Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αναθεώρηση μεθόδου χειρισμού ερωτήσεων ορισμού για συστήματα ερωταποκρίσεων και μεγαλύτερης κλίμακας πειραματική αξιολόγησή της

Λάμπουρας Γεράσιμος

Α.Μ.: 3020076

Επιβλέπων Καθηγητής: Ίων Ανδρουτσόπουλος

Αθήνα 2006

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Η εξάπλωση του διαδικτύου αποτελεί φαινόμενο της εποχής μας . Οι εταιρίες διακινούν τα προϊόντα τους μέσω ηλεκτρονικών καταστημάτων, οι πολίτες ενημερώνονται μέσω ιστοσελίδων, οι οδηγοί βρίσκουν τη θέση και την διαδρομή που θέλουν και εκατομμύρια άνθρωποι επικοινωνούν καθημερινά μέσω e-mail. Κάθε μέρα διακινούνται έτσι τεράστιες ποσότητες πληροφοριών. Ο χειρισμός όμως αυτού του όγκου πληροφοριών είναι δύσκολος.

Μέρος του προβλήματος αυτού αντιμετωπίζεται με τις υπάρχουσες μηχανές αναζήτησης, οι οποίες ψάχνοντας το διαδίκτυο εντοπίζουν τις ιστοσελίδες που αφορούν το θέμα που ενδιαφέρει το χρήστη. Τα τελευταία χρόνια εμφανίστηκαν στον Παγκόσμιο Ιστό και ηλεκτρονικές εγκυκλοπαίδειες, όπως η Wikipedia, στη συγγραφή της οποίας μπορεί να συμμετάσχει ο οποιοσδήποτε.[1] Η εξεύρεση όμως πληροφοριών σε ιστοσελίδες, είτε αυτές είναι ιστοσελίδες εγκυκλοπαιδειών είτε άλλου είδους (π.χ. άρθρα ηλεκτρονικών εφημερίδων), εξακολουθεί να παρουσιάζει δυσκολίες, επειδή οι υπάρχουσες μηχανές αναζήτησης επιστρέφουν ιστοσελίδες (ή παραγράφους τους) και όχι ακριβείς απαντήσεις σε ερωτήματα των χρηστών. Αν, για παράδειγμα, ένας χρήστης εισαγάγει την ερώτηση «Τι είναι η θαλασσαιμία;» σε μια υπάρχουσα μηχανή αναζήτησης, θα λάβει ιστοσελίδες που περιέχουν αυτόν τον όρο, συμπεριλαμβανομένων πολλών ιστοσελίδων που τον περιέχουν χωρίς να τον ορίζουν, αντί για ένα σύντομο ορισμό της θαλασσαιμίας.

Αυτό το κενό επιζητούν να συμπληρώσουν τα συστήματα ερωταποκρίσεων (Question Answering Systems), που φιλοδοξούν να αποτελέσουν την επόμενη γενιά στις μηχανές αναζήτησης. Λαμβάνουν ερωτήσεις σε φυσική γλώσσα και αντί για την επιστροφή μιας λίστας ιστοσελίδων (ή γενικότερα εγγράφων μιας συλλογής), επιστρέφουν μια λίστα με σύντομες πιθανές απαντήσεις στο ερώτημα του χρήστη. Η έρευνα σε αυτόν το τομέα έχει γίνει εντατική τα τελευταία χρόνια, και έχουν ανακοινωθεί πολλά ελπιδοφόρα αποτελέσματα στα πλαίσια του Question Answering Track του TREC (Text Retrieval Conference).[2] Η αυξανόμενη ανάγκη των χρηστών για αυτά τα συστήματα και για την βελτιστοποίηση των επιδόσεων τους προκαλεί το ενδιαφέρον πολλών ερευνητών.

Στα πλαίσια προηγούμενων εργασιών των Μηλιαράκη [1], Γαλάνη [2] και Γιακουμή [3] έγινε έρευνα πάνω στα θέματα που αφορούν τον τομέα των συστημάτων ερωταποκρίσεων. Ως αποτέλεσμα κατασκευάστηκε ένα σύστημα το οποίο χρησιμοποιεί

1 Βλ. http://www.wikipedia.org/.
2 Βλ. http://trec.nist.gov/.

τεχνικές μηχανικής μάθησης και εξάγει τις απαντήσεις του από τις ιστοσελίδες που επιστρέφει μια μηχανή αναζήτησης του διαδικτύου. Το σύστημα επικεντρώνεται στην απάντηση ερωτήσεων ορισμού (π.χ. «Τι είναι ο μυστικισμός;»). Συγκεκριμένα το σύστημα χρησιμοποιεί μια Μηχανή Διανυσμάτων Υποστήριξης (ΜΔΥ Support Vector Machine), μια μορφή επιβλεπόμενης μάθησης, που κατατάσσει τις πιθανές απαντήσεις σε ορισμούς και μη-ορισμούς. Το σύστημα συμπεριλαμβάνει, ακόμη, έναν αλγόριθμο που παράγει αυτομάτως παραδείγματα εκπαίδευσης για τη ΜΔΥ από κείμενα ηλεκτρονικών εγκυκλοπαιδειών, μετατρέποντας έτσι ουσιαστικά τη συνολική μέθοδο μάθησης σε μη επιβλεπόμενη.

Στη διάρκεια της παρούσας εργασίας, επανεξετάστηκε το παραπάνω σύστημα και οι μέθοδοι στις οποίες βασίζεται και έγινε μεγαλύτερης κλίμακας πειραματική αξιολόγησή του, με κύριο στόχο τη βελτιστοποίηση των παραμέτρων του.

1.2 Διάρθρωση της εργασίας

Η εργασία ξεκινάει (κεφάλαιο 2) με μια γενική επισκόπηση των Συστημάτων Ερωταποκρίσεων και των τεχνικών Μηχανικής Μάθησης που χρησιμοποιούνται στην εργασία . Στο κεφάλαιο 3 περιγράφουμε με λεπτομέρεια τη λειτουργία του συστήματός μας και τις μεθόδους στις οποίες βασίζεται. Η πειραματική αξιολόγηση του συστήματος περιγράφεται στο κεφάλαιο 4. Η εργασία τελειώνει (κεφάλαιο 5) με μια σύνοψη των συμπερασμάτων της εργασίας και προτάσεις περαιτέρω βελτίωσης του συστήματος.

2. ΣΥΣΤΗΜΑΤΑ ΕΡΩΤΑΠΟΚΡΙΣΕΩΝ

2.1 Σκοπός των Συστημάτων Ερωταποκρίσεων

Τα Συστήματα Ερωταποκρίσεων για τον Παγκόσμιο Ιστό (και γενικότερα τις συλλογές εγγράφων) προϋποθέτουν και συμπληρώνουν τις υπάρχουσες μηχανές αναζήτησης. Ο χρήστης δίνει στο σύστημα μια ερώτηση σε φυσική γλώσσα. Με τη βοήθεια μιας μηχανής αναζήτησης ανακτώνται ιστοσελίδες (ή γενικότερα έγγραφα) που είναι πιθανώς σχετικά με την ερώτηση. Κατόπιν, μετά από μια σειρά βημάτων επεξεργασίας που περιγράφονται στη συνέχεια, το σύστημα επιστρέφει στο χρήστη μια σειρά από αποσπάσματα των ανακτηθέντων ιστοσελίδων (ή εγγράφων) ως πιθανές απαντήσεις.

2.2 Κατηγορίες ερωτήσεων

Χωρίζουμε τις ερωτήσεις που συνήθως χειρίζονται τα συστήματα ερωταποκρίσεων σε τρεις βασικές κατηγορίες, χρησιμοποιώντας ως κριτήριο τον τύπο της απάντησης που απαιτείται.

- Ερωτήσεις που επιδέχονται καθορισμένη απάντηση (factual questions). Αυτές χωρίζονται περαιτέρω στις εξής υποκατηγορίες:
  - Ερωτήσεις προσώπου, π.χ. «Ποίος ζωγράφισε την "Μόνα Λίζα";».
  - Ερωτήσεις οργανισμού, π.χ. «Ποια εταιρία παράγει το I-Pod;».
  - Ερωτήσεις χρόνου, π.χ. «Πότε πέθανε ο Μότσαρτ;».
  - Ερωτήσεις τόπου, π.χ. «Πού βρίσκεται ο Πύργος της Πίζας;».
  - Ερωτήσεις ποσότητας, π.χ. «Πόσα χρόνια διάρκεσε ο Εκατονταετής πόλεμος;».
  - Ερωτήσεις ορισμού, π.χ. «Τι είναι ο μυστικισμός;».
- Ερωτήσεις γνώμης (opinion questions), π.χ. «Τι θεωρείτε επηρέασε το αποτέλεσμα των φετινών εκλογών;»
- Ερωτήσεις περίληψης (summary questions), π.χ. «Ποια είναι η βασική ιστορία που παρουσιάζεται στο βιβλίο «Όνειρο σε Κύκλο»;»

Η παρούσα εργασία εστιάζεται στις ερωτήσεις ορισμού.[3] Επίσης επικεντρωνόμαστε στην αγγλική γλώσσα, λόγω του μεγαλύτερου αριθμού ιστοσελίδων και ηλεκτρονικών εγκυκλοπαιδειών που διατίθενται σε αυτή τη γλώσσα, αν και οι μέθοδοι που χρησιμοποιούμε μπορούν να εφαρμοστούν και σε κείμενα γραμμένα σε άλλες γλώσσες [4].

Οι ερωτήσεις ορισμού στα αγγλικά έχουν συνήθως τη μορφή: «What/Who is/are/were <ονομαστική φράση> ?». Για παράδειγμα:

«Who was Socrates?»

«What is osteoporosis?»

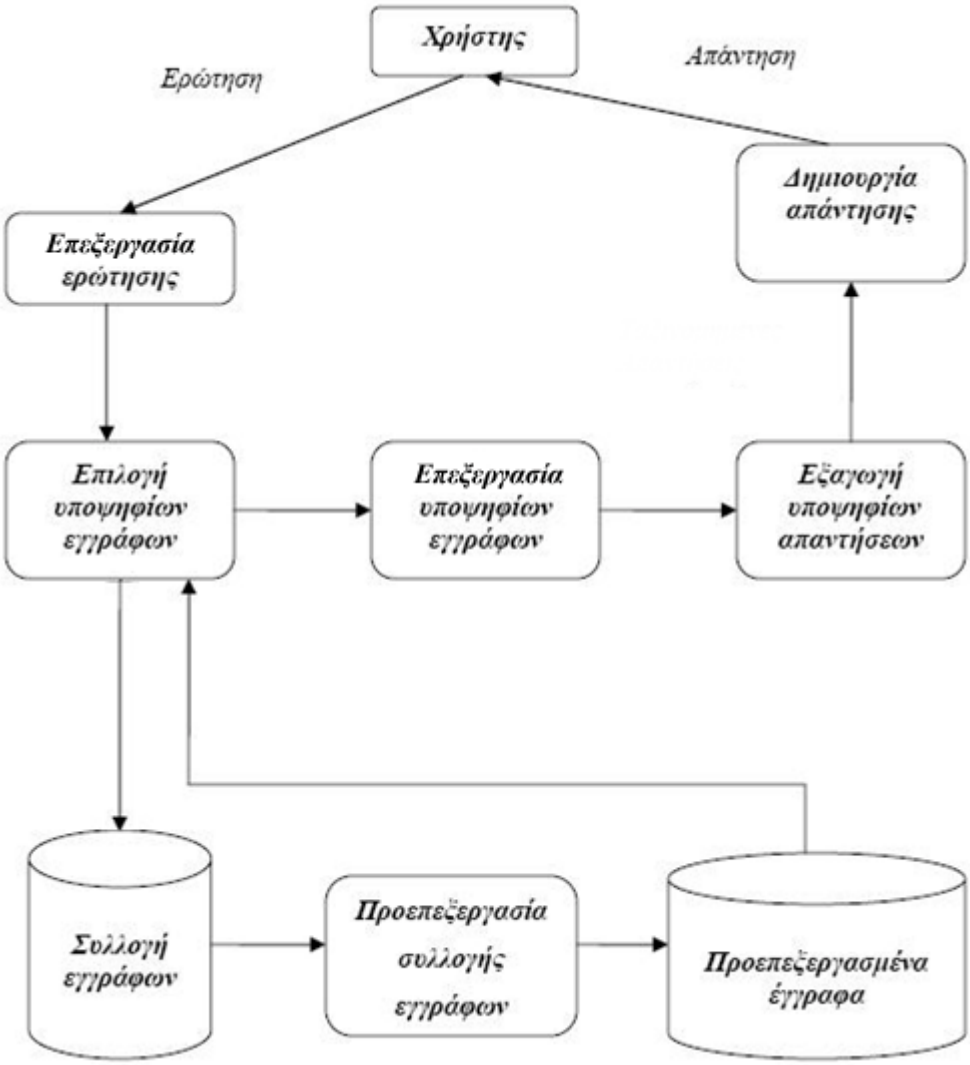
Η ονομαστική φράση που εμφανίζεται παραπάνω είναι και ο «όρος-στόχος» του οποίου επιζητούμε τον ορισμό. Ο όρος-στόχος μπορεί να αποτελείται από μία ή και περισσότερες λέξεις (π.χ. «Who was Duke Ellington?»). Στο υπόλοιπο της εργασίας, όπου χρησιμοποιούμε τη λέξη «όρος» ή τη φράση «όρος-στόχος» θα εννοούμε την ονομαστική φράση της οποίας ζητείται ο ορισμός.

2.3 Συνήθης αρχιτεκτονική Συστήματος Ερωταποκρίσεων

Στο παρακάτω διάγραμμα φαίνεται η συνήθης αρχιτεκτονική ενός συστήματος ερωταποκρίσεων. Το σύστημα δέχεται τις ερωτήσεις από το χρήστη, τις επεξεργάζεται, συλλέγει σχετικές ιστοσελίδες (ή έγγραφα) από τον Παγκόσμιο Ιστό (ή μια συλλογή εγγράφων), τις επεξεργάζεται, εντοπίζει μέσα σε αυτές υποψήφιες απαντήσεις και τέλος επιστρέφει στο χρήστη μία ή περισσότερες υποψήφιες απαντήσεις που θεωρεί πιθανότερο ότι αποτελούν αποδεκτές απαντήσεις.. Σημειώστε ότι αυτή η αρχιτεκτονική αναφέρεται γενικά στα συστήματα ερωταποκρίσεων. Αργότερα θα παρουσιάσουμε ακριβώς την δομή του δικού μας συστήματος.

Ας δούμε κάθε βήμα πιο αναλυτικά:

3 Βλ. [5]για μεθόδους αυτόματης κατάταξης των ερωτήσεων σε κατηγορίες.



Επεξεργασία Ερώτησης:

Ο χρήστης δίνει τις ερωτήσεις στο σύστημα σε φυσική γλώσσα. Το σύστημα πρέπει να αναλύσει την κάθε ερώτηση και να εξαγάγει από αυτή όσες πληροφορίες χρειάζεται για τα επόμενα στάδια. Τέτοιες πληροφορίες είναι ο τύπος της ερώτησης (Προσώπου, Χρόνου, Ορισμού, Γνώμης κ.τ.λ.), ο όρος-στόχος, λέξεις-κλειδιά που περιέχονται στην ερώτηση, κ.τ.λ.

Επιλογή Υποψήφιων Εγγράφων:

Το σύστημα χρησιμοποιεί μια μηχανή αναζήτησης ιστοσελίδων (ή γενικότερα μια μηχανή ανάκτησης πληροφοριών για συλλογές εγγράφων) για να εντοπίσει ιστοσελίδες (ή, γενικότερα, έγγραφα μιας προεπεξεργασμένης συλλογής εγγράφων) που είναι πιθανόν να περιέχουν την απάντηση στην ερώτηση του χρήστη. Στην περίπτωση των ερωτήσεων ορισμού, η αναζήτηση γίνεται δίνοντας ως ερώτημα στη μηχανή αναζήτησης τον όρο-στόχο.

Επεξεργασία Υποψήφιων Εγγράφων:

Στη συνέχεια το σύστημα επεξεργάζεται τις ιστοσελίδες (ή έγγραφα) που προέκυψαν από το προηγούμενο βήμα. Στην περίπτωση των ερωτήσεων προσώπων ή οργανισμών, για παράδειγμα, ενδέχεται να απαιτείται ο εντοπισμός όλων των ονομάτων προσώπων ή οργανισμών, αντίστοιχα, μέσα στις ιστοσελίδες [6]. Στην περίπτωση των ερωτήσεων ορισμού δεν απαιτείται ιδιαίτερη επεξεργασία των ιστοσελίδων του προηγούμενου βήματος.

Εξαγωγή Υποψηφίων Απαντήσεων:

Κατόπιν το σύστημα εντοπίζει στις επεξεργασμένες ιστοσελίδες (ή έγγραφα) υποψήφιες απαντήσεις (π.χ. ονόματα προσώπων, στη περίπτωση των ερωτήσεων προσώπων). Στην περίπτωσή μας, εντοπίζονται ακολουθίες 250 χαρακτήρων που περιέχουν στο κέντρο τους τον όρο-στόχο. Κάθε μία τέτοια ακολουθία καλείται «παράθυρο» του όρου-στόχου και θεωρείται υποψήφια απάντηση.

Δημιουργία Απάντησης:

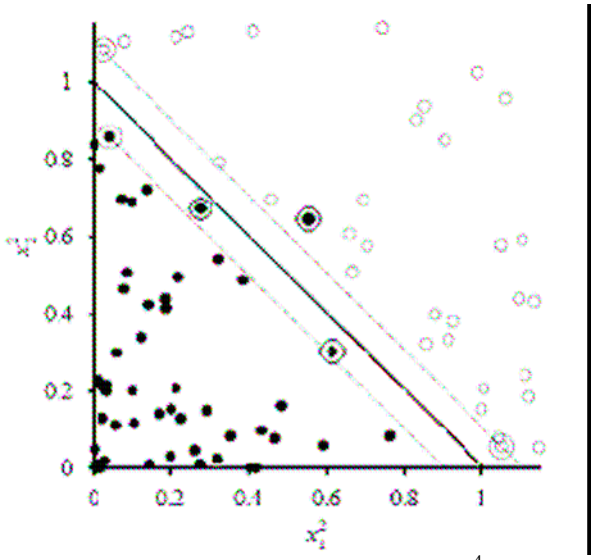
Το σύστημα χρησιμοποιεί στη συνέχεια κάποιον αλγόριθμο για να κρίνει ποιες από τις υποψήφιες απαντήσεις είναι πιθανότερο να αποτελούν ορθές απαντήσεις στην ερώτηση του χρήστη. Ο αλγόριθμος αυτός μπορεί, στη γενικότερη περίπτωση, να συνθέτει πολλές υποψήφιες απαντήσεις, προκειμένου να παραγάγει μια μεγαλύτερη απάντηση. Στη δική μας περίπτωση χρησιμοποιείται ένας αλγόριθμος μηχανικής μάθησης, ο οποίος εκπαιδεύεται στο να κατατάσσει τις υποψήφιες απαντήσεις σε «ορισμούς» και «μη ορισμούς». Τελικά επιλέγονται οι *m* υποψήφιες απαντήσεις για τις οποίες ο αλγόριθμος μάθησης είναι περισσότερο βέβαιος ότι ανήκουν στην κατηγορία των «ορισμών» και επιστρέφονται στο χρήστη. Στα πειράματά μας το *m* παίρνει τις τιμές 1 και 5.

2.4 Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί σημαντικό τομέα της Τεχνητής Νοημοσύνης. Μελετά την κατασκευή προγραμμάτων που μπορούν να βελτιώνουν αυτόματα τις επιδόσεις τους με την συλλογή εμπειρικών δεδομένων. Στην παρούσα εργασία χρησιμοποιούμε τεχνικές επιβλεπόμενης μάθησης. Ένα σύστημα επιβλεπόμενης μηχανικής μάθησης πρώτα εκπαιδεύεται σε μια συλλογή από χειρωνακτικά καταταγμένα παραδείγματα εκπαίδευσης που ονομάζεται «σύνολο δεδομένων εκπαίδευσης». Κατόπιν, οι επιδόσεις του εκπαιδευμένου συστήματος αξιολογούνται με μια συλλογή νέων παραδειγμάτων, για τα οποία το σύστημα δεν γνωρίζει τις σωστές τους κατηγορίες. Τα παραδείγματα αυτά αποτελούν το «σύνολο δεδομένων αξιολόγησης».

2.4.1 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) [10, 11, 12] αναπαριστούν τα δεδομένα (εκπαίδευσης και αξιολόγησης) ως διανύσματα ενός διανυσματικού χώρου και προσπαθούν να βρουν ένα επίπεδο ή υπερεπίπεδο (όταν έχουμε περισσότερες από τρεις διαστάσεις) που να διαχωρίζει τα θετικά από τα αρνητικά δεδομένα εκπαίδευσης. Στην περίπτωσή μας, οι κατηγορίες «θετικό» και «αρνητικό» αντιστοιχούν στις κατηγορίες των «ορισμών» και «μη ορισμών». Αφού εντοπίσουν το καταλληλότερο υπερεπίπεδο, μπορούν εύκολα να κατατάξουν τα διανύσματα αξιολόγησης, των οποίων η κατηγορία είναι άγνωστη, ανάλογα με τη μεριά («επάνω» ή «κάτω») του υπερεπιπέδου διαχωρισμού στην οποία εμφανίζονται. Υπολογίζοντας την απόσταση κάθε διανύσματος αξιολόγησης από το υπερεπίπεδο, το σύστημα επιστρέφει επίσης ένα βαθμό βεβαιότητας για την απόφασή του.



Υπερεπίπεδο με μέγιστο περιθώριο[4]

Σημειώνουμε, επίσης, ότι οι Μηχανές Διανυσμάτων Υποστήριξης συχνά απεικονίζουν τα δεδομένα εκπαίδευσης και αξιολόγησης σε ένα νέο διανυσματικό χώρο, περισσοτέρων διαστάσεων, προκειμένου να γίνει πιο εφικτός ο (γραμμικός) διαχωρισμός των παραδειγμάτων εκπαίδευσης μέσω ενός υπερεπιπέδου. Τα εσωτερικά γινόμενα στο νέο χώρο διαστάσεων υπολογίζονται μέσω ενός «πυρήνα» (kernel). Δεν θα προβούμε σε περαιτέρω ανάλυση της διαδικασίας επιλογής του καλύτερου υπερεπιπέδου και γενικότερα της

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

λειτουργίας των Μηχανών Διανυσμάτων Υποστήριξης, αφού αυτή έχει αναλυθεί αρκετά σε άλλες εργασίες. Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε, για παράδειγμα, στην εργασία του Λουκαρέλλι [6].

Στην περίπτωσή μας, κάθε υποψήφια απάντηση μπορεί να είναι είτε ορισμός (θετική) είτε μη-ορισμός (αρνητική). Προκειμένου να γίνει δυνατός ο διαχωρισμός των υποψηφίων απαντήσεων στις δύο κατηγορίες μέσω Μηχανών Διανυσμάτων Υποστήριξης, κάθε υποψήφια απάντηση («παράθυρο» του όρου-στόχου) μετατρέπεται σε ένα διάνυσμα με ένα μεγάλο πλήθος ιδιοτήτων (attributes). Παραδείγματα τέτοιων ιδιοτήτων εμφανίζονται στον παρακάτω πίνακα. Το (πολύ μεγαλύτερο) σύνολο των ιδιοτήτων που χρησιμοποιούνται στο σύστημα της εργασίας θα παρουσιαστεί σε επόμενο κεφάλαιο.

	Κατάταξη του παραθύρου μέσα στο κείμενο	Εμφάνιση κόμματος μετά τον όρο-στόχο	Κατηγορία
Παράθυρο 1	1	0	Ορισμός (1)
Παράθυρο 2	2	1	Μη-Ορισμός (0)
Παράθυρο 3	3	0	Μη-Ορισμός (0)

Έτσι οι παραπάνω υποψήφιες απαντήσεις θα μετατρέπονταν στα ακόλουθα διανύσματα, στα οποία είναι επίσης σημειωμένες οι ορθές κατηγορίες τους. Στην περίπτωση που ένα διάνυσμα ανήκει στο σύνολο αξιολόγησης, η ορθή κατηγορία του δεν είναι ορατή στο σύστημα.

Απάντηση 1 <1,0> (ορθή κατηγορία: 1)

Απάντηση 2 <2,1> (ορθή κατηγορία: 0)

Απάντηση 3 <3,0> (ορθή κατηγορία: 0)

3. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΚΑΙ ΤΩΝ ΜΕΘΟΔΩΝ ΤΟΥ

3.1 Τρόποι λειτουργίας του συστήματος

Όπως αναφέραμε και παραπάνω, το σύστημα χρησιμοποιεί μηχανική μάθηση.

Ουσιαστικά εκτελείται με δύο τρόπους (modes), η μία εκ των οποίων είναι η «Εκπαίδευση» και η άλλη η «Χρήση» (απάντηση νέων ερωτήσεων). Και στους δύο τρόπους το σύστημα λαμβάνει ως είσοδο ένα σύνολο ερωτήσεων και διάφορες παραμέτρους. Με βάση αυτές τις παραμέτρους το σύστημα μπαίνει σε κατάσταση εκπαίδευσης ή χρήσης. Προφανώς η εκπαίδευση πρέπει να γίνει πριν τη χρήση. Κατά την εκπαίδευση οι ερωτήσεις που δίνουμε στο σύστημα χρησιμοποιούνται για την εκπαίδευση της Μηχανής Διανυσμάτων Υποστήριξης, ενώ κατά τη χρήση το σύστημα προσπαθεί να τις απαντήσει.

3.2 Επεξεργασία ερώτησης - Εξαγωγή παραθύρων

Σε πρώτο στάδιο, τόσο κατά την εκπαίδευση όσο και κατά τη χρήση, το σύστημα επεξεργάζεται τις ερωτήσεις και εξάγει από αυτές τους όρους-στόχους. Για κάθε όρο-στόχο, το σύστημα αναζητάει σχετικά έγγραφα στον παγκόσμιο ιστό. Συγκεκριμένα, στην υλοποίησή μας χρησιμοποιείται η μηχανή αναζήτησης Altavista, η οποία επιστρέφει τις πιο σχετικές σελίδες με φθίνουσα σειρά συσχέτισης. Από αυτές τις σελίδες κρατάμε μόνο τις 10 πρώτες, δηλαδή τις 10 πιο σχετικές.

Σκοπός μας, όμως, είναι να δείξουμε ότι το σύστημά μας καταφέρνει να βρει ορισμούς όρων οι οποίοι δεν καλύπτονται από τις διάφορες ηλεκτρονικές εγκυκλοπαίδειες του διαδικτύου. Προκειμένου να προσομοιώσουμε την αναζήτηση ορισμών αυτού του είδους, το σύστημα αγνοεί τελείως όσες σελίδες επιστρέφει η μηχανή αναζήτησης και προέρχονται από τέτοιες εγκυκλοπαίδειες. Για το σκοπό αυτό έχει δημιουργηθεί μια λίστα με τις διευθύνσεις των πιο γνωστών εγκυκλοπαιδειών, ώστε να αγνοηθούν όσες σελίδες προέρχονται από αυτές. Μερικές από τις σελίδες που εμφανίζονται στη λίστα αυτή είναι οι εξής:

http://www.britannica.com

http://en.wikipedia.org

http://www.howstuffworks.com

http://www.encyclopedia.com

http://www.worldbook.com

...

Κάθε μία από τις σελίδες που απομένουν υφίσταται επεξεργασία, η οποία αφαιρεί όλα τα περιττά στοιχεία από το κείμενό της. Κυρίως, αφού οι σελίδες είναι γραμμένες σε HTML, αφαιρούνται όλες οι ετικέτες (tags) της HTML που αφορούν τη μορφή και τη δομή της σελίδας. Έπειτα, από το «καθαρό» κείμενο που έχει προκύψει, εξάγονται όλα τα παράθυρα που περιέχουν τον όρο-στόχο στο κέντρο τους.

Προφανώς ο ορισμός ενός όρου είναι πιθανότερο να υπάρχει στις πρώτες εμφανίσεις αυτού σε ένα κείμενο. Έτσι κρατάμε μόνο τα πέντε πρώτα παράθυρα από κάθε κείμενο. Συνολικά λοιπόν συλλέγουμε πέντε παράθυρα ανά σελίδα, άρα πενήντα παράθυρα ανά όρο-στόχο.

3.3 Κατασκευή παραθύρων εκπαίδευσης

Αυτό το στάδιο συμβαίνει μόνο κατά τη λειτουργία της εκπαίδευσης. Τα παράθυρα που έχουμε συλλέξει πρέπει τώρα να μετατραπούν σε διανύσματα και να δοθούν στον ταξινομητή για την εκπαίδευσή του, αφού πρώτα σημειωθούν με την ορθή τους κατηγορία (ορισμοί ή μη-ορισμοί). Η σημείωση των παραθύρων αποτελεί ένα από τα βασικά προβλήματα που πρέπει να αντιμετωπίσουμε κατά την εκπαίδευση.

3.3.1 Απλές μέθοδοι

Η απλούστερη μέθοδος επισημείωσης των παραθύρων εκπαίδευσης είναι προφανώς η χειρωνακτική. Αν και έχει το πλεονέκτημα της μέγιστης ακρίβειας στη σημείωση και, άρα, του λιγότερου θορύβου στα δεδομένα εκπαίδευσης., η μέθοδος αυτή κάνει το κόστος εκπαίδευσης και επανεκπαίδευσης ενός συστήματος πολύ μεγάλο. Αν, για παράδειγμα, ένα σύστημα εκπαιδεύεται σε 200 ερωτήσεις, τότε πρέπει να σημειωθούν (να καταταγούν ως ορισμοί ή μη-ορισμοί) χειρωνακτικά 10.000 παράθυρα! Είναι προφανές πως η διαδικασία αυτή είναι υπερβολικά επίπονη και χρονοβόρα.

Μια άλλη μέθοδος είναι η σύγκριση των παραθύρων με πρότυπα (patterns) απαντήσεων. Για κάθε όρο-στόχο κατασκευάζουμε πρότυπα από τις πιθανές απαντήσεις και στην συνέχεια βλέπουμε αν τα παράθυρα ταιριάζουν στα πρότυπα. Αυτή η μέθοδος χρησιμοποιήθηκε και σε προηγούμενη εργασία (Μηλιαράκη [1]), η οποία εκμεταλλεύτηκε τις ερωτήσεις εκπαίδευσης και τα πρότυπα απαντήσεών τους που παρέχουν οι διοργανωτές του

QA Track του TREC. Για τους όρους «Archimedes» και «Galaxy», για παράδειγμα, θα κατασκευάζαμε τα παρακάτω πρότυπα:

Archimedes

Greek (mathematician | astronomer | philosopher | physicist | engineer)

Galaxy

(collection | group | assemblance) of stars

Πρότυπα, όμως, όπως τα παραπάνω κατασκευάζονται χειρωνακτικά. Αν και ο κόπος/χρόνος που χρειάζεται για την κατασκευή τους είναι σαφώς λιγότερος από ότι στην χειρωνακτική κατάταξη κάθε παραθύρου ξεχωριστά, δεν είναι λίγος. Τα πρότυπα πρέπει, επίσης, να είναι έτσι κατασκευασμένα ώστε να προβλέπουν όλες τις ορθές απαντήσεις αλλά και όλες τις πιθανές διατυπώσεις κάθε ορθής απάντησης. Αυτό είναι ευκολότερο στην περίπτωση που οι απαντήσεις πρέπει να βρεθούν σε μια περιορισμένη συλλογή εγγράφων, όπως οι συλλογές που συνήθως χρησιμοποιούνται στο QA Track του TREC, και πολύ δυσκολότερο όταν οι απαντήσεις αναζητούνται σε ολόκληρο τον Παγκόσμιο Ιστό, όπου η ποικιλία διατυπώσεων των ορθών απαντήσεων είναι γενικά πολύ μεγαλύτερη.

3.3.2 Αυτόματη κατασκευή παραθύρων εκπαίδευσης

Η χειρωνακτική σημείωση παραθύρων εκπαίδευσης είναι χρονοβόρα διαδικασία και έτσι περιορίζει και το πλήθος των δεδομένων εκπαίδευσης. Για αυτό έχουν προταθεί μέθοδοι αυτόματης σημείωσης των παραθύρων εκπαίδευσης. Μία από αυτές είναι η μέθοδος του Γαλάνη [2], η οποία επεκτάθηκε από το Γιακουμή [3] ώστε να λαμβάνει υπόψη της και ν-γράμματα λέξεων, αντί μόνο μεμονωμένες λέξεις.

3.3.2.1 Μέθοδος του Γαλάνη

Η βασική ιδέα είναι απλή. Ας υποθέσουμε ότι έχουμε στην διάθεσή μας έναν όρο-στόχο (που έχουμε εξαγάγει από μία ερώτηση) καθώς και ένα παράθυρο κειμένου που τον περιέχει, το οποίο λάβαμε με τον τρόπο που περιγράψαμε παραπάνω. Αν είχαμε στην διάθεσή μας και έναν ορισμό του όρου-στόχου θα μπορούσαμε να υπολογίσουμε (με κάποιο μέτρο ομοιότητας) την ομοιότητα μεταξύ του παραθύρου και του ορισμού και να αποφασίσουμε αν το παράθυρο είναι ορισμός με βάση την ομοιότητά του με τον ορισμό που ήδη διαθέτουμε.

Ένα απλό μέτρο ομοιότητας είναι να μετρηθούν οι κοινές λέξεις του παραθύρου και του ορισμού που έχουμε. Παράδειγμα χρήσης αυτού του μέτρου είναι το παρακάτω.

Όρος-στόχος:

Archimedes

Παράθυρο κειμένου 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 ήδη:

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

Παρατηρούμε πως παρότι και τα δύο κείμενα περιέχουν ορισμούς του όρου-στόχου «Archimedes», οι κοινές λέξεις τους είναι λίγες και δεν φανερώνουν πραγματική ομοιότητα, αφού θα μπορούσαν να είναι κοινές μεταξύ οποιονδήποτε κειμένων που αναφέρονται στον Αρχιμήδη. Το μέτρο αποτυγχάνει, επίσης, να εντοπίσει τη λέξη «mathematician» ως κοινή λόγω της διαφορετικής της κατάληξης σε καθένα από τα κείμενα.

Ένα άλλο πρόβλημα είναι πως ένας ορισμός μπορεί να διατυπωθεί με πολλούς τρόπους. Συγκρίνοντας το παράθυρο με έναν μόνο ορισμό περιοριζόμαστε μόνο στις κοινές λέξεις που έχει το παράθυρο με τη συγκεκριμένη διατύπωση του ορισμού. Το πρόβλημα αυτό είναι ακόμα εντονότερο όταν οι όροι-στόχοι έχουν παραπάνω από μία δυνατές σημασίες. Για να γίνει πιο κατανοητό το πρόβλημα, παρουσιάζουμε παρακάτω διαφορετικούς ορισμούς του όρου «Γαλαξίας», που λάβαμε από ηλεκτρονικές εγκυκλοπαίδειες/γλωσσάρια.

Όρος-στόχος:

Galaxy

Ορισμοί από το διαδίκτυο:

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 that their light takes millions of years to reach the Earth.

Τα προβλήματα αυτά εντόπισε και αντιμετώπισε ο Γαλάνης [2] με τον παρακάτω αλγόριθμο, ο οποίος υπολογίζει την ομοιότητα μεταξύ ενός παραθύρου του όρου-στόχου και ενός συνόλου ορισμών για τον ίδιο όρο-στόχο που έχουμε ήδη στη διάθεσή μας. Το παράθυρο του όρου-στόχου είναι μια υποψήφια απάντηση (ενδεχομένως ένας ορισμός του όρου-στόχου) που έχουμε βρει σε μια ιστοσελίδα, ενώ το σύνολο των ήδη διαθέσιμων ορισμών (του ιδίου όρου-στόχου) προέρχεται από ηλεκτρονικές εγκυκλοπαίδειες. Κατά την εκπαίδευση, ο Γαλάνης περιορίζεται σε όρους-στόχους για τους οποίους έχουμε πολλούς ορισμούς από ηλεκτρονικές εγκυκλοπαίδειες και χρησιμοποιεί τους ορισμούς των εγκυκλοπαιδειών για να σημειώσει αυτόματα (ως ορισμούς ή μη ορισμούς) παράθυρα των αντιστοίχων όρων-στόχων που προέρχονται από ιστοσελίδες. Κατά αυτόν τον τρόπο επιτυγχάνει να εκπαιδεύσει τη Μηχανή Διανυσμάτων Υποστήριξης (ΜΔΥ) στο να εντοπίζει παράθυρα ιστοσελίδων που αποτελούν ορισμούς. Κατόπιν χρησιμοποιεί τη ΜΔΥ για να εντοπίσει παράθυρα ορισμών όρων-στόχων για τους οποίους δεν διαθέτουμε ορισμούς από εγκυκλοπαίδειες. Για την εξεύρεση ορισμών εγκυκλοπαιδειών κατά την εκπαίδευση, χρησιμοποιείται η λειτουργία define της μηχανής αναζήτησης Google («define: <όρος>), η οποία επιστρέφει ορισμούς από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια.

Ο αλγόριθμος του Γαλάνη εκτελεί αρχικά τα εξής βήματα προεπεξεργασίας:

- Αφαίρεση από το παράθυρο και τους ορισμούς εγκυκλοπαιδειών 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 το British National Corpus (βλ. http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bncreadme.html). Η αφαίρεση αυτών των λέξεων γίνεται διότι κατά τη σύγκριση δύο κειμένων η εύρεση κοινών λέξεων που είναι πολύ συχνές δεν φανερώνει ομοιότητα.

- Εφαρμογή ενός αλγορίθμου που αποκόπτει την κατάληξη κάθε λέξης αφήνοντας προς σύγκριση μόνο τη ρίζα της (stemmer, π.χ. το «invented» γίνεται «invent»). Ο αλγόριθμος αποκοπής που χρησιμοποιήθηκε είναι εκείνος του Porter (βλ. http://www.tartarus.org/~martin/PorterStemmer).

- Διαγραφή από κάθε παράθυρο των ειδικών συμβόλων (!@&^%$#., κ.λ.π.).

-

Στη συνέχεια υπολογίζεται για κάθε παράθυρο (υποψήφια απάντηση) ένας αριθμός (score). Αυτός ο αριθμός προκύπτει μετά από τη σύγκριση του παραθύρου με κάθε έναν από τους ορισμούς του όρου-στόχου που έχουμε στην διάθεσή μας και δείχνει την ομοιότητα του παραθύρου με αυτούς. Συγκεκριμένα, όσο μεγαλύτερος ο αριθμός τόσο μεγαλύτερη η ομοιότητα με τους ορισμούς και άρα τόσο μεγαλύτερη και η πιθανότητα το παράθυρο να είναι πράγματι ορισμός. Αντίθετα αν το παράθυρο δεν έχει επαρκή ομοιότητα με τους διαθέσιμους ορισμούς, ο αριθμός που θα του αντιστοιχηθεί θα είναι χαμηλός. Στη συνέχεια περιγράφεται ο τρόπος υπολογισμού αυτού του αριθμού (score).

Σε κάθε λέξη του παραθύρου δίνουμε ένα βάρος w, το οποίο υπολογίζουμε ως εξής.

$$w = fdef * idf$$

Όπου:

w, το βάρος της λέξης

fdef,το ποσοστό των ορισμών που περιέχουν την λέξη

idf (inverse document frequency), η αντίστροφη συχνότητα εγγράφων της λέξης.

Το fdef ορίζεται ως εξής:

$$fdef = cdef / defs$$

Όπου:

cdefs, αριθμός ορισμών που διαθέτουμε για τον όρο-στόχο και που περιέχουν τη λέξη

defs, ολικός αριθμός ορισμών που διαθέτουμε για τον όρο-στόχο.

Το idf ορίζεται ως εξής:

$$idf = 1 + \log\left(N / df\right)$$

Όπου:

N, ο ολικός αριθμός των εγγράφων του British National Corpus (BNC)

df, ο αριθμός των εγγράφων του British National Corpus που περιέχουν την λέξη

Παρατηρούμε ότι δεν έχει το ίδιο βάρος κάθε λέξη του παραθύρου. Αν η λέξη εμφανίζεται σε μεγάλο ποσοστό των ορισμών, υπάρχει μεγαλύτερη πιθανότητα η εμφάνιση της λέξης αυτής σε κάποιο παράθυρο κειμένου να φανερώνει ότι είναι και αυτό ορισμός (fdef). Επίσης όσο πιο σπάνια είναι μια λέξη γενικά τόσο πιο απίθανο είναι η εμφάνιση της και στον ορισμό και στο παράθυρο να οφείλεται σε σύμπτωση (idf).

Τελικά το score κάθε παραθύρου υπολογίζεται από τον παρακάτω τύπο.

$$score = \frac{\sum_{i=1}^{n} w_i}{n}$$

Όπου:

n, ο αριθμός των λέξεων του παραθύρου. Λέξεις που εμφανίζονται στο παράθυρο πολλές φορές υπολογίζονται μόνο μία φορά

w_i**,** το βάρος της λέξης i.

Παράθυρα εκπαίδευσης των οποίων το score υπερβαίνει ένα άνω κατώφλι σημειώνονται ως ορισμοί, ενώ παράθυρα εκπαίδευσης των οποίων το score είναι μικρότερο ενός κάτω

κατωφλίου σημειώνονται ως μη ορισμοί. Τα παράθυρα εκπαίδευσης των οποίων το score βρίσκεται μεταξύ των δύο κατωφλίων αγνοούνται κατά την εκπαίδευση της ΜΔΥ, γιατί δεν είμαστε επαρκώς σίγουροι για την ορθή κατηγορία τους. Η χρήση και ο τρόπος επιλογής των δύο κατωφλίων επεξηγούνται στη συνέχεια, μετά την παρουσίαση της επέκτασης της μεθόδου του Γαλάνη που πρότεινε ο Γιακουμής.

3.3.2.2 Μέθοδος του Γαλάνη με επέκταση σε ν-γράμματα

Ενώ η μέθοδος του Γαλάνη συγκρίνει μεμονωμένες λέξεις του παραθύρου και των διαθέσιμων ορισμών, ο Γιακουμής πρόσθεσε και την σύγκριση ν-γραμμάτων (n-gram) λέξεων, δηλαδή ακολουθιών ν λέξεων.

Οι συγκρίσεις που γίνονται εξαρτώνται από την τιμή της παραμέτρου ν. Στην περίπτωση του ν = 1, η νέα μέθοδος κάνει πάλι συγκρίσεις μεταξύ μεμονωμένων λέξεων και είναι πολύ παρόμοια με την αρχική μέθοδο του Γαλάνη, αλλά έχει το πλεονέκτημα ότι επιστρέφει μια κανονικοποιημένη τιμή στο [0, 1]. Ουσιαστική διαφορά υπάρχει για ν > 1. Για παράδειγμα για ν = 3 η νέα μέθοδος θα συγκρίνει όλες τις μεμονωμένες λέξεις, όλες τις ακολουθίες λέξεων μήκους 2 και όλες τις ακολουθίες λέξεων μήκους 3.

Όπως και στην απλή μέθοδο του Γαλάνη, υπολογίζουμε το βάρος για κάθε λέξη. Μόνο που αυτή τη φορά πρέπει να βρούμε και το βάρος κάθε ν-γράμματος. Το βάρος κάθε ν-γράμματος ορίζεται ως εξής.

$$w = fdef * avgidf$$

Όπου:

w, το βάρος του ν-γράμματος

fdef, το ποσοστό των ορισμών που περιέχουν το ν-γραμμα

avgidf , ο μέσος όρος των idf των λέξεων των ν-γραμμάτων.

Ο υπολογισμός του score γίνεται τώρα ως εξής:

$$score = \frac{\sum_{n=1}^{m} \frac{\sum_{\gamma \in grams_C(n) \cap \gamma \in grams_W(n)} w_\gamma}{\sum_{\gamma \in grams_C(n)} w_\gamma}}{m}$$

Όπου:

m, το μέγιστο μήκος n-γραμμάτων που εξετάζουμε

γ, ένα n-γράμμα

grams$_C$(n), το σύνολο των n-γραμμάτων (μήκους n) που έχουμε από ορισμούς

grams$_W$(n), το σύνολο των n-γραμμάτων (μήκους n) που έχει το παράθυρο

w_i, το βάρος της λέξης i

Τελικά, είτε χρησιμοποιούμε την αρχική μέθοδο του Γαλάνη είτε την επέκτασή της του Γιακουμή, σε κάθε παράθυρο εκπαίδευσης του όρου-στόχου θα έχει δοθεί ένα score. Μας μένει να σημειώσουμε τα παράθυρα εκπαίδευσης ως ορισμούς ή μη-ορισμούς με βάση αυτό το score. Για το σκοπό αυτό χρησιμοποιούμε (όπως και στις εργασίες των Γαλάνη και Γιακουμή) δύο κατώφλια t_- (κάτω κατώφλι) και t_+ (άνω κατώφλι) για τα οποία θα ισχύουν τα εξής.

- Τα παράθυρα εκπαίδευσης που θα έχουν score μεγαλύτερο του άνω κατωφλιού θα είναι παράθυρα ορισμού με μεγάλη πιθανότητα.
- Τα παράθυρα εκπαίδευσης που θα έχουν score μικρότερο του κάτω κατωφλιού θα είναι παράθυρα μη-ορισμού με μεγάλη πιθανότητα.

Τα δύο αυτά κατώφλια θα χωρίσουν τα παράθυρα σε τρεις κατηγορίες. Πρώτον, σε παράθυρα για τα οποία είμαστε σχεδόν βέβαιοι ότι είναι ορισμοί. Δεύτερον, σε παράθυρα για τα οποία είμαστε σχεδόν βέβαιοι ότι δεν είναι ορισμοί. Τρίτον, σε παράθυρα για τα οποία δεν μπορούμε να αποφασίσουμε με βεβαιότητα για την κατηγορία τους. Για την εκπαίδευση της ΜΔΥ θα χρησιμοποιήσουμε μόνο τις δύο πρώτες κατηγορίες, ενώ τα παράθυρα της τρίτης θα αγνοηθούν.

Αν τα περισσότερα παράθυρα εκπαίδευσης καταταχθούν στην τρίτη κατηγορία, τότε πιθανόν να έχουν μείνει πολύ λίγα παράθυρα που να μην επαρκούν για την εκπαίδευση της ΜΔΥ. Σε μια τέτοια περίπτωση πρέπει να χρησιμοποιηθούν περισσότερες ερωτήσεις/κείμενα εκπαίδευσης, ώστε ο αριθμός των παραθύρων των δύο πρώτων κατηγοριών να αυξηθεί. Εναλλακτικά, μπορεί κανείς να μειώσει το άνω κατώφλι και να αυξήσει το κάτω, ώστε να αγνοούνται λιγότερα παράθυρα εκπαίδευσης, διακινδυνεύοντας όμως να αυξηθεί ο αριθμός των παραθύρων εκπαίδευσης που σημειώνονται λάθος. Επίσης προσοχή πρέπει να δοθεί στην ισορροπία μεταξύ του πλήθους των παραθύρων που κατατάσσονται στην πρώτη κατηγορία και τη δεύτερη. Αν μια κατηγορία αποκτήσει πολύ περισσότερα παραδείγματα εκπαίδευσης από την άλλη, υπάρχει ο κίνδυνος η ΜΔΥ να μάθει να ταξινομεί όλα τα παράθυρα σε εκείνη.

Στην εργασία του Γιακουμή παρουσιάζονται πειράματα που είχαν σκοπό να προσδιορίσουν τις καλύτερες τιμές των δύο κατωφλίων. Χρησιμοποιώντας όμως τις τιμές που προκύψαν από εκείνα τα πειράματα παρατηρήσαμε ότι τα περισσότερα παράθυρα κατατάσσονται ως μη-ορισμοί. Αυτό είχε επιπτώσεις στην εκπαίδευση της ΜΔΥ η οποία στη συνέχεια κατέτασσε όλα τα προς αξιολόγηση παράθυρα σαν μη-ορισμούς. Για να αποφύγουμε αυτό το φαινόμενο, εμείς χρησιμοποιήσαμε το άνω κατώφλι που προέκυψε από κείνα τα πειράματα, δηλαδή 0,03 αλλά όχι το κάτω κατώφλι 0,016. Έπειτα δοκιμάσαμε διάφορες τιμές κάτω κατωφλίων με σταθερό το άνω κατώφλι μέχρι να επιτευχθεί η αναλογία μη ορισμών/ορισμών να είναι 60%-40%. Τελικά τα κατώφλια που χρησιμοποιήθηκαν είναι τα $t_+ = 0,03$ και $t_- = 0,005$.

3.4 Αναπαράσταση των παραθύρων ως διανύσματα

Όπως προαναφέρθηκε, για κάθε ερώτηση που δίνεται στο σύστημα (κατά την εκπαίδευση ή τη χρήση του) συλλέγουμε σελίδες από το διαδίκτυο και από αυτές εξάγουμε τα παράθυρα που περιέχουν τον όρο-στόχο της ερώτησης. Κάθε παράθυρο θα πρέπει σε αυτό το στάδιο να μετατραπεί σε ένα διάνυσμα, ώστε είτε να προστεθεί το διάνυσμα στα δεδομένα εκπαίδευσης της ΜΔΥ είτε να ρωτηθεί η ΜΔΥ για την κατηγορία (ορισμός ή μη ορισμός) του.

Κάθε παράθυρο παριστάνεται ως ένα διάνυσμα που αποτελείται από χαρακτηριστικά (features), δηλαδή τιμές ιδιοτήτων (attributes). Οι πρώτες 22 ιδιότητες έχουν επιλεγεί χειρωνακτικά από τη Μηλιαράκη [1], βάσει πειραμάτων σε δεδομένα των διαγωνισμών TREC.

3.4.1 Χειρωνακτικά επιλεγμένες ιδιότητες

Οι 22 χειρωνακτικά επιλεγμένες ιδιότητες είναι οι εξής. Οι πρώτες τρεις είναι αριθμητικές ιδιότητες με ακέραιες τιμές. Οι υπόλοιπες είναι δυαδικές και δείχνουν η κάθε μία αν το παράθυρο περιέχει (τιμή 1) ή όχι (τιμή 0) μία συγκεκριμένη φράση.

6. Η **κατάταξη (ranking) του κειμένου** από το οποίο προέρχεται το παράθυρο. Δηλαδή η σειρά με την οποία κατέταξε / επέστρεψε τη σελίδα η μηχανή αναζήτησης.

   Έχει παρατηρηθεί ότι συνήθως οι ζητούμενοι ορισμοί βρίσκονται στα πρώτα κείμενα που επιστρέφονται παρά στα τελευταία.

7. Η **θέση του παραθύρου** μέσα στο έγγραφο. Δηλαδή αν πρόκειται για την πρώτη, δεύτερη κ.τ.λ εμφάνιση του όρου στο κείμενο.

   Είναι συνηθέστερο ένας όρος να ορίζεται στην αρχή ενός κειμένου.

8. Το **πλήθος των κοινών λέξεων του παραθύρου.**

   Τα παράθυρα ορισμού ενός όρου-στόχου έχουν συνήθως κοινές λέξεις μεταξύ τους. Βρίσκοντας τις κοινές λέξεις όλων των παραθύρων του όρου-στόχου που περιέχονται στα έγγραφα που επέστρεψε η μηχανή αναζήτησης, δημιουργούμε ένα κεντροειδές, δηλαδή έναν «μέσο-όρο» των παραθύρων. Όσο λιγότερο απέχει ένα παράθυρο από το κεντροειδές, με άλλα λόγια όσο περισσότερες από τις κοινές λέξεις έχει, τόσο μεγαλύτερη η πιθανότητα να είναι ορισμός.

   Κατά τον υπολογισμό του κεντροειδούς αφαιρούνται και πάλι οι 100 πιο συχνές λέξεις της αγγλικής γλώσσας. Στα πειράματα, η λίστα των κοινών λέξεων που δημιουργήθηκε για κάθε όρο είχε μέγεθος 20.

9. Η φράση **"such <…> as *όρος*"**

   Παράδειγμα : *"such antibiotics as amoxicillin"*

10. Η φράση **"*όρος* and other <…>"**

    Παράδειγμα : "*broken bones and other injuries"*

11. Η φράση **"*όρος* or other <…>"**

    Παράδειγμα : *"cats or other animals"*

12. Η φράση **"especially *όρος*"**

    Παράδειγμα : *"some plastics especially Teflon"*

13. Η φράση **"including *όρος*"**

    Παράδειγμα : *"some amphibians including frog"*

14. **Παρενθέσεις μετά** τον όρο

    Παράδειγμα : *"sodium chloride (salt)"*

15. **Παρενθέσεις πριν** τον όρο

Παράδειγμα : *"(Vitamin B1) thiamine"*

16. Η φράση **"*όρος* is a"**

Ακριβέστερα αναζητείται η πληρέστερη φράση της μορφής "*όρος* is/are/was/were a/an/the <…>"

Παράδειγμα : *"Galileo was a great astronomer"*

17. **Κόμμα μετά** τον όρο

Παράδειγμα : *"amoxicillin, an antibiotic"*

18. Η φράση **"*όρος* which is/was/are/were <…>"**

Παράδειγμα : *"tsunami which is a giant wave"*

19. Η φράση **"*όρος* like <…>"**

Παράδειγμα : *"antibiotics like amoxicillin"*

20. Η φράση **"*όρος* , <…> , is/was/are/were"**

Παράδειγμα : *"amphibians, like frogs, are animals that can live both on land and in water"*

21. Η φράση **"*όρος* or <…>"**

Παράδειγμα : *"autism or some other type of disorder"*

22. Ένα από τα ρήματα **"can", "refer", "have"** μετά τον όρο **(3 ιδιότητες)**

Παράδειγμα : *"Amphibians can live both on land and in water"*

23. Ένα από τα ρήματα **"called", "known", "defined"** πριν τον όρο **(3 ιδιότητες)**

Παράδειγμα : *" The giant wave known as tsunami"*

3.4.2 Αυτόματα επιλεγμένες ιδιότητες

Οι υπόλοιπες ιδιότητες είναι επίσης δυαδικές. Αφορούν και αυτές φράσεις, οι οποίες στην περίπτωση αυτή είναι ν-γράμματα μήκους ενός έως και τριών λέξεων που προηγούνται ή ακολουθούν αμέσως τον όρο-στόχο. Η διαφορά αυτών των ιδιοτήτων από τις προηγούμενες είναι πως οι φράσεις στις οποίες αντιστοιχούν επιλέγονται αυτόματα από το σύστημα. Η διαδικασία επιλογής επηρεάζεται πολύ από το περιεχόμενο των ερωτήσεων και μπορεί έτσι να προκύψουν διαφορετικές ιδιότητες αν το σύστημα εκπαιδευθεί σε ερωτήσεις ιατρικού ή γεωγραφικού περιεχομένου.

Το πλήθος των ιδιοτήτων αυτού του είδους είναι παράμετρος του συστήματος, η τιμή της οποίας προσδιορίζεται κατά την εκπαίδευση. Η επιλογή των φράσεων γίνεται από τα παράθυρα εκπαίδευσης, όπως στην εργασία της Μηλιαράκη [1].

- Δημιουργείται μια κενή λίστα φράσεων.

- Από όλα τα παράθυρα εκπαίδευσης (για όλους τους όρους-στόχους εκπαίδευσης) εξάγονται όλες οι φράσεις μήκους ένα έως και τρία που προηγούνται ή ακολουθούν τον όρο-στόχο. Δηλαδή συνολικά εξάγονται έξι φράσεις ανά παράθυρο.

  Για κάθε φράση σημειώνεται αν εμφανίστηκε πριν ή μετά τον όρο-στόχο.

- Ελέγχουμε αν κάθε φράση εμφανίζεται ήδη στη λίστα. Αν μια φράση προηγείται του όρου-στόχου, τότε ελέγχουμε μόνο ανάμεσα στις φράσεις που υπάρχουν στην λίστα και προηγούνται και αυτές του όρου-στόχου. Αντίστοιχα αν η φράση ακολουθεί τον όρο-στόχο.

  Αν υπάρχει τότε απλά αυξάνουμε ένα μετρητή εμφανίσεών της κατά 1, αλλιώς εισάγεται στην λίστα η φράση με την σημείωση αν είναι πριν ή μετά τον όρο και ο μετρητής της αρχικοποιείται στην τιμή 1.

- Ως αποτέλεσμα έχουμε μία λίστα με όλες τις φράσεις που έχουν εμφανιστεί αμέσως πριν ή μετά από οποιονδήποτε όρο-στόχο εκπαίδευσης, καθώς και πόσες φορές έχει εμφανιστεί η καθεμία πριν ή μετά από όρο-στόχο.

  Αυτή η λίστα όπως είναι προφανές θα είναι πάρα πολύ μεγάλη. Για αυτόν το λόγο φροντίζουμε να διαγράψουμε από τη λίστα κάθε φράση που εμφανίζεται λιγότερες φορές από κάποιο κατώφλι. Το κατώφλι εξαρτάται από το πλήθος των παραθύρων που χρησιμοποιήθηκαν και άρα από το πλήθος των ερωτήσεων που δόθηκαν για εκπαίδευση.

- Μετά υπολογίζουμε την ακρίβεια (precision) κάθε φράσης που έχει απομείνει στην λίστα. Η ακρίβεια υπολογίζεται ως ο λόγος των παραθύρων όπου εμφανίζεται η φράση και είναι ορισμοί δια τα συνολικά παράθυρα στα οποία εμφανίζεται η φράση.

  Η ακρίβεια μας δείχνει κατά πόσο η εμφάνιση της φράσης σηματοδοτεί με βεβαιότητα ότι το παράθυρο είναι ορισμός.

  Εδώ παρατηρούμε ότι αν μια φράση εμφανίζεται μόνο μια φορά σε ένα παράθυρο και αν αυτό το παράθυρο είναι και ορισμός τότε η ακρίβεια αυτής της φράσης είναι 1. Αυτή όμως η φράση δεν δίνει σημαντική πληροφορία στο σύστημά μας, γιατί είναι πολύ σπάνια η εμφάνιση της. Αυτός είναι ένας ακόμα λόγος που διαγράψαμε τις φράσεις που εμφανίζονται πάρα πολύ λίγες φορές στη λίστα, στο προηγούμενο βήμα.

- Στο τέλος επιλέγουμε τον αριθμό των φράσεων που θέλουμε και έχουν την μεγαλύτερη ακρίβεια, ανεξάρτητα αν ακολουθούν ή προηγούνται του όρου, και τις μετατρέπουμε σε ιδιότητες του διανύσματος που ελέγχουν αν η αντίστοιχη φράση βρίσκεται αμέσως πριν ή μετά (ανάλογα με το πού είχε βρεθεί η φράση) από τον όρο-στόχο.

3.5 Εκπαίδευση και Ταξινόμηση

Όπως αναφέραμε και στην αρχή, πρώτα δίνουμε στο σύστημα τις ερωτήσεις εκπαίδευσης. Έχοντας όλα τα παράθυρα που προέκυψαν από αυτές κωδικοποιημένα σε διανύσματα, τα δίνουμε στη ΜΔΥ για να εκπαιδευθεί. Το σύστημα της παρούσας εργασίας χρησιμοποιεί την υλοποίηση libSVM των ΜΔΥ, με πυρήνα RBF.[5] Για την εξεύρεση των καλύτερων τιμών των παραμέτρων της ΜΔΥ χρησιμοποιείται η μέθοδος grid search που παρέχουν οι κατασκευαστές του libSVM.

Μετά την εκπαίδευση, μπορούμε να δώσουμε στο σύστημα ένα σύνολο ερωτήσεων ορισμού που θέλουμε να απαντήσει. Το σύστημα θα κάνει την ίδια διαδικασία για κάθε ερώτηση, παράγοντας πάλι διανύσματα τα οποία θα ταξινομηθούν ως ορισμοί ή μη ορισμοί από την εκπαιδευμένη πλέον ΜΔΥ. Για κάθε ερώτηση, το σύστημα θα επιλέξει τα παράθυρα που τα αντίστοιχά τους διανύσματα θεωρήθηκαν ως τα πιο πιθανά να είναι ορισμοί και θα τα επιστρέψει στον χρήστη. Ο αριθμός των επιστρεφομένων παραθύρων ανά ερώτηση είναι επίσης παράμετρος του συστήματος.

5 Βλ. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4. ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

4.1 Μέτρα Αξιολόγησης

Ακολουθώντας τους κανονισμούς των διαγωνισμών TREC 2000 [8] και TREC 2001 [9] για τις ερωτήσεις ορισμού, το σύστημά μας επιστρέφει μια λίστα με τις πέντε απαντήσεις (παράθυρα του όρου-στόχου) που έκρινε ως πιθανότερο να περιέχουν αποδεκτό ορισμό του όρου-στόχου. Στην περίπτωση που τουλάχιστον ένα από αυτά τα παράθυρα περιέχει ορισμό, τότε θεωρούμε ότι το σύστημα κατάφερε και απάντησε σωστά.

Για να μπορούμε να κρίνουμε και το κατά πόσο το σύστημα επιστρέφει σωστές απαντήσεις στις υψηλές θέσεις της λίστας, χρησιμοποιούμε τη Μέση Αντίστροφη Κατάταξη (Mean Reciprocal Rank). Για να την υπολογίσουμε αντιστοιχούμε σε κάθε ερώτηση ένα βαθμό. Αυτός ισούται με 1 δια τη θέση της πρώτης σωστής απάντησης στη λίστα (1 – 5). Αν δεν εμφανίζεται σωστή απάντηση τότε ο βαθμός παίρνει την τιμή 0. Παίρνοντας τον μέσο όρο των βαθμών προκύπτει η αριθμητική τιμή της Μέσης Αντίστροφης Κατάταξης. Όπως είναι προφανές αυτή κυμαίνεται από 0 έως 1, και όσο υψηλότερη είναι τόσο λιγότερες απαντήσεις απαιτείται να επιστρέψει το σύστημα μέχρι να επιστραφεί η σωστή.

Τέλος υπολογίζεται και το ποσοστό επιτυχίας του συστήματος αν του επιτραπεί να δώσει μόνο μία απάντηση ανά ερώτηση, αυτήν που θεωρεί ως πιο πιθανή. Ο σκοπός αυτής της μέτρησης είναι κυρίως η σύγκριση με μεθόδους προηγούμενων εργασιών όπως αυτή του Γαλάνη (2004).

4.2 Κριτές Αξιολόγησης

Χρησιμοποιήσαμε τρεις διαφορετικούς ανθρώπους-κριτές για να αξιολογήσουν τις απαντήσεις του συστήματος, ώστε να μην βασίζονται τα αποτελέσματα στην κρίση ενός μόνο ατόμου. Πιο συγκεκριμένα, οι τρεις κριτές αξιολογήσανε τις απαντήσεις από τέσσερα διαφορετικά συστήματα (TREC, Baseline 1, Baseline 2 και το σύστημα της παρούσας εργασίας εκπαιδευμένο σε 250, 500 και 750 ερωτήσεις εκπαίδευσης), τα οποία περιγράφονται παρακάτω. Αφού δείξαμε ότι οι κριτές συμφωνούν επαρκώς μεταξύ τους, τα υπόλοιπα πειράματα και αξιολογήσεις έγιναν με έναν μόνο κριτή.

Για να υπολογίσουμε τους βαθμούς συμφωνίας μεταξύ των κριτών χρησιμοποιήσαμε το στατιστικό Κ [7]. Το στατιστικό Κ υπολογίζεται με βάση τον παρακάτω τύπο.

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Όπου:

P(A), Η παρατηρούμενη συμφωνία μεταξύ των κριτών

P(E), Η αναμενόμενη συμφωνία μεταξύ των κριτών

Το P(E) ουσιαστικά παριστάνει την πιθανότητα οι κριτές να συμφωνούνε κατά τύχη. Ο υπολογισμός του K γίνεται σε πέντε βήματα.

- Εκτιμούμε την πιθανότητα που έχει ο κάθε κριτής να κατατάξει ένα παράθυρο σαν ορισμό (P_+) ή όχι (P_-). Η εκτίμηση γίνεται διαιρώντας το πλήθος των παραθύρων που κατάταξε σαν ορισμούς (μη-ορισμού αντίστοιχα) με το σύνολο των παραθύρων.
- Για κάθε ζευγάρι κριτών, η πιθανότητα αυτό να συμφωνεί ή να διαφωνεί κατά τύχη δίνεται από το γινόμενο των $P_+(A)$ $P_+(B)$ και $P_-(A)$ $P_-(B)$ αντίστοιχα.
- Το P(E) κάθε ζευγαριού είναι το άθροισμα $P_+(A)$ $P_+(B)$ + $P_-(A)$ $P_-(B)$.
- Υπολογίζουμε το K του κάθε ζευγαριού με βάση τον παραπάνω τύπο.
- Η συνολική συμφωνία των κριτών για το σύστημα δίνεται από των μέσο όρο των K των ζευγαριών.

Το στατιστικό K παίρνει τιμές στο διάστημα [-1, 1] με το 0 να δηλώνει πως η συμφωνία είναι τελείως τυχαία και τα δύο άκρα απόλυτη διαφωνία και συμφωνία.

4.3 Ερωτήσεις εκπαίδευσης και αξιολόγησης

Οι ερωτήσεις εκπαίδευσης του συστήματος της παρούσας εργασίας επελέγησαν τυχαία από το ευρετήριο της ηλεκτρονικής εγκυκλοπαίδειας.[6] Όπως προαναφέρθηκε, για κάθε ερώτηση εκπαίδευσης συλλέξαμε 50 παράθυρα του όρου-στόχου μέσω της Altavista, εξαιρώντας ιστοσελίδες που προέρχονταν από ηλεκτρονικές εγκυκλοπαίδειες.

Η αξιολόγηση και των τεσσάρων συστημάτων έγινε με 100 νέες ερωτήσεις ορισμού, κοινές και για τα τέσσερα συστήματα, που συλλέξαμε από το ευρετήριο της ίδιας εγκυκλοπαίδειας. Η συλλογή των ερωτήσεων ορισμού έγινε ώστε οι ερωτήσεις να περιλαμβάνουν ένα μεγάλο εύρος θεμάτων, από ιατρικούς όρους και ιστορικά πρόσωπα μέχρι όρους φυσικής και καθημερινούς όρους. Όπως και κατά την εκπαίδευση, κατά την αξιολόγηση συλλέξαμε για κάθε όρο-στόχο 50 παράθυρα μέσω της Altavista, εξαιρώντας και πάλι ιστοσελίδες που προέρχονταν από ηλεκτρονικές εγκυκλοπαίδειες.

6 Βλ. http://www.encyclopedia.com

4.4 Συστήματα Σύγκρισης

Οι επιδόσεις του συστήματος της παρούσας εργασίας συγκρίθηκαν με εκείνες τριών απλούστερων συστημάτων. Τα δύο πρώτα, που θα ονομάσουμε Baseline 1 και Baseline 2, δεν χρησιμοποιούν μηχανική μάθηση. Το πρώτο απλά επιστρέφει ως απαντήσεις τα πρώτα παράθυρα του όρου-στόχου των πέντε κορυφαίων ιστοσελίδων (1 παράθυρο από κάθε ιστοσελίδα) που επέστρεψε η μηχανή αναζήτησης. Το δεύτερο επιλέγει τυχαία 5 παράθυρα του όρου-στόχου από το σύνολο των 50 παραθύρων του όρου-στόχου που ανακτώνται από τις ιστοσελίδες που επέστρεψε η μηχανή αναζήτησης και τα επιστρέφει. Αντίστοιχα όταν επιτρέπεται μόνο μία απάντηση το Baseline 1 επιστρέφει το πρώτο παράθυρο της κορυφαίας ιστοσελίδας και το Baseline 2 ένα μόνο τυχαίο παράθυρο. Και οι δύο αυτές μέθοδοι δεν κάνουν καμία ανάλυση στον όρο-στόχο ή στα παράθυρα. Χρησιμεύουν απλά για να αποδείξουμε ότι η μέθοδος μας είναι καλύτερη από το να χρησιμοποιεί κανείς μόνη της τη μηχανή αναζήτησης και ότι είναι καλύτερα από το να επιλέγει παράθυρα στην τύχη.

Το τρίτο σύστημα είναι ουσιαστικά το ίδιο με εκείνο του προηγούμενου κεφαλαίου, αλλά είναι εκπαιδευμένο πάνω στα δεδομένα (ερωτήσεις και κείμενα) του QA Track των διαγωνισμών TREC των ετών 2000 και 2001, χρησιμοποιώντας τα πρότυπα απαντήσεων (patterns) που παρέχουν οι διοργανωτές του συνεδρίου. Δηλαδή το σύστημα εκπαιδεύεται στις 136 ερωτήσεις ορισμού εκείνων των ετών του διαγωνισμού και τα παράθυρα εκπαίδευσης δεν ανασύρονται από σελίδες του διαδικτύου, αλλά από τα κείμενα που παρέχει ο διαγωνισμός. (Οι διοργανωτές παρέχουν, επίσης, για κάθε ερώτηση τα έγγραφα που επέστρεψε από τη συλλογή εγγράφων του διαγωνισμού μια μηχανή αναζήτησης.) Σε όλα τα πειράματα, όλες οι άλλες τιμές των παραμέτρων του τρίτου συστήματος είναι ίδιες με εκείνες του συστήματος της παρούσας εργασίας.

Τα αποτελέσματα των τριών συστημάτων παρουσιάζονται στον παρακάτω πίνακα. Στην περίπτωση του συστήματος TREC, χρησιμοποιήθηκαν 300 ιδιότητες, αριθμός που είχε οδηγήσει στα καλύτερα αποτελέσματα των πειραμάτων της Μηλιαράκη και μέγιστο μήκος ν-γραμμάτων m = 3 (δηλαδή οι αυτόματα επιλεγμένες ιδιότητες αντιστοιχούσαν σε φράσεις μήκους 1, 2 ή 3 λέξεων).

	Επιτυχία όταν επιστρέφεται 1 απάντηση / ερώτηση (%)	**Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**	**Συμφωνία Κριτών**
Baseline 1	6,67 %	24 %	0,090667	0,784323
Baseline 2	13,33 %	34 %	0,149667	0,73749
TREC	19,67 %	42 %	0,258167	0,769459

4.5 Πειράματα με μεταβλητό αριθμό ερωτήσεων εκπαίδευσης

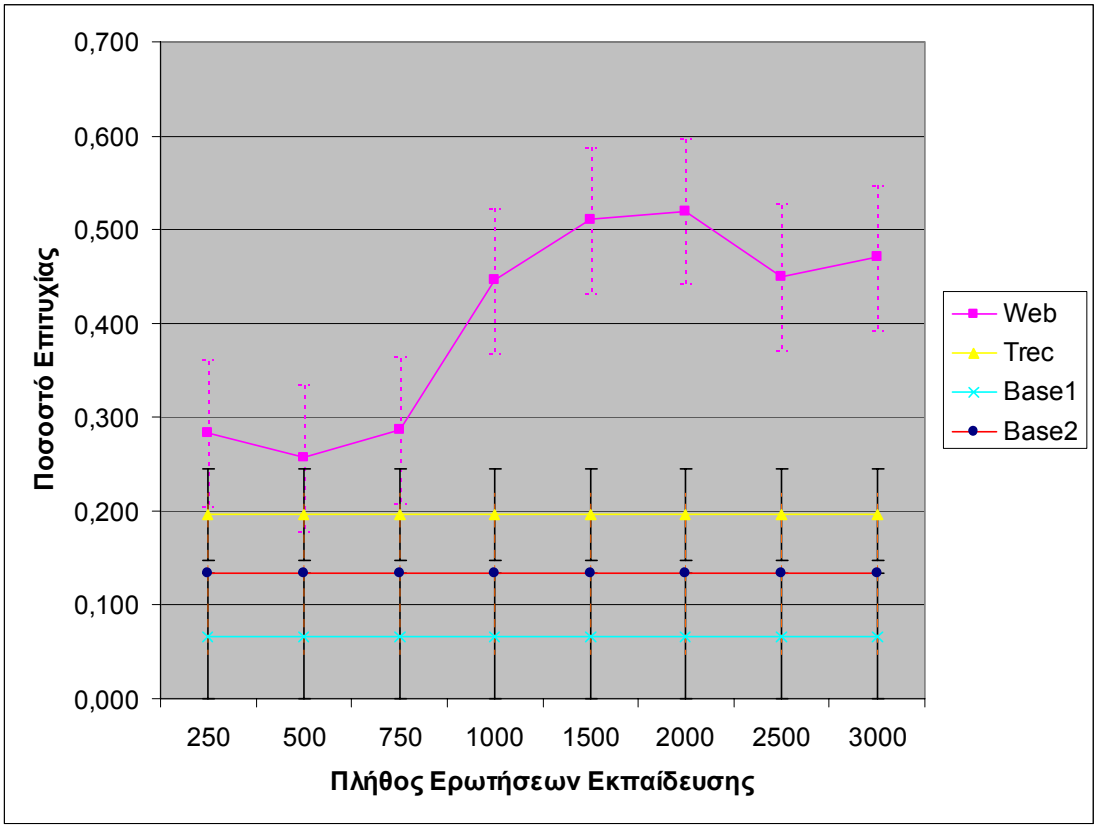
Αρχίσαμε τα πειράματά μας προσπαθώντας να βρούμε τη βέλτιστη τιμή του πλήθους ερωτήσεων εκπαίδευσης. Στη προσπάθεια μας αυτή κάναμε πειράματα με 250, 500, 750, 1000, 1500, 2000, 2500 και 3000 ερωτήσεις εκπαίδευσης. Περαιτέρω πειράματα με πάνω από 3000 ερωτήσεις ήταν δυστυχώς ανέφικτα, λόγω του απαιτούμενου χρόνου και υπολογιστικής δύναμης. Για την αξιολόγηση χρησιμοποιούνται πάντα οι ίδιες 100 ερωτήσεις. Η εκπαίδευση έγινε με 300 αυτόματα επιλεγμένες ιδιότητες, όπως στην περίπτωση του συστήματος TREC, και μέγιστο μήκος ν-γραμμάτων m = 3.

Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

Ερωτήσεις Εκπαίδευσης	**Επιτυχία όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**	**Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**	**Συμφωνία Κριτών**
250	28,33 %	46 %	0,26	0,540889
500	25,67 %	55 %	0,399	0,648497
750	28,67 %	61 %	0,422167	0,850556
1000	45 %	66 %	0,567833	-
1500	51 %	65 %	0,565667	-
2000	52 %	68 %	0,574	-
2500	45 %	68 %	0,528833	-
3000	47 %	66 %	0,542833	-

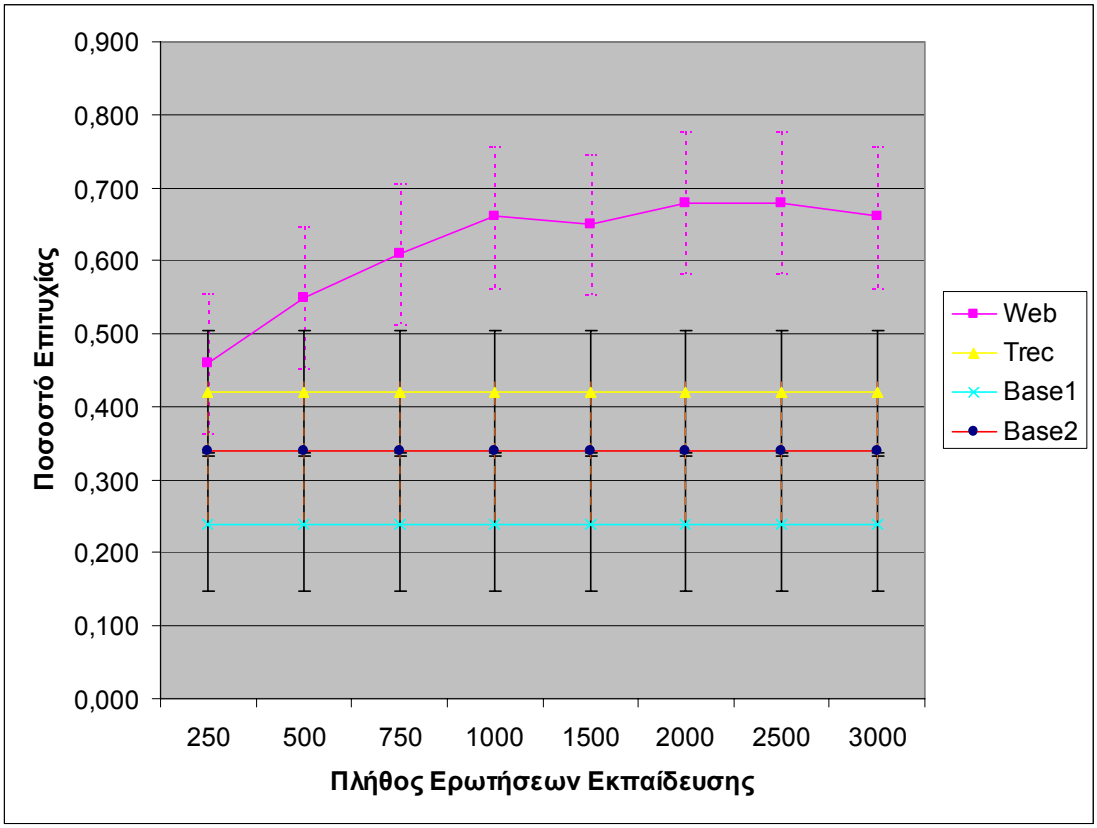
Το παρακάτω σχήμα συγκρίνει διαγραμματικά τα παραπάνω συστήματα, όταν επιτρέπεται μία απάντηση ανά ερώτηση. Σε αυτό έχουμε προσθέσει και τα τρία συστήματα Baseline 1, Baseline 2 και TREC, των οποίων οι επιδόσεις είναι ανεξάρτητες του αριθμού των ερωτήσεων εκπαίδευσης (τα Baseline 1 και 2 δεν εκπαιδεύονται, ενώ το σύστημα TREC είναι πάντα εκπαιδευμένο στις 136 ερωτήσεις του διαγωνισμού).

Βλέπουμε πως η συμφωνία των κριτών αυξάνεται καθώς αυξάνονται οι ερωτήσεις εκπαίδευσης. Αυτό μας οδηγεί στο συμπέρασμα ότι τα παράθυρα που επιστρέφει το σύστημα είναι λιγότερο αμφιλεγόμενα ως ορισμοί. Με βάση αυτή την ανοδική τάση της συμφωνίας αποφασίσαμε τα πειράματα για 1000 ερωτήσεις και άνω να γίνονται από ένα κριτή.



Στο διάγραμμα φαίνονται και τα διαστήματα εμπιστοσύνης 95% της κάθε μέτρησης. Έτσι μπορούμε να συμπεράνουμε ότι για τιμές ερωτήσεων από 250 έως και 750 το σύστημά μας είναι καλύτερο από τα υπόλοιπα, αλλά ειδικά σε σύγκριση με το σύστημα TREC η διαφορά δεν είναι στατιστικά σημαντική. Μετά τις 1000 ερωτήσεις εκπαίδευσης, όμως, το σύστημά μας δείχνει στατιστικά σημαντική υπεροχή έναντι των υπολοίπων, ενώ οι καλύτερες επιδόσεις του ήταν για 1500-2000 ερωτήσεις εκπαίδευσης. Μετά τις 2000 ερωτήσεις εκπαίδευσης, το ποσοστό επιτυχίας δείχνει σημάδια κορεσμού.

Παρακάτω ακολουθεί το αντίστοιχο διάγραμμα όταν το σύστημα επιστρέφει 5 παράθυρα για κάθε ερώτηση.



Η καμπύλη του συστήματος μας είναι πιο ομαλή εδώ αν και η πορεία που ακολουθεί είναι ίδια με αυτήν του προηγούμενου διαγράμματος. Η απόδοση αυξάνεται μέχρι τις 1000 ερωτήσεις και μετά παραμένει σταθερή.

4.6 Πειράματα με μεταβλητό αριθμό αυτόματα επιλεγμένων ιδιοτήτων

Θεωρώντας ότι τα καλύτερα αποτελέσματα επιτυγχάνονται χρησιμοποιώντας 2000 ερωτήσεις εκπαίδευσης, προχωρήσαμε σε πειράματα με μεταβαλλόμενο αριθμό αυτόματα επιλεγμένων ιδιοτήτων, κρατώντας σταθερό τον αριθμό των ερωτήσεων εκπαίδευσης στις 2000. Όπως και στα προηγούμενα πειράματα, λόγοι χρόνου και υπολογιστικής ισχύος μας απέτρεψαν από πειράματα με μεγαλύτερα σύνολα αυτόματα επιλεγμένων ιδιοτήτων.

Αυτόματα Επιλεγμένες Ιδιότητες	**Επιτυχία όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**	**Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
0	34 %	58 %	0,439833
50	33 %	58 %	0,445

150	41 %	61 %	0,496
300	50 %	68 %	0,574
450	37 %	63 %	0,496333
600	35 %	58 %	0,441167

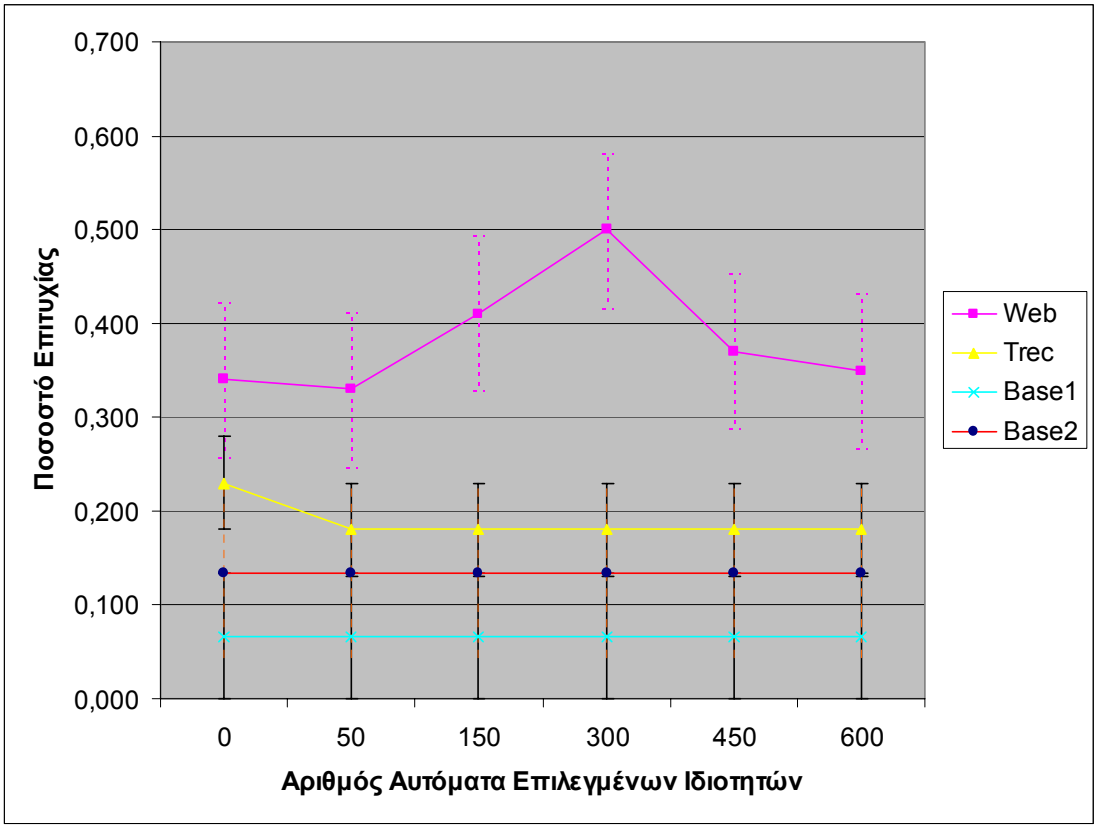
Τα πιο ικανοποιητικά αποτελέσματα τα λάβαμε για 300 ιδιότητες. Παρατηρούμε επίσης πως οι πρώτες 50 ιδιότητες δεν φαίνεται να συνεισφέρουν σημαντική πληροφορία στο σύστημα.

Ο επόμενος πίνακας δείχνει τα αντίστοιχα αποτελέσματα στην περίπτωση του συστήματος TREC.

Αυτόματα Επιλεγμένες Ιδιότητες	**Επιτυχία όταν επιστρέφεται 1 απάντηση / ερώτηση (%)**	**Επιτυχία όταν επιστρέφονται 5 απαντήσεις / ερώτηση (%)**	**MRR**
0	23 %	55 %	0,340833
50	18 %	54 %	0,309333
150	18 %	49 %	0,293833
300	18 %	49 %	0,293833
450	18 %	49 %	0,293833
600	18 %	49 %	0,293833

Η σταθερότητα που εμφανίζεται σε τιμές άνω των 150 αυτόματα επιλεγμένων ιδιοτήτων οφείλεται στο γεγονός ότι το σύστημα δεν βρίσκει παραπάνω από 56 ιδιότητες για τόσο μικρό πλήθος ερωτήσεων.

Τα παραπάνω φαίνονται ακόμα πιο καθαρά στο παρακάτω διάγραμμα. Και αυτή τη φορά έχουμε προσθέσει τα συστήματα Baseline 1, Baseline 2 και TREC για σύγκριση. Εύκολα φαίνεται ότι σε όλες τις τιμές το σύστημα μας είναι σημαντικά καλύτερο από τα υπόλοιπα.



Για να διερευνήσουμε γιατί οι πρώτες 50 ιδιότητες δεν φαίνεται να συνεισφέρουν σημαντική πληροφορία στο σύστημα της παρούσας εργασίας, να παρουσιάζουμε στον παρακάτω πίνακα τις 50 πρώτες ιδιότητες. Στους πίνακες οι ιδιότητες εμφανίζονται με φθίνουσα ακρίβεια από πάνω προς τα κάτω και από αριστερά στα δεξιά.

Πριν από τον όρο – στόχο

admiralty and	page communicable disease
purple	communicable disease
spiny	disease
convention on the	what is it?
alaskan	is it?
tennis	it?
the chinese	definition
the mineral	infections
mineral	to greek mythology
respiratory syndrome	europe greek mythology
syndrome	greek mythology
printouts	font-size: % }

Μετά από τον όρο – στόχο

Marcasite
encarta articles
Encarta
chemistry:
Alloys
dictionary home upgrade
dictionary home
what are
Noun
is called
symbol:
Species

enchantedlearning.com	
australian	% }
cmos	y z misc
encarta right-click dictionary	z misc
right-click dictionary	misc
lexicon on *	mythology
on *	

what is it?
include
is usually

Παρατηρούμε αρχικά ότι μεγάλο πλήθος των ιδιοτήτων αφορούν πολύ συγκεκριμένα θέματα όπως οι "to greek mythology", "europe greek mythology", "greek mythology", "mythology" που αφορούν μυθολογία και οι "respiratory syndrome", "syndrome", "page communicable disease", "communicable disease", "disease", "infections" που αφορούν ασθένειες.

Το σύστημά μας επιλέγει ιδιότητες με βάση την ακρίβειά τους. Όλες οι παραπάνω φράσεις που εμφανίζονται στο πίνακα έχουν ακρίβεια 1, δηλαδή τη μέγιστη δυνατή τιμή. Σε όσες ερωτήσεις συναντηθήκανε αυτές περιείχαν ορισμό. Ενώ όμως τέτοιες ιδιότητες είναι χρήσιμες για την εύρεση ορισμών συγκεκριμένης θεματολογίας, όπως μυθολογία ή ιατρική, δεν βοηθούν ερωτήσεις άλλου περιεχομένου. Φαίνεται, επίσης, ότι πολλές από αυτές τις φράσεις είναι αρκετά σπάνιες και έτυχε να εμφανίζονται μόνο σε παράθυρα ορισμού, κάτι που τους δίνει πολύ υψηλή ακρίβεια.

Για να καταφέρει να καλύψει περισσότερα θέματα το σύστημα χρειάζεται περισσότερες ιδιότητες. Ας δούμε τις επιπλέον 100 ιδιότητες που επιλέγονται αν ζητηθούν 150.

Πριν από τον όρο – στόχο

y z	mamma.com for
x y z	references
the term	searched for '
term	for '
sign in above.	species of
in above.	french
above.	chinese
start --> get-->	forms of
--> get-->	z
get-->	most
] general name	causes

Μετά από τον όρο – στόχο

cite / print	what
cite /	turtle
cite	chemical
dictionary	part
animal printouts	hall
animal	is found
is most	because
are found	what is
of a	articles
type in	since
animal printouts	description

general name	what is a
of an	return to top
types	infectious diseases
renin	known as the
are the	called the
is called	symptoms of
'	

label	
is used	are
symbol	*
help definition of	overview
help definition	common
may be	was a
is that	search help printer-friendly
Occurs	search help
are	there are
many	were
came	can be

Βλέπουμε ότι τώρα προστέθηκαν περισσότερο γενικές φράσεις («known as», «are», «causes», «general name» κ.τ.λ.), πολλές από τις οποίες συνοδεύουν συχνά ορισμούς. Ο λόγος που το σύστημα εξακολουθεί να μην παρουσιάζει μεγάλη βελτίωση με τις 150 αυτόματα επιλεγόμενες ιδιότητες, σε σχέση με τη χρήση μόνο των χειρωνακτικά επιλεγμένων ιδιοτήτων, είναι απλός. Οι ιδιότητες που επιλέχθηκαν τώρα είναι σε μεγάλο βαθμό παρόμοιες με τις 22 χειρωνακτικά επιλεγμένες ιδιότητες του συστήματος. Έτσι δεν προσθέτουν κάποια επιπλέον πληροφορία που το σύστημα δεν έχει ήδη.

Ενδιαφέρον είναι ότι αυτές οι ιδιότητες εμφανίστηκαν εφόσον τελείωσαν οι ιδιότητες ακρίβειας 1. Ίσως θα ήταν φρονιμότερο να μην επιλέγονταν οι ιδιότητες με τη μέγιστη τιμή 1, αφού ως επί το πλείστον προκύπτει ότι είναι πολύ ειδικές και δεν βοηθάνε το σύστημα συνολικά. Μία τέτοια αλλαγή δεν θα οδηγούσε μάλλον σε μεγάλη αύξηση στην επιτυχία του συστήματος, αλλά θα επέτρεπε να επιτύχει το σύστημα το ίδιο ποσοστό επιτυχίας με με λιγότερες ιδιότητες.

Για 300 συνολικά ιδιότητες, έχουμε τις επιπλέον 150:

Πριν από τον όρο – στόχο

common	top	. the
text articles	print more on	overview
full text articles	] [	history of
encyclopedia plants	has	glossary
a-z list >	research	form of
list >	pure	by the
reference encyclopedia plants	archive	trees
as the	as	in a

Μετά από τον όρο – στόχο

help	may	like
there are	type	index
definition	this page	(-
became	or	common
#	are the	by alphabet :
around	can be	by alphabet
consists	does not	association
there	is not	from infoplease:

}	no	called	most	photo	click here to
known as	the word	tips more on	club	has a	. this
plants	western	[	is about	and other	to the
introduction	of a	some	which	what is	topics
as a	red	general	may	it is	resulted in
definition of	commercial	query of	pictures	von	resulted
gallery	free newsletter!	an	standard	the	faq
enchantedlearning.com	newsletter!	properties of	an	is	can
diseases	name:	are	from fact monster:	is a	is available
to top	: the	his	from fact	a	that
or	is the	and the	world	general information	is one of
Edit] noun	name	encyclopedia	museum	help learn more	had
] noun	items found for	chemistry	is an	help learn	forms
Noun	more on	>	" in all	they	should be
Eastern	word	(e.g., edison):	" in	publications	is in
>>	articles	edison):	also	belongs to	would be
site map encyclopedia	#	):	were	belongs	for a
map encyclopedia	subject:	v	is the	at the	does
yellow-bellied	called	and	can	plant	advertisement
is a	in ,	know about	have	on	general
, the	%	in .	society	test	lesson

Εύκολα παρατηρούμε ότι υπάρχουν πολλές ιδιότητες που δικαιολογούν την άνοδο της απόδοσης στις 300 ιδιότητες (“called”, “belongs to”, “is one of ” κ.τ.λ.). Επίσης έχουν αρχίσει να εμφανίζονται μερικές ιδιότητες-σκουπίδια όπως οι “(e.g., edison):”, “Eastern”, “newsletter!”, “advertisement” κ.α.

Οι παρουσία τους είναι λογική, αφού έχουν ήδη επιλεχθεί όλες οι ιδιότητες με μεγάλη ακρίβεια. Οι ιδιότητες με λιγότερο από 0,5 ακρίβεια δεν έχουν σημαντική πληροφορία για το σύστημα και καταλήγουν να «θολώνουν» την κρίση του.

Το ίδιο φαινόμενο παρατηρείται σε μεγαλύτερο βαθμό στις 450 και 600 ιδιότητες, αφού οι επιπλέον ιδιότητες έχουν ακρίβεια που κυμαίνεται από 0,3 σε 0,5. Αυτό εξηγεί και την πτώση στην απόδοση αυτών των συστημάτων.

4.7 Παρατηρήσεις πάνω στις ερωτήσεις αξιολόγησης

Κατά την διάρκεια της αξιολόγησης των συστημάτων παρατηρήσαμε ότι υπήρχαν κάποιες ερωτήσεις που κανένα σύστημα δεν κατάφερε να απαντήσει και άλλες στις οποίες

σχεδόν όλα τα συστήματα είχαν επιτυχία. Παραδείγματα ερωτήσεων που δεν απαντήθηκαν από κανένα σύστημα είναι οι εξής:

What is gasoline?

What is the gastrointestinal system?

Who was Geiger?

What is a generator?

Μερικές από τις παραπάνω ερωτήσεις (π.χ. «generator», «geography», «gasoline») φαίνεται πως ζητούν ορισμούς πολύ κοινών όρων. Η τεχνική μας βασίζεται κυρίως στην υπόθεση ότι όταν κάπου σε ένα κείμενο πρωτο-αναφέρεται ο όρος-στόχος, κοντά του εμφανίζεται και ένας σύντομος ορισμός του. Αυτό δεν ισχύει για κοινούς όρους, αφού ο συγγραφέας δεν θεωρεί ότι πρέπει να τους ορίσει.

Αντιθέτως, ερωτήσεις όπως οι παρακάτω, που αφορούν σχετικά άγνωστους όρους ή ζητούν να οριστούν ονόματα προσώπων, είχαν απαντηθεί σε ποσοστό άνω του 80% από το σύστημα μας:

What is galactose?

Who was Galois?

What is gametophyte?

Who was Goebbels?

What is Zither?

Το συμπέρασμα είναι θετικό, αφού το σύστημά μας θέλουμε να βρίσκει ορισμούς κυρίως όρων που δεν είναι ευρέως γνωστοί.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΤΑΣΕΙΣ

Σκοπός της εργασίας ήταν κυρίως η επανεξέταση των τεχνικών χειρισμού ερωτήσεων ορισμού που είχαν προτείνει οι Μηλιαράκη [1], Γαλάνης [2] και Γιακουμής [3] και η μεγαλύτερης κλίμακας πειραματική αξιολόγησή τους, με παράλληλο εντοπισμό των καλύτερων τιμών των παραμέτρων τους.. Μετά από πληθώρα πειραμάτων με μεταβλητούς αριθμούς ερωτήσεων εκπαίδευσης και αυτόματα επιλεγμένων ιδιοτήτων, καταλήξαμε ότι οι καταλληλότερες παράμετροι εκπαίδευσης του συστήματος είναι 2000 ερωτήσεις εκπαίδευσης και 300 αυτόματα επιλεγμένες ιδιότητες. Επίσης, επιβεβαιώσαμε ότι η μέθοδος αυτόματης δημιουργίας παραδειγμάτων εκπαίδευσης, όπως προτάθηκε από το Γαλάνη και βελτιώθηκε από το Γιακουμή, οδηγεί σε καλύτερες επιδόσεις από ότι η εκπαίδευση σε δεδομένα των διαγωνισμών TREC, όταν στόχος είναι ο εντοπισμός ορισμών σε ιστοσελίδες.

Θα προτείναμε ίσως μερικά ακόμα πειράματα πάνω στην θεματολογία των ερωτήσεων εκπαίδευσης. Επίσης ίσως θα ήταν πιο αποτελεσματική η χρήση κατωφλιών στην αυτόματη επιλογή των ιδιοτήτων. Αν οι ιδιότητες με μεγάλη ακρίβεια δείχνουν ότι η θετική τιμή τους σε ένα διάνυσμα συνιστά ορισμό, αντίστοιχα οι ιδιότητες με πολύ χαμηλή ακρίβεια θα ήταν απόδειξη μη-ορισμού. Τέλος, οι ιδιότητες με μέτρια τιμή ακρίβειας θα έπρεπε να αγνοούνται αφού ουσιαστικά προσθέτουν μόνο θόρυβο, που παραπλανάει το σύστημα.

ΑΝΑΦΟΡΕΣ

[1] **Miliaraki S. and Androutsopoulos I.**, "*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.

[2] **Androutsopoulos I. and Galanis D.**, "*An Unsupervised Learning Method to Identify Single-Snippet Answers to Definition Questions on the Web*", Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural LanguageProcessing (HLT/EMNLP), pages 323–330, Vancouver, 2005.

[3] **Γιακουμής Ε.**, "*Βελτιώσεις και περαιτέρω αξιολόγηση μεθόδου χειρισμού ερωτήσεων ορισμού για συστήματα ερωταποκρίσεων φυσικής γλώσσας*", πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005

[4] **Καρακατσιώτης Γ.**, "*Ανάπτυξη συστήματος χειρισμού ερωτήσεων ορισμού προσώπων για αρχεία εφημερίδων*", πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005

[5] **Μαυροειδής Δ.**, "*Αυτόματη κατάταξη ερωτήσεων φυσικής γλώσσας σε κατηγορίες*", πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005

[6] **Λουκαρέλλι Γ.**, "*Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα*", εργασία μεταπτυχιακού διπλώματος ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005

[7] **Eugenio B. Di and Glass. M.** "*The kappa statistic: A second look.*" Comput. Linguistics, 30(1):95–101, 2004

[8] **Voorhees Ellen M.**, "*Overview of the TREC-9 Question Answering Track*", National Institute of Standards and Technology, In Proceedings of TREC-9, 2000

[9] **Voorhees Ellen M.**, "*Overview of the TREC2001 Question Answering Track*",National Institute of Standards and Technology, In Proceedings of TREC-10 2001

[10] **Cristianini N. and Shawe-Taylor J.**, "*An Introduction to Support Vector Machines"*, Cambridge University Press, 2000.

[11] **Cortes C. and Vapnik V. P.**, "*Support-vector networks"*, Machine Learning, 20(3) (1995), pp. 273-297.

[12] **Vapnik V. P.**, "*Statistical Learning Theory"*, (1998).