ερωτήσεων, που διαχωρίζει τις ερωτήσεις ορισμού από τις υπόλοιπες και εντοπίζει μέσα σε αυτές τον όρο-στόχο.

2. Ανάκτηση εγγράφων

Οι όροι της ερώτησης (στην περίπτωσή μας, ο όρος-στόχος) δίνονται ως λέξεις-κλειδιά σε μια μηχανή αναζήτησης, η οποία ανασύρει από τη συλλογή εγγράφων (π.χ. τα αρχεία μιας εφημερίδας ή ολόκληρο τον Παγκόσμιο Ιστό) έγγραφα που πιθανώς σχετίζονται με την ερώτηση του χρήστη και τα επιστρέφει με κάποια σειρά κατάταξης (ranked list). Επειδή η μηχανή αναζήτησης ενδέχεται να επιστρέψει μεγάλο πλήθος εγγράφων, το σύστημα επιλέγει μόνο τα X κορυφαία έγγραφα της σειράς κατάταξης. Στην παρούσα εργασία X = 5.

3. Προεπεξεργασία εγγράφων

Τα έγγραφα που προέκυψαν από τη συλλογή πιθανόν να χρειάζονται κάποια επεξεργασία, όπως αφαίρεση ετικετών (tags) HTML ή μετατροπή τους σε κάποια άλλη επιθυμητή μορφή, ώστε να γίνει η μετέπειτα επεξεργασία τους από το σύστημα. Επίσης, ενδέχεται να απαιτείται ο εντοπισμός μέσα στα έγγραφα ονομάτων συγκεκριμένων κατηγοριών (π.χ. ονόματα τοποθεσιών στην περίπτωση ερωτήσεων τοποθεσιών), η συντακτική ανάλυση των εγγράφων (σε συστήματα που μετρούν πόσο μοιάζει το συντακτικό δέντρο της ερώτησης με το συντακτικό δέντρο κάθε υποψήφιας απάντησης) κ.λ.π.

4.Εξαγωγή υποψήφιων απαντήσεων

Από τα ανακτηθέντα έγγραφα επιλέγονται εκείνα τα μέρη τους που ενδέχεται να αποτελούν και απαντήσεις στην ερώτηση (π.χ. προτάσεις που περιέχουν ονόματα τοποθεσιών, στην περίπτωση ερωτήσεων τοποθεσιών). Στην παρούσα εργασία, επιλέγονται ως υποψήφιες απαντήσεις όλα τα τμήματα κειμένων μήκους 250 χαρακτήρων που περιέχουν στο μέσο τους τον όρο-στόχο. Στο εξής αναφερόμαστε στα τμήματα αυτά ως «παράθυρα» του όρου-στόχου.

5. Αξιολόγηση υποψηφίων απαντήσεων

Οι υποψήφιες απαντήσεις αξιολογούνται και παράγεται μια σειρά κατάταξης (ranked list) που δείχνει ποιες υποψήφιες απαντήσεις το σύστημα θεωρεί καταλληλότερες, πιθανόν μαζί με το βαθμό βεβαιότητας του συστήματος για κάθε απάντηση. Στην περίπτωσή μας, η ΜΔΥ διαχωρίζει τα παράθυρα του όρου-στόχου στις κατηγορίες «ορισμός» και «μη ορισμός». Ακριβέστερα, επιστρέφει για κάθε ένα παράθυρο ένα βαθμό βεβαιότητας, που δείχνει πόσο σίγουρη είναι πως το παράθυρο ανήκει στην κατηγορία «ορισμός». Καλύτερες απαντήσεις θεωρούνται τα παράθυρα με υψηλότερο βαθμό βεβαιότητας.

6. Επιλογή απάντησης

Στο στάδιο αυτό το σύστημα επιλέγει την απάντηση ή τις απαντήσεις (αν π.χ. επιτρέπονται ως 5 απαντήσεις ανά ερώτηση) που θεωρεί καταλληλότερες και τις επιστρέφει στον χρήστη. Αν χρησιμοποιείται και κάποια τεχνική παραγωγής φυσική γλώσσας, η απάντηση ενδέχεται να αποτελεί σύνθεση πολλών από τις κορυφαίες υποψήφιες απαντήσεις.

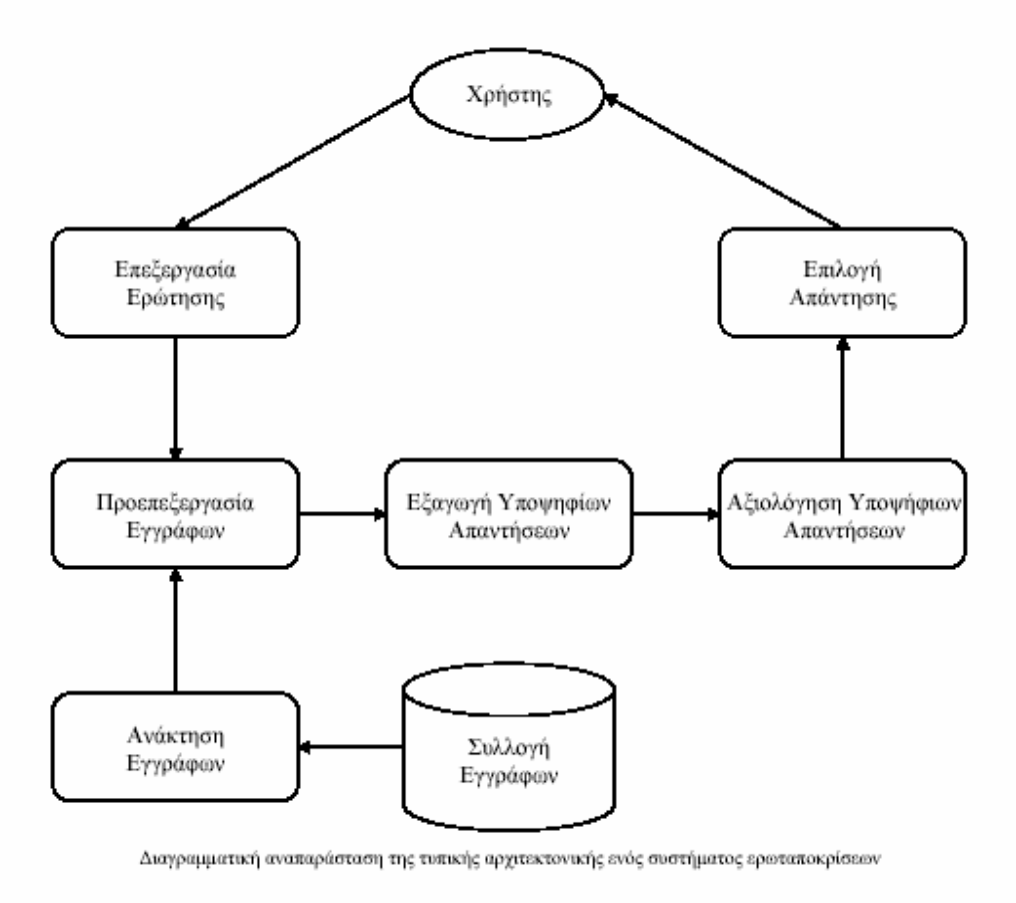
Ακολουθούν τρία παραδείγματα παραθύρων (υποψηφίων απαντήσεων) για την ερώτηση «Who was Archimedes?». Από αυτά μόνο το δεύτερο είναι αποδεκτό ως ορισμός του όρου-στόχου.

*Ερώτηση: Who was **Archimedes**?*

*estimating the value of pi. for grades 6-8, 9-12. find it at: www.pbs.org/nova/teachers/activities/3010_archimed.html **archimedes** and the palimpsest learn about archimedes, his hidden manuscript, and the nova program that features him.*

*nova | infinite secrets | library resource kit | who was archimedes? | pbs who was archimedes? by [author] infinite secrets homepage **archimedes** of syracuse was one of the greatest mathematicians in history.*

*comes from his writings and those of his contemporaries. born in syracuse, sicily (then part of greece), in about 287 b.c., **archimedes** traveled to egypt at the age of 18 to study at the great library of alexandria. upon completing his studies, he*


Χρήστης
Επεξεργασία Ερώτησης
Επιλογή Απάντησης
Προεπεξεργασία Εγγράφων
Εξαγωγή Υποψηφίων Απαντήσεων
Αξιολόγηση Υποψήφιων Απαντήσεων
Ανάκτηση Εγγράφων
Συλλογή Εγγράφων
Διαγραμματική αναπαράσταση της τυπικής αρχιτεκτονικής ενός συστήματος ερωταποκρίσεων

Εικόνα 1

3. Μηχανική Μάθηση και Μηχανές Διανυσμάτων Υποστήριξης[2]

Η μηχανική μάθηση είναι ένας από τους παλαιότερους τομείς της τεχνητής νοημοσύνης [26]. Σκοπός της είναι, σε γενικές γραμμές, να κατασκευάσει συστήματα τα οποία να αποκτούν αυτόματα νέες γνώσεις από εμπειρικά δεδομένα του παρελθόντος. Στην εργασία αυτή ασχολούμαστε με μεθόδους επιβλεπόμενης μηχανικής μάθησης για το διαχωρισμό σε κατηγορίες (παράθυρα που αποτελούν ή όχι αποδεκτούς ορισμούς του όρου-στόχου). Αυτού του είδους η μάθηση είναι δυνατόν να χωριστεί σε τρία στάδια. Πρώτον, επισημειώνονται, συνήθως χειρωνακτικά, παραδείγματα εκπαίδευσης με τη ορθή τους κατηγορία. Τα παραδείγματα παριστάνονται στη συνέχεια με τη μορφή διανυσμάτων ιδιοτήτων. Το στάδιο της επισημείωσης των παραδειγμάτων εκπαίδευσης είναι το βασικό μειονέκτημα της επιβλεπόμενης μηχανικής μάθησης στις περισσότερες εφαρμογές επεξεργασίας φυσικής γλώσσας, καθώς σε πολλές εφαρμογές ο όγκος των κειμένων που πρέπει να επισημειωθεί είναι μεγάλος, με αποτέλεσμα η διαδικασία αυτή να είναι κουραστική και ιδιαίτερα χρονοβόρα.

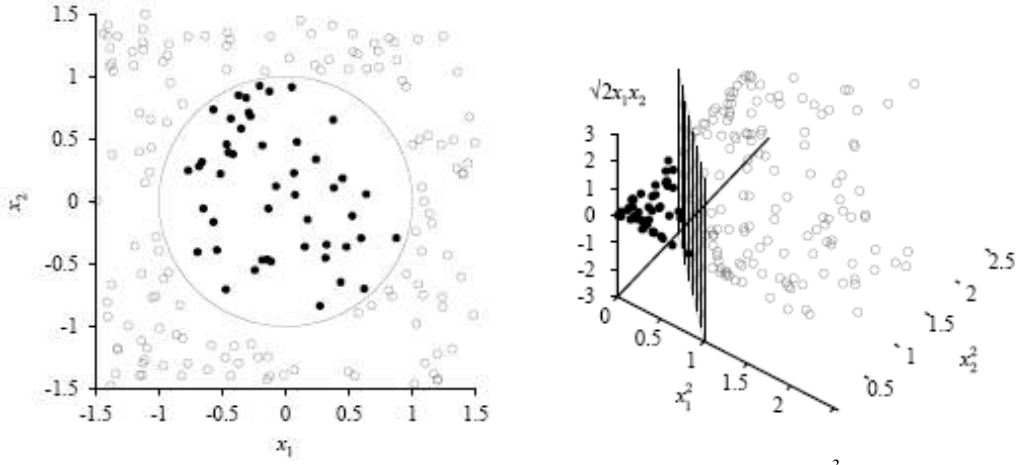
Στο δεύτερο στάδιο χρησιμοποιείται ένας αλγόριθμος μάθησης, ο οποίος επεξεργάζεται τα παραδείγματα εκπαίδευσης προκειμένου να κατασκευάσει έναν ταξινομητή που θα είναι σε θέση να διαχωρίζει παραδείγματα διαφορετικών κατηγοριών με βάση τις τιμές των ιδιοτήτων τους. Έχουν προταθεί πολλοί αλγόριθμοι, όπως ο C4.5 (δημιουργία δέντρων απόφασης), τα νευρωνικά δίκτυα, το μοντέλο μέγιστης εντροπίας, οι Μηχανές Διανυσμάτων Υποστήριξης κλπ.

Το τελευταίο στάδιο αφορά την κατηγοριοποίηση νέων περιπτώσεων, για τις οποίες δεν είναι γνωστή η ορθή κατηγορία. Οι νέες περιπτώσεις αναπαρίστανται επίσης με τη μορφή διανυσμάτων ιδιοτήτων και κατατάσσονται στις κατηγορίες χρησιμοποιώντας τον ταξινομητή του προηγούμενου σταδίου.

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines, SVMs [17, 16, 33]) είναι μία σχετικά καινούρια μέθοδος επιβλεπόμενης μηχανικής μάθησης, η οποία μπορεί να εφαρμοστεί και σε προβλήματα κατηγοριοποίησης και η οποία έχει επιτύχει εξαιρετικά αποτελέσματα σε πολλές εφαρμογές. Στην απλούστερή τους μορφή, που χρησιμοποιούμε εδώ, οι ΜΔΥ μαθαίνουν να διαχωρίζουν περιπτώσεις δύο κατηγοριών. Ουσιαστικά προβάλλουν, με τη χρήση μίας συνάρτησης μετασχηματισμού, τα διανύσματα ιδιοτήτων σε ένα χώρο περισσοτέρων διαστάσεων και στη συνέχεια προσπαθούν να βρουν ένα γραμμικό διαχωριστή, δηλαδή ένα υπερεπίπεδο, που να διαχωρίζει τις δύο κατηγορίες με μέγιστο περιθώριο (margin) στο νέο διανυσματικό χώρο.

Η μετάβαση στο νέο χώρο περισσοτέρων διαστάσεων διευκολύνει την εύρεση γραμμικού διαχωριστή. Για παράδειγμα, στο παρακάτω σχήμα αριστερά φαίνεται μία περίπτωση, όπου δεν υπάρχει γραμμικός διαχωριστής (ευθεία) στο επίπεδο (εδώ υπάρχουν δύο μόνο ιδιότητες). Χρησιμοποιώντας όμως τη συνάρτηση μετασχηματισμού $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2} \cdot x_1 \cdot x_2 \right\rangle$ και μεταβαίνοντας στις τρεις διαστάσεις παρατηρούμε ότι υπάρχει ένα επίπεδο που διαχωρίζει τα διανύσματα (σχήμα στα δεξιά). Στην περίπτωση περισσοτέρων ιδιοτήτων, ο διαχωριστής θα είναι ένα υπερεπίπεδο.

[2] Το κείμενο αυτής της ενότητας προέρχεται από την εργασία του Γιώργου Λουκαρέλλι (Λουκαρέλλι 2005) και περιλαμβάνεται σε αυτήν την εργασία έπειτα από συνεννόηση με τον επιβλέποντα καθηγητή του, κ.Ίωνα Ανδρουτσόπουλο.



Μετασχηματισμός από τις δύο διαστάσεις στις τρεις [3]

Γενικά, η εξίσωση του υπερεπιπέδου διαχωρισμού θα είναι της ακόλουθης μορφής, όπου F η συνάρτηση μετασχηματισμού:

$$\vec{w} \cdot \vec{F}(\vec{x}) + b = 0$$

Το υπερεπίπεδο διαχωρισμού τοποθετείται στο μέσον της απόστασης δύο παράλληλων υπερεπιπέδων, τα οποία διαχωρίζουν πλήρως τα παραδείγματα εκπαίδευσης και εφάπτονται με τουλάχιστον ένα παράδειγμα εκπαίδευσης, διαφορετικής κατηγορίας για το κάθε ένα από τα δύο υπερεπίπεδα. Τα $\vec{w}$ ($\vec{w} \in R^l$, όπου l ο αριθμός των ιδιοτήτων στο νέο χώρο) και b μπορούν να επιλεγούν (με scaling) ώστε τα δύο παράλληλα εφαπτόμενα υπερεπίπεδα να έχουν εξισώσεις:

$$\vec{w} \cdot \vec{F}(\vec{x}) + b = \pm 1$$

οπότε η απόσταση μεταξύ των δύο εφαπτόμενων υπερεπιπέδων είναι $2/\left\|\vec{w}\right\|$. Η απόσταση αυτή είναι το «περιθώριο» του υπερεπιπέδου διαχωρισμού, που τοποθετείται στο μέσον της απόστασης των δύο εφαπτόμενων υπερεπιπέδων. Όπως αναφέρθηκε ήδη, ο στόχος των ΜΔΥ είναι να βρουν το υπερεπίπεδο διαχωρισμού με το μέγιστο περιθώριο. Οπότε, προκύπτει τα παρακάτω πρόβλημα βελτιστοποίησης:

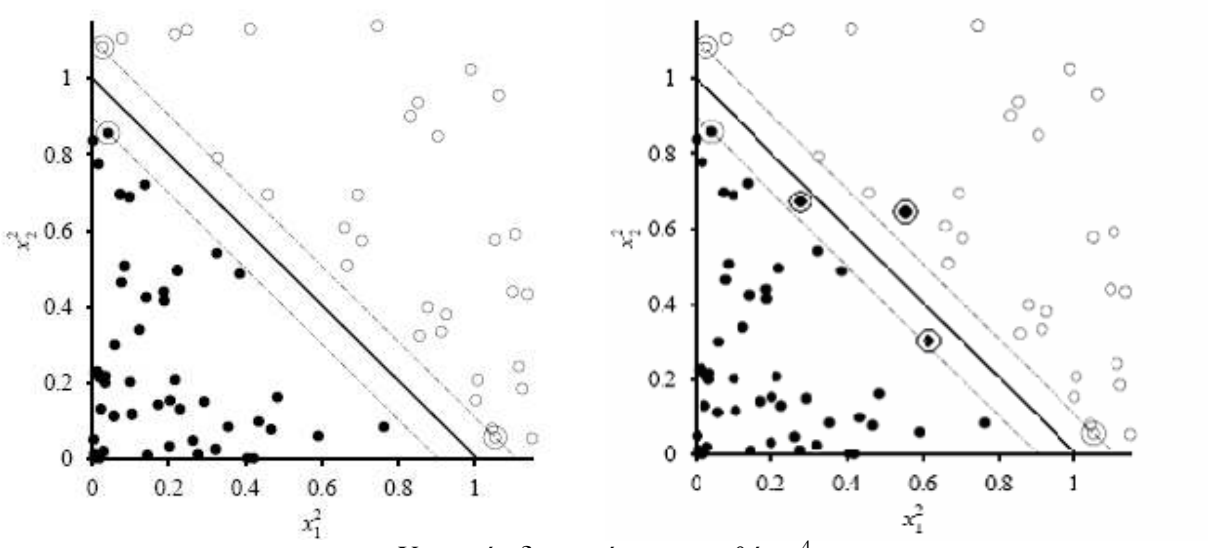
$$\min \left\|\vec{w}\right\|^2 / 2$$
$$(\vec{w} \cdot \vec{F}(\vec{x_j}) + b) \cdot y_j \geq 1$$

όπου x_j με $1 \leq j \leq n$ είναι το διάνυσμα του j -οστού παραδείγματος εκπαίδευσης και $y_j \in \{1,-1\}$ είναι η κατηγορία του j -οστού διανύσματος εκπαίδευσης.

Οι περιορισμοί του παραπάνω προβλήματος βελτιστοποίησης επιβάλλουν όλα τα διανύσματα εκπαίδευσης να βρίσκονται έξω ή το πολύ στα όρια του περιθωρίου και από τη σωστή πλευρά

[3] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, 2002.

του υπερεπιπέδου, ανάλογα με την κατηγορία τους, όπως φαίνεται στο παρακάτω σχήμα αριστερά. Οι περιορισμοί αυτοί, όμως, είναι πολύ αυστηροί. Για παράδειγμα, ενδέχεται να μην είναι δυνατή η εύρεση γραμμικού διαχωριστή που να διαχωρίζει πλήρως τα παραδείγματα εκπαίδευσης, παρά τη μετάβαση στο νέο χώρο διαστάσεων. Ή ενδέχεται να προτιμούμε ένα υπερεπίπεδο διαχωρισμού που έχει μεγαλύτερο περιθώριο αλλά κατατάσσει λανθασμένα ή εντός του περιθωρίου κάποια παραδείγματα εκπαίδευσης (όπως στο παρακάτω σχήμα στα δεξιά) από κάποιο άλλο που ικανοποιεί όλους τους περιορισμούς αλλά έχει μικρότερο περιθώριο.



Υπερεπίπεδο με μέγιστο περιθώριο[4]

Για τους λόγους αυτούς, είναι δυνατόν οι περιορισμοί να χαλαρώσουν, με αποτέλεσμα να κατασκευαστεί ένα ανεκτικότερο πρόβλημα βελτιστοποίησης, το οποίο ορίζεται ως εξής:

$$\min \left\| \vec{w} \right\|^2 / 2 + C \cdot \sum_j \xi_j$$
$$(\vec{w} \cdot \vec{F}(\vec{x_j}) + b) \cdot y_j \geq 1 - \xi_j$$
$$\xi_j \geq 0$$

όπου το ξ_j είναι το σφάλμα για κάθε διάνυσμα εκπαίδευσης (το πόσο απέχουμε από το να ικανοποιείται ο αντίστοιχος περιορισμός) και C το κόστος (ανοχή) που δίνεται στο συνολικό σφάλμα. Στην περίπτωση αυτή, όπως φαίνεται στο παραπάνω σχήμα δεξιά, υπάρχει καλύτερη δυνατότητα γενίκευσης και ο διαχωριστής είναι πιο ανεκτικός σε λάθη επισημείωσης των δεδομένων εκπαίδευσης.

Τελικά, επιλύνοντας το παραπάνω πρόβλημα ελαχιστοποίησης, προκύπτει $\vec{w}$ της μορφής:

$$\vec{w} = \sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j})$$

[4] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, (2002)

Οπότε η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j})\right) \cdot \vec{F}(\vec{x}) + b = 0$$

ή

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x})\right) + b = 0$$

όπου οι τιμές των a_j είναι διάφορες του μηδενός μόνο για τα «διανύσματα υποστήριξης», δηλαδή τα διανύσματα εκπαίδευσης που βρίσκονται πάνω στα δύο εφαπτόμενα υπερεπίπεδα και (στην περίπτωση που ανεχόμαστε σφάλματα) τα διανύσματα που κατατάσσονται λανθασμένα ή εντός του περιθωρίου. Τα παραδείγματα που δεν είναι διανύσματα υποστήριξης ουσιαστικά αγνοούνται.

Η συνάρτηση μετασχηματισμού συμμετέχει μόνο σε εσωτερικά γινόμενα $\vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$. Ορίζουμε, λοιπόν, ως πυρήνα της ΜΔΥ τη συνάρτηση:

$$K(\vec{x_j}, \vec{x_i}) = \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$$

Σύμφωνα με το θεώρημα του Mercer, κάθε συνάρτηση $K(\vec{x_i}, \vec{x_j})$ για την οποία ο πίνακας $K_{ij} = K(\vec{x_i}, \vec{x_j})$ είναι θετικά ορισμένος[5] υπολογίζει το εσωτερικό γινόμενο των $\vec{x_i}, \vec{x_j}$ σε κάποιο νέο διανυσματικό χώρο, δηλαδή μπορεί να χρησιμοποιηθεί ως πυρήνας μιας ΜΔΥ. Το ενδιαφέρον είναι ότι σε πολλές περιπτώσεις είναι δυνατόν να υπολογιστούν οι τιμές του πυρήνα χωρίς να υπολογιστεί πρώτα η τιμή των $\vec{F}(\vec{x_j})$ και $\vec{F}(\vec{x_i})$, δηλαδή χωρίς να υπολογίσουμε τις (συνήθως πολύ περισσότερες) ιδιότητες των διανυσμάτων στο νέο χώρο, κάτι που επιτρέπει τη χρήση πυρήνων που υπολογίζουν εσωτερικά γινόμενα σε νέους χώρους πολύ μεγάλου αριθμού διαστάσεων. Για παράδειγμα, στην περίπτωση του αρχικού παραδείγματος μετασχηματισμού με $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2} \cdot x_1 \cdot x_2 \right\rangle$ ο πυρήνας έχει τη μορφή $K(\vec{x_i}, \vec{x_j}) = F(\vec{x_i}) \cdot F(\vec{x_j}) = (\vec{x_i} \cdot \vec{x_j})^2$, δηλαδή οι τιμές του μπορούν να υπολογισθούν με βάση μόνο τις τιμές των ιδιοτήτων του αρχικού χώρου. Τελικά, αντικαθιστώντας το εσωτερικό γινόμενο $\vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$ με τη συνάρτηση πυρήνα $K(\vec{x_j}, \vec{x_i})$ η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot K(\vec{x_j}, \vec{x})\right) + b = 0$$

Παραδείγματα πυρήνων που χρησιμοποιούνται είναι τα εξής:

---

[5] Ο πίνακας $A \in R^{n \times n}$ είναι ένας θετικά ορισμένος πίνακας αν για όλα τα μη μηδενικά διανύσματα $x \in R^n$ ισχύει $x^T \cdot A \cdot x > 0$, όπου x^T είναι το ανάστροφο διάνυσμα.

- γραμμικός: $K(\vec{x_i}, \vec{x_j}) = \vec{x_i} \cdot \vec{x_j}$
- πολυωνυμικός: $K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d$
- ακτινωτής βάσης (radial base function – RBF):
  $K(\vec{x_i}, \vec{x_j}) = \exp\left(-\gamma \cdot \left\| \vec{x_i} - \vec{x_j} \right\|^2 \right), \gamma > 0$
- σιγμοειδής: $K(\vec{x_i}, \vec{x_j}) = \tanh(\vec{x_i} \cdot \vec{x_j} + r)$

όπου τα γ, r και d είναι παράμετροι κάθε πυρήνα. Ο γραμμικός πυρήνας (που δεν προκαλεί μετάβαση σε νέο διανυσματικό χώρο) είναι ειδική περίπτωση του πυρήνα ακτινωτής βάσης (Keerthi και Lin 2003). Επίσης, ο σιγμοειδής πυρήνας συμπεριφέρεται όπως ο πυρήνας ακτινωτής βάσης για συγκεκριμένες παραμέτρους (Lin και Lin 2003).

Τέλος, η απόφαση για την κατηγορία ενός καινούριου διανύσματος, δεδομένου ενός εκπαιδευμένου ταξινομητή, λαμβάνεται με βάση το πρόσημο της ακόλουθης παράστασης:

$$sign\left(\sum_j a_j \cdot y_j \cdot K(\vec{x}_j, \vec{x}) + b \right)$$

Στην εργασία έχει χρησιμοποιηθεί το LIBSVM [6] με πυρήνα ακτινωτής βάσης (RBF) με C =1 και $\gamma = 0.00310559$. Όλα τα δεδομένα (και τα δεδομένα εκπαίδευσης και αξιολόγησης) έχουν κανονικοποιηθεί στο διάστημα [-1 , 1].[7]

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[7] Το LIBSVM παρέχει αυτή τη δυνατότητα, βλ. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

4. Αυτόματη κατασκευή παραθύρων εκπαίδευσης

4.1 Συλλογές δεδομένων διαγωνισμών TREC

Οι διοργανωτές του Question Answering Track του TREC παρέχουν κάθε χρόνο μία συλλογή ερωτήσεων, καθώς και μια συλλογή κειμένων μέσα στην οποία πρέπει να εντοπιστούν οι απαντήσεις των ερωτήσεων. Η εργασία της Μηλιαράκη (2003), στην οποία βασίζεται η παρούσα εργασία, χρησιμοποίησε τις ερωτήσεις ορισμού που περιλαμβάνονταν στις συλλογές ερωτήσεων του TREC των ετών 2000 και 2001 (συνολικά 160 ερωτήσεις ορισμού). Για κάθε ερώτηση εκείνων των ετών, τα δεδομένα του TREC περιλαμβάνουν επίσης τα 50 κορυφαία κείμενα που επέστρεψε μια μηχανή αναζήτησης από τη συλλογή κειμένων του αντίστοιχου έτους χρησιμοποιώντας ως όρους αναζήτησης τους όρους της ερώτησης. Κάθε ένα από αυτά τα κείμενα συνοδεύεται από έναν αριθμό (1-50), ο οποίος δηλώνει την σειρά (κατάταξη) με την οποία επιστράφηκε το κείμενο από την μηχανή αναζήτησης. Για κάθε ερώτηση, παρέχεται επίσης μια λίστα προτύπων (patterns) απαντήσεων γραμμένων στη γλώσσα Perl, που μπορούν να χρησιμοποιηθούν για την αξιολόγηση των απαντήσεων που επιστρέφει ένα σύστημα ως ορθών ή λανθασμένων. Τα πρότυπα έχουν κατασκευαστεί από τους διοργανωτές των διαγωνισμών TREC λαμβάνοντας υπόψη τους όλες τις σωστές απαντήσεις που επέστρεψαν τα συστήματα των διαγωνισμών για τις αντίστοιχες ερωτήσεις.

Στην προσέγγιση της Μηλιαράκη (2003), η οποία χρησιμοποιεί μια ΜΔΥ για να διαχωρίζει τα παράθυρα των όρων-στόχων που αποτελούν ορισμούς από εκείνα που δεν αποτελούν ορισμούς, τα πρότυπα των ερωτήσεων ορισμού χρησιμοποιούνται για τη δημιουργία των παραδειγμάτων εκπαίδευσης της ΜΔΥ. Για κάθε μία ερώτηση ορισμού που χρησιμοποιείται κατά την εκπαίδευση του συστήματος, εντοπίζονται τα παράθυρά της μέσα στη συλλογή κειμένων (ενότητα 2.2) και αντί να κατατάσσονται αυτά χειρωνακτικά στις δύο κατηγορίες, όποιο παράθυρο εκπαίδευσης ικανοποιεί τουλάχιστον ένα πρότυπο της αντίστοιχης ερώτησης θεωρείται ότι ανήκει στην κατηγορία «ορισμός» και διαφορετικά στην κατηγορία «μη ορισμός». Με αντίστοιχο τρόπο, τα πρότυπα ερωτήσεων του TREC οι οποίες δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση της ΜΔΥ μπορούν να χρησιμοποιηθούν για την αξιολόγηση του συστήματος.

Παρακάτω δίνονται κάποια πρότυπα για τις ερωτήσεις «Who was Galileo?» και «What is cholesterol?»

Who was Galileo ?
astronomer
the Italian sunspots expert
What is cholesterol ?
steroidlike compound
(fatty|waxy) substance
fat(ty|s)?

Η δημιουργία των προτύπων απαντήσεων γίνεται χειρωνακτικά και δεν είναι μια εύκολη διαδικασία. Σκοπός της παρούσας εργασίας είναι η εξεύρεση μιας εναλλακτικής μεθόδου δημιουργίας παραδειγμάτων εκπαίδευσης, που δεν θα απαιτεί χειρωνακτική κατάταξη των παραδειγμάτων εκπαίδευσης σε κατηγορίες ούτε χειρωνακτική δημιουργία προτύπων απαντήσεων.

4.2 Υπάρχουσες μέθοδοι δημιουργίας παραδειγμάτων εκπαίδευσης

Όπως αναφέρθηκε στις προηγούμενες ενότητες, η μέθοδος χειρισμού ερωτήσεων ορισμού στην οποία βασίζεται η παρούσα εργασία προϋποθέτει μια συλλογή παραδειγμάτων (παραθύρων) εκπαίδευσης, στην οποία πρέπει να φαίνεται η κατηγορία (ορισμός ή μη- ορισμός) στην οποία ανήκει κάθε παράδειγμα. Για την κατάταξη των παραθύρων εκπαίδευσης αναφέρθηκαν ήδη δύο μέθοδοι: a) Η χρήση προτύπων απαντήσεων, όπως και στα δεδομένα των διαγωνισμών TREC (ενότητα 4.1). b) Η χειρωνακτική κατάταξη κάθε παραθύρου εκπαίδευσης ξεχωριστά. Οι δύο προαναφερθείσες λύσεις παρουσιάζουν η κάθε μία πλεονεκτήματα και μειονεκτήματα. Για την (a), το βασικότερο πλεονέκτημα είναι ότι δεν απαιτείται η χειρωνακτική κατάταξη κάθε παραθύρου ξεχωριστά αλλά αρκεί η δημιουργία προτύπων απαντήσεων για κάθε ερώτηση. Η δημιουργία προτύπων απαντήσεων, όμως, δεν είναι μια εύκολη διαδικασία, αφού τα πρότυπα απαντήσεων πρέπει να προβλέπουν όλες τις πιθανές διατυπώσεις των ορθών απαντήσεων. Επίσης, στην περίπτωση που θέλουμε να εκπαιδεύσουμε το σύστημα με νέες ερωτήσεις ορισμού και κείμενα (π.χ. ιατρικά κείμενα και ερωτήσεις ορισμού ιατρικών όρων, αντί για κείμενα και ερωτήσεις γενικής φύσης), πρέπει να δημιουργηθούν νέα πρότυπα απαντήσεων για τις νέες ερωτήσεις εκπαίδευσης και τα νέα κείμενα. Στη μέθοδο (b), το πλεονέκτημα είναι ότι επειδή η κατάταξη γίνεται χειρωνακτικά, ο θόρυβος στα δεδομένα εκπαίδευσης (τα λάθη στην κατάταξη των παραδειγμάτων) είναι λιγότερος, από ό,τι στη μέθοδο (a). Η επανεκπαίδευση, όμως, του συστήματος με νέες ερωτήσες και κείμενα απαιτεί τη χειρωνακτική κατάταξη ενός νέου μεγάλου όγκου παραθύρων (των παραθύρων των νέων ερωτήσεων εκπαίδευσης), που είναι μια διαδικασία επίπονη και χρονοβόρα.

Όλα τα παραπάνω κάνουν επιθυμητή τη δημιουργία μιας νέας μεθόδου κατάταξης παραθύρων εκπαίδευσης, η οποία θα είναι σε μεγάλο βαθμό αυτόματη. Ένας επιπλέον στόχος είναι τα παράθυρα εκπαίδευσης να προέρχονται από γενικής φύσεως ιστοσελίδες, ώστε να είναι δυνατόν η ΜΔΥ να εκπαιδευθεί στο να εντοπίζει ορισμούς όρων σε γενικής φύσεως ιστοσελίδες. Σε προηγούμενη εργασία (Γαλάνης 2004) έγινε μια πρώτη απόπειρα δημιουργίας μιας τέτοιας μεθόδου, την οποία επιχειρεί να βελτιώσει η παρούσα εργασία.

4.3 Η μέθοδος του Γαλάνη

Έστω ότι έχουμε στην διάθεσή μας μια ερώτηση ορισμού από την οποία έχει εξαχθεί ο όρος-στόχος. Έστω επίσης ότι έχουμε και ένα παράθυρο του όρου-στόχου το οποίο προέρχεται από μια ιστοσελίδα. Αν είχαμε στην διάθεσή μας ένα τουλάχιστον ορισμό του όρου-στόχου (π.χ. από μια ηλεκτρονική εγκυκλοπαίδεια), τότε θα μπορούσαμε με ένα μέτρο ομοιότητας να συγκρίνουμε τα δύο κομμάτια κειμένου και να αποφασίσουμε αν το παράθυρο είναι και αυτό ορισμός του όρου-στόχου, βάσει της ομοιότητάς του με τον ορισμό που διαθέτουμε. Για να υλοποιηθεί αυτή η ιδέα πρέπει να κατασκευαστεί ένα μέτρο ομοιότητας που θα αποφασίζει αν το παράθυρο είναι επαρκώς όμοιο με τον ορισμό που διαθέτουμε.

Μια αρχική προσέγγιση είναι να μετρηθεί ο αριθμός των κοινών λέξεων που έχουν τα δύο κείμενα (το κείμενο του ορισμού που διαθέτουμε και το παράθυρο εκπαίδευσης του οποίου δεν γνωρίζουμε την κατηγορία), όπως φαίνεται παρακάτω:
Ερώτηση: *Who was **Archimedes**?*
Παράθυρο του όρου-στόχου από ιστοσελίδα:

*nova | infinite secrets | library resource kit | who was archimedes? | pbs who was **archimedes**?*
*by [author] infinite secrets homepage **archimedes** of syracuse was one of the greatest*

mathematicians in history.

Ορισμός που διαθέτουμε από ηλεκτρονική εγκυκλοπαίδεια:

A Greek mathematician living from approximately 287 BC to 212 BC in Syracuse. He invented much plane geometry, studying the circle, parabola and three-dimensional geometry of the sphere as well as studying physics. See also Archimedean solid.

Κοινές λέξεις:

of, the, in, Syracuse

Παρότι τα δύο κείμενα είναι και τα δύο ορισμοί του όρου *Archimedes*, οι κοινές λέξεις που βρίσκει ο αλγόριθμος δεν φαίνεται να φανερώνουν επαρκώς την ομοιότητά τους. Επίσης, δεν εντοπίζονται λέξεις που εμφανίζονται και στα δύο κείμενα αλλά με διαφορετικούς τύπους (π.χ. με διαφορετικές καταλήξεις, όπως «mathematicians» και «mathematician»). Ένα επιπλέον πρόβλημα είναι πως για τον ορισμό ενός όρου είναι δυνατό να χρησιμοποιηθεί μια μεγάλη ποικιλία διαφορετικών εκφράσεων και λέξεων. Επομένως, συγκρίνοντας το παράθυρο εκπαίδευσης με τον ορισμό που διαθέτουμε κινδυνεύουμε να χαρακτηρίσουμε το παράθυρο εκπαίδευσης ως μη αποδεκτό ορισμό του όρου-στόχου, ενώ ενδέχεται να είναι αποδεκτός ορισμός αλλά να χρησιμοποιεί διαφορετική διατύπωση από τον ορισμό που διαθέτουμε. Ως παράδειγμα των μεγάλων διαφορών διατύπωσης που μπορεί να εμφανιστούν μεταξύ αποδεκτών ορισμών, παρατίθενται παρακάτω διαφορετικοί ορισμοί του όρου «galaxy» από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια.

*What is a **galaxy**?*

A large aggregation of stars, bound together by gravity. There are three major classifications of galaxies-spiral, elliptical, and irregular.
a very large cluster of stars (tens of millions to trillions of stars) gravitationally bound together . an organized system of many hundreds of millions of stars, often mixed with gas and dust. The universe contains billions of galaxies.
a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.
A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant the their light takes millions of years to reach the Earth.

Με σκοπό την εξάλειψη των παραπάνω προβλημάτων, ο Γαλάνης (2004) υιοθέτησε τις ακόλουθες βελτιώσεις:

- Αφαίρεση από τα συγκρινόμενα κείμενα (παράθυρα και ορισμούς που διαθέτουμε) των 100 συχνότερων λέξεων που εμφανίζονται σε αγγλικά κείμενα (π.χ. “the”, “be”, “of”, “and”, “a”, “in”, “to”, “have”, “it”, “to”, “for”, “i”, “that”, “you”, “he”, “on”, “with”, “do”, “at”, “by”, “not”, “this”). Οι 100 συχνότερες λέξεις έχουν προκύψει από τοBritish National Corpus (http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bncreadme.html). Η αφαίρεση γίνεται διότι κατά την σύγκριση δύο κειμένων η εύρεση κοινών λέξεων που είναι πολύ συχνές στα Αγγλικά (γνωστών ως «stop-words») δεν φανερώνει ομοιότητα.

- Εφαρμογή ενός αλγορίθμου που αποκόπτει την κατάληξη κάθε λέξης αφήνοντας μόνο την ρίζα της (stemming). Για παράδειγμα, το «approximately» γίνεται «approxim» και το «invented» γίνεται «invent». Χρησιμοποιήθηκε ο αλγόριθμος του Porter (http://www.tartarus.org/~martin/PorterStemmer).
- Διαγραφή από κάθε παράθυρο των σημείων στίξεως και άλλων ειδικών χαρακτήρων (!@&^%$#. κ.λ.π.).
- Σύγκριση κάθε παραθύρου εκπαίδευσης με περισσότερους από έναν γνωστούς ορισμούς του όρου-στόχου. Για να επιτευχθεί αυτό χρησιμοποιήθηκε η επιλογή «define:» της μηχανής αναζήτησης Google, η οποία επιστρέφει ορισμούς που προέρχονται από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια εάν χρησιμοποιηθεί μια έκφραση αναζήτησης της μορφής «define: <όρος-στόχος>». Κατά την εκπαίδευση του συστήματος, επιλέγονται όροι-στόχοι για τους οποίους η επιλογή define του Google επιστρέφει πολλούς (π.χ. πάνω από 5) ορισμούς.

Συγκεκριμένα, η μέθοδος του Γαλάνη υπολογίζει, μετά την αφαίρεση των stop-words και των ειδικών χαρακτήρων και μετά την αποκοπή των καταλήξεων, ένα μέτρο ομοιότητας sim(W,C), όπου W το παράθυρο και C η συλλογή των ορισμών του ιδίου όρου-στόχου που έχουμε στη διάθεσή μας από την επιλογή define του Google.

$$sim(W,C) = 1/|W| * \sum_{i=1} |W| sim(w_i, C)$$

Όπου |W| είναι ο αριθμός των λέξεων του W και $sim(w_i, C)$ είναι η ομοιότητα της ι-οστης λέξης του W με το C. Η ομοιότητα αυτή βρίσκεται από τον τύπο :

$$sim(w_i, C) = fdef(w_i, C) * idf(w_i)$$

Όπου $fdef(w_i, C)$ είναι το ποσοστό των ορισμών στο C που περιέχουν το w_i και $idf(w_i)$ (inverse document frequency) η αντίστροφη συχνότητα της λέξης w_i .Το $idf(w_i)$ ορίζεται αναλυτικά στην επόμενη ενότητα.

Στόχος είναι κάθε παράθυρο εκπαίδευσης W το οποίο μοιάζει αρκετά με τους ορισμούς C του όρου-στόχου που διαθέτουμε να έχει υψηλό sim(W,C), ενώ αντίθετα κάθε παράθυρο που δεν μοιάζει αρκετά να έχει χαμηλό sim(W,C). Με αυτό τον τρόπο ελπίζουμε ότι θα καταστεί δυνατό να γίνει ο διαχωρισμός των παραθύρων εκπαίδευσης που ανήκουν στην κατηγορία «ορισμός» από τα παράθυρα εκπαίδευσης που ανήκουν στην κατηγορία «μη ορισμός». Ειδικότερα, η μέθοδος του Γαλάνη χρησιμοποιεί δύο κατώφλια t_- και t_+, με $t_- \leq t_+$, τα οποία επιλέγονται πειραματικά. Τα παράθυρα εκπαίδευσης W με $sim(W,C) \geq t_+$ σημειώνονται ως παραδείγματα ορισμών, τα παράθυρα με $sim(W,C) \leq t_-$ σημειώνονται ως παραδείγματα μη ορισμών, ενώ τα παράθυρα με $t_- \leq sim(W,C) \leq t_+$ δεν χρησιμοποιούνται κατά την εκπαίδευση του συστήματος, γιατί δεν είμαστε αρκετά βέβαιοι αν αποτελούν παραδείγματα ορισμών ή μη ορισμών.

4.4 Η μέθοδος του κεντροειδούς

Στηριζόμενοι σε μια προηγούμενη προσπάθεια αυτόματης εύρεσης ορισμών (Cui κ.ά. 2004), πειραματιστήκαμε με μια εναλλακτική μέθοδο διαχωρισμού των παραθύρων εκπαίδευσης σε ορισμούς και μη ορισμούς, που χρησιμοποιεί την έννοια του κεντροειδούς (centroid). Για κάθε όρο-στόχο εκπαίδευσης *sch_term*, η μέθοδος κατασκευάζει πρώτα ένα κεντροειδές, ένα ψευδο-κείμενο που παριστάνει το «μέσο όρο» όλων των παραθύρων του *sch_term*. Το κεντροειδές

κατασκευάζεται ως εξής. Υπολογίζεται πρώτα το βάρος $weight_{sch\_term}(w)$ κάθε μίας λέξεως *w* που εμφανίζεται σε όλα τα διαθέσιμα παράθυρα του *sch_term*, με τον ακόλουθο τύπο:

$$weight_{sch\_term(w)} = \frac{\log(Co(w, sch\_term)+1)}{\log(sf(w)+1)+\log(sf(sch\_term)+1)} * idf(w)$$

όπου *sf(sch_term)* είναι ο αριθμός εμφάνισης του *sch_term*, σε όλο το σώμα κειμένων, *sf(w)* είναι ο αριθμός εμφανίσεων της λέξης *w* στα έγγραφα από τα οποία προέρχονται τα παράθυρα του *sch_term* (δεν μετριέται απλά ο αριθμός των παραθύρων που την περιέχουν), και *Co(w, sch_term)* ο αριθμός παραθύρων του *sch_term* που περιέχουν την *w*. (Ο τύπος είναι ο ίδιος με εκείνον που χρησιμοποιούν οι Cui κ.ά., με τη διαφορά ότι εκείνοι μετρούν προτάσεις σε ένα σώμα κειμένων, αντί για παράθυρα σε έγγραφα που επέστρεψε μια μηχανή αναζήτησης για το *sch_term*.) Πριν την εφαρμογή της παραπάνω συνάρτησης έχουν ήδη αφαιρεθεί οι συχνές λέξεις (stop-words) και έχουν αποκοπεί οι καταλήξεις (stemming) από τις εναπομείναντες λέξεις. Το idf(w) (inverse document frequency) υπολογίζεται στην παρούσα εργασία, όπως και στην εργασία του Γαλάνη, ως εξής:

$$idf(w) = 1 + \log\frac{N}{df(w)}$$

όπου *N* ο συνολικός αριθμός των εγγράφων του British National Corpus (BNC), και *df(w)* ο αριθμός των εγγράφων του British National Corpus που περιέχουν τη λέξη *w*.

Μετά τον υπολογισμό των βαρών $weight_{sch\_term}(w)$, οι λέξεις *w* των οποίων το βάρος ξεπερνά το μέσο όρο των βαρών συν την τυπική απόκλιση περιλαμβάνονται στο κεντροειδές του *sch_term*. Στη συνέχεια, κάθε παράθυρο εκπαίδευσης του *sch_term* σημειώνεται ως παράθυρο ορισμού ή μη ορισμού, ανάλογα με το πόσο μοιάζει με το κεντροειδές του *sch_term*. Η μέτρηση της ομοιότητας γίνεται με το μέτρο του συνημιτόνου (cosine similarity) ως εξής:.

$$score = \frac{Vector_{παραθύρου} * Vector_{κεντροειδούς}}{|Vector_{παραθύρου}| * |Vector_{κεντροειδούς}|}$$

Όπου Vector-παραθύρου είναι ένα διάνυσμα με όλες τις λέξεις του προς αξιολόγηση παραθύρου, Vectort-κεντροειδούς είναι το διάνυσμα που περιέχει ολες τις λέξεις του κεντροειδούς.| Vector-παραθύρου | και | Vectort-κεντροειδούς | είναι τα μέτρα των παραπάνω διανυσμάτων.

Με τη μέθοδο που περιγράφτηκε παραπάνω, για κάθε διαθέσιμο παράθυρο εκπαίδευσης (που δεν γνωρίζουμε αν ανήκει στην κατηγορία ορισμός ή στην κατηγορία μη ορισμός) παράγουμε ένα αριθμό, ο οποίος δηλώνει την ομοιότητα του παραθύρου με το κεντροειδές του όρου-στόχου του παραθύρου. Όπως και στη μέθοδο του Γαλάνη, για την κατάταξη των διαθεσίμων παραθύρων εκπαίδευσης στις δύο κατηγορίες χρειάζονται δύο αριθμητικά κατώφλια t_- και t_+, με $t_- \le t_+$. Τα παράθυρα των οποίων η ομοιότητα με το αντίστοιχο κεντροειδές υπερβαίνει το t_+ κατατάσσονται ως παράθυρα ορισμού, τα παράθυρα των οποίων η ομοιότητα με το κεντροειδές είναι μικρότερη του t_- κατατάσσονται ως μη ορισμοί και τα υπόλοιπα παράθυρα δεν χρησιμοποιούνται κατά την εκπαίδευση της ΜΔΥ. Αν τα περισσότερα διαθέσιμα παράθυρα εκπαίδευσης ανήκουν στο τρίτο είδος, τότε τα παράθυρα που απομένουν για την εκπαίδευση της

ΜΔΥ είναι πολύ λίγα και κινδυνεύουμε η ΜΔΥ να εκπαιδευθεί σε ένα δείγμα παραθύρων που δεν θα είναι αντιπροσωπευτικό του συνολικού πληθυσμού παραθύρων που θα χρειαστεί να κατατάξει μετά την εκπαίδευσή της. Επίσης, όσο αυξάνουν τα παράθυρα του τρίτου είδους, τόσο μεγαλώνει ο αριθμός των ερωτήσεων, ορισμών από εγκυκλοπαίδειες και παραθύρων ιστοσελίδων που πρέπει να συγκεντρωθούν αρχικά, ώστε να απομείνει ικανός αριθμός παραθύρων εκπαίδευσης της ΜΔΥ. Επίσης, αν τα παράθυρα του πρώτου είδους είναι πολύ λιγότερα από εκείνα του δεύτερου ή αντίστροφα, τότε και πάλι δημιουργείται πρόβλημα κατά την εκπαίδευση της ΜΔΥ, επειδή η ΜΔΥ είναι πιθανόν να μάθει να κατατάσσει όλα τα παράθυρα είτε ως ορισμούς είτε ως μη-ορισμούς.

Για να συγκριθεί η μέθοδος του κεντροειδούς με τη μέθοδο του Γαλάνη, επαναλάβαμε ένα πείραμα της εργασίας του Γαλάνη, αυτή τη φορά με τη μέθοδο του κεντροειδούς. Ο Γαλάνης χρησιμοποίησε 130 όρους-στόχους, για τους οποίους διέθετε ορισμούς από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια, και τις αντίστοιχες ιστοσελίδες που επέστρεψε για κάθε όρο-στόχο η μηχανή αναζήτησης Altavista. Για κάθε όρο-στόχο, κράτησε τις 10 κορυφαίες ιστοσελίδες και τα 5 πρώτα παράθυρα των ιστοσελίδων, αφού τα παρακάτω παράθυρα είναι απίθανο να είναι ορισμοί. Στη συνέχεια επέλεξε τυχαία 400 από τα παράθυρα που προέκυψαν και τα κατέταξε στις δύο κατηγορίες (ορισμοί και μη-ορισμοί) τόσο χειρωνακτικά όσο και με το μέτρο ομοιότητας της εργασίας του. Θεωρώντας ότι οι σωστές κατηγορίες ήταν εκείνες της χειρωνακτικής κατάταξης, υπολόγισε την **ακρίβεια** (precision) και την **ανάκληση** (recall) του μέτρου ομοιότητας για τα παράθυρα ορισμού και μη-ορισμού ως εξής:

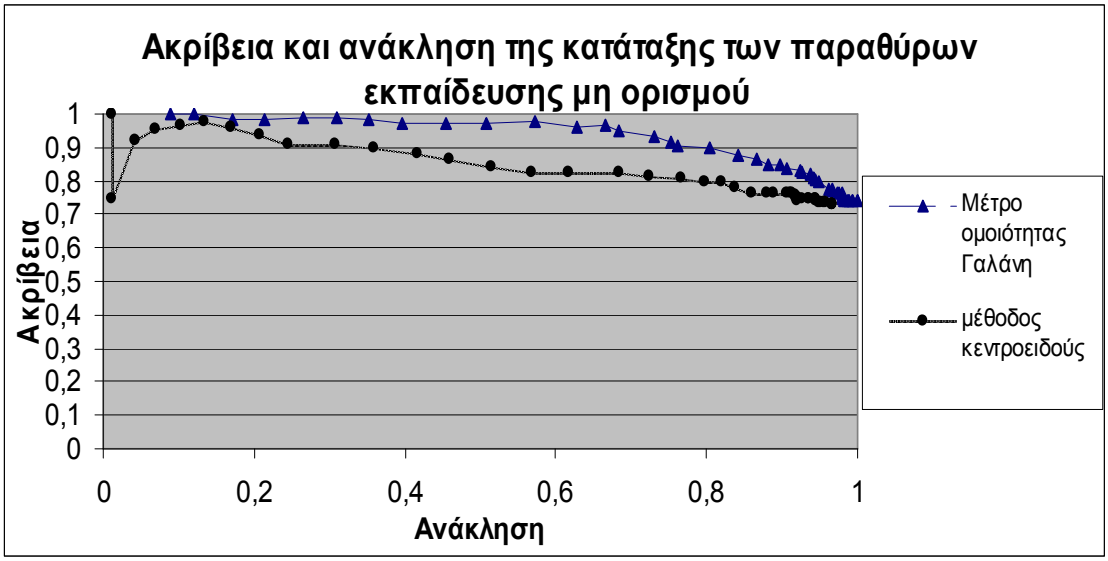
$$\text{Ακρίβεια}_{\text{ορισμών}} = \frac{TP}{TP + FP}$$

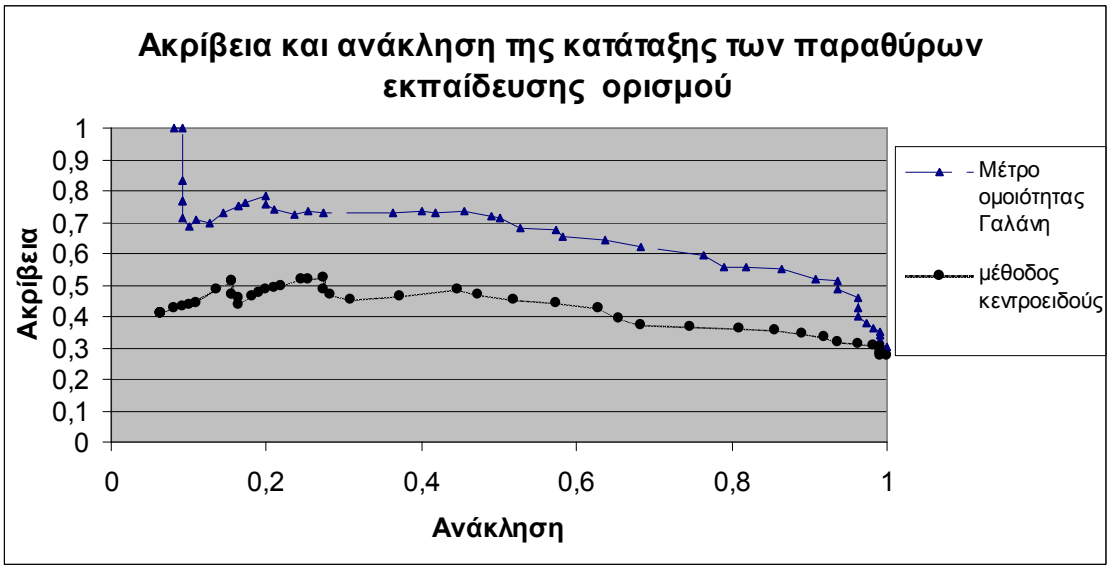
$$\text{Ανάκληση}_{\text{ορισμών}} = \frac{TP}{TP + FN}$$

$$\text{Ακρίβεια}_{\text{μη-ορισμών}} = \frac{TN}{TN + FN}$$

$$\text{Ανάκληση}_{\text{μη-ορισμών}} = \frac{TN}{TN + FP}$$

όπου TP (true positives) τα παράθυρα που κατατάσσονται σωστά ως ορισμοί, FP (false positives) τα παράθυρα που κατατάσσονται λανθασμένα ως ορισμοί, TN (true negatives) τα παράθυρα που κατατάσσονται σωστά ως μη ορισμοί και FN (false negatives) τα παράθυρα που κατατάσσονται λανθασμένα ως μη ορισμοί. Τα τέσσερα παραπάνω μεγέθη υπολογίστηκαν θέτοντας $t_- = t_+$ και μεταβάλλοντας την τιμή του κοινού κατωφλιού. Στα παρακάτω διαγραμματα φαινονται τα αποτελέσματα. Η συνάρτηση ομοιότητας του Γαλάνη είναι εμφανέστατα καλύτερη από τη μέθοδο του κεντροειδούς.





4.5 Επέκταση της μεθόδου του Γαλάνη με n-γράμματα

Μια δεύτερη εναλλακτική μέθοδος που δοκιμάσαμε για τον αυτόματο διαχωρισμό των παραθύρων εκπαίδευσης στις δύο κατηγορίες χρησιμοποιεί, όπως και η μέθοδος του Γαλάνη, μια συνάρτηση που υπολογίζει την ομοιότητα των παραθύρων εκπαίδευσης με ορισμούς που διαθέτουμε από εγκυκλοπαίδειες και γλωσσάρια. Η συνάρτηση είναι παρόμοια με αυτή του Γαλάνη αλλά λαμβάνει υπόψη της n-γράμματα (n-grams) λέξεων, δηλαδή ακολουθίες συνεχόμενων λέξεων μήκους n, ενώ η συνάρτηση του Γαλάνη λαμβάνει υπόψη της μόνο μεμονωμένες λέξεις. Η νέα συνάρτηση ομοιότητας $NGramSim_m(W,C)$ είναι η ακόλουθη:

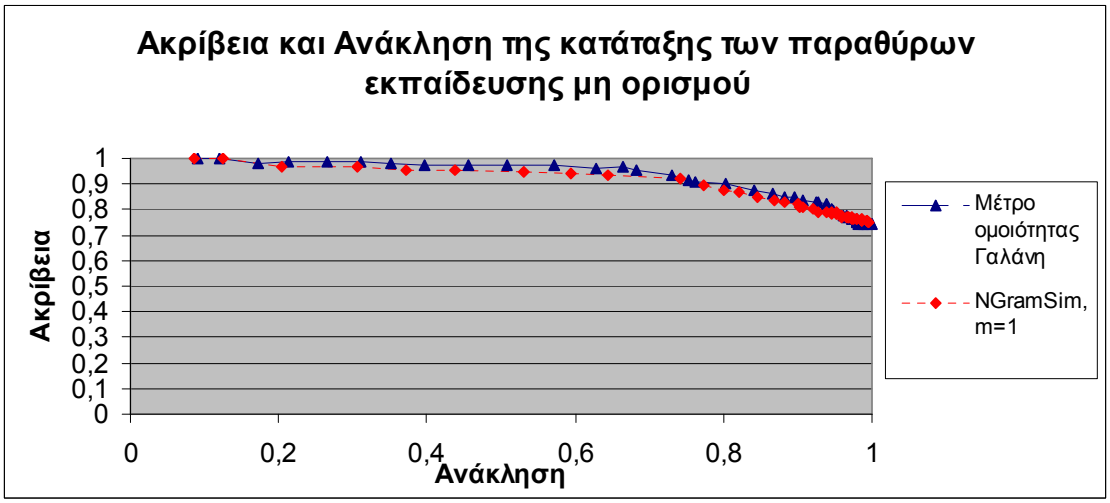
$$NGramSim\_m(W,C) = \frac{1}{m} * \sum_{n=1}^{m} \frac{\sum_{\gamma \in n-grams(C) \cap \gamma \in n-grams(W)} weight(\gamma)}{\sum_{\gamma \in n-grams(C)} weight(\gamma)}$$

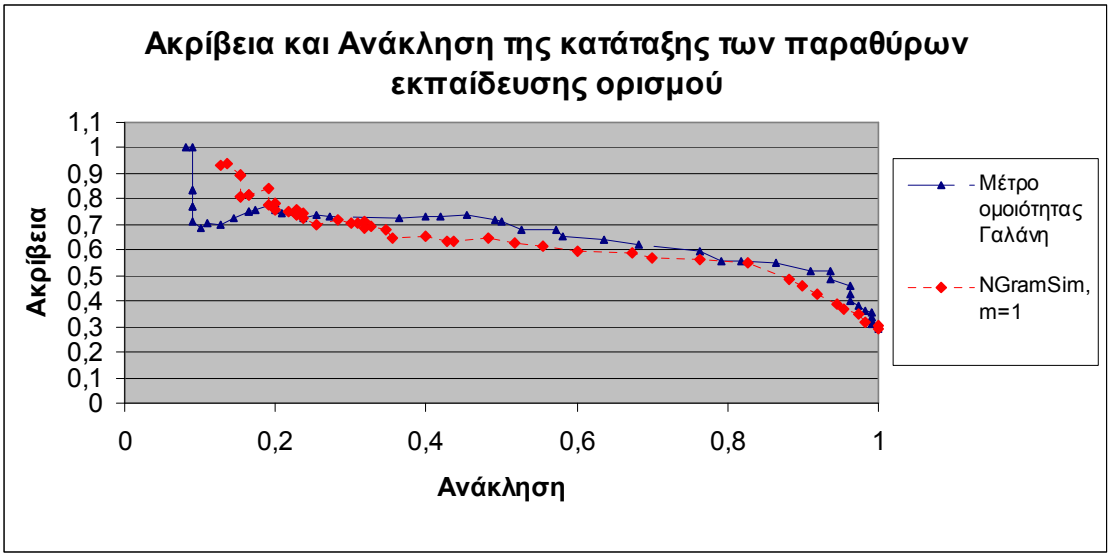
όπου m είναι το μέγιστο μήκος n-γραμμάτων που εξετάζουμε (η μέγιστη τιμή του n). Όπως και στη συνάρτηση του Γαλάνη, W είναι το παράθυρο εκπαίδευσης (μετά την αφαίρεση των συχνών λέξεων και την αποκοπή των καταλήξεων) και C το σύνολο των ορισμών του ιδίου όρου-στόχου που διαθέτουμε από εγκυκλοπαίδειες και γλωσσάρια. Με n-grams(C) συμβολίζουμε το σύνολο των n-γραμμάτων που εμφανίζονται στους ορισμούς του C και με n-grams(W) το σύνολο των n-γραμμάτων του παραθύρου W. Το weight(γ) είναι το βάρος κάθε n-γράμματος γ, που υπολογίζεται ως εξής:

weight(γ) = fdef(γ) * avgidf(γ)

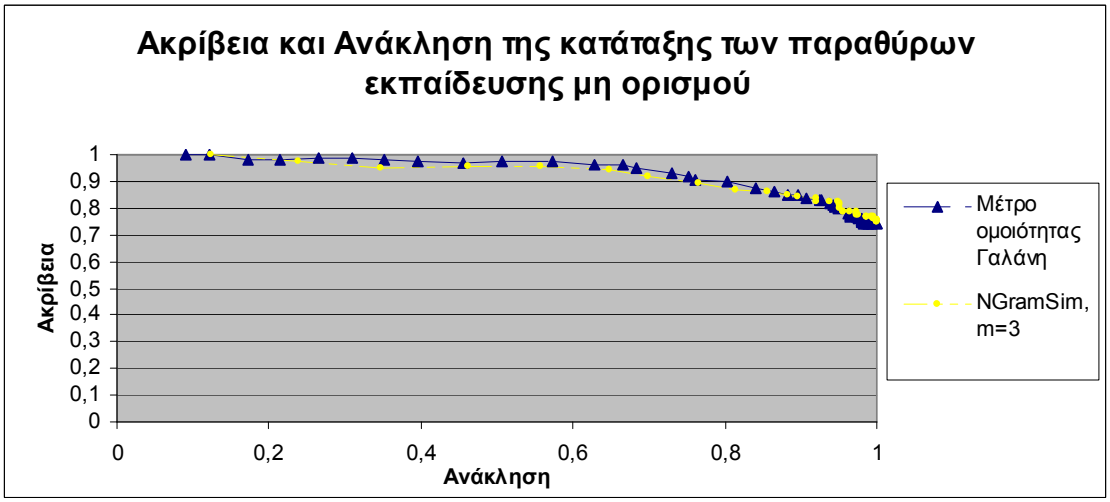
όπου fdef(γ) ο αριθμός των ορισμών του C στους οποίους εμφανίζεται το γ και avgidf(γ) ο μέσος όρος των idf(w) των λέξεων w του γ.

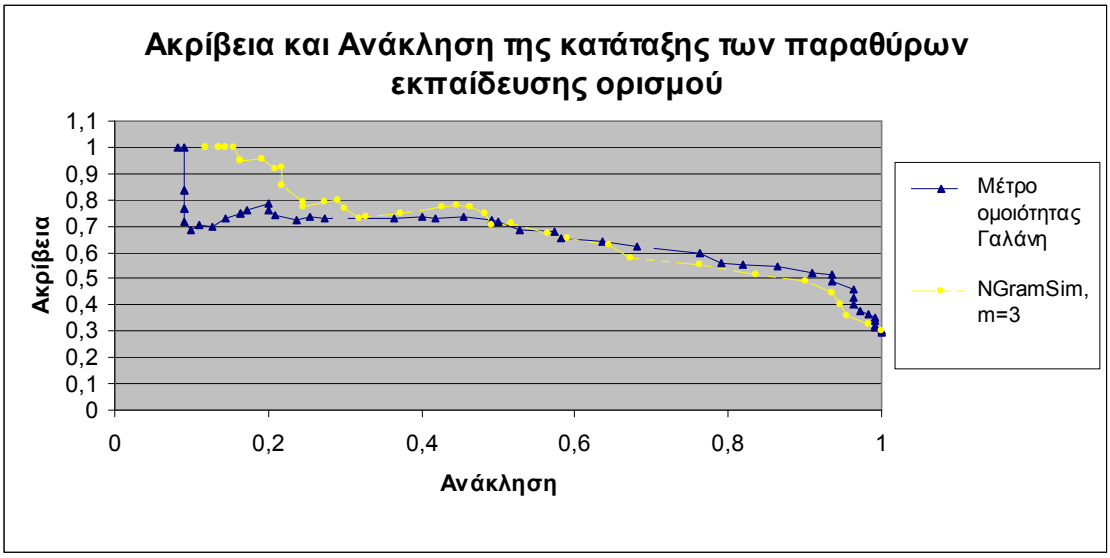
Όπως και με τη μέθοδο του κεντροειδούς, αξιολογήσαμε τη νέα συνάρτηση ομοιότητας στα ίδια 400 παράθυρα που είχε χρησιμοποιήσει και ο Γαλάνης για την αξιολόγησης της δικής του συνάρτησης ομοιότητας (βλ. ενότητα 4.4). Για m = 1, τα αποτελέσματα φαίνονται στα παρακάτω δύο διαγράμματα. Βλέπουμε ότι οι δυο συναρτήσεις έχουν σχεδόν την ίδια επίδοση, με την αρχική συνάρτηση του Γαλάνη να διατηρεί μια ελαφρά υπεροχή.



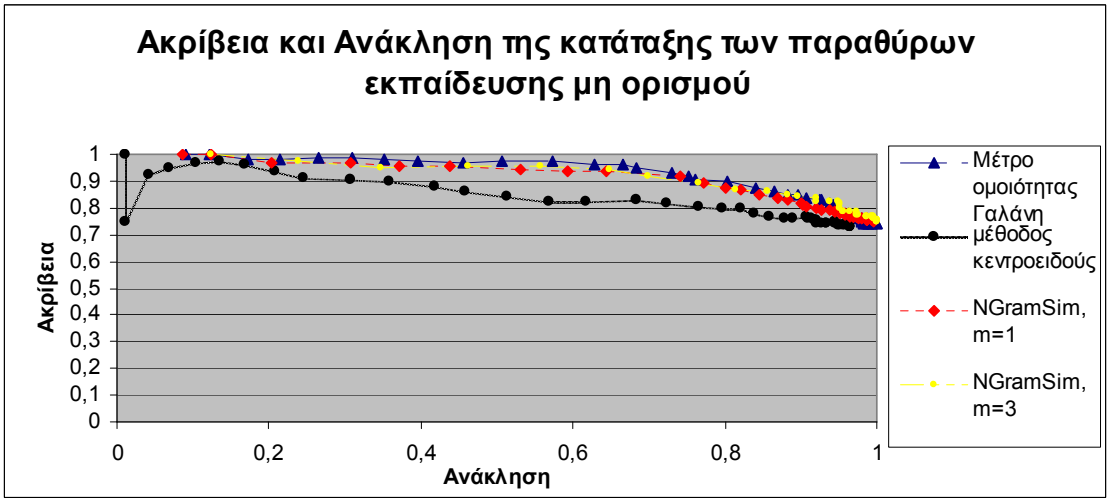


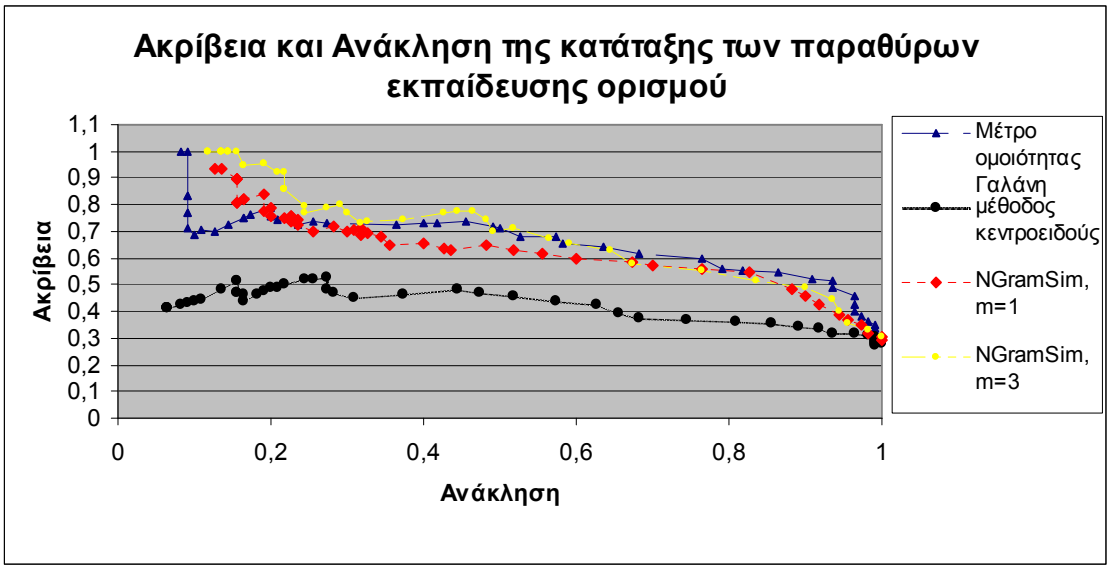
Για m = 3, τα αποτελέσματα φαίνονται στα παρακάτω δύο διαγράμματα. Όπως βλέπουμε, η νέα συνάρτηση είναι καλύτερη στην κατηγορία των ορισμών, ενώ στην κατηγορία των μη ορισμών δεν υπάρχει ουσιαστική διαφορά από τη συνάρτηση του Γαλάνη.





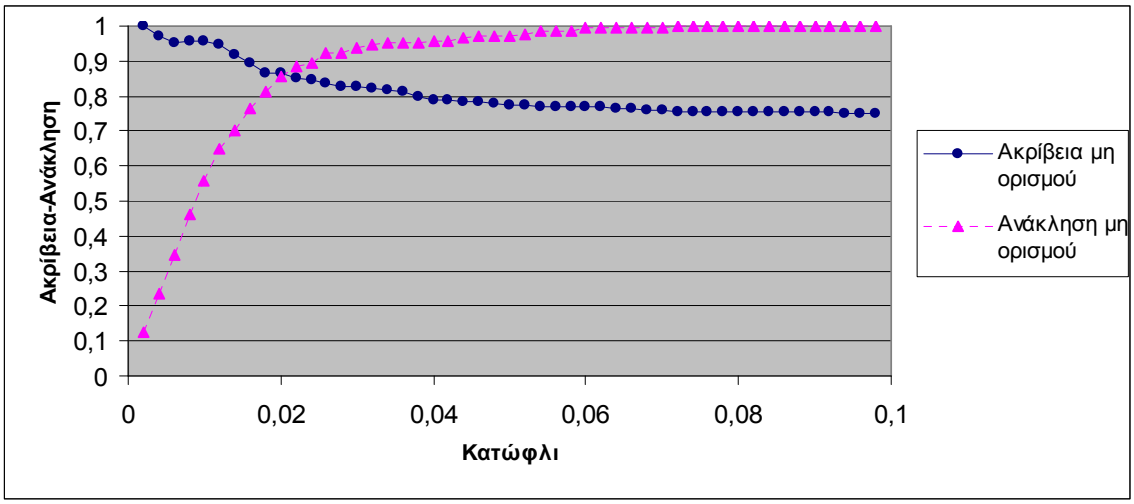
Στα παρακάτω δύο διαγράμματα φαίνονται συγκεντρωμένα τα αποτελέσματα όλων των αυτόματων μεθόδων διαχωρισμού των παραδειγμάτων εκπαίδευσης που δοκιμάσαμε.

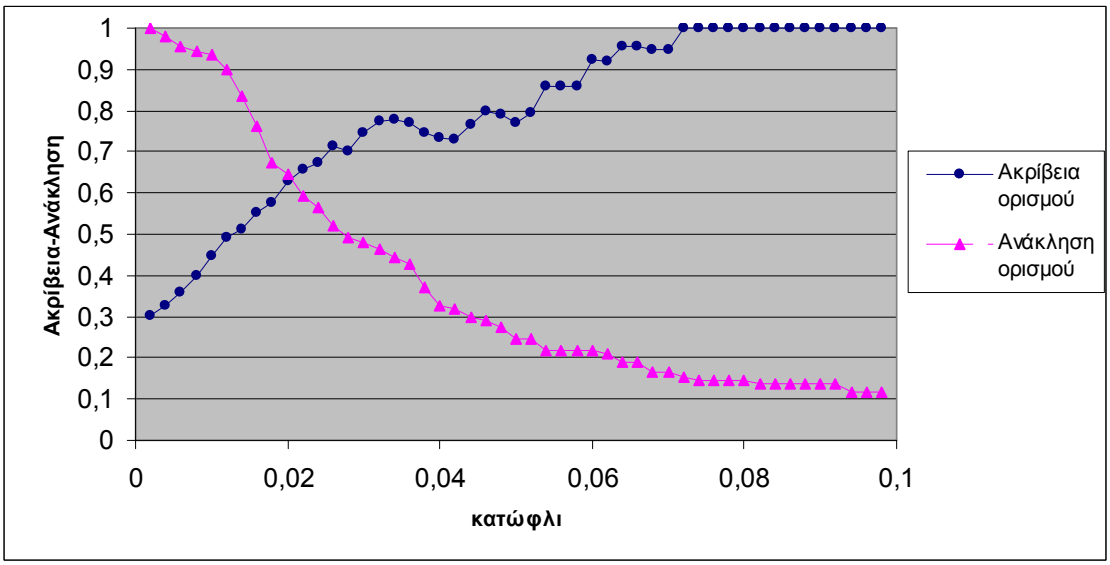




Όπως φαίνεται από τα διαγράμματα, η καλύτερη μέθοδος είναι η NGramSim με m = 3, η οποία και χρησιμοποιείται στο υπόλοιπο αυτής της εργασίας. Απομένει να επιλέξουμε τα κατώφλια t_+ και t_- που θα χρησιμοποιήσουμε με αυτή τη μέθοδο.

Από τα παρακάτω σχήματα, όπου χρησιμοποιείται η NGramSim με m = 3 και $t_+ = t_-$, γίνεται φανερό πως η συγκεκριμένη συνάρτηση ομοιότητας που επιλέξαμε επιτυγχάνει να διαχωρίσει τα παράθυρα εκπαίδευσης με ικανοποιητικό τρόπο, έτσι ώστε όσο μεγαλύτερη είναι η τιμή της συνάρτησης ομοιότητας, τόσο μεγαλύτερη να είναι η πιθανότητα να πρόκειται για παράθυρο ορισμού. Το ίδιο συμβαίνει και με τα παράθυρα μη-ορισμού, αφού όσο μικρότερη είναι η τιμή της συνάρτησης ομοιότητας τόσο μεγαλύτερη είναι η πιθανότητα να πρόκειται για παράθυρο μη ορισμού. Το πρόβλημα όμως που δημιουργείται είναι ότι όταν μεγαλώνει η ακρίβεια (precision) για οποιαδήποτε από τις δύο κατηγορίες, η αντίστοιχη ανάκληση (recall) μικραίνει. Αυτό σημαίνει ότι κρατώντας ως παράθυρα εκπαίδευσης της ΜΔΥ παράθυρα με υψηλή τιμή της συνάρτησης ομοιότητας, μπορούμε να είμαστε σχεδόν βέβαιοι ότι πρόκειται πραγματικά για παράθυρα ορισμού αλλά κρατάμε λίγα παραδείγματα εκπαίδευσης αυτής της κατηγορίας. Αντίστοιχα, κρατώντας παράθυρα με χαμηλή τιμή της συνάρτησης ομοιότητας, μπορούμε να είμαστε σχεδόν βέβαιοι ότι πρόκειται πραγματικά για παράθυρα μη-ορισμού αλλά και πάλι κρατάμε λίγα παραδείγματα της κατηγορίας αυτής. Επομένως, τα t_+ και t_- πρέπει να επιλεγούν έτσι, ώστε να είμαστε σχεδόν σίγουροι για τις κατηγορίες των παραδειγμάτων εκπαίδευσης (υψηλή ακρίβεια) διατηρώντας όσο το δυνατόν περισσότερα παράθυρα (υψηλή ανάκληση). Βασιζόμενοι στο πρώτο από τα παρακάτω δύο διαγράμματα, χρησιμοποιήσαμε στο υπόλοιπο της εργασίας $t_- = 0.016$, γιατί για εκείνη την τιμή του t_- έχουμε ικανοποιητική ακρίβεια μη ορισμών (περίπου 0.9) και οποιαδήποτε προσπάθεια για περαιτέρω βελτίωση της ακρίβειας με χρήση μικρότερου t_- ρίχνει πολύ απότομα την ανάκληση. Για το t_+ επιλέξαμε, βάσει του δευτέρου διαγράμματος, την τιμή 0.03 γιατί συνδυάζει σχετικά ικανοποιητική ακρίβεια (περίπου 0.7), ενώ και πάλι οποιαδήποτε προσπάθεια για περαιτέρω βελτίωση της ακρίβειας με χρήση μεγαλύτερου t_+ ρίχνει πολύ την ανάκληση.





5 Διανυσματική αναπαράσταση των παραθύρων

Όπως προαναφέρθηκε, η παρούσα εργασία βασίζεται στην προσέγγιση της Μηλιαράκη (2003), η οποία χρησιμοποιεί μια ΜΔΥ για να κατατάσσει τα παράθυρα των όρων-στόχων σε ορισμούς και μη-ορισμούς. Στην παρούσα εργασία, η ΜΔΥ εκπαιδεύεται με παράθυρα στα οποία η «ορθή» κατηγορία (ορισμός ή μη ορισμός) έχει σημειωθεί χρησιμοποιώντας τη μέθοδο NGramSim της προηγούμενης ενότητας. Κατά την εκπαίδευση και τη χρήση της ΜΔΥ, τα παράθυρα παριστάνονται ως διανύσματα ιδιοτήτων. Ακολουθώντας την προσέγγιση της Μηλιαράκη, χρησιμοποιούμε 22 χειρωνακτικά επιλεγμένες ιδιότητες, στις οποίες προσθέτουμε m ιδιότητες που παράγονται αυτόματα από τα δεδομένα εκπαίδευσης. Στα πειράματά μας, το m κυμαινόταν από 100 ως 500.

5.1 Χειρωνακτικά επιλεγμένες ιδιότητες

Οι χειρωνακτικά επιλεγμένες ιδιότητες που χρησιμοποιούμε είναι οι ίδιες με αυτές της εργασίας της Μηλιαράκη, δηλαδή οι ακόλουθες:
1. Η **κατάταξη (ranking) του κειμένου**, δηλαδή η σειρά με την οποία επέστρεψε η μηχανή αναζήτησης το έγγραφο από το οποίο προέρχεται το παράθυρο. Η ιδιότητα αυτή είναι χρήσιμη, διότι συνήθως οι ζητούμενοι ορισμοί βρίσκονται στα κορυφαία κείμενα που επιστρέφει η μηχανή αναζήτησης.
2. Η **θέση του παραθύρου** μέσα στο έγγραφο. Η ιδιότητα δείχνει αν πρόκειται για το πρώτο, δεύτερο, κ.ο.κ. παράθυρο του εγγράφου. Χρησιμοποιείται διότι συνήθως τα παράθυρα ορισμού εντοπίζονται στην αρχή των κειμένων.
3. Το **πλήθος των κοινών λέξεων του παραθύρου.** Η ιδιότητα αυτή δείχνει τι ποσοστό από τις 20 λέξεις που εμφανίζονται πιο συχνά στα παράθυρα του συγκεκριμένου όρου-στόχου περιλαμβάνονται στο συγκεκριμένο παράθυρο. Ουσιαστικά οι 20 αυτές λέξεις συνιστούν ένα απλοϊκό κεντροειδές (βλ. ενότητα 4.4 και η ιδιότητα αυτή μετρά με απλοϊκό τρόπο την ομοιότητα του συγκεκριμένου παραθύρου με το κεντροειδές. Κατά τον υπολογισμό της τιμής αυτής της ιδιότητας, αγνοούνται οι κοινές λέξεις των αγγλικών (stop-words). Χρησιμοποιείται πάλι η λίστα συχνών λέξεων του BNC της ενότητας 4.3.

Οι υπόλοιπες χειρωνακτικά επιλεγμένες ιδιότητες είναι δυαδικές, με τιμή 1 αν η αντίστοιχη φράση περιέχεται στο παράθυρο και τιμή 0 αν δεν περιέχεται.

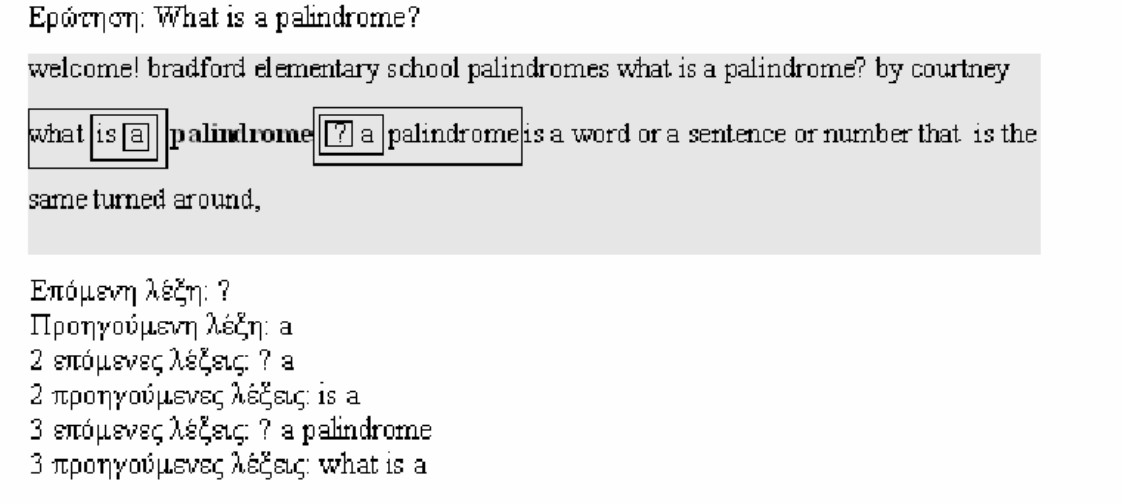
4. Η φράση **"such <…> as *όρος-στόχος*"**
Παράδειγμα : *"such antibiotics as amoxicillin"*
5. Η φράση **"*όρος-στόχος* and other <…>"**
Παράδειγμα : "*broken bones and other injuries"*
6. Η φράση **"*όρος-στόχος* or other <…>"**
Παράδειγμα : *"cats or other animals"*
7. Η φράση **"especially *όρος-στόχος*"**
Παράδειγμα : *"some plastics especially Teflon"*
8. Η φράση **"including *όρος-στόχος*"**
Παράδειγμα : *"some amphibians including frog"*
9. **Παρενθέσεις μετά** τον όρο-στόχο
Παράδειγμα : *"sodium chloride (salt)"*
10. **Παρενθέσεις πριν** τον όρο-στόχο
Παράδειγμα : *"(salt) sodium chloride"*
11. Η φράση **"*όρος-στόχος* is a"**
Ακριβέστερα αναζητείται η πληρέστερη φράση της μορφής "*όρος* is/are/was/were a/an/the <…>"
Παράδειγμα : *"Galileo was a great astronomer"*
12. Η φράση "***όρος-στόχος* , a/an/the <…>** "
Παράδειγμα : *"amoxicillin, an antibiotic"*
13. Η φράση **"*όρος-στόχος*, which is/was/are/were <…>"**
Παράδειγμα : *"tsunami which is a giant wave"*
14. Η φράση **"*όρος-στόχος* like <…>"**
Παράδειγμα : *"antibiotics like amoxicillin"*
15. Η φράση **"*όρος-στόχος*, <…>, is/was/are/were"**
Παράδειγμα : *"amphibians, like frogs, are animals that can live both on land and in water"*
16. Η φράση **"*όρος-στόχος* or <…>"**
Παράδειγμα : *"autism or some other type of disorder"*
17. Ένα από τα ρήματα **"can", "refer", "have"** μετά τον όρο **(3 ιδιότητες)**

Παράδειγμα : *"Amphibians can live both on land and in water"*
18. Ένα από τα ρήματα **"called", "known as","defined"** πριν τον όρο **(3 ιδιότητες)**
Παράδειγμα : *" The giant wave known as tsunami"*

5.2 Αυτόματα επιλεγόμενες ιδιότητες

Οι αυτόματα επιλεγόμενες ιδιότητες είναι όλες δυαδικές. Η κάθε μία δείχνει αν το παράθυρο περιέχει (τιμή 1) ή όχι (τιμή 0) κάποιο συγκεκριμένη φράση. Οι ιδιότητες αυτές επιλέγονται όπως στην εργασία της Μηλιαράκη, δηλαδή με τη διαδικασία που συνοψίζεται παρακάτω.

- Δημιουργείται ένα κενό σύνολο φράσεων.
- Για κάθε παράθυρο εκπαίδευσης εξάγονται οι φράσεις που αποτελούνται από: την επόμενη λέξη του όρου-στόχου, τις 2 επόμενες λέξεις του όρου-στόχου, τις 3 επόμενες λέξεις του όρου-στόχου, την προηγουμένη λέξη του όρου-στόχου, τις 2 προηγούμενες λέξεις του όρου-στόχου, τις 3 προηγούμενες λέξεις του όρου-στόχου. Η διαδικασία εξαγωγής των φράσεων φαίνεται στο παρακάτω σχήμα, το οποίο προέρχεται από την εργασία του Γαλάνη. Οι φράσεις που αποτελούνται από λέξεις που έπονται του όρου-στόχου θεωρούνται διαφορετικές από τις φράσεις που αποτελούνται από λέξεις που προηγούνται του όρου-στόχου.
- Για κάθε φράση που εξάγεται, ελέγχεται αν υπάρχει ήδη στο σύνολο φράσεων. Αν υπάρχει τότε απλά αυξάνεται ένας μετρητής εμφανίσεων της κατά 1, αλλιώς εισάγεται στο σύνολο φράσεων και ο μετρητής της αρχικοποιείται στην τιμή 1.



- Όταν ολοκληρωθεί η επεξεργασία των παραθύρων εκπαίδευσης, διαγράφονται από το σύνολο φράσεων οι φράσεις που έχουν αριθμό εμφανίσεων μικρότερο από ένα κατώφλι.
- Ύστερα υπολογίζεται η ακρίβεια (precision) για κάθε φράση του συνόλου φράσεων, επιλέγονται οι m φράσεις με τις υψηλότερες τιμές ακρίβειας και για κάθε μία από αυτές δημιουργείται μία δυαδική ιδιότητα.

Η ακρίβεια μιας φράσης υπολογίζεται στα παράθυρα εκπαίδευσης, ως ο λόγος του αριθμού παραθύρων ορισμού που περιέχουν τη φράση εμφανιζόμενη ακριβώς πριν η μετά τον

όρο-στόχο (ανάλογα με το αν η φράση αποτελείται από λέξεις που προηγούνταν ή έπονταν του όρου-στόχου) δια το συνολικό αριθμό παραθύρων (ορισμού και μη-ορισμού) που την περιέχουν εμφανιζόμενη ακριβώς πριν η μετά τον όρο-στόχο αντίστοιχα.

6. Πειράματα-Αξιολόγηση συστήματος

6.1 Μέτρα αξιολόγησης

Για την αξιολόγηση ενός συστήματος ερωταποκρίσεων συνήθως χρησιμοποιούνται 2 μέτρα: ο αριθμός των ερωτήσεων που απάντησε σωστά το σύστημα και η μέση αντίστροφη κατάταξη (**M**ean **R**eciprocal **R**ank) των απαντήσεων. Τα μέτρα αυτά εξηγούνται παρακάτω.

Ένα σύστημα ερωταποκρίσεων, όπως ήδη έχει αναφερθεί, αξιολογεί όλες τις υποψήφιες απαντήσεις μίας ερώτησης (στην περίπτωσή μας, τα παράθυρα του όρου-στόχου που περιλαμβάνονται στα έγγραφα που επέστρεψε η μηχανή αναζήτησης). Ακολουθώντας τους κανονισμούς που ίσχυαν για τις ερωτήσεις ορισμού στους διαγωνισμούς TREC 2000 και 2001, το σύστημα αυτής της εργασίας επιστρέφει στο χρήστη μία ταξινομημένη λίστα με τα 5 παράθυρα του όρου-στόχου (υποψήφιες απαντήσεις) που θεωρεί πως έχουν την μμεγαλύτερη πιθανότητα να περιέχουν αποδεκτούς σύντομους ορισμούς του όρου-στόχου. Αν τουλάχιστον ένα από τα 5 αυτά παράθυρα περιέχει όντως αποδεκτό ορισμό, τότε θεωρείται ότι σύστημα απάντησε σωστά την ερώτηση. Προκειμένου να είναι δυνατή η σύγκριση με τα αποτελέσματα της εργασίας του Γαλάνη (2004), υπολογίζουμε επίσης σε πόσες από τις ερωτήσεις το σύστημα κατάφερε να επιστρέψει παράθυρο με αποδεκτό ορισμό στην πρώτη θέση της λίστας των 5 παραθύρων.

Η μέση αντίστροφη κατάταξη (MRR) μετρά το κατά πόσο ένα σύστημα επιστρέφει σωστές απαντήσεις στις υψηλές θέσεις της λίστας των απαντήσεων που παρουσιάζει στο χρήστη (στην περίπτωσή μας, το κατά πόσον υπάρχουν αποδεκτά παράθυρα ορισμών στις κορυφαίες θέσεις της λίστας των 5 παραθύρων). Ο υπολογισμός του MRR γίνεται ως εξής. Σε κάθε ερώτηση αξιολόγησης, το σύστημα λαμβάνει ένα βαθμό, ο οποίος είναι ίσος με 1 δια τη θέση της πρώτης σωστής απάντησης (1-5). Αν καμία από τις επιστρεφόμενες απαντήσεις δεν είναι σωστή, ο βαθμός που λαμβάνει το σύστημα για την ερώτηση είναι 0. Οι βαθμοί που έλαβε το σύστημα σε όλες τις ερωτήσεις αξιολόγησης αθροίζονται και διαιρούνται με το συνολικό αριθμό των ερωτήσεων αξιολόγησης. Το αποτέλεσμα είναι η τιμή του MRR.

.

6.2 Δεδομένα εκπαίδευσης και αξιολόγησης

Στα πειράματα των επομένων ενοτήτων χρησιμοποιήσαμε ως δεδομένα εκπαίδευσης 900 όρους-στόχους που συλλέχθηκαν τυχαία από το ευρετήριο μιας ηλεκτρονικής εγκυκλοπαίδειας (http://www.encyclopedia.com), καθώς και τα 5 πρώτα παράθυρα αυτών των όρων-στόχων που περιλαμβάνονταν στις 10 κορυφαίες ιστοσελίδες που επέστρεψε για τους όρους-στόχους η μηχανή αναζήτησης AltaVista Τα παράθυρα αυτά σημειώθηκαν ως παραδείγματα ορισμών ή μη ορισμών χρησιμοποιώντας το μέτρο NGramSim με m = 3 (βλ. ενότητα 4.5) και τα διανύσματά τους (βλ. κεφάλαιο 5) χρησιμοποιήθηκαν για την εκπαίδευση της ΜΔΥ.

Ως δεδομένα αξιολόγησης χρησιμοποιήσαμε 200 όρους-στόχους, επιλεγμένους τυχαία από το ευρετήριο της ίδιας ηλεκτρονικής εγκυκλοπαίδειας. Κατά την αξιολόγηση του συστήματος, για κάθε έναν από τους 200 όρους-στόχους συλλέξαμε τα κορυφαία 10 έγγραφα που επέστρεψε η μηχανή αναζήτησης, και από κάθε έγγραφο τα 5 πρώτα παράθυρα του όρου-στόχου. Στη συνέχεια η ΜΔΥ κατέταξε τα παράθυρα του όρου-στόχου στις δύο κατηγορίες (ορισμοί και μη ορισμοί) και θεωρήσαμε ως απάντηση του συστήματος για το συγκεκριμένο όρο-

στόχο τα 5 ή 1 (ανάλογα με το μέτρο αξιολόγησης, βλ. ενότητα 6.1) παράθυρα για τα οποία η ΜΔΥ ήταν περισσότερο βέβαιη ότι ανήκαν στην κατηγορία των ορισμών. Τα παράθυρα είχαν καταταγεί και χειρωνακτικά στις δύο κατηγορίες και οι σωστές κατηγορίες των παραθύρων θεωρήθηκε πως ήταν αυτές που είχαν σημειώσει οι κριτές.

Οι ορισμοί όρων που περιλαμβάνονται σε ηλεκτρονικές εγκυκλοπαίδειες ή λεξικά μπορούν να βρεθούν εύκολα, για παράδειγμα με το μηχανισμό define του Google (βλ. ενότητα 4.3). Υπάρχουν, όμως, πάντα όροι-στόχοι (π.χ. ονόματα προσώπων της επικαιρότητας, πρόσφατοι τεχνικοί όροι) που δεν περιλαμβάνονται σε εγκυκλοπαίδειες ή λεξικά αλλά μπορούν να βρεθούν σε γενικής φύσεως ιστοσελίδες (π.χ. το απόσπασμα «He said that gasohol, a mixture of gasoline and ethanol, has been great for his business.» από την ιστοσελίδα μιας ηλεκτρονικής εφημερίδας) και σκοπός αυτής της εργασίας ήταν να αναπτύξει τεχνικές για τον εντοπισμό ορισμών αυτού του είδους. Από την άλλη πλευρά, όμως, η αξιολόγηση του συστήματος ήταν δύσκολο να γίνει με όρους-στόχους για τους οποίους δεν υπάρχουν ορισμοί σε ηλεκτρονικές εγκυκλοπαίδειες ή λεξικά, γιατί σε αυτήν την περίπτωση πολλοί από τους όρους θα ήταν άγνωστοι στους κριτές και δεν θα είχαμε ορισμούς-υποδείγματα να τους δώσουμε. Για το λόγο αυτό οι 200 όροι-στόχοι της αξιολόγησης επιλέχθηκαν και αυτοί από το ευρετήριο μιας ηλεκτρονικής εγκυκλοπαίδειας. Προκειμένου, όμως, να αξιολογήσουμε το κατά πόσον το σύστημά μας καταφέρνει να ανακαλύψει ορισμούς που δεν περιλαμβάνονται σε ηλεκτρονικές εγκυκλοπαίδειες και λεξικά, αγνοήσαμε από τις ιστοσελίδες που επέστρεψε για τους 200 όρους-στόχους η μηχανή αναζήτησης εκείνες που προέρχονταν από ηλεκτρονικές εγκυκλοπαίδειες και λεξικά (π.χ. Wikipedia,), φροντίζοντας, όμως, πάντα να κρατάμε 10 ιστοσελίδες ανά όρο-στόχο.

6.3 Πειράματα με μεταβλητό αριθμό αυτόματα αποκτηθέντων ιδιοτήτων

Στα πειράματα αυτά χρησιμοποιήθηκαν τα δεδομένα εκπαίδευσης και αξιολόγησης της προηγούμενης ενότητας. Κάθε παράθυρο παριστανόταν με τις 22 χειρωνακτικά επιλεγμένες ιδιότητες (ενότητα 5.1) και m αυτόματα επιλεγμένες ιδιότητες (ενότητα 5.2), όπου το m είχε τις τιμές 100, 200, 300, 400, και 500. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

Αυτόματα αποκτηθείσες ιδιότητες	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρέπονταν 5 απαντήσεις ανά ερώτηση (σε παρένθεση το MRR)	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρεπόταν μία απάντηση ανά ερώτηση
100	41% (0.259)	18%
200	51% (0.325)	23%
300	64% (0.407)	30%
400	63% (0.403)	30%
500	62% (0.411)	31%

Για λόγους σύγκρισης, στον παρακάτω πίνακα φαίνονται τα αντίστοιχα αποτελέσματα όταν το σύστημα εκπαιδεύεται στα δεδομένα των TREC 2000 και 2001 (ενότητα 4.1). Στην περίπτωση αυτή, για την εκπαίδευση του συστήματος χρησιμοποιούνται 160 όροι-στόχοι, τα κορυφαία 10 έγγραφα από τα 50 που παρέχουν οι διοργανωτές του TREC, και από κάθε έγγραφο τα 5 πρώτα παράθυρα. Η αξιολόγηση γίνεται όπως προηγουμένως, δηλαδή με τους 200 όρους-στόχους της ενότητας 6.3 και τα παράθυρα των ιστοσελίδων που επιστρέφει για τους 200 όρους-στόχους η μηχανή αναζήτησης.

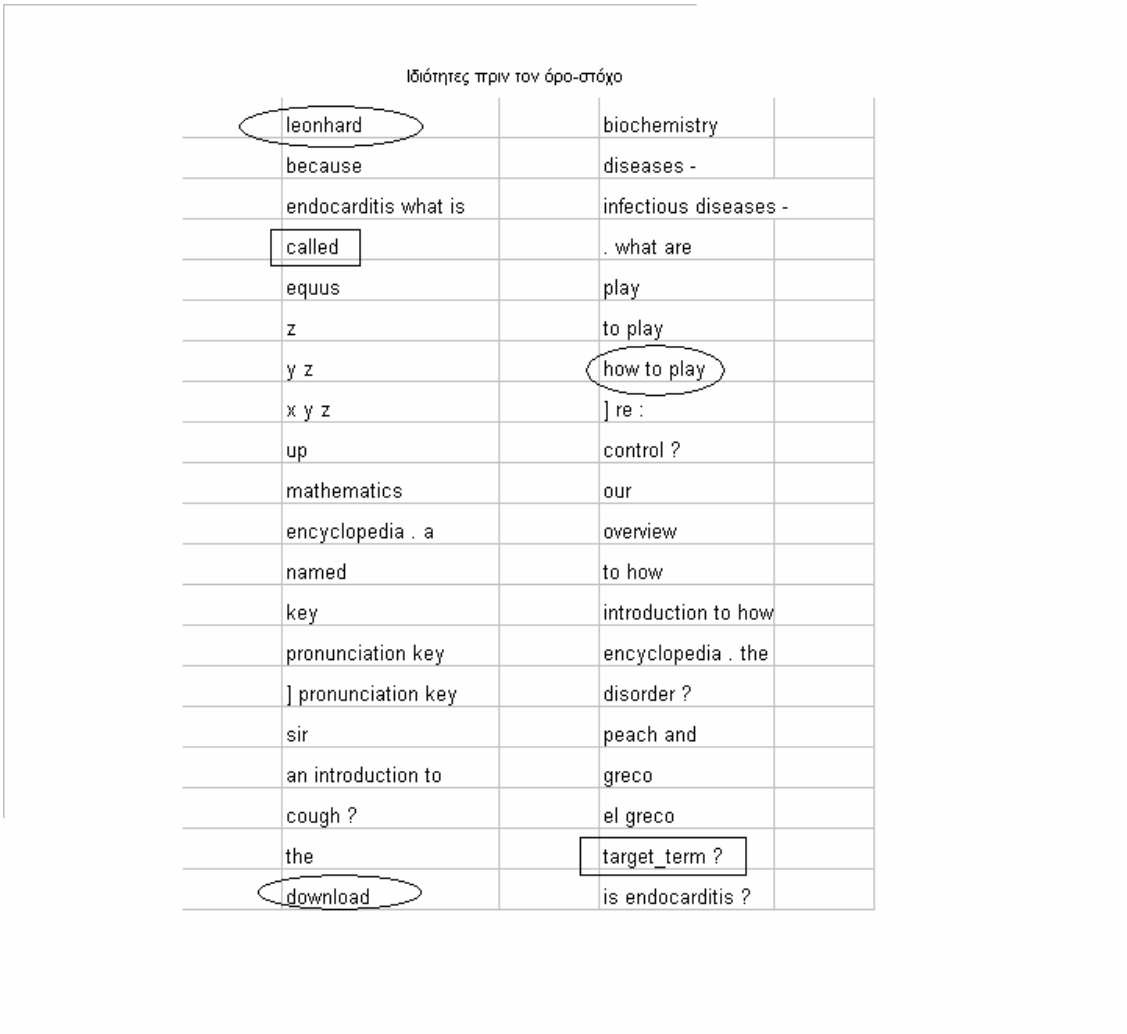
Αυτόματα αποκτηθείσες ιδιότητες	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρέπονταν 5 απαντήσεις ανά ερώτηση (σε παρένθεση το MRR)	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρεπόταν μία απάντηση ανά ερώτηση
100	54% (0,34)	20%
200	51% (0,31)	19%
300	57% (0,34)	20%
400	57% (0,35)	20%
500	54% (0,32)	18%

Από τον πρώτο πίνακα βλέπουμε πως το σύστημα μας έχει τα καλύτερα αποτελέσματα όταν χρησιμοποιούμε 300 ιδιότητες, αφού έχει τα υψηλοτέρα ποσοστά αλλά και ικανοποιητικό MRR. Όταν επιτρέπουμε 5 απαντήσεις ανά ερώτηση βλέπουμε ότι στις 100 και 200 ιδιότητες το σύστημα που είναι εκπαιδευμένο με τα δεδομένα του TREC επιτυγχάνει καλύτερα αποτελέσματα από το σύστημα μας αλλά από τις 300 και μετά είναι σταθερά πίσω. Αυτό οφείλεται ουσιαστικά γιατί κατά την εκπαίδευση του το σύστημα "TREC" χρησιμοποιεί επιβλεπόμενη μάθηση και για αυτό το λόγο είναι λογικό οι πρώτες ιδιότητες να είναι πολύ χρήσιμες. Καθώς όμως μεγαλώνει ο αριθμός τους αναγκάζεται να χρησιμοποιήσει ιδιότητες που συνεισφέρουν θόρυβο. Εν αντίθεση το σύστημα μας ακριβώς επειδή χρησιμοποιεί μη επιβλεπόμενη μάθηση χρειάζεται ένα μεγαλύτερο αριθμό ιδιοτήτων ώστε να συμπεριλαμβάνονται μέσα οι χρήσιμες ιδιότητες.Και εδώ βέβαια βλέπουμε πως μετά τις 300 ιδιότητες δεν μπορεί να βελτιωθεί άλλο. Στα αποτελέσματα από το πείραμα με τη μία απάντηση ανά ερώτηση δεν παρατηρείται κάτι διαφορετικό.

Στα επόμενα σχήματα φαίνονται οι 100 αυτόματα επιλεγμένες ιδιότητες. Μέσα σε τετράγωνο βρίσκονται οι λέξεις-ιδιότητες που είναι σημαντικές ενώ μέσα σε κύκλο αυτές που συνεισφέρουν θόρυβο. Σε αυτό το παράδειγμα οι ιδιότητες είναι χωρισμένες σε 40 ιδιότητες πριν τον όρο-στόχο και 60 μετά από αυτόν. Ο αριθμός των ιδιοτήτων πριν και μετά τον όρο-στόχο καθορίζεται αποκλειστικά κατά τον έλεγχο της ακρίβειας (ενότητα 5.2)

Ιδιότητες μετά τον όρο-στόχο

and his	born :	refers to any
is a word	is a condition	cream
was born in	was the	(yeast
, is	? birth	is a substance
is the study	were	plants
& # 160	is a highly	is the practice
what is endocarditis	is a type	orchestra
n	, also	pages (
insurance	, also known	is used to
(also	as "	trees
history	as " a	is an inflammation
is also	baily	clearinghouse
refers	baily '	about ?
refers to	baily ' s	condenser
born	kill	condenser microphon
el	kill pine	breaker
was born	kill pine flooring	el greco
is found	bazaar	? endocarditis
? whooping	is the science	is an infection
? whooping cough	? birth control	and pseudoephedrine



Σημειώνουμε ότι αντίστοιχα πειράματα είχαν γίνει στην εργασία του Γαλάνη αλλά μόνο για 200 αυτόματα αποκτηθείσες ιδιότητες και με τη χρήση του δικού του μέτρου ομοιότητας (ενότητα 4.3) για την κατάταξη των παραθύρων εκπαίδευσης στις δύο κατηγορίες.

6.4 Πειράματα με μεταβλητό αριθμό ερωτήσεων εκπαίδευσης

Στα πειράματα αυτά η αξιολόγηση του συστήματος γινόταν όπως ακριβώς στην προηγούμενη ενότητα (200 όροι-στόχοι) αλλά η εκπαίδευσή του έγινε χρησιμοποιώντας μεταβαλλόμενο ποσοστό των δεδομένων εκπαίδευσης. Πιο συγκεκριμένα, χρησιμοποιήθηκαν κατά την εκπαίδευση του συστήματος οι m πρώτοι από τους 900 όρους-στόχους και τα αντίστοιχα παράθυρά τους, με το m να λαμβάνει τις τιμές: 500, 600, 700, 800, και 900. Σε όλα τα πειράματα αυτής της ενότητας χρησιμοποιήθηκαν οι 22 χειρωνακτικά επιλεγμένες και 300 αυτόματα επιλεγμένες ιδιότητες, επειδή για 300 αυτόματα επιλεγμένες ιδιότητες είχαμε τα καλύτερα αποτελέσματα στα πειράματα της προηγούμενης ενότητας. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

Αριθμός ερωτήσεων του σώματος εκπαίδευσης	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά όταν επιτρέπονταν 5 απαντήσεις ανά ερώτηση (σε παρένθεση το MRR)	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρεπόταν μία απάντηση ανά ερώτηση
500 ερωτήσεις	66% (0.465)	32%
600 ερωτήσεις	64% (0.418)	27%
700 ερωτήσεις	64% (0.453)	31%
800 ερωτήσεις	63% (0.439)	29%
900 ερωτήσεις	64% (0.407)	26%

Όπως παρατηρούμε το σύστημά μας έχει την καλύτερη απόδοση όταν χρησιμοποιούμε 500 ερωτήσεις στο σώμα εκπαίδευσης, πετυχαίνοντας ταυτόχρονα υψηλό ποσοστό σωστών απαντήσεων αλλά και ικανοποιητικό MRR. Το ενδιαφέρον όμως σε αυτά τα αποτελέσματα είναι τα σχετικά χαμηλά ποσοστά του συστήματος όταν επιτρέπεται μια μόνο απάντηση ανά ερώτηση. Παρόλο που το σύστημα μας έχει καλύτερη μέθοδο κατασκευής παραδειγμάτων και η μηχανική μάθηση χρησιμοποιείται πιο αποτελεσματικά (LIBSVM , RBF) από την υλοποίηση του Γαλάνη παρατηρούμε πως τα αποτελέσματα του συστήματος του (Γαλανης 2004) είναι αρκετά καλύτερα. Αυτή η διαφορά οφείλεται στην απουσία ιστοσελίδων από εγκυκλοπαίδειες στο σώμα αξιολόγησης, η οποία μπορεί να μειώνει τα ποσοστά του συστήματος αλλά κάνει την αξιολόγηση πιο αντικειμενική, μια και σκοπός μας είναι να δημιουργήσουμε ένα σύστημα το οποίο ανεξαρτήτως από την ύπαρξη έτοιμων ορισμών θα μπορεί να τον εξάγει από τα κείμενα.

6.5 Σύγκριση με μεθόδους που δεν χρησιμοποιούν μηχανική μάθηση

Στον παρακάτω πίνακα παρατίθενται, πάλι για λόγους σύγκρισης, τα αποτελέσματα από την αξιολόγηση (με τους 200 όρους-στόχους) τριών απλοϊκών μεθόδων που δεν χρησιμοποιούν μηχανική μάθηση. Η πρώτη μέθοδος («centroid») χρησιμοποιεί τη συνάρτηση ομοιότητας κεντροειδούς (ενότητα 4.4) για να αξιολογήσει τα παράθυρα καθενός από τους 200-όρους στόχους και επιστρέφει τα 5 ή 1 (ανάλογα με τον τρόπο αξιολόγησης) παράθυρα που μοιάζουν περισσότερο με το κεντροειδές του όρου-στόχου. Η δεύτερη μέθοδος («baseline1») επιστρέφει τυχαία 5 ή 1 (ανάλογα με τον τρόπο αξιολόγησης) από τα παράθυρα του όρου-στόχου. Η τρίτη μέθοδος («baseline2») επιστρέφει το πρώτο παράθυρο καθεμιάς από τις 5 κορυφαίες ιστοσελίδες που επέστρεψε η μηχανή αναζήτησης ή το πρώτο παράθυρο της κορυφαίας ιστοσελίδας (ανάλογα με τον τρόπο αξιολόγησης).

Για την σύγκριση θα χρησιμοποιήσουμε 3 διαφορετικά συστήματα. Το πρώτο θα είναι ένα baseline σύστημα το οποίο στη τύχη θα επιλέγει 5 παράθυρα κειμένου. Το επόμενο θα είναι ένα σύστημα εκπαιδευμένο με παράθυρα τα οποία έχουμε κατατάξει με τη μέθοδο centrality. Και τέλος με το σύστημα (Μηλιαρακη 2003) στην νέα πλέον εφαρμογή SVM.

Σύστημα	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά όταν επιτρέπονταν 5 απαντήσεις ανά ερώτηση (σε παρένθεση το MRR)	Ποσοστό ερωτήσεων που απαντήθηκαν σωστά, όταν επιτρεπόταν μία απάντηση ανά ερώτηση
centroid	40% (0,33)	17%
baseline1	41%(0,23)	13%
baseline2	23% (0,12)	5%

Παρατηρούμε ότι τα αποτελέσματα είναι σαφώς χειρότερα από εκείνα που επιτυγχάνουμε με τη χρήση της ΜΔΥ και ότι η μέθοδος centroid τα πηγαίνει χειρότερα από τη μέθοδο baseline 1 ενώ καλύτερα από την baseline 2 όταν επιτρέπονται 5 απαντήσεις ανά ερώτηση. Στο πείραμα στο οποίο επιτρεπόταν μόνο μια απάντηση ανά ερώτηση βλέπουμε πως η μέθοδος centroid είναι καλύτερη από τις δυο baseline και αυτό οφείλεται στο υψηλότερο MRR που πετυχαίνει.

.

7. Συμπεράσματα

Η εργασία αυτή ήταν μια προσπάθεια βελτίωσης και περαιτέρω αξιολόγησης ενός συστήματος ερωταποκρίσεων που χρησιμοποιεί μηχανική μάθηση. Χρησιμοποιήθηκαν 3 νέες μέθοδοι για την αυτόματη παραγωγή παραδειγμάτων εκπαίδευσης και αξιολογήθηκαν σε σχέση με αυτή του Γαλάνη (ενότητα 4.3). Μια από αυτές (ενότητα 4.5) αποδείχθηκε πως ήταν καλύτερη από αυτή του Γαλάνη. Επομένως συνεχίσαμε τα πειράματα μας για την εύρεση του βέλτιστου αριθμού ιδιοτήτων για την ΜΔΥ. Ύστερα από πειράματα με μεταβλητό αριθμό αυτόματα επιλεγμένων ιδιοτήτων καταλήξαμε σε 322 ιδιότητες(300 μέσω της ακρίβειας και 22 χειρωνακτικές). Στη συνέχεια έγιναν πειράματα για να βρεθεί ο βέλτιστος αριθμός όρων-στόχων του σώματος εκπαίδευσης. Αποδείχθηκε πειραματικά πως ο βέλτιστος αριθμός είναι 500 όροι-στόχοι.

Από τα πειράματα που έγιναν για να αξιολογηθεί το σύστημα είδαμε πως όταν επιτρέπουμε μια μόνο απάντηση ανά ερώτηση τα αποτελέσματα που πετυχαίνουμε είναι χειρότερα από αυτά του Γαλάνη. Αυτό οφείλεται στην απουσία από το σώμα αξιολόγησης ιστοσελίδες από ηλεκτρονικές εγκυκλοπαίδειες. Σε σχέση με τα 3 baseline συστήματα είδαμε ότι το δικό μας είναι πολύ καλύτερο.

Αναφορές

Alin Dobra, “Support Vector Machine Learning”, CS478 Machine Learning May 2, 2000

Androutsopoulos Ion and Galanis Dimitrios, “An Unsupervised Learning Method to Identify Single-Snippet Answers toDefinition Questions on the Web”, 2004

Cui Hang, Kan Min-Yen, Chua Tat-Seng, “Unsupervised Learning of Soft Patterns for Generating Definitions from Online News, In Proceedings of WWW-2004, pp 90–99, New York, NY.

Cui Hang, Kan Min-Yen, Chua Tat-Seng, Xiao Jing, “A comparative study on Sentence Retrieval for Definitional Question Answering”, IR4QA, A SIGIR 2004 Workshop, Sheffield, UK, 2004

Eugenio B. Di and Glass. M. “The kappa statistic: A second look.” Comput. Linguistics, 30(1):95–101, 2004

Hirschamn L., Gaizauskas R., “Natural language question answering: the view from here”, Cambridge University Press, 2001

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

Prager John, Brown Eric, Radev Dragomir R., Krzysztof Czuba, “One Search Engine or Two for Question-Answering”, TREC9 QA-Track Notebook Paper, NIST, 2000

Prager John, Dragomir Radev, Brown Eric, Coden Anni, Samn Valerie, “The use of Predictive Annotation for Question-Answering in TREC8”, in Proceedings of TREC8, 1999

Prager John, Dragomir Radev, Krzysztof Czuba, “Answering What-Is Questions by Virtual Annotation”, In Proceedings, HLT-2001, San Diego, CA, pp. 26-30, March 2001

Radev Dragomir R., Prager John, Samn Valerie, “Ranking suspected answers to natural language questions using predictive annotation”, ANLP'00, Seattle, WA, May 2000

Reiter E., Dale R., “Building Natural Language Generation Systems”, Cambridge University Press, 2000

S. Miliaraki and I. Androutsopoulos, "Learning to Identify Single-Snippet Answers to Definition Questions". Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 1360-1366, 2004.

Schölkopf Bernhard, Smola Alex, “Learning with Kernels”, MIT Press, Cambridge, MA, 2002

Simmons R. F., “Answering English questions by computer : A survey”, Communications Association for Computing Machinery (ACM), 8(1): 53-70, 1965

Voorhees Ellen M., “Overview of the TREC2001 Question Answering Track”,National Institute of Standards and Technology, In Proceedings of TREC-10 2001

Voorhees Ellen M., “Overview of the TREC-9 Question Answering Track”, National Institute of Standards and Technology, In Proceedings of TREC-9, 2000

Voorhees Ellen M., “The TREC-8 Question Answering Track Report”, National Institute of Standards and Technology, In Proceedings of TREC-8, 1999

Xu Jinxi, Weischedel Ranlp, Licuana, Ana “Evaluation of an Extraction-Based Approach to Answering Definitional Questions”, Association for Computing Machinery (ACM),, in Proceedings, pp 418-424, 2004