



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Χειρισμός Ερωτήσεων Ορισμού σε Συστήματα
Ερωταποκρίσεων**

Μηλιαράκη Σπυριδούλα

A.M. 3990065

Επιβλέπων Καθηγητής : Ίων Ανδρουτσόπουλος

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ, 2003**

ΠΕΡΙΕΧΟΜΕΝΑ

Περιεχόμενα.....	2
Περίληψη.....	4
1. Εισαγωγή.....	5
1.1 Αντικείμενο της εργασίας.....	5
1.2 Διάρθρωση της εργασίας.....	7
2. Θεωρητικό υπόβαθρο.....	8
2.1 Τύποι Ερωτήσεων.....	8
2.2 Δημιουργία απαντήσεων.....	8
2.3 Αρχιτεκτονική συστημάτων ερωταποκρίσεων.....	9
3. Περιγραφή τεχνικών εργασίας.....	11
3.1 Τύπος ερωτήσεων.....	11
3.2 Δημιουργία απαντήσεων.....	11
3.3 Στάδια τεχνικής ερωταποκρίσεων εργασίας.....	13
3.4 Τρόπος αξιολόγησης της τεχνικής.....	14
3.5 Διάκριση των διαφορετικών προσεγγίσεων της εργασίας.....	15
4. Υλοποίηση τεχνικής με χρήση Wordnet.....	16
4.1 Περιγραφή τεχνικής.....	16
4.2 Υλοποίηση τεχνικής.....	17
4.3 Πειράματα.....	19
5. Χρήση Μηχανικής Μάθησης.....	23
5.1 Θεωρητικό υπόβαθρο.....	23
5.2 Περιγραφή χρήσης Μηχανικής Μάθησης.....	25
5.3 Περιγραφή αλγορίθμου Support Vector Machines.....	26
5.4 Χρήση περιορισμένου αριθμού ιδιοτήτων.....	28
5.4.1 Σύστημα ερωταποκρίσεων Joho & Sanderson.....	28
5.4.2 Ιδιότητες Μηχανικές Μάθησης.....	30

5.4.3 Αποτελέσματα.....	32
5.5 Επιλογή ιδιοτήτων μέσω Πληροφοριακού Κέρδους.....	34
5.5.1 Περιγραφή.....	34
5.5.2 Αποτελέσματα.....	35
5.6 Επιλογή ιδιοτήτων μέσω Ακρίβειας.....	37
5.6.1 Περιγραφή.....	37
5.6.2 Αποτελέσματα.....	38
5.7 Υπόλοιπα πειράματα.....	40
6. Σύγκριση αποτελεσμάτων.....	41
7. Συμπεράσματα.....	42
Αναφορές.....	43

ΠΕΡΙΛΗΨΗ

Τα συστήματα ερωταποκρίσεων φιλοδοξούν να αποτελέσουν την επόμενη γενιά των μηχανών αναζήτησης για συλλογές κειμένων και ιστοσελίδες. Επιτρέπουν στους χρήστες να θέτουν ερωτήματα σε φυσική γλώσσα (π.χ. «Που βρίσκεται το Οικονομικό Πανεπιστήμιο Αθηνών;», «Πόσα ξόδεψε για διαφήμιση η εταιρεία Microsoft το 2002;») και επιχειρούν να τα απαντήσουν εντοπίζοντας ή συνθέτοντας τμήματα των διαθέσιμων κειμένων ή ιστοσελίδων. Η εργασία εστιάζεται στο χειρισμό ερωτήσεων ορισμού (π.χ. «Ποιος ήταν ο Γαλιλαίος;», «Τι είναι η χοληστερίνη;»). Δημιουργήθηκε μια τεχνική ευρέσεως απαντήσεων για ερωτήσεις ορισμού, η οποία συνδυάζει τη χρήση μηχανικής μάθησης (Support Vector Machines) με προηγούμενες προσεγγίσεις που βασίζονταν στη χρήση του λεξικού WordNet και προτύπων φράσεων. Η τεχνική αξιολογήθηκε στα Αγγλικά χρησιμοποιώντας τις ερωτήσεις ορισμού και τις συλλογές κειμένων από τους διαγωνισμούς TREC9 (Text Retrieval Conference 2000) και TREC10 (2001), με αποτελέσματα που δείχνουν ότι έχει σημαντικά υψηλότερα ποσοστά επιτυχίας από προϋπάρχουσες μεθόδους.

1. ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Καθώς το διαδίκτυο γίνεται μέρος της καθημερινότητάς μας, ο μέσος χρήστης απαιτεί να βρίσκει τις πληροφορίες που επιθυμεί εύκολα και γρήγορα μέσα από έναν μεγάλο όγκο πληροφοριών. Είναι έκδηλη η ανάγκη δημιουργίας ενός συστήματος που θα επιτρέπει στον χρήστη να θέτει ερωτήσεις, όπως στην καθημερινή ζωή του, διατυπωμένες σε φυσική γλώσσα και να λαμβάνει απαντήσεις. Οι σημερινές μηχανές αναζήτησης απέχουν από το να δίνουν απαντήσεις και αρκούνται συχνά στο να επιστρέφουν ταξινομημένες λίστες εγγράφων που απλά περιέχουν λέξεις-κλειδιά. Η πλειοψηφία των χρηστών προτιμούν να τους δοθεί απευθείας η απάντηση παρά να χρειαστεί να βρουν οι ίδιοι την απάντηση μέσα σε ένα έγγραφο. Ένα **σύστημα ερωταποκρίσεων (Question Answering System)** έχει σκοπό να επιστρέφει σύντομες απαντήσεις και όχι ολόκληρα έγγραφα που τυχόν να τις περιέχουν.

Είναι αξιοσημείωτο ότι ήδη από το 1965, ο Simmons στο άρθρο του «*Answering English Questions by Computer*» [Simmons, 1965], αναφέρει περίπου 15 συστήματα ερωταποκρίσεων που είχαν δημιουργηθεί κατά τα 5 προηγούμενα χρόνια. Τα συστήματα αυτά περιλάμβαναν συστήματα που προσπαθούσαν να βρουν απαντήσεις από πηγές κειμένων όπως εγκυκλοπαίδειες. Τα επόμενα χρόνια, με ιδιαίτερη έμφαση στη δεκαετία του '80, δόθηκε μεγάλη βαρύτητα στα συστήματα ερωταποκρίσεων για βάσεις δεδομένων [Androutsopoulos κ.ά, 2000]. Σταδιακά και κυρίως προς τα τέλη του '90, η εμφάνιση του παγκόσμιου ιστού έστρεψε πάλι το ενδιαφέρον σε συλλογές κειμένων.

Μεγάλη ώθηση δόθηκε στην επιστροφή προς συστήματα ερωταποκρίσεων για συλλογές κειμένων με την παρουσίαση του **QA Track του TREC** (Text Retrieval Conference) το 1999 [Voorhees, 1999].

Η πλειοψηφία των συστημάτων που συμμετέχουν τα τελευταία χρόνια στο TREC QA Track, ακολουθεί μια παρόμοια γενική αρχιτεκτονική. Συγκεκριμένα, χρησιμοποιούνται καθιερωμένες τεχνικές ανάκτησης πληροφοριών για την εύρεση υποψήφιων εγγράφων που περιέχουν λέξεις-κλειδιά της ερώτησης, στην συνέχεια βρίσκεται η κατηγορία της ερώτησης (π.χ. χρόνου, τοποθεσίας, προσώπου) και τέλος εντοπίζονται τα τμήματα στα υποψήφια έγγραφα που περιέχουν συγκεκριμένες εκφράσεις (π.χ. χρονικές εκφράσεις στην περίπτωση

ερωτήσεων χρόνου, ονόματα προσώπων στην περίπτωση ερωτήσεων προσώπου κ.ο.κ.) κοντά στις λέξεις-κλειδιά.

Η εργασία εστιάζεται σε μια συγκεκριμένη κατηγορία ερωτήσεων, τις **ερωτήσεις ορισμού** (π.χ. «Τι είναι η θαλασσαιμία;», «Ποιος ήταν ο Γαλιλαίος;»). Αυτή η κατηγορία έχει ιδιαίτερο ενδιαφέρον διότι οι ερωτήσεις της δεν μπορούν να απαντηθούν με την τυπική αντιμετώπιση της εύρεσης συγκεκριμένων εκφράσεων (π.χ. εύρεση ονομάτων προσώπων), καθώς το μόνο συμπέρασμα που προκύπτει μέσω της κατηγορίας τους είναι ότι πρέπει να βρεθούν ονομαστικές φράσεις.

Η τεχνική εύρεσης απαντήσεων για ερωτήσεις ορισμού που δημιουργήθηκε κατά την εργασία, συνδυάζει τη χρήση μηχανικής μάθησης με προϋπάρχουσες μεθόδους που βασίζονταν στη χρήση του θησαυρού WordNet [Miller κ.ά., 1993] και προτύπων φράσεων. Η τεχνική αξιολογήθηκε χρησιμοποιώντας τις ερωτήσεις ορισμού και τις συλλογές κειμένων από τους διαγωνισμούς TREC9 (Text Retrieval Conference 2000) και TREC10 (2001).

Η ενασχόληση μόνο με τις ερωτήσεις ορισμού παρουσιάζει πρακτικό ενδιαφέρον. Επισημαίνεται ότι το σύνολο ερωτήσεων του διαγωνισμού TREC9 αποτελείται κατά 6% από ερωτήσεις ορισμού ενώ το σύνολο του TREC10 αποτελείται κατά 27% από ερωτήσεις ορισμού και θεωρείται ότι αποτυπώνει καλύτερα τις πραγματικές ερωτήσεις των χρηστών [Voorhees, 2001].

Χρησιμοποιήθηκε ο αλγόριθμος μάθησης των **Μηχανών Διανυσμάτων Υποστήριξης** (Support Vector Machines). Για κάθε ερώτηση ορισμού, εξάγονται τμήματα κειμένου από την αντίστοιχη συλλογή κειμένων, που θεωρείται πιθανόν να περιέχουν τον ορισμό, δηλαδή μια απάντηση στην ερώτηση. Ο αλγόριθμος εκπαιδεύεται στα χαρακτηριστικά των τμημάτων κειμένων που περιέχουν κάποιον ορισμό και των τμημάτων κειμένου που δεν περιέχουν ορισμό. Με βάση την εκπαίδευσή του, ο αλγόριθμος μαθαίνει να ταξινομεί τα τμήματα κειμένου στις 2 παραπάνω κατηγορίες (περιέχουν ή δεν περιέχουν ορισμό).

Τα αποτελέσματα των πειραμάτων που έγιναν στη διάρκεια της εργασίας δείχνουν ότι η τεχνική έχει σημαντικά υψηλότερα ποσοστά επιτυχίας από προϋπάρχουσες μεθόδους. Επομένως, η **Μηχανική Μάθηση** μπορεί να αποτελέσει καλή λύση στα συστήματα ερωταποκρίσεων, τουλάχιστον για ερωτήσεις ορισμού.

1.2 Διάρθρωση της εργασίας

Η υπόλοιπη εργασία ξεκινάει με μια παρουσίαση του απαραίτητου θεωρητικού υποβάθρου για τα Συστήματα Ερωταποκρίσεων στο Κεφάλαιο 2. Ακολουθεί στο Κεφάλαιο 3 η συνολική περιγραφή όλων των μεθόδων που χρησιμοποιήθηκαν για την υλοποίηση της τεχνικής ευρέσεως απάντησης σε ερωτήσεις ορισμού. Στο Κεφάλαιο 4 παρουσιάζεται η πρώτη μέθοδος η οποία χρησιμοποιεί το Wordnet και τα αποτελέσματα των πειραμάτων της μεθόδου αυτής. Το επόμενο κεφάλαιο, το Κεφάλαιο 5 παρουσιάζει την τεχνική που χρησιμοποιεί Μηχανική Μάθηση και τα αποτελέσματα των πειραμάτων. Τέλος, στα Κεφάλαιο 6 και 7 συνοψίζονται και συγκρίνονται τα αποτελέσματα των διαφορετικών προσεγγίσεων της εργασίας και αναφέρονται συμπεράσματα και μελλοντικές προσεγγίσεις.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Τύποι ερωτήσεων

Μπορούμε να διακρίνουμε 3 βασικούς τύπους ερωτήσεων βάσει του τύπου απάντησης που πρέπει να δοθεί [Hirschman, 2001]:

- **ερωτήσεις που επιδέχονται καθορισμένη απάντηση** (factual questions) που χωρίζονται σε υποκατηγορίες ερωτήσεων όπως:
 - ✓ τοποθεσίας : π.χ. « Που βρίσκεται το Οικονομικό Πανεπιστήμιο Αθηνών;»
 - ✓ χρόνου : π.χ. « Πότε αρχίζουν οι Ολυμπιακοί Αγώνες της Αθήνας το 2004;»
 - ✓ ονόματος προσώπου : π.χ. « Ποιος είναι ο πρωθυπουργός της Ελλάδας;»
 - ✓ ποσότητας : π.χ. « Πόσα χρόνια διήρκεσε ο Πελοποννησιακός πόλεμος;»
 - ✓ ορισμού : π.χ. « Τι είναι η νικοτίνη;» κ.ά.
- **ερωτήσεις γνώμης** (opinion) π.χ. « Ποιες θα είναι οι συνέπειες στην παγκόσμια οικονομία στην περίπτωση ενιαίου νομίσματος;»
- **ερωτήσεις περίληψης** (summary) π.χ. « Ποια η ιστορία που περιγράφεται στο βιβλίο “Εγκλημα και τιμωρία”;»

Η εργασία εστιάστηκε μόνο σε μια υποκατηγορία των ερωτήσεων που επιδέχονται καθορισμένη απάντηση, τις **ερωτήσεις ορισμού**. Η εργασία επικεντρώθηκε στα Αγγλικά, δεδομένου ότι το σύνολο των ερωτήσεων και οι συλλογές κειμένου που χρησιμοποιήθηκαν διατίθενται μόνο στα Αγγλικά.

Οι ερωτήσεις ορισμού στα Αγγλικά είναι της μορφής

«What/Who is/are <ονομαστική φράση> ?»

Π.χ.

- «Who was Socrates?»
- «What is osteoporosis?»

2.2 Δημιουργία απαντήσεων

Οι απαντήσεις ενός συστήματος ερωταποκρίσεων μπορεί να είναι μικρού ή μεγάλου μήκους (ενδεικτικά μήκη απαντήσεων αποτελούν τα 50 και 250 bytes). Επιπλέον, τα συστήματα που

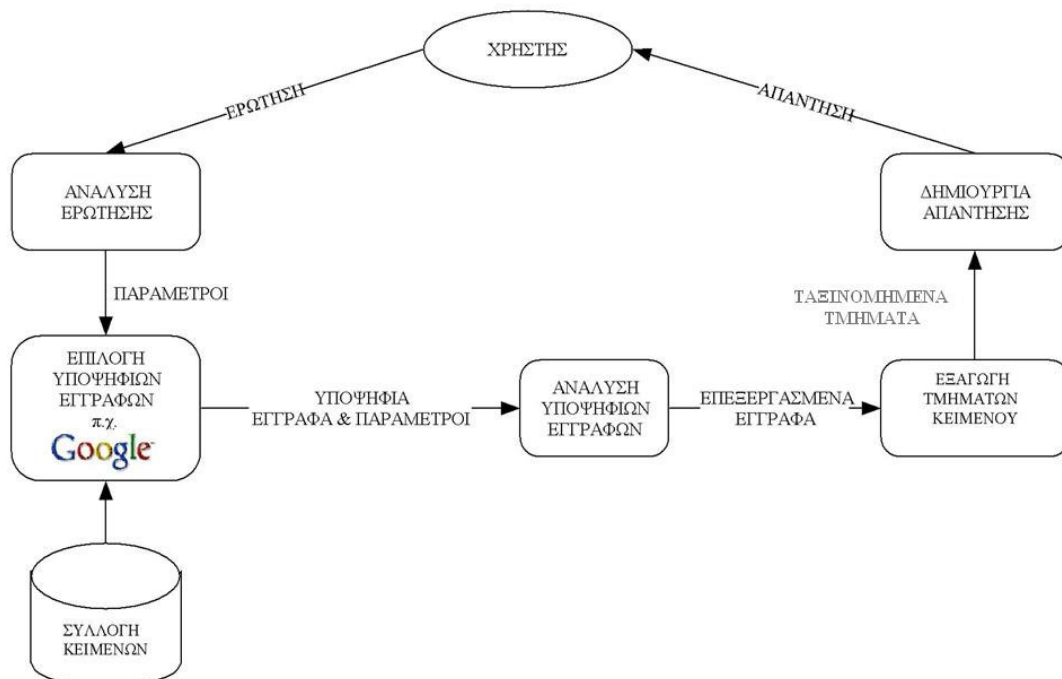
συμμετέχουν στους διαγωνισμούς του TREC επιτρέπεται συνήθως να επιστρέψουν μέχρι 5 απαντήσεις για κάθε ερώτηση [Voorhees, 1999].

Υπάρχουν διάφορες μεθοδολογίες για τη δημιουργία απάντησης. Ένα σύστημα ερωταποκρίσεων μπορεί να **εξάγει και να επιστρέφει κομμάτια από τα κείμενα** που περιέχουν την απάντηση ή να **παράγει μόνο του την απάντηση** (generation). Η παραγωγή της απάντησης πραγματοποιείται μεταξύ άλλων και με τη χρήση τεχνικών παραγωγής φυσικής γλώσσας (natural language generation) [Reiter & Dale, 2000].

Όλες οι τεχνικές που υλοποιήθηκαν κατά τη διάρκεια της εργασίας περιορίζονται στο να εξάγουν την απάντηση από τα κείμενα. Όταν η απάντηση προέρχεται από πολλές διαφορετικές προτάσεις, η συνοχή της εξαγόμενης απάντησης είναι μειωμένη και μπορούν να χρησιμοποιηθούν τεχνικές παραγωγής φυσικής γλώσσας για να ενωθούν τα επιμέρους τμήματα σε μια απάντηση με μεγαλύτερη συνοχή. Η εργασία όμως δεν ασχολείται με θέματα σύνδεσης τμημάτων.

2.3 Αρχιτεκτονική Συστημάτων ερωταποκρίσεων

Η γενική αρχιτεκτονική ενός συστήματος ερωταποκρίσεων φαίνεται στο σχήμα 2.1.



Σχήμα 2.1 Γενική αρχιτεκτονική ενός συστήματος ερωταποκρίσεων

Ακολουθεί ανάλυση κάθε σταδίου επεξεργασίας:

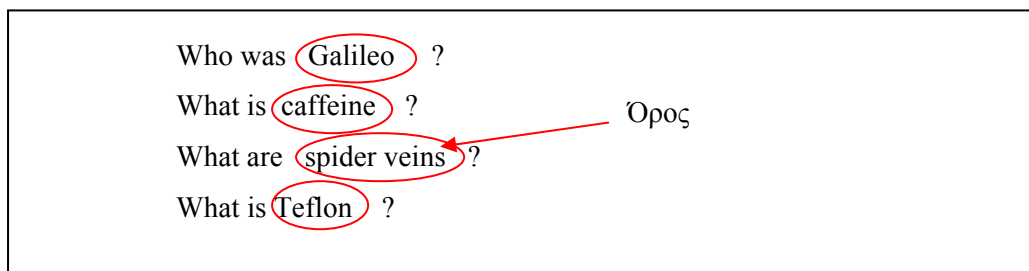
1. **Ανάλυση Ερώτησης:** Η ερώτηση που δίνεται από τον χρήστη σε φυσική γλώσσα πρέπει να αναλυθεί και να προκύψουν κάποιοι παράμετροι που θα την περιγράψουν και θα είναι χρήσιμες στα υπόλοιπα στάδια επεξεργασίας. Είδη παραμέτρων θα μπορούσαν να αποτελούν ο τύπος απάντησης που πρέπει να δοθεί, η κατηγορία της ερώτησης ή κάποιες λέξεις-κλειδιά.
2. **Επιλογή υποψήφιων εγγράφων:** Με την χρήση κάποιας μηχανής αναζήτησης (π.χ. Google) επιλέγονται τα έγγραφα στα οποία θα γίνει η αναζήτηση της απάντησης στην ερώτηση. Πρέπει να έχει αποφασιστεί το πλήθος των εγγράφων που θα χρησιμοποιηθούν δεδομένου ότι η πλειοψηφία των μηχανών αναζήτησης επιστρέφουν μεταβλητό πλήθος εγγράφων. Στα συστήματα που συμμετείχαν στο διαγωνισμό του TREC το 2001, είχαν δοθεί τα 50 πρώτα έγγραφα για κάθε ερώτηση όπως τα επέστρεψε η μηχανή αναζήτησης PRISE [Voorhees, 2001].
3. **Ανάλυση υποψήφιων εγγράφων:** Μετά την επιλογή των υποψήφιων εγγράφων, ακολουθεί το στάδιο της ανάλυσής τους. Τα έγγραφα μπορούν να αναλυθούν με ένα εργαλείο αναγνώρισης ονομάτων οντοτήτων, το οποίο αναγνωρίζει και ταξινομεί συμβολοσειρές ως ονόματα ατόμων, εταιρειών, τοποθεσιών, κ.ά., μια μονάδα εντοπισμού χρονικών εκφράσεων κ.τ.λ. Σε αυτό το στάδιο μπορεί επιπλέον να χρησιμοποιηθεί ένας διαχωριστής προτάσεων (sentence splitter) ή ακόμη και ένας συντακτικός αναλυτής.
4. **Εξαγωγή τμημάτων κειμένου:** Από τα επεξεργασμένα έγγραφα εξάγονται τμήματα κειμένου που πιθανόν να περιέχουν την απάντηση στην ερώτηση. Αν έχει χρησιμοποιηθεί διαχωριστής προτάσεων, κάθε τμήμα κειμένου μπορεί να είναι μια ολοκληρωμένη πρόταση. Διαφορετικά μπορεί να είναι ένα «παράθυρο», δηλαδή μια συμβολοσειρά συγκεκριμένου μήκους. Τα τμήματα κειμένου ταξινομούνται ανάλογα με την πιθανότητα να αποτελούν σωστές απαντήσεις.
5. **Δημιουργία απάντησης:** Δημιουργείται η απάντηση ή οι απαντήσεις που θα επιστραφούν στον χρήστη τελικά είτε με παραγωγή είτε με εξαγωγή. Μαζί με τις απαντήσεις συνήθως παρέχονται και σύνδεσμοι προς τα κείμενα από τα οποία προέρχονται οι απαντήσεις καθώς και βαθμοί βεβαιότητας αναλόγως της πιθανότητας που έχει το κάθε κείμενο να περιέχει τη σωστή απάντηση.

3. ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΠΡΟΣΕΓΓΙΣΗΣ ΤΗΣ ΕΡΓΑΣΙΑΣ

3.1 Τύπος ερωτήσεων

Η τεχνική ευρέσεως απάντησης της εργασίας αφορά μόνο την κατηγορία των ερωτήσεων ορισμού, όπως έχει ήδη αναφερθεί. Χρησιμοποιήθηκε το σύνολο των ερωτήσεων ορισμού από τον διαγωνισμό **TREC9** (Text REtrieval Conference 2000) και τον διαγωνισμό **TREC10** (2001).

Για να πραγματοποιηθεί η αξιολόγηση της τεχνικής, απομονώθηκαν οι ερωτήσεις ορισμού, δηλαδή οι ερωτήσεις με μορφή “What/Who is/are <ονομαστική φράση>?” και τελικά το σύνολο των ερωτήσεων περιείχε 160 ερωτήσεις (136 από τον TREC10 και 24 από τον TREC9).



Σχήμα 3.1 Παραδείγματα επιλεγθέντων ερωτήσεων και των όρων τους

Για μεγαλύτερη ευκολία δεν χρησιμοποιήθηκαν ολόκληρες οι ερωτήσεις ορισμού αλλά για κάθε ερώτηση απομονώθηκε χειρωνακτικά και χρησιμοποιήθηκε μόνο ο **όρος** (ονομαστική φράση) του οποίου ζητείται ο ορισμός. Έτσι αποφεύχθηκε η υλοποίηση της επεξεργασίας εξαγωγής από κάθε ερώτηση του όρου που ζητείται να οριστεί.

3.2 Δημιουργία απαντήσεων

Η αναζήτηση των απαντήσεων γίνεται στις συλλογές κειμένων που διατίθενται από τους διαγωνισμούς TREC9 και TREC10 και αναμένεται να περιέχουν τις απαντήσεις. Συγκεκριμένα σε κάθε ερώτηση αντιστοιχούν **50 κείμενα** τα οποία αποτελούνται κυρίως από

άρθρα εφημερίδων [Voorhees, 2001] και έχουν επιστραφεί από μια μηχανή αναζήτησης. Δεν χρησιμοποιείται κάποιος διαχωριστής απαντήσεων, επομένως επιστρέφονται κομμάτια κειμένου που δεν αποτελούνται υποχρεωτικά από ολόκληρες προτάσεις. Αυτά τα κομμάτια κειμένου θα αναφέρονται στην συνέχεια ως **παράθυρα κειμένου**.

Ερώτηση : What is cholesterol?

Απαντήσεις :

1. [And not only should you know your number but you should know what it means . "Cholesterol is a waxy substance that is ferried through the bloodstream by substances called lipoproteins ; depending on the fatty proteins attached to it ,]
2. [blood cholesterol , a fatty substance that can clog arteries contribute to heart disease and heart attacks . Besides cholesterol , other factors that have been identified as factors in heart disease are high blood pressure , obesity , smoking and]
3. [to get people interested in knowing how much of the fatty substance is in their arteries has made cholesterol a popular topic from cocktail parties to medical-screening booths in shopping malls . Researchers are releasing new studies]

Ερώτηση : What is strep throat?

Απαντήσεις :

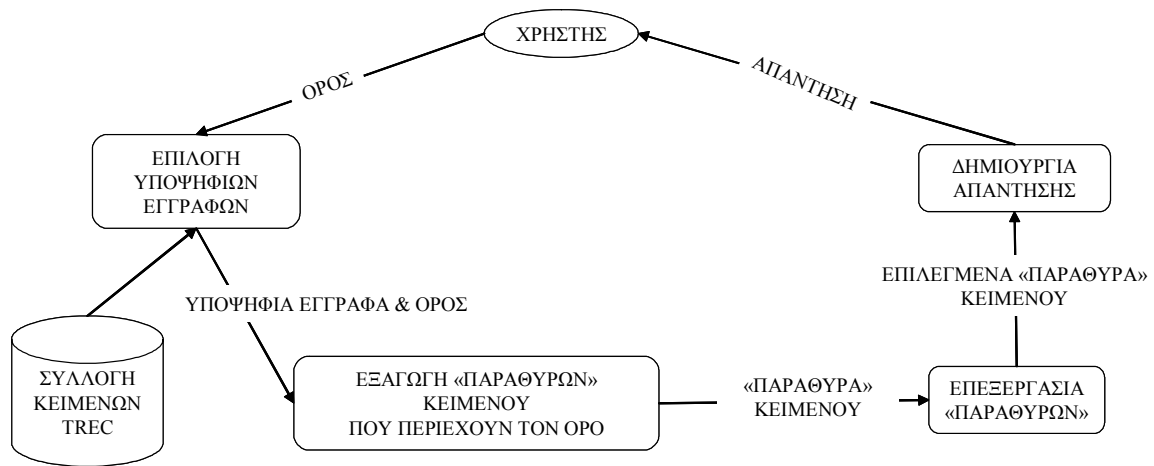
1. [changes in the pattern of this disease the agency said Most strep infections are mild such as the typical case of strep throat and most cases can be cleared up with antibiotics such as penicillin Schwartz said Very rarely does it spread to]
2. [one person in 10 is infected at any given time Usually there are no symptoms sometimes it produces a sore throat strep throat But streps can also cause a terrifying variety of diseases depending on which toxins they release They were responsible for]
3. [the bloodstream and cause death Glomerulonephritis can appear by itself or in conjunction with a chronic disease called lupus strep throat or a variety of infections The researchers gave rats a form of glomerulonephritis by making the rats disease]

Σχήμα 3.2 Παράδειγμα απαντήσεων όπως επιστρέφονται μέσω της τεχνικής

Το μήκος των απαντήσεων που επιστρέφονται είναι **250 bytes** και συνολικά επιστρέφονται μέχρι **5 απαντήσεις για κάθε ερώτηση**. Το σχήμα 3.2 δείχνει παραδείγματα των ερωτήσεων και των παραθύρων που επιστρέφει το σύστημα που αναπτύχθηκε στη διάρκεια της εργασίας. Οι υπογραμμίσεις έχουν προστεθεί χειρωνακτικά.

3.3 Στάδια τεχνικής ερωταποκρίσεων εργασίας

Όλες οι τεχνικές που υλοποιήθηκαν κατά την εργασία αποτελούνται από συγκεκριμένα στάδια επεξεργασίας, τα οποία φαίνονται στο σχήμα 3.3 που ακολουθεί.



Σχήμα 3.3 Στάδια επεξεργασίας των τεχνικών της εργασίας

Αρχικά, δίνεται ο όρος της ερώτησης ορισμού που επιθυμούμε να απαντηθεί. Συγκριτικά με την γενική αρχιτεκτονική ενός συστήματος ερωταποκρίσεων, όπως περιγράφηκε στο κεφάλαιο 2.3, το στάδιο ανάλυσης της ερώτησης παραλείπεται και θεωρείται ότι ο όρος της κάθε ερώτησης έχει ήδη εξαχθεί.

Το πρώτο στάδιο αφορά την επιλογή των υποψήφιων εγγράφων που θα αποτελέσουν το σύνολο των κειμένων μέσα στα οποία θα αναζητηθεί η απάντηση. Τα υποψήφια έγγραφα που χρησιμοποιήθηκαν για κάθε ερώτηση κατά την εργασία είναι το σύνολο των κειμένων που προτείνεται από τους αντίστοιχους διαγωνισμούς TREC. Συγκεκριμένα, σε κάθε ερώτηση αντιστοιχούν **50 κείμενα** όπως περιγράφηκαν στην ενότητα 3.2. Τα κείμενα αυτά είναι

ταξινομημένα με βάση την κατάταξη (**rank**) του κάθε κειμένου που κυμαίνεται από 1 έως 50 και η οποία αντιστοιχεί στην σειρά με την οποία επέστρεψε το κάθε κείμενο η μηχανή αναζήτησης που χρησιμοποιήθηκε.

Στην συνέχεια, γίνεται αναζήτηση του όρου μέσα στο σύνολο των υποψήφιων εγγράφων. Το στάδιο εξαγωγής των παραθύρων κειμένου, δημιουργεί παράθυρα συγκεκριμένου μήκους με μοναδικό κριτήριο να περιέχουν τον ζητούμενο όρο. Το σύνολο αυτών των παραθύρων θα αποτελέσουν τις **υποψήφιες απαντήσεις**.

Μια ενδεχόμενη προσθήκη σε αυτό το στάδιο, θα μπορούσε να είναι η εύρεση των παραθύρων που είναι σχεδόν όμοια μεταξύ τους, ώστε το σύνολο των υποψήφιων απαντήσεων να αποτελείται από όσο το δυνατόν διαφορετικά παράθυρα κειμένου.

Ακολουθεί το βασικότερο στάδιο όπου γίνεται η επεξεργασία των παραθύρων κειμένου, ώστε να επιλεγούν οι απαντήσεις που θα επιστραφούν. Αυτό είναι το στάδιο στο οποίο διαφοροποιούνται οι διάφορες προσεγγίσεις που χρησιμοποιήθηκαν κατά την εργασία. Με διαφορετικά κριτήρια αναλόγως της προσέγγισης, τα οποία θα αναλυθούν στην συνέχεια, κάθε παράθυρο κειμένου **βαθμολογείται** ως προς την πιθανότητα να περιέχει τον ορισμό δηλαδή την απάντηση.

Η απάντηση που δημιουργείται στο τελευταίο στάδιο αποτελείται από τα 5 παράθυρα κειμένου που έχουν την μεγαλύτερη πιθανότητα να περιέχουν την απάντηση.

3.4 Τρόπος αξιολόγησης της τεχνικής

Το υλικό που παρέχει το TREC αποτελείται από τις ερωτήσεις, τα 50 κείμενα που επέστρεψε για κάθε ερώτηση η μηχανή αναζήτησης και τα **πρότυπα απαντήσεων (answer patterns)** για κάθε μία ερώτηση. Τα πρότυπα απαντήσεων είναι στην γλώσσα Perl και αντιστοιχούν στη μορφή μιας σωστής απάντησης. Επισημαίνεται ότι υπάρχουν ερωτήσεις που δεν έχουν απάντηση μέσα στο σύνολο κειμένων και επομένως δεν έχουν πρότυπα απάντησης.

Ερώτηση :	What are polymers?
Πρότυπα απαντήσεων :	plastics? polypropylene/polyam blend long chains of.*molecules
	} Πρότυπα απαντήσεων

Σχήμα 3.4 Παράδειγμα προτύπων απαντήσεων στην γλώσσα Perl

Κατά την αξιολόγηση της κάθε προσέγγισης, μία απάντηση θεωρήθηκε σωστή εάν ταίριαζε με τουλάχιστον ένα από τα αντίστοιχα πρότυπα απαντήσεων από το TREC. Θεωρήθηκε επιτυχία να υπάρχει τουλάχιστον μία σωστή απάντηση στο σύνολο των 5 απαντήσεων που επιστρέφονται. Για την διεξαγωγή της αξιολόγησης χρειάστηκε να οργανωθούν όλοι οι όροι μαζί με τα αναγνωριστικά τους, η λίστα των προτύπων απαντήσεων και το σύνολο των κειμένων.

3.5 Διάκριση των διαφορετικών προσεγγίσεων της εργασίας

Αρχικά υλοποιήθηκε και αξιολογήθηκε μια μέθοδος επιλογής παραθύρων των Prager, Radev και Czuba [Prager κ.ά., 2001] που χρησιμοποιεί τον ιεραρχικό θησαυρό Wordnet [Miller κ.ά., 1993]. Στην συνέχεια υλοποιήθηκε και αξιολογήθηκε μια προσέγγιση που χρησιμοποιεί **Μηχανική Μάθηση (Machine Learning)** με 3 διαφορετικούς τρόπους :

1. Χρήση χειρωνακτικά επιλεγμένων **ιδιοτήτων**
2. Χρήση ιδιοτήτων που επιλέγονται από το σώμα εκπαίδευσης και επιλογή μέσω του **Πληροφοριακού Κέρδους (Information Gain)**
3. Χρήση ιδιοτήτων που επιλέγονται από το σώμα εκπαίδευσης και επιλογή μέσω της **Ακρίβειας (Precision)**

Επισημαίνεται ότι για την υλοποίηση της προσέγγισης με Μηχανική Μάθηση χρησιμοποιήθηκε ο αλγόριθμος μάθησης των **Μηχανών Διανυσμάτων Υποστήριξης** (Support Vector Machines). Στα επόμενα κεφάλαια περιγράφονται αναλυτικά οι μέθοδοι που υλοποιήθηκαν και αξιολογήθηκαν.

4. Υλοποίηση τεχνικής με χρήση Wordnet

4.1 Περιγραφή τεχνικής

Οι Prager, Radev και Czuba [Prager κ.ά., 2001] έχουν παρουσιάσει μια τεχνική (Predictive annotation) που αποτελεί μια μεθοδολογία για την απάντηση ερωτήσεων αναζήτησης γεγονότων (fact-seeking questions). Η τεχνική βασίζεται στην αναγνώριση του τύπου της απάντησης, μέσω ανάλυσης της ερώτησης. Για παράδειγμα, ερωτήσεις της μορφής «Who ...» αναζητούν άτομα ή οργανισμούς, ερωτήσεις της μορφής «Where ...» αναζητούν μέρη και ερωτήσεις της μορφής «When...» αναζητούν χρόνους κ.ο.κ. Οπότε ο στόχος είναι να βρεθεί ο σωστός τύπος απάντησης στην συλλογή κειμένων. Αυτό επιτυγχάνεται με εμπλουτισμό των κειμένων με ψευδο-λεκτικές μονάδες (QA-tokens) όπως «PERSON\$», «PLACE\$», «MONEY\$», «THING\$» κ.α. που σηματοδοτούν εκφράσεις των αντίστοιχων τύπων.

Ενώ η τεχνική σχεδιάστηκε με στόχο να δίνει απαντήσεις σε όλους τους τύπους ερωτήσεων, χρειαζόταν ξεχωριστή προσέγγιση για τις ερωτήσεις ορισμού. Η μειωμένη επιτυχία στις ερωτήσεις ορισμού οφειλόταν στο ότι δεν φανερώνουν τον επιθυμητό τύπο απάντησης οπότε δεν αρκεί να αναζητήσει κανείς κάποια συγκεκριμένη ψευδο-λεκτική μονάδα. Επιπλέον η ψευδο-λεκτική μονάδα που προέκυπτε ότι έπρεπε να βρεθεί ήταν μια ονομαστική φράση (QA token THINGS) το οποίο είναι γενικό και εμφανίζεται συχνά στην συλλογή κειμένων. Έτσι προέκυψε μια νέα τεχνική (Virtual annotation) [Prager κ.ά. 2002] που αποτελεί επέκταση της προηγούμενης τεχνικής τους για ερωτήσεις ορισμού (What-Is questions) και η οποία υλοποιήθηκε κατά την εργασία. Η τελευταία τεχνική θα αναφέρεται ως «**Μέθοδος P&C**».

Η τεχνική στηρίζεται στην χρήση των υπερωνύμων ενός όρου. **Υπερώνυμο** ενός όρου αποτελεί μία λέξη με γενικότερη σημασία από τον όρο. Π.χ. η λέξη «disease» αποτελεί υπερώνυμο της λέξης «pneumonia». Συνήθως μια πρόταση που ορίζει έναν όρο περιέχει τόσο τον όρο όσο και κάποιο υπερώνυμο του.

Παράδειγμα : Ερώτηση: «What is nicotine?»
Απάντηση: «Nicotine is a poison.»

υπερώνυμο του όρου
nicotine

Στην ερώτηση «What is nicotine?», μια σωστή απάντηση θα ήταν η πρόταση «Nicotine is a poison», όπου η λέξη poison αποτελεί ένα υπερώνυμο της λέξης nicotine.

Δεν αρκεί όμως να αναφερθεί ως απάντηση μόνο ένα υπερώνυμο, διότι συχνά δεν αποτελεί επαρκή απάντηση από μόνο του. Για παράδειγμα, δεν είναι επαρκές να δοθεί η απάντηση «Ο κροκόδειλος είναι ένα ερπετό» στην ερώτηση «Τι είναι ο κροκόδειλος;». Είναι καλύτερα να βρεθεί ένα παράθυρο κειμένου που να περιέχει τις λέξεις «κροκόδειλος» και «ερπετό» και που λογικά θα περιέχει και επιπλέον πληροφορίες (π.χ. «Ο κροκόδειλος, το γνωστό ερπετό που ζει στους βάλτους...»).

4.2 Υλοποίηση τεχνικής

Υπάρχουν θησαυροί όπως το Wordnet [Miller κ.ά., 1993] που περιέχουν οργανωμένα τα ουσιαστικά, τα ρήματα και τα επίθετα της Αγγλικής γλώσσας. Με την χρήση του Wordnet βρίσκουμε για έναν όρο τα υπερώνυμα του. Ακολουθεί ένα παράδειγμα με τα υπερώνυμα του όρου «computer».

Επίπεδο	Υπερώνυμα
0	{computer, computing machine, computing device, data processor, electronic computer, information processing system}
1	{machine}
2	{device}
3	{instrumentality, instrumentation}
4	{artifact, artefact}
5	{object, physical object}
6	{entity, physical thing}
7	{whole, whole thing, unit}
8	{object, physical object}
9	{entity, physical thing}

Σχήμα 4.1 Υπερώνυμα της λέξης “computer” όπως τα επιστρέφει το Wordnet

Κάθε λέξη έχει πολλά υπερώνυμα, κάποια από αυτά είναι πιο ειδικά και άλλα πιο γενικά. Όπως φαίνεται στο σχήμα 4.1, δύο υπερώνυμα της λέξης «computer» είναι οι λέξεις

«machine» και «object». Όμως, δεν θα ήταν χρήσιμο όταν κάποιος ρωτάει «What is a computer?» να του δοθεί η απάντηση «Computer is an object». Πρέπει επομένως να βρεθεί κάποιο μέτρο καταλληλότητας βάσει του οποίου να βρίσκεται το καταλληλότερο υπερώνυμο για κάποιον όρο.

Ο αλγόριθμος εύρεσης του καταλληλότερου υπερωνύμου (Wordnet look-up algorithm) μετράει τις ταυτόχρονες εμφανίσεις του όρου με κάθε ένα από τα υπερώνυμα του μέσα σε παράθυρα συγκεκριμένου μήκους και διαιρεί το πλήθος αυτό με το επίπεδο του υπερωνύμου.

Το μέτρο αυτό αναφέρεται ως Level Adapted Count (LAC), όπου :

$$LAC = \frac{\text{Πλήθος ταυτόχρονων εμφανίσεων όρου και υπερωνύμου}}{\text{Επίπεδο υπερωνύμου}}$$

Στην υλοποίηση της μεθόδου P&C, επιλέγεται το υπερώνυμο με το υψηλότερο LAC καθώς και όσα υπερώνυμα έχουν LAC εντός κατωφλίου 20%. Π.χ. Εάν το «καλύτερο» υπερώνυμο έχει LAC 30 και υπάρχει ένα άλλο υπερώνυμο με LAC 25 τότε θα προταθούν και τα δύο διότι το LAC 25 είναι εντός του 20% του LAC 30.

Τα υπερώνυμα των πολύ υψηλών επιπέδων είναι συνήθως άχρηστα, γιατί αντιστοιχούν σε πολύ γενικές έννοιες. Για να μειωθεί η πιθανότητα επιλογής υπερωνύμου πολύ υψηλού επιπέδου, η αναζήτηση γίνεται μέχρι κάποιο ανώτατο επίπεδο. Επισημαίνεται ότι στο σύνολο των υπερωνύμων ενός όρου συμπεριλαμβάνονται και τα συνώνυμα του, τα οποία θεωρούνται υπερώνυμα μηδενικού επιπέδου.

Η τεχνική όπως υλοποιήθηκε, επιστρέφει παράθυρα κειμένου που θεωρεί ότι περιέχουν ορισμό του όρου της ερώτησης. Μια μελλοντική βελτίωση θα μπορούσε να είναι η επιστροφή αυτοτελών προτάσεων ή περιόδων με την χρήση ενός εργαλείου αναγνώρισης τέλους περιόδου.

Συγκεκριμένα δοθέντος ενός όρου (question term) γίνονται τα παρακάτω βήματα :

1. Εύρεση των 50 κειμένων που επέστρεψε η μηχανή αναζήτησης για τον όρο.
2. Εύρεση όλων των υπερωνύμων και των συνωνύμων του όρου μέσω του Wordnet.

3. Εύρεση του πλήθους ταυτόχρονων εμφανίσεων του κάθε υπερωνύμου και του όρου μέσα σε παράθυρα 250 bytes στο σύνολο των 50 κειμένων.
4. Υπολογισμός κριτηρίου LAC (Level-Adapted Count) για κάθε υπερώνυμο.
5. Εύρεση του υπερωνύμου με το μεγαλύτερο LAC καθώς και όλων των υπολοίπων υπερωνύμων με LAC εντός ενός προκαθορισμένου κατωφλίου 20%.
6. Εύρεση όλων των παραθύρων κειμένου μέσα στα 50 κείμενα που περιέχουν ταυτόχρονα τον όρο και ένα από τα υπερώνυμα του προηγούμενου βήματος.
7. Επιστροφή των παραθύρων του προηγούμενου βήματος ταξινομημένων βάσει της κατάταξης (rank) των κειμένων μέσα στα οποία βρέθηκαν. Η κατάταξη ενός κειμένου είναι η σειρά (1-50) με την οποία επέστρεψε το κείμενο η μηχανή αναζήτησης.

4.3 Πειράματα

Στην περίπτωση που ένας όρος δεν συμπεριλαμβανόταν στο Wordnet τότε η τεχνική δεν επέστρεφε κανένα παράθυρο, σε αντίθεση με την πρωτότυπη τεχνική όπου σε περίπτωση όπου δεν επιστρεφόταν κανένα παράθυρο γινόταν χρήση της αρχικής μεθόδου Predictive Annotation.

Επιτυχία θεωρείται η επιστροφή μιας σωστής απάντησης μέσα στα 5 κορυφαία παράθυρα. Ακολουθούν τα συνολικά αποτελέσματα για τις 160 ερωτήσεις. Ο χρόνος εκτέλεσης των πειραμάτων ήταν σχετικά περιορισμένος, δηλαδή λίγα δευτερόλεπτα ανά ερώτηση.

Αποτελέσματα	Υλοποίηση εργασίας		Υλοποίηση P&C	
	ερωτήσεις	ποσοστό	ερωτήσεις	ποσοστό
Συνολική επιτυχία	80 / 160	50.00%	80 / 154	51.95%
Επιτυχία σε ερωτήσεις για τις οποίες υπάρχει απάντηση στα κείμενα	80 / 137	58.39%	80 / 133	60.15%
Επιτυχία σε ερωτήσεις των οποίων ο όρος υπάρχει στο Wordnet	80 / 130	61.54%	71 / 125	56.08%

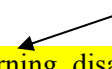

Πίνακας 4.1 Αποτελέσματα υλοποίησης της τεχνικής P&C

Στον πίνακα 4.1, τα αποτελέσματα χωρίζονται σε 3 κατηγορίες, στα ποσοστά επιτυχίας όλων των ερωτήσεων, στα ποσοστά επιτυχίας των ερωτήσεων που έχουν απάντηση εντός των κειμένων και τέλος στα ποσοστά των ερωτήσεων που ο όρος τους περιλαμβάνεται στο λεξικό Wordnet. Στην δεύτερη κατηγορία δεν περιλαμβάνονται οι ερωτήσεις που δεν έχουν απάντηση και οι ερωτήσεις που δεν έχουν κάποιο πρότυπο απάντησης ώστε να μπορεί να ελεγχθεί αν ένα παράθυρό τους αποτελεί ορισμό.

Τα ποσοστά επιτυχίας της υλοποίησης της παρούσας εργασίας δεν είναι άμεσα συγκρίσιμα από τα αντίστοιχα ποσοστά που αναφέρονται στην εργασία των P&C. Αυτό οφείλεται αρχικά στην έκδοση Wordnet που χρησιμοποιήθηκε, η οποία περιείχε περισσότερους όρους από την αντίστοιχη που χρησιμοποιήθηκε στην μέθοδο P&C καθώς και στην διαφορά μεταξύ των συνόλων ερωτήσεων. Ακόμη, σε κάποιες ερωτήσεις η υλοποίηση των P&C επέστρεφε παράθυρα μήκους 50 bytes (TREC10), αισθητά μικρότερες από τις απαντήσεις μήκους των 250 bytes που επιστρέφει η υλοποίηση της παρούσας εργασίας.

Η τεχνική είχε υψηλά ποσοστά στις ερωτήσεις που οι όροι τους περιλαμβάνονταν στο Wordnet. Συνολικά όμως τα ποσοστά δεν ήταν αρκετά υψηλά και δεν θα μπορούσε να αποτελέσει καλή λύση χωρίς την προσθήκη επιπλέον κριτηρίων. Βέβαια ο πιθανός συνδυασμός της τεχνικής με άλλες τεχνικές, όπως έγινε στην εργασία, έχει καλύτερες προοπτικές.

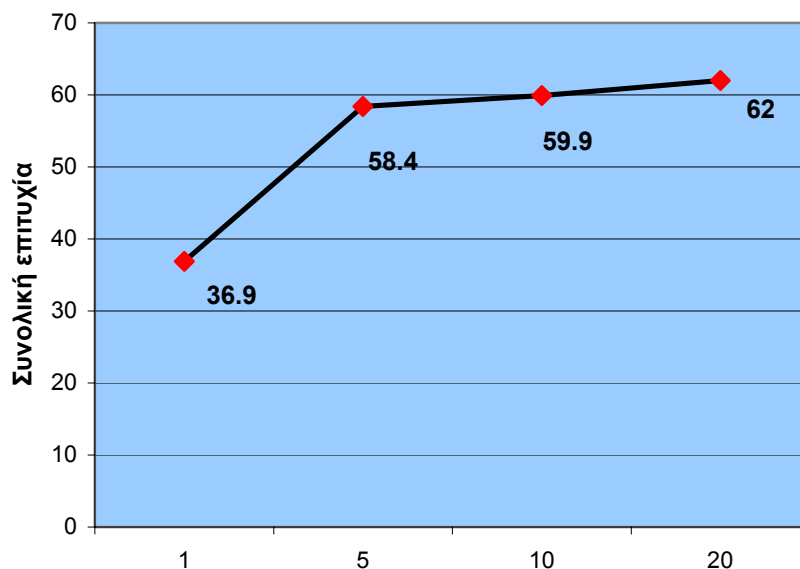
Ακολουθούν παραδείγματα σωστών απαντήσεων που προέκυψαν κατά τα πειράματα:

- Ερώτηση : “What is autism?”
 - Επιλεγμένα υπερώνυμα: syndrome
 - Απάντηση : {represent a wide range of **physical and learning disabilities**, including cerebral palsy, Down's **syndrome**, blindness, deafness, orthopedic disabilities, **autism** and spina bifida. On Monday, the first day of the workshop, the youngsters broke up into three groups}
- Ορισμός**
- 
- Ερώτηση : “What is a shaman?”
 - Επιλεγμένα υπερώνυμα : priest
 - Απάντηση : {out but they traveled along the Silk Road for thousands of years Basilov s favorite subject in his rich field is shamanism A **shaman is a kind of priest** Basilov likes it because I was able to find something new in my field Always I had a}
- Ορισμός**
- 

Η τεχνική όμως, δεν λειτούργησε όπως αναμενόταν σε κάποιες περιπτώσεις. Υπήρχαν περιπτώσεις όπου τυχαία εμφανιζόταν ο όρος και το υπερώνυμο του ταυτόχρονα αλλά το υπερώνυμο είχε διαφορετική σημασία. Στο παρακάτω παράδειγμα τόσο ο όρος “Caldera” είναι όνομα αλλά και το υπερώνυμο “opening” χρησιμοποιείται όχι με την σημασία του ουσιαστικού “opening” αλλά του ρήματος “opening up” :

- Ερώτηση : “What is a caldera?”
- Επιλεγμένα υπερώνυμα : opening
- Απάντηση : {Fedecamaras. President Caldera pointed out that all sectors of society must join Solidarity's efforts so it can make progress. Caldera also talked about the importance of opening up opportunities for foreign investment by providing greater legal guarantees. He said}

Επιπλέον, λόγω της απουσίας κάποιου ειδικού μηχανισμού ταξινόμησης των παραθύρων που η μέθοδος θεωρεί πως περιέχουν ορισμούς, (χρησιμοποιείται μόνο η κατάταξη των εγγράφων από όπου προέρχονται, όπως επιστρέφεται από τη μηχανή αναζήτησης) συχνά τα παράθυρα με τις σωστές απαντήσεις δεν συμπεριλαμβάνονται στα πέντε κορυφαία. Συγκεκριμένα, όπως φαίνεται στο σχήμα 4.2, αν θεωρούσαμε σωστή μια απάντηση που περιέχεται στα 10 κορυφαία (αντί για τα 5 κορυφαία), τότε η επιτυχία των όρων που έχουν ορισμό θα αυξανόταν κατά **1.5%**, ενώ αν ελέγχαμε τις 20 πρώτες απαντήσεις τότε θα αυξανόταν κατά **3.6%**.



Σχήμα 4.2 Το πλήθος απαντήσεων αντιστοιχεί στο πόσες απαντήσεις επιστρέφουμε και θεωρείται επιτυχία αν περιλαμβάνεται σε αυτές τουλάχιστον ένας ορισμός

Επομένως, μέσω του σχήματος 4.2 φαίνεται ότι η πλειοψηφία των απαντήσεων βρίσκεται στα πρώτα παράθυρα αλλά υπάρχουν αρκετές απαντήσεις και στα υπόλοιπα παράθυρα, δηλαδή στα παράθυρα κειμένων με χαμηλότερη κατάταξη.

Ένα επιπλέον μέτρο αξιολόγησης που χρησιμοποιείται για αξιολόγηση στους διαγωνισμούς TREC είναι η μέση αντίστροφη κατάταξη (**Mean Reciprocal Rank**) ή **MRR** που λαμβάνει τιμές μεταξύ 0-1. Κάθε ερώτηση λαμβάνει έναν βαθμό ίσο με $1 / \text{σειρά της πρώτης απάντησης που ήταν σωστή}$. Αν η ερώτηση δεν είχε καμία σωστή απάντηση στις 5 πρώτες τότε ο βαθμός ισούται με 0. Συνολικά το MRR είναι ο μέσος όρος των ξεχωριστών βαθμών των ερωτήσεων.

Mean Reciprocal Rank	Υλοποίηση εργασίας	Υλοποίηση P&C
Συνολικά	0.412	0.412
Ερωτήσεις των οποίων ο όρος υπάρχει στο Wordnet	0.508	0.490

Πίνακας 4.2 Mean Reciprocal Rank

5. ΧΡΗΣΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

5.1 Θεωρητικό υπόβαθρο

Με αφετηρία τα αισιόδοξα αποτελέσματα της υλοποίησης της μεθόδου των P&C, ξεκίνησε το βασικό στάδιο της εργασίας όπου μελετήθηκε το πρόβλημα χειρισμού ερωτήσεων ορισμού με χρήση μηχανικής μάθησης.

Η Μηχανική Μάθηση (Machine Learning) είναι ένας από τους παλαιότερους ερευνητικούς τομείς της Τεχνητής Νοημοσύνης. Αντικείμενο της είναι η δημιουργία συστημάτων με δυνατότητα μάθησης. Συγκεκριμένα η έννοια της μάθησης, όπως παρουσιάζεται στην καθημερινή ζωή, έγκειται στην δυνατότητα αυτόματης απόκτησης περισσότερης γνώσης και στην χρησιμοποίηση της γνώσης αυτής για την βελτίωση της εκτέλεσης κάποιων λειτουργιών. Η Μηχανική Μάθηση έχει χρησιμοποιηθεί για την επίλυση ποικίλων προβλημάτων και παραδείγματα εφαρμογών της είναι η αναγνώριση χειρόγραφων χαρακτήρων, η αναγνώριση φωνής ακόμη και η οδήγηση αυτοκινήτου [Pomerleau, 1989].

Ακολουθεί ένας ορισμός της έννοιας της Μηχανικής Μάθησης:

*Ένα πρόγραμμα **μαθαίνει** από την εμπειρία **E** που αποκτά κατά την εκτέλεση ενός συνόλου διεργασιών **A**, εφόσον η απόδοσή του **A** βελτιώνεται με την αξιοποίηση της εμπειρίας **E**. [Mitchell 1997].*

Το πρόβλημα εύρεσης απάντησης σε μια ερώτηση ορισμού πρέπει να ξαναδιατυπωθεί ώστε να επιλυθεί με την χρήση της Μηχανικής Μάθησης. Συγκεκριμένα, το πρόβλημα πλέον είναι η αναγνώριση εκείνων των παραθύρων κειμένου, από το σύνολο των παραθύρων της ερώτησης ορισμού, που αποτελούν ορισμό. Κάθε παράθυρο κειμένου ανήκει είτε στην **κατηγορία «ορισμού»**, δηλαδή περιέχει ορισμό, είτε στην **κατηγορία «μη-ορισμού»**.

Με βάση τον προηγούμενο ορισμό χρειάζεται να προσδιοριστούν η εμπειρία E, το σύνολο των διεργασιών Δ και η απόδοση του προγράμματος A. Η διεργασία Δ είναι η αναγνώριση των παραθύρων κειμένου που αποτελούν ορισμό, η εμπειρία E είναι ένα σύνολο παραθύρων κειμένου των οποίων η κατηγορία είναι γνωστή και η απόδοση A είναι το ποσοστό των παραθύρων των οποίων αναγνωρίζεται η σωστή κατηγορία.

Το σύνολο των παραθύρων κειμένου που χρησιμοποιούνται ως εμπειρία αποτελούν το **σώμα εκπαίδευσης**, ενώ το σύνολο των παραθύρων κειμένου που χρησιμοποιούνται για την μέτρηση της απόδοσης αποτελούν το **σώμα αξιολόγησης**. Τα σώματα εκπαίδευσης και αξιολόγησης παριστάνονται σε διανυσματική μορφή. Συγκεκριμένα κάθε παράθυρο κειμένου παριστάνεται με ένα διάνυσμα των τιμών κάποιων **ιδιοτήτων (attributes)**.

Στόχος είναι ένας αλγόριθμος που να μπορεί να κατατάσσει κάθε παράθυρο κειμένου σε μια κατηγορία βασισμένος στις τιμές των ιδιοτήτων του διανύσματος που το παριστάνει. Στο σχήμα 5.1 παρουσιάζεται ένα υποθετικό παράδειγμα του τρόπου αναπαράστασης των παραθύρων κειμένου με χρήση 3 ιδιοτήτων. Τα παράθυρα κειμένου κατά την εργασία παριστάθηκαν με αντίστοιχο τρόπο αλλά με περισσότερες ιδιότητες.

Παράθυρο	Κατάταξη κειμένου	Εμφάνιση λέξης is μετά τον όρο	Εμφάνιση παρένθεσης μετά τον όρο	Ορισμός;
{κείμενο1}	1(1)	Ναι (1)	Όχι (1)	Ναι(1)
{κείμενο2}	5(5)	Όχι (0)	Όχι (0)	Ναι(1)
{κείμενο3}	2(2)	Όχι (0)	Ναι (1)	Όχι(0)
{κείμενο4}	22 (22)	Ναι (1)	Όχι (1)	Όχι(0)

Διανυσματική αναπαράσταση εμπειρίας:

{ <1, 1, 1, **1**>, <5, 0, 0, **1**>, <2, 0, 1, **0**>, <22, 1, 1, **0**> }

επιθυμητές απαντήσεις

Σχήμα 5.1 Υποθετικό παράδειγμα αναπαράστασης παραθύρων κειμένου

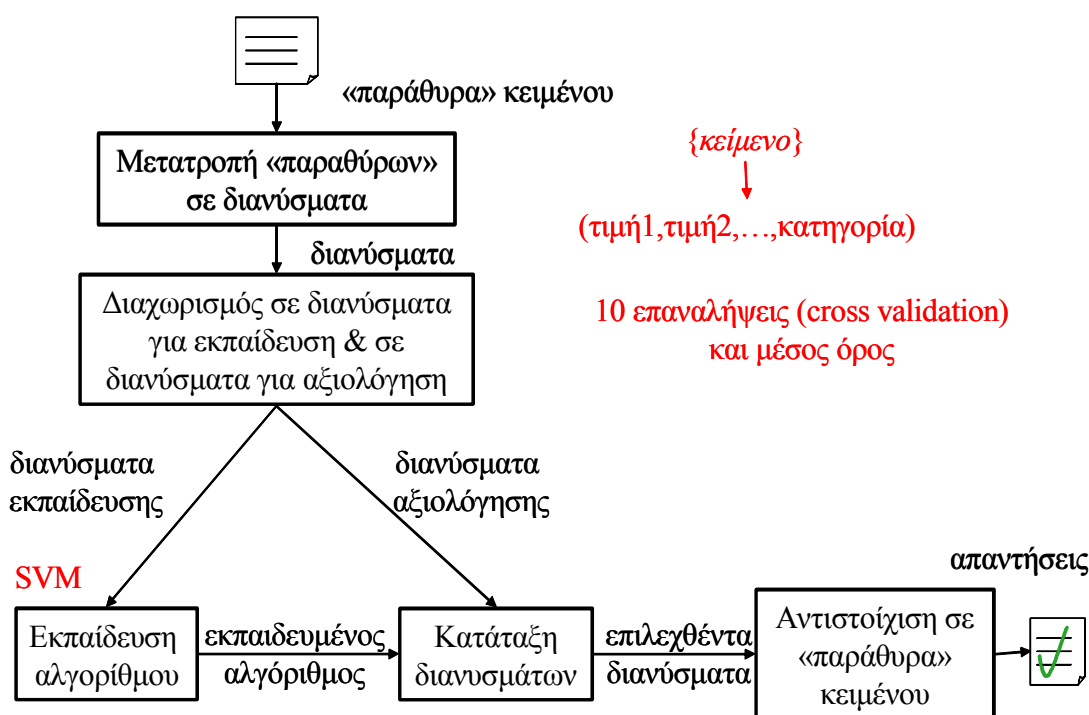
5.2 Περιγραφή χρήσης Μηχανικής Μάθησης

Στο σχήμα 5.2 φαίνεται αναλυτικά το στάδιο της επεξεργασίας των παραθύρων κειμένου όταν χρησιμοποιείται η Μηχανική Μάθηση.

Αρχικά τα παράθυρα κειμένου μετατρέπονται στα αντίστοιχα διανύσματα. Στην συνέχεια πρέπει να επιλεγεί το σύνολο των διανυσμάτων που θα αποτελέσουν τα σώματα εκπαίδευσης και αξιολόγησης. Η απλούστερη προσέγγιση θα ήταν να χρησιμοποιηθεί ένα μέρος του συνόλου για εκπαίδευση και ένα μέρος για αξιολόγηση. Επιλέχθηκε όμως η μέθοδος της «διασταυρωμένης επικύρωσης 10 επαναλήψεων» (**ten-fold cross validation**). Με την μέθοδο αυτή, τα παράθυρα κειμένου του 90% του συνόλου των ερωτήσεων χρησιμοποιούνται για εκπαίδευση και τα παράθυρα του υπόλοιπου 10% για αξιολόγηση. Γίνονται 10 επαναλήψεις όπου κάθε φορά αλλάζουν τα σώματα αξιολόγησης και εκπαίδευσης. Η μέθοδος αυτή, έχει το πλεονέκτημα ότι πραγματοποιούνται περισσότερα πειράματα, χρησιμοποιώντας τελικά ολόκληρο το σώμα και για εκπαίδευση και για αξιολόγηση. Επίσης, επειδή τα πειράματα επαναλαμβάνονται πολλές φορές, υπάρχει μια ένδειξη της διασποράς των αποτελεσμάτων.

Ο αλγόριθμος εκπαιδεύεται με τα διανύσματα εκπαίδευσης και στην συνέχεια κατατάσσει τα διανύσματα αξιολόγησης, δίνοντας και ένα βαθμό βεβαιότητας για το κατά πόσον κάθε διάνυσμα αξιολόγησης αντιστοιχεί σε παράθυρο ορισμού.

Τέλος, επιλέγονται τα πέντε διανύσματα «ορισμού» με τους μεγαλύτερους βαθμούς βεβαιότητας και αντιστοιχούνται στα παράθυρα κειμένου από τα οποία προέκυψαν. Αν μεταξύ αυτών των παραθύρων υπάρχει τουλάχιστον ένας σωστός ορισμός, η απόκριση του συστήματος θεωρείται σωστή. Η προηγούμενη διαδικασία επαναλαμβάνεται 10 φορές, κάθε φορά χρησιμοποιώντας ένα ξεχωριστό μέρος (10%) του σώματος για αξιολόγηση, ώστε τελικά να επιλεγούν από μία φορά όλα τα διανύσματα για αξιολόγηση. Το ποσοστό επιτυχίας του αλγορίθμου είναι ο μέσος όρος των ποσοστών επιτυχίας όλων των επαναλήψεων.



Σχήμα 5.2 Επεξεργασία «παραθύρων» κειμένου με χρήση Μηχανικής Μάθησης

5.3 Περιγραφή αλγορίθμου Support Vector Machines (SVMs)

Χρησιμοποιήθηκε η υλοποίηση του αλγορίθμου από το εργαλείο **Weka** που διατίθεται ελεύθερα [Witten & Frank, 2000]. Επιλέχθηκε ο αλγόριθμος των Μηχανών Διανυσμάτων Υποστήριξης διότι έχει το πλεονέκτημα να χειρίζεται πολύ καλά μεγάλο πλήθος ιδιοτήτων. Επιπλέον, παρουσιάζει ανεκτικότητα στο πλήθος των διανυσμάτων εκπαίδευσης, ιδιαίτερα όταν αυτό διαφέρει μεταξύ των 2 κατηγοριών, κάτι που παρατηρείται στην εργασία δεδομένου ότι τα διανύσματα που παριστάνουν παράθυρα κειμένου δεν κατανέμονται ίσα στις κατηγορίες ορισμού και μη-ορισμού αλλά τα παράθυρα μη-ορισμού αποτελούν την πλειοψηφία.

Ακολουθεί μια γενική περιγραφή της κεντρικής ιδέας του αλγορίθμου SVMs [Schölkopf κ.ά. 2002]:

Έστω ότι έχουμε αντικείμενα που ανήκουν σε 2 κατηγορίες (π.χ. παράθυρα ορισμού και παράθυρα μη ορισμού). Τότε θέλουμε να μπορούμε να κατατάσσουμε κάθε νέο αντικείμενο, του οποίου δεν γνωρίζουμε την κατηγορία, σε μία από τις 2 κατηγορίες.

Έστω ότι μας δίνονται τα παρακάτω εμπειρικά δεδομένα:

$$(x_1, y_1), \dots, (x_m, y_m) \in X \times \{\pm 1\}.$$

Τα x_i αποτελούν **αντικείμενα** του συνόλου X και τα y_i αποτελούν τις **ετικέτες** τους. Υπάρχουν μόνο 2 κατηγορίες αντικειμένων και για διευκόλυνση συμβολίζονται με τις τιμές +1 και -1. Στην μάθηση, θέλουμε όταν μας δίνεται ένα νέο αντικείμενο $x \in X$, να προβλέψουμε το αντίστοιχο $y \in \{\pm 1\}$. Δηλαδή, διαλέγουμε ένα y τέτοιο ώστε το ζεύγος (x, y) να είναι με κάποιο τρόπο παρόμοιο με τα εμπειρικά δεδομένα, τα οποία αποτελούν τα δεδομένα εκπαίδευσης.

Επομένως, χρειάζονται κάποια **μέτρα ομοιότητας** στο X και στο $\{\pm 1\}$. Ο χαρακτηρισμός της ομοιότητας των y_i είναι εύκολος διότι 2 ετικέτες μπορούν να είναι είτε όμοιες είτε διαφορετικές (δυαδική ταξινόμηση – binary classification). Η επιλογή μέτρου ομοιότητας για τα αντικείμενα x αποτελεί ένα δυσκολότερο πρόβλημα που αποτελεί και την βάση στον τομέα της μηχανικής μάθησης.

Ας θεωρήσουμε ένα μέτρο ομοιότητας της μορφής

$$k : X \times X \rightarrow \mathfrak{R}$$

$$(x, x') \mapsto k(x, x'),$$

αυτό αποτελεί μια συνάρτηση η οποία, δεδομένου 2 αντικειμένων x και x' , επιστρέφει έναν πραγματικό αριθμό χαρακτηρίζοντας την ομοιότητα τους. Αυτή η συνάρτηση καλείται **πυρήνας**.

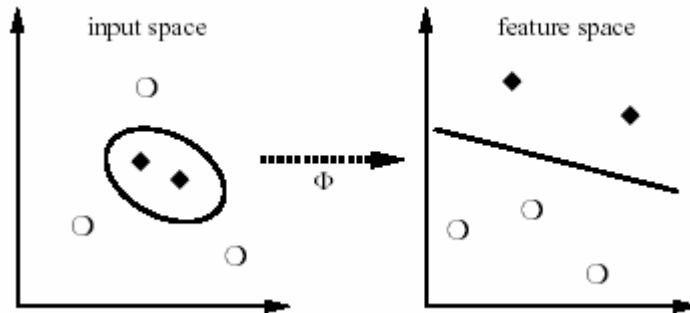
Τα αντικείμενα του συνόλου X παριστάνονται σαν διανύσματα σε έναν διανυσματικό χώρο μεγαλύτερων διαστάσεων, H . Χρησιμοποιείται η απεικόνιση

$$\Phi : X \rightarrow H$$

$$x \mapsto \mathbf{x} := \Phi(x)$$

Η ιδέα του αλγορίθμου είναι αφού απεικονιστούν τα δεδομένα εκπαίδευσης, δηλαδή τα αντικείμενα του συνόλου X , σε ένα διανυσματικό χώρο H που αποτελεί μετασχηματισμό του αρχικού μέσω της απεικόνισης Φ , να βρεθεί ένα **υπερεπίπεδο** (hyperplane) Π που να διαχωρίζει κατά βέλτιστο τρόπο τα διανύσματα των 2 κατηγοριών. Το υπερπίπεδο δεν εξαρτάται από το σύνολο των διανυσμάτων εκπαίδευσης αλλά από ένα υποσύνολο που ονομάζονται **διανύσματα υποστήριξης**.

Για να κατατάξουμε ένα νέο αντικείμενο x , το απεικονίζουμε στον διανυσματικό χώρο H , και ελέγχουμε που τοποθετείται σε σχέση με το υπερεπίπεδο.



Σχήμα 5.3 Η βασική ιδέα του αλγορίθμου SVMs [Schölkopf κ.ά. 2002] με την δημιουργία ενός υπερεπιπέδου

5.4 Χρήση περιορισμένου αριθμού ιδιοτήτων

Η πρώτη φάση των πειραμάτων έγινε με περιορισμένο αριθμό ιδιοτήτων. Πολλές από τις ιδιότητες βασίστηκαν στο σύστημα ερωταποκρίσεων των Joho και Sanderson [Joho, Sanderson, 2000]. Ακολουθεί μια περιγραφή του συστήματος ερωταποκρίσεων που δημιούργησαν.

5.4.1 Σύστημα ερωταποκρίσεων Joho & Sanderson

Οι Joho & Sanderson δημιούργησαν ένα σύστημα που βρίσκει **περιγραφικές φράσεις όρων** σε ελεύθερο κείμενο. Οι προτάσεις που περιέχουν τον συγκεκριμένο όρο ταξινομούνται με χρήση μιας τεχνικής που βασίζεται σε ταίριασμα προτύπων, μέτρηση λέξεων και στην θέση των προτάσεων. Σε αντίθεση με πολλές μεθόδους δεν γίνεται χρήση τεχνικών συντακτικής ανάλυσης και η επιτυχία της απλούστερης αυτής μεθόδου οφείλεται πιθανόν στον μεγάλο όγκο του ελεύθερου κειμένου που χρησιμοποιείται για αναζήτηση.

Η εύρεση περιγραφικών φράσεων από ένα σύστημα δεν μπορεί να καλύψει το γενικό πρόβλημα της απόκρισης ερωτήσεων αλλά είναι ικανή να απαντήσει ερωτήσεις της μορφής

«Who is ...» «What is ...». Ως βασικά πλεονεκτήματα της ενασχόλησης μόνο με τις ερωτήσεις ορισμού αναφέρονται η ευκολία προσαρμογής αυτών των τεχνικών σε διάφορους τομείς και η συχνή εμφάνιση περιγραφών σε κείμενα.

Δεδομένου ενός όρου του οποίου ζητείται ο ορισμός, το σύστημα ανακτά όλα τα κείμενα που περιέχουν τον όρο και τα ξεχωρίζει σε προτάσεις. Αυτές ταξινομούνται βάσει 3 κριτηρίων:

1. Ύπαρξη μιας **φράσης-κλειδί (key phrase)** στην πρόταση,
2. Υψηλός αριθμός **«συχνών λέξεων» (common words)**,
3. Η **θέση της πρότασης (ordinal position)** μέσα στο κείμενο.

Οι φράσεις-κλειδιά αποτελούν τη βάση του συστήματος. Οι φράσεις προέρχονται από την μελέτη της Hearst [Hearst, 1998], η οποία μελέτησε το πρόβλημα εντοπισμού της σχέσης «is-a», σχέση μεταξύ όρου και υπερωνύμου, μέσα σε μια συλλογή. Οι φράσεις εντοπίστηκαν με ημι-αυτόματο τρόπο, χρησιμοποιώντας τα υπερώνυμα του Wordnet.

Ακολουθούν μερικά παραδείγματα:

- όρος (περιγραφική φράση) ή (περιγραφική φράση) όρος
 - π.χ. “*A tsunami (giant wave) in China ...*”
- όρος (is|was|are|were) (a|an|the) περιγραφική φράση
 - π.χ. “*Galileo is an astronomer...*”

Έπειτα χρησιμοποιήθηκε η υπόθεση ότι οι περιγραφές ενός όρου σε διάφορα κείμενα είναι ίδιες ή παρόμοιες. Το σύστημα μετράει τις συχνότητες των διαφορετικών λέξεων στις προτάσεις που περιέχουν τον όρο και βρίσκει τις 20 συχνότερες. Στην συνέχεια δίνεται σε κάθε πρόταση ένας βαθμός ανάλογα με τον αριθμό των συχνών λέξεων που περιέχει. Η λογική είναι ότι έτσι θα αποφευχθούν οι τυχαίες προτάσεις που δεν αποτελούν ορισμό.

Τέλος, εάν ένας όρος εμφανίζεται πολλές φορές σε ένα κείμενο είναι πιθανότερο μια περιγραφή του να βρίσκεται στην αρχή του κειμένου παρά στο τέλος. Έτσι η θέση της πρότασης, μεταξύ των προτάσεων ενός κειμένου που περιέχουν τον όρο, αποτέλεσε κριτήριο ταξινόμησης.

Πρόβλημα αποτέλεσε ο συνδυασμός των παραπάνω τριών κριτηρίων. Οι Joho και Sanderson κατέληξαν εμπειρικά στον παρακάτω τύπο που αξιολογεί κάθε πρόταση που περιέχει τον όρο του οποίου ζητείται ο ορισμός:

$$\alpha * \Phi KB + \beta * \Sigma \Lambda + \gamma * (\delta - \text{ΑΠ})$$

όπου ΦKB είναι το βάρος της φράσης-κλειδί που περιέχει η πρόταση, $\Sigma \Lambda$ είναι το πλήθος των συχνών λέξεων που περιέχει η πρόταση και ΑΠ είναι ο τακτικός αριθμός της πρότασης μεταξύ όλων των προτάσεων που περιέχουν τον όρο στο κείμενο από το οποίο προέρχεται η αξιολογούμενη πρόταση.

Βασικό πρόβλημα της μεθόδου είναι η εξάρτησή της από τις παραμέτρους-βάρη α, β, γ και δ των οποίων οι βέλτιστες τιμές είναι δύσκολο να εντοπιστούν. Στην εργασία των Joho και Sanderson, οι τιμές των παραμέτρων ορίστηκαν μέσω πειραματικής διερεύνησης, ώστε να βελτιστοποιούν το ποσοστό επιτυχίας σε ένα συγκεκριμένο σώμα κειμένων. Με τον τρόπο αυτό όμως, τόσο η εκπαίδευση όσο και η αξιολόγηση του συστήματος έγιναν ουσιαστικά στο ίδιο σώμα κειμένων. Αντίθετα, στο σύστημα της παρούσας εργασίας, όπου κάποια από τα παραπάνω κριτήρια χρησιμοποιήθηκαν ως ιδιότητες, τα βάρη των ιδιοτήτων υπολογίστηκαν από τον ίδιο τον αλγόριθμο SVM με βάση μόνο το σώμα εκπαίδευσης κάθε μιας από τις 10 επαναλήψεις της διασταυρωμένης επικύρωσης. Επομένως, η επιλογή των βαρών έγινε με τις τεχνικές βελτιστοποίησης που ενσωματώνουν τα SVM και τα αποτελέσματα των πειραμάτων είναι πιο αξιόπιστα.

5.4.2 Ιδιότητες μηχανικής μάθησης

Ακολουθούν οι ιδιότητες που χρησιμοποιήθηκαν στην πρώτη προσέγγιση μηχανικής μάθησης της εργασίας, στις οποίες όπως έχει ήδη αναφερθεί, συμπεριλαμβάνονται διάφορες χαρακτηριστικές φράσεις από το σύστημα των Joho & Sanderson.

Συγκεκριμένα επιλέχθηκαν οι παρακάτω ιδιότητες για την διανυσματική αναπαράσταση των παραθύρων :

1. Η κατάταξη κειμένου όπου περιέχεται το παράθυρο (αριθμητική τιμή 1 -50)
Θεωρείται χρήσιμη ιδιότητα, δεδομένης της εμπειρικής γνώσης ότι είναι πιο πιθανό ένας ορισμός να βρίσκεται στα κορυφαία κείμενα που επιστρέφει μια μηχανή αναζήτησης κείμενα παρά στα τελευταία.
2. Η θέση του παραθύρου μέσα στο κείμενο (αριθμητική τιμή >0)
Ένας όρος μπορεί να εμφανίζεται πολλές φορές σε κάθε κείμενο. Είναι πιθανότερο ο ορισμός να δίνεται στις αρχικές εμφανίσεις του όρου παρά στις τελευταίες. Η ιδιότητα

αυτή αντιπροσωπεύει την σειρά που εμφανίστηκε ο όρος στο κείμενο μέσα στο αντίστοιχο παράθυρο.

3. Το αποτέλεσμα της **μεθόδου P&C** (τιμές 0,1,2)

Η τιμή 0 αντιπροσωπεύει την περίπτωση όπου δεν υπάρχει κάποιο από τα επιλεγμένα υπερώνυμα στο παράθυρο, η τιμή 1 την περίπτωση όπου υπάρχει κάποιο υπερώνυμο και η τιμή 2 την περίπτωση να μην έχει επιστρέψει η τεχνική κάποιο υπερώνυμο.

4. Το **πλήθος των «συχνών λέξεων»** του παραθύρου (τιμές 0-20)

Η παραπάνω ιδιότητα βασίζεται στο εμπειρικό συμπέρασμα ότι τείνουν οι διαφορετικοί ορισμοί ενός όρου να έχουν κοινά στοιχεία δηλαδή κοινές λέξεις. Επομένως όσες περισσότερες λέξεις περιέχει ένα παράθυρο από τις «συχνές» (βλέπε ενότητα 5.4.1), τόσο μεγαλύτερη πιθανότητα υπάρχει να αποτελεί παράθυρο ορισμού. Για να αποκλειστούν συχνές λέξεις της αγγλικής γλώσσας όπως η λέξη “the” χρησιμοποιήθηκε μια stop-list που περιέχει τις 100 συχνότερες λέξεις του British National Corpus. Το κριτήριο των «συχνών λέξεων» προέρχεται από το σύστημα ερωταποκρίσεων των Joho & Sanderson.

5. Η φράση **“such <...> as όρος”**

Παράδειγμα : “*such antibiotics as amoxicillin*”

6. Η φράση **“όρος and other <...>”**

Παράδειγμα : “*broken bones and other injuries*”

7. Η φράση **“όρος or other <...>”**

Παράδειγμα : “*cats or other animals*”

8. Η φράση **“especially όρος”**

Παράδειγμα : “*some plastics especially Teflon*”

9. Η φράση **“including όρος”**

Παράδειγμα : “*some amphibians including frog*”

10. **Παρενθέσεις μετά** τον όρο

Παράδειγμα : “*sodium chloride (salt)*”

11. **Παρενθέσεις πριν** τον όρο

Παράδειγμα : “*(Vitamin B1) thiamine*”

12. 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*”

13. **Κόμμα μετά** τον όρο

Παράδειγμα : “*amoxicillin, an antibiotic*”

14. Η φράση **“όρος which is/was/are/were <...>”**

Παράδειγμα : “*tsunami which is a giant wave*”

15. Η φράση **“<... >like όρος”**

Παράδειγμα : “*antibiotics like amoxicillin*”

16. Η φράση **“όρος , <...> , is/was/are/were”**

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

17. Η φράση **“όρος or <...>”**

Παράδειγμα : “*autism or some other type of disorder*”

18. Ένα από τα ρήματα **“can”, “refer”, “have”** μετά τον όρο (**3 ιδιότητες**)

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

19. Ένα από τα ρήματα **“called”, “known as”, “defined”** πριν τον όρο (**3 ιδιότητες**)

Παράδειγμα : “*The giant wave known as tsunami*”

Το σύνολο των ιδιοτήτων είναι 24, συμπεριλαμβανομένης και της κατηγορίας του παραθύρου.

5.4.3 Αποτελέσματα

Τα πειράματα έγιναν με τον **αλγόριθμο μηχανικής μάθησης SVM (Support Vector Machines)**. Για την πραγματοποίηση των πειραμάτων, οι ερωτήσεις χωρίστηκαν σε 10 μέρη, το καθένα εκ των οποίων περιείχε 16 ερωτήσεις. Σε κάθε μια από τις 10 επαναλήψεις της διασταυρωμένης επικύρωσης, ο αλγόριθμος είχε ως σώμα εκπαίδευσης τις 144 ερωτήσεις και η αξιολόγηση γινόταν με τις υπολειπόμενες 16 ερωτήσεις.

Τα αποτελέσματα δεν ήταν υψηλά, συγκεκριμένα υπήρχε επιτυχία 84.3% αλλά αυτό μεταφραζόταν σε 99.9% επιτυχία στην αναγνώριση παραθύρων μη-ορισμού και μόλις 0.9% επιτυχία στην αναγνώριση παραθύρων ορισμού.

Αλγόριθμος	«παράθυρα» ορισμών	«παράθυρα» μη-ορισμών	Ποσοστό επιτυχίας		
			συνολικά	ορισμοί	μη-ορισμοί
SVM	15.7	84.3	84.3	0.9	99.9

Πίνακας 5.1 Αποτελέσματα μηχανικής μάθησης – αρχικό στάδιο

Αυτό οφειλόταν στο μειωμένο πλήθος των ορισμών στο σύνολο της εκπαίδευσης. Συγκεκριμένα μόλις 16% επί του συνόλου ήταν παράθυρα ορισμού ενώ το υπόλοιπο 84% ήταν παράθυρα μη ορισμού. Ο αλγόριθμος επομένως **έτεινε να κατατάσσει όλα τα παράθυρα σαν παράθυρα μη-ορισμού**. Επειδή η κατανομή των παραθύρων αποτελούσε την πραγματική εικόνα των κειμένων θα ήταν λάθος να δημιουργηθεί ένα σώμα εκπαίδευσης με 50% παράθυρα ορισμού και 50% παράθυρα μη-ορισμού για να εξομαλυνθούν τα αποτελέσματα.

Έπρεπε επομένως τα αποτελέσματα να μην στηρίζονται μεμονωμένα στην κατηγορία που κατατάσσει ο αλγόριθμος ένα παράθυρο αλλά να λαμβάνεται υπόψη το πόσο σίγουρος είναι ο αλγόριθμος για την απόφασή του αυτή. Η υλοποίηση του SVM του Weka μπορεί να επιστρέψει την βεβαιότητα με την οποία κατατάσσει ένα διάνυσμα σε μια κατηγορία. Για παράδειγμα μπορεί να κατατάσσει ένα παράθυρο κειμένου ως ορισμό με πιθανότητα 60% και ως μη ορισμό με πιθανότητα 40%. Οπότε για κάθε όρο ελέγχουμε όλες τις πιθανότητες του κάθε διανύσματος να αποτελεί «ορισμό» και επιλέγουμε τα παράθυρα με τις 5 μεγαλύτερες πιθανότητες. Αν μέσα σε αυτά υπάρχει τουλάχιστον ένα παράθυρο που αποτελεί πραγματικά ορισμό τότε αυτό θεωρείται επιτυχής εύρεση παραθύρου ορισμού για τον συγκεκριμένο όρο. Με βάση τα παραπάνω έγιναν τα επόμενα πειράματα και ακολουθούν τα αποτελέσματα.

Αποτελέσματα	Μηχανική μάθηση		
	ερωτήσεις	ποσοστό	τυπική απόκλιση
Συνολική επιτυχία	101 / 160	63.13%	10.63%
Επιτυχία σε ερωτήσεις για τις οποίες υπάρχει απάντηση στα κείμενα	101 / 137	73.72%	12.38%

Πίνακας 5.2 Αποτελέσματα μηχανικής μάθησης ελέγχοντας τις 5 καλύτερες ερωτήσεις

Όπως προηγουμένως τα ποσοστά επιτυχίας χωρίζονται σε ποσοστά επιτυχίας για όλες τις ερωτήσεις και σε ποσοστά επιτυχίας στις ερωτήσεις για τις οποίες υπάρχει απάντηση στα κείμενα. Αρνητικό αποτέλεσμα αποτελεί η **μεγάλη τυπική απόκλιση** μεταξύ των 10 επαναλήψεων της διασταυρωμένης επικύρωσης, η οποία όμως πιθανόν οφείλεται στο μικρό μέγεθος του σώματος εκπαίδευσης. Για παράδειγμα, επειδή κάθε μέρος αποτελείται από 16

ερωτήσεις αρκεί μία ερώτηση να έχει λάθος απάντηση και το ποσοστό από 100% να πέσει σε 93.75%, δηλαδή κατά 6.25%, που αποτελεί αρκετά μεγάλο ποσοστό. Σε σχέση με την τεχνική των P&C παρατηρείται **αύξηση των ποσοστών μέχρι και 13%**. Αυτό έδειξε ότι πέρα από την σημαντικότητα της ιδιότητας που προκύπτει από την τεχνική P&C, έπαιξαν σημαντικό ρόλο και οι υπόλοιπες ιδιότητες.

Ήδη με την χρήση μικρού πλήθους ιδιοτήτων τα αποτελέσματα δείχνουν ότι η Μηχανική Μάθηση μπορεί να αποδειχθεί χρήσιμη μέθοδος για την απάντηση ερωτήσεων ορισμού.

5.5 ΕΠΙΛΟΓΗ ΙΔΙΟΤΗΤΩΝ ΜΕΣΩ ΠΛΗΡΟΦΟΡ. ΚΕΡΔΟΥΣ

5.5.1 Περιγραφή

Οι ιδιότητες στα προηγούμενα πειράματα ήταν λίγες (23) και είχαν προταθεί από ερευνητές που προσπάθησαν να εντοπίσουν χειρωνακτικά συχνούς τρόπους διατύπωσης ορισμών. Η δεύτερη προσέγγιση μηχανικής μάθησης της εργασίας χρησιμοποιεί μεγαλύτερο αριθμό ιδιοτήτων, ελπίζοντας ότι αυτό θα βελτιώσει το ποσοστό επιτυχίας του αλγορίθμου μάθησης. Στις προηγούμενες 23 ιδιότητες προστίθενται επιπλέον ιδιότητες που προκύπτουν αυτόματα από το σώμα εκπαίδευσης και αντιπροσωπεύουν, όπως και οι περισσότερες από τις προηγούμενες 23 ιδιότητες, την εμφάνιση ή όχι συγκεκριμένων χαρακτηριστικών φράσεων.

Οι επιπλέον ιδιότητες επιλέγονται ως εξής. Εντοπίζονται στο σώμα εκπαίδευσης όλες οι φράσεις μιας, δύο ή τριών λέξεων που εμφανίζονται πριν ή μετά τον όρο του οποίου ζητείται ο ορισμός. Οι φράσεις αυτές αποτελούν τις υποψήφιες ιδιότητες. Στη συνέχεια χρειάζεται ένα κριτήριο, βάσει του οποίου θα επιλεγούν από τις υποψήφιες ιδιότητες εκείνες που θα αποτελέσουν τις τελικές ιδιότητες. Μια πρώτη προσέγγιση αποτέλεσε η χρήση του **Πληροφοριακού κέρδους (Information Gain)**.

Συγκεκριμένα, ας θεωρήσουμε μια υποψήφια ιδιότητα που προκύπτει με τον προηγούμενο τρόπο σαν μια τυχαία μεταβλητή X και την τυχαία μεταβλητή D , όπου αν $D=1$ τότε το παράθυρο αποτελεί ορισμό ενώ αν $D=0$ τότε το παράθυρο δεν αποτελεί «ορισμό». Τότε η

εντροπία της D δείχνει πόσο αβέβαιοι είμαστε για την τιμή της D ή ισοδύναμα πόση πληροφορία χρειάζεται να μας δοθεί για να μην υπάρχει καθόλου αβεβαιότητα για την τιμή της D . Τότε, **πληροφοριακό κέρδος** $IG(D, X)$ της τυχαίας μεταβλητής X αποτελεί η μέση μείωση της εντροπίας της D αν γνωρίζω την τιμή της X .

*Π.χ. Αν στα παράθυρα ορισμού η φράση «**known as**» εμφανίζεται πολύ συχνά αμέσως μετά από τον όρο, του οποίου ζητάμε τον ορισμό, και ταυτοχρόνως αν εμφανίζεται σπάνια στα παράθυρα μη ορισμού αμέσως μετά τον όρο, τότε θα έχει ως ιδιότητα μεγάλο Πληροφοριακό Κέρδος.*

Επομένως μέσω του μέτρου του Πληροφοριακού Κέρδους θα προκύψουν ενδεχομένως ιδιότητες που θα βοηθήσουν στο να αυξηθεί η βεβαιότητα και επομένως και η επιτυχία.

5.5.2 Αποτελέσματα

Για να μειωθεί το μεγάλο πλήθος φράσεων που προέκυπταν ως υποψήφιες ιδιότητες, απορρίπτονταν εξαρχής όσες φράσεις εμφανίζονταν λιγότερες από 10 φορές (κατώφλι τιμής 10) στα παράθυρα εκπαίδευσης. Στην συνέχεια υπολογιζόταν το Πληροφοριακό Κέρδος όσων φράσεων παρέμεναν και επιλεγόντουσαν όλες οι υποψήφιες ιδιότητες με Πληροφοριακό Κέρδος μεγαλύτερο του μηδενός. Με χρήση κατωφλίου με τιμή 10 προέκυπταν περίπου 90 έως 100 ιδιότητες σε κάθε επανάληψη της διασταυρωμένης επικύρωσης. Συνολικά επομένως περίπου 120 ιδιότητες, διατηρώντας τις ιδιότητες που χρησιμοποιήθηκαν προηγουμένως.

Για να είναι ξεκάθαρος ο τρόπος επιλογής των ιδιοτήτων επισημαίνεται ότι η διαδικασία επιλογής ιδιοτήτων επαναλαμβανόταν σε κάθε επανάληψη της διασταυρωμένης επικύρωσης, χρησιμοποιώντας το σώμα εκπαίδευσης της συγκεκριμένης επανάληψης. Ο τρόπος μέτρησης των ποσοστών επιτυχίας παραμένει όπως προηγουμένως.

Αποτελέσματα	Μηχανική μάθηση		
	ερωτήσεις	ποσοστό	τυπική απόκλιση
Συνολική επιτυχία	106 / 160	66.25%	10.53%
Επιτυχία σε ερωτήσεις για τις οποίες υπάρχει απάντηση στα κείμενα	106 / 137	77.37%	11%

Πίνακας 5.3 Αποτελέσματα μηχανικής μάθησης επιλέγοντας ιδιότητες μέσω Π.Κ.

Τα ποσοστά **καλύτερευαν κατά 3%** ή κατά 5 σωστές ερωτήσεις. Είναι αξιοσημείωτο επιπλέον ότι μειώθηκε η τυπική απόκλιση κάτι που πιθανόν να οφείλεται στην ύπαρξη περισσότερων ιδιοτήτων. Ακολουθούν στο σχήμα 5.4 παραδείγματα φράσεων που προέκυψαν με τον τρόπο που περιγράφηκε. Σε **πράσινο** πλαίσιο (κυκλικό πλαίσιο) είναι οι φράσεις που πράγματι φαίνεται να αποτελούν καλή ένδειξη ότι ένα «παράθυρο» κειμένου αποτελεί ορισμό, ενώ σε **κόκκινο** πλαίσιο (ορθογώνιο πλαίσιο) είναι φράσεις που προέκυψαν λόγω τυχαιότητας και δεν φαίνονται να αποτελούν καλές ιδιότητες. Οι φράσεις χωρίζονται σε αυτές που παρουσιάζονται μετά τον όρο και σε αυτές που παρουσιάζονται πριν από τον όρο.

Μετά τον όρο

has been	Entertainment	. It
, but	will be	, the
and other	, which	at the
, said	will	. But
may be	said	. In
is not	for	, is
. He	to	as a
'	was	, who
can be	of	,
, an	is a	Technology Inc
, or	in the	is the
is an	. The	, as

Πριν τον όρο

strains of
.
or
called
carinii
60
Fender
as
total

Σχήμα 5.4 Παραδείγματα φράσεων που προέκυψαν ως ιδιότητες μέσω Π.Κ.

Επισημαίνεται ότι η πλειοψηφία των φράσεων που έχουν μεγάλο Πληροφοριακό Κέρδος εμφανίζονται μετά τον όρο. Οι φράσεις «and other», «is a/an/the» και «called» αποτελούν παραδείγματα φράσεων που είχαν περιληφθεί χειρωνακτικά στις 23 αρχικές ιδιότητες και που επιλέχθηκαν και με την χρήση του Πληροφοριακού Κέρδους.

5.6 ΕΠΙΛΟΓΗ ΙΔΙΟΤΗΤΩΝ ΜΕΣΩ ΑΚΡΙΒΕΙΑΣ

5.6.1 Περιγραφή

Όπως και προηγουμένως για να αυξηθούν τα ποσοστά επιτυχίας πρέπει να αυξηθούν οι ιδιότητες. Με την προηγούμενη χρήση του μέτρου του Πληροφοριακού Κέρδους επιλεγόταν ο μέγιστος αριθμός ιδιοτήτων και προέκυπταν έτσι, το πολύ 100 ιδιότητες. Φάνηκε ενδιαφέρον να βρεθεί κάποιο άλλο κριτήριο επιλογής των τελικών ιδιοτήτων ώστε να προκύπτουν περισσότερες ιδιότητες. Υποψήφια μέτρα αποτέλεσαν η **ακρίβεια (precision)**, η **ανάκληση(recall)** και η **ορθότητα(accuracy)** που ορίζονται ως εξής :

Ακρίβεια μιας φράσης είναι τα παράθυρα ορισμού όπου εμφανίζεται η φράση δια τα συνολικά παράθυρα όπου εμφανίζεται η φράση.

Ανάκληση μιας φράσης είναι τα παράθυρα ορισμού όπου εμφανίζεται η φράση διά τα συνολικά παράθυρα ορισμού.

Ορθότητα μιας φράσης είναι τα παράθυρα ορισμού όπου εμφανίζεται η φράση και τα παράθυρα μη-ορισμού όπου δεν εμφανίζεται διά τα συνολικά παράθυρα.

Με βάση τους παραπάνω ορισμούς, προκύπτει ότι η ακρίβεια δείχνει πόσο βέβαιοι είμαστε ότι κάποιο παράθυρο που περιέχει την φράση αποτελεί όντως παράθυρο ορισμού, η ανάκληση μετράει το πλήθος των παραθύρων ορισμού που βρίσκονται μέσω της φράσης και τέλος η ορθότητα μετράει το πλήθος των σωστών συμπερασμάτων στα οποία καταλήγουμε μέσω της φράσης στο σύνολο των παραθύρων. Προτιμήθηκε η χρήση του μέτρου της ακρίβειας.

Επομένως, επιλέγονται οι τελικές ιδιότητες από το σύνολο των υποψήφιων ιδιοτήτων με χρήση της ακρίβειας. Συγκεκριμένα θεωρήθηκε καλύτερο να επιλέγονται τόσο οι φράσεις με υψηλή ακρίβεια όσο και φράσεις με χαμηλή ακρίβεια. Στόχος ήταν να βρεθούν όχι μόνο φράσεις χαρακτηριστικές ορισμών αλλά και φράσεις που να δείχνουν ότι το παράθυρο δεν αποτελεί ορισμό.

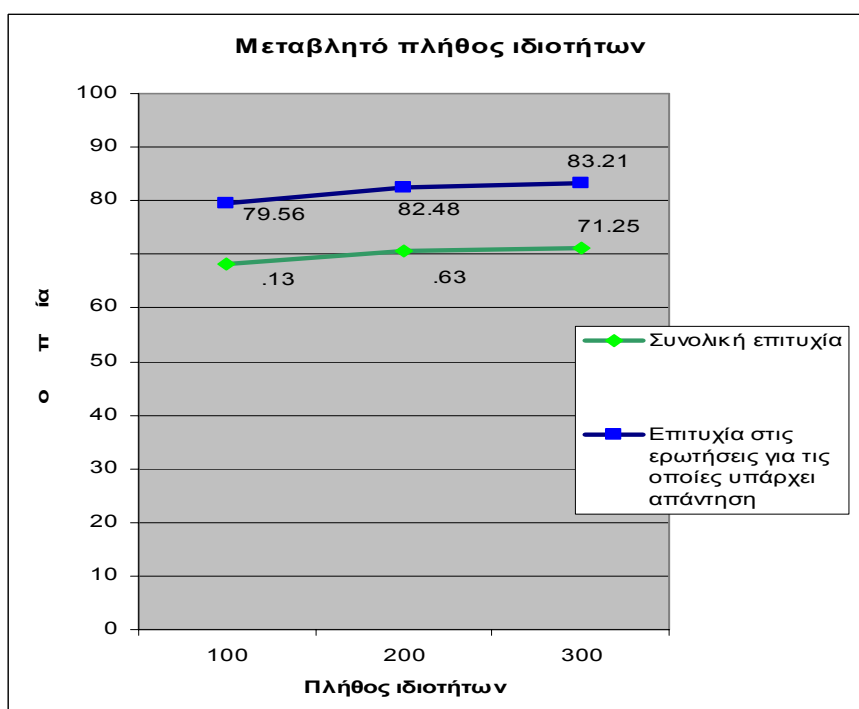
5.6.2 Αποτελέσματα

Με αυτόν τον τρόπο προκύπτει μεγάλο πλήθος υποψήφιων ιδιοτήτων. Για παράδειγμα με κατώφλι τιμής 10, δηλαδή διατήρηση των φράσεων που εμφανίζονται τουλάχιστον 10 φορές στα παράθυρα εκπαίδευσης, προκύπτουν περίπου 700 ιδιότητες. Λόγω των μεγάλων υπολογιστικών απαιτήσεων, γίνανε πειράματα με μέγιστο πλήθος ιδιοτήτων 300. Τα πειράματα γίνανε για 3 πλήθη ιδιοτήτων, συγκεκριμένα για 100, 200 και 300 ιδιότητες.

Αποτελέσματα	100 ιδιότητες		200 ιδιότητες		300 ιδιότητες	
	ερωτ.	ποσοστό	ερωτ.	ποσοστό	ερωτ.	ποσοστό
Συνολική επιτυχία	109 / 160	68.13%	113 / 160	70.63%	114 / 160	71.25%
Επιτυχία σε ερωτήσεις που υπάρχει απάντηση στα κείμενα	109 / 137	79.56%	113 / 137	82.48%	114 / 137	83.21%

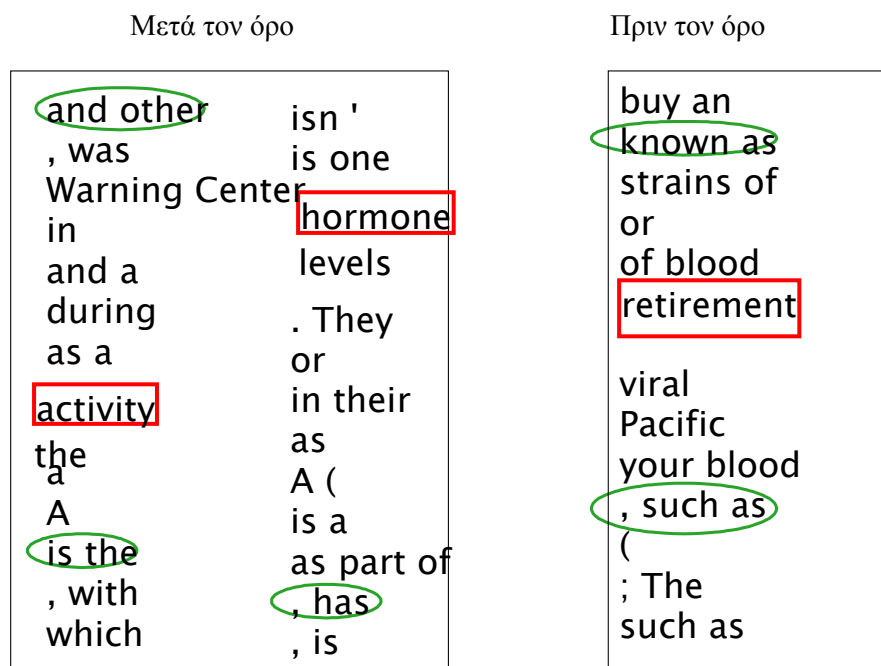
Πίνακας 5.4 Αποτελέσματα μηχανικής μάθησης επιλέγοντας ιδιότητες μέσω ακρίβειας

Τα ποσοστά επιτυχίας αυξήθηκαν κατά 11 ερωτήσεις δηλαδή κατά περίπου 7%. Οι 200 ιδιότητες μπορούν να θεωρηθούν σχετικά επαρκείς δεδομένου ότι επιλέγοντας 300 ιδιότητες αυξάνονται οι σωστές ερωτήσεις μόλις κατά μία όπως φαίνεται και στο σχήμα 5.5.



Σχήμα 5.5
Ποσοστά επιτυχίας με χρήση διαφορετικού πλήθους ιδιοτήτων

Ακολουθούν στο σχήμα 5.6 παραδείγματα φράσεων που προέκυψαν με τον τρόπο που περιγράφηκε. Σε πράσινο πλαίσιο (κυκλικό πλαίσιο) είναι οι φράσεις που πράγματι φαίνεται να αποτελούν καλή ένδειξη ότι ένα παράθυρο κειμένου αποτελεί ορισμό, ενώ σε κόκκινο πλαίσιο (ορθογώνιο πλαίσιο) είναι φράσεις που προέκυψαν λόγω τυχαιότητας και δεν φαίνονται να αποτελούν καλές ιδιότητες. Οι φράσεις χωρίζονται σε αυτές που παρουσιάζονται μετά τον όρο και σε αυτές που παρουσιάζονται πριν από τον όρο.



Σχήμα 5.6 Παραδείγματα φράσεων που προέκυψαν ως ιδιότητες μέσω ακρίβειας

Προκύπτουν φράσεις που φανερώνουν ορισμό, κάποιες από τις οποίες είχαν περιληφθεί και στις αρχικές 23 ιδιότητες, όπως οι φράσεις «and other», «is the», «has», «known as» και «such as». Παρά το γεγονός ότι επιλέξαμε και φράσεις με χαμηλή ακρίβεια δεν φαίνεται να υπάρχουν φράσεις που να δείχνουν ότι ένα παράθυρο κειμένου δεν αποτελεί ορισμό.

5.7 ΥΠΟΛΟΙΠΑ ΠΕΙΡΑΜΑΤΑ

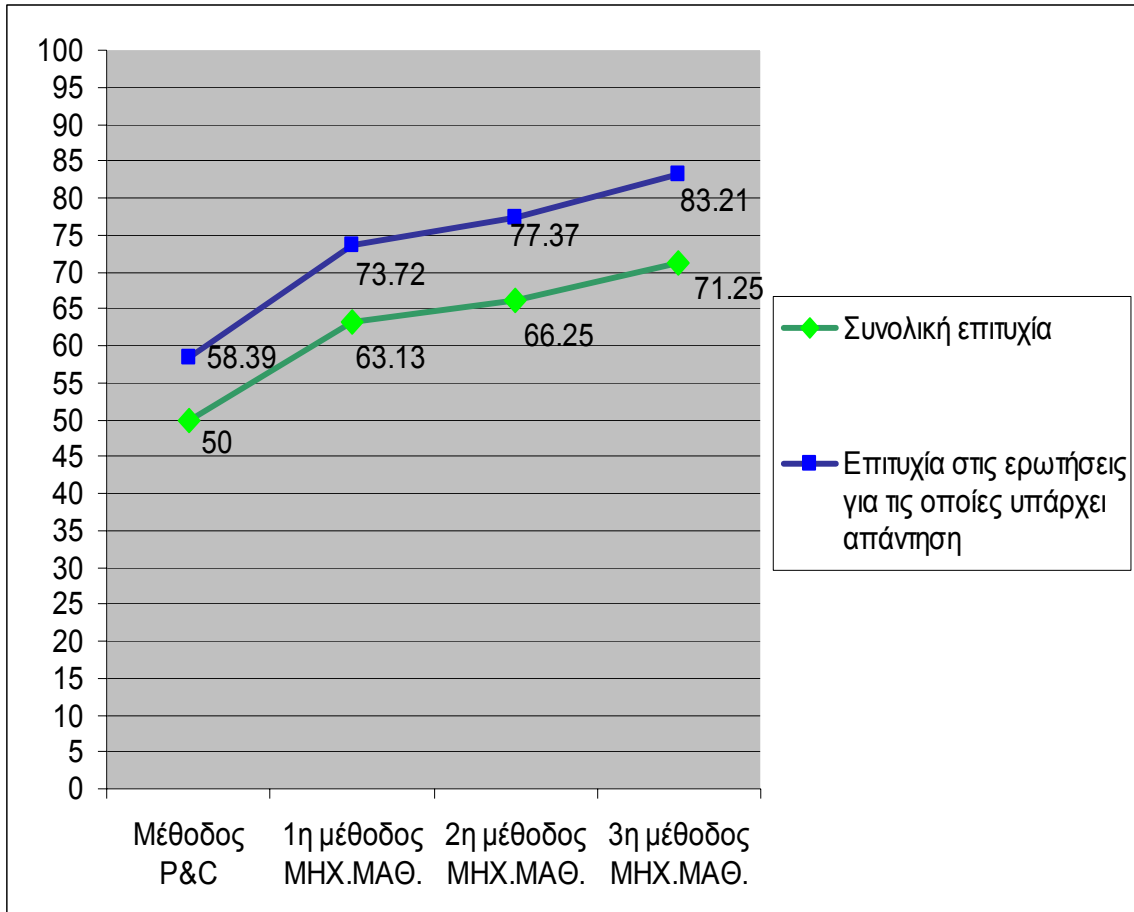
Στα πειράματα όπου επιλέγονταν οι ιδιότητες με το κριτήριο της ακρίβειας έγινε προσπάθεια να βρεθούν ιδιότητες δηλαδή φράσεις που να είναι χαρακτηριστικές παραθύρων κειμένου που δεν αποτελούν ορισμό. Βάσει όμως των αποτελεσμάτων δεν προέκυψαν τέτοιες ιδιότητες. Επαναλήφθηκαν επομένως τα πειράματα επιλέγοντας τις 100 ιδιότητες με τις **υψηλότερες ακρίβειες**. Το ποσοστό επιτυχίας αυξήθηκε κατά 1.8% και έφτασε στο **81%**.

Στην συνέχεια επαναλάβαμε τα ίδια πειράματα δίχως να χρησιμοποιήσουμε την ιδιότητα του αποτελέσματος της μεθόδου των P&C. Ήταν ενδιαφέρον, δεδομένων των ποσοστών επιτυχίας που είχε η μέθοδος μεμονωμένη, να φανεί αν τα υψηλά ποσοστά οφείλονταν στην χρήση της. Τα ποσοστά επιτυχίας όμως παρέμειναν στα ίδια επίπεδα.

Άρα συμπεραίνεται ότι η χρήση μόνο των φράσεων για ιδιότητες επαρκεί για να απαντηθεί ένα σύνολο ερωτήσεων. Επομένως, μπορεί να **απαλλαγθεί η τεχνική από την χρήση λεξικών** όπως του Wordnet και τα ποσοστά επιτυχίας να παραμείνουν υψηλά. Το τελευταίο συμπέρασμα είναι σημαντικό ιδιαίτερα για γλώσσες όπου αντίστοιχα υπολογιστικά λεξικά είναι δυσεύρετα.

6. ΣΥΓΚΡΙΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Ακολουθεί στο παρακάτω σχήμα ένα γράφημα με όλα τα αποτελέσματα από τις μεθόδους που μελετήθηκαν. Τα ποσοστά επιτυχίας μεγιστοποιούνται όταν χρησιμοποιείται Μηχανική Μάθηση και οι ιδιότητες επιλέγονται με το κριτήριο της ακρίβειας.



Σχήμα 6.1 Ποσοστά επιτυχίας των μεθόδων που μελετήθηκαν κατά την εργασία

7. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

Όσον αφορά τον τρόπο υλοποίησης της τεχνικής θα μπορούσαν να υπάρξουν διάφορες εναλλακτικές προσεγγίσεις καθώς και προσθήκες.

Αρχικά, επιλέξαμε την γλώσσα των Αγγλικών διότι οι συλλογές κειμένων του TREC διατίθενται μόνο στα Αγγλικά. Μπορούν να γίνουν αντίστοιχα πειράματα στα Ελληνικά, αν βρεθεί μια κατάλληλη συλλογή κειμένων και ένα αντίστοιχο σύνολο ερωτήσεων.

Επιπλέον, με την ενσωμάτωση ενός διαχωριστή προτάσεων, οι απαντήσεις που επιστρέφονται θα αποτελούνται από ολόκληρες προτάσεις και επομένως θα είναι περισσότερο κατανοητές και ευανάγνωστες.

Η Μηχανική Μάθηση μπορεί να αποτελέσει καλή λύση στα συστήματα ερωταποκρίσεων καθώς τα ποσοστά επιτυχίας είναι σημαντικά υψηλότερα από προϋπάρχουσες μεθόδους τουλάχιστον για τις **ερωτήσεις ορισμού**. Θα ήταν ενδιαφέρον να μελετηθεί η χρήση της μεθόδου για άλλες κατηγορίες ερωτήσεων όπως οι ερωτήσεις τοποθεσίας («Που βρίσκεται το Μέγαρο Μουσικής;») και οι χρονικές ερωτήσεις («Πότε εφευρέθηκε το τηλέφωνο;»).

Τέλος, το σύστημα μπορεί να υλοποιηθεί με χρήση μιας μηχανής αναζήτησης (front-end), όπως το Google, από την οποία θα λαμβάνει τα κείμενα που επιστρέφει για μια ερώτηση και θα εξάγει από αυτά τις απαντήσεις.

ΑΝΑΦΟΡΕΣ

- Androutsopoulos I, Ritchie G.D., *"Database Interfaces"*, In R. Dale, H. Moisl, and H. Somers (Eds.), *Handbook of Natural Language Processing*, chapter 9, pp. 209-240, Marcel Dekker Inc., 2000
- Harabagiu Sanda, Moldovan Dan, *"The Oxford Handbook of Computational Linguistics Specific"*, chapter 31, pp. 560-583
- Hearst, M.A., *"Automated Discovery of Wordnet Relations"* in *Wordnet: an Electronic Lexical Database*, Christiane Fellbaum Ed, MIT Press, Cambridge MA, 1998
- Hirschmann L., Gaizauskas R. *"Natural language question answering: the view from here"*, Cambridge University Press, 2001
- Joho Hideo, Sanderson Mark, *"Retrieving Descriptive Phrases from Large Amounts of Free Text"*, Proceedings of CIKM, 2000
- Joho Hideo, Ying Ki Liu, Sanderson Mark, *"Large scale testing of a descriptive phrase finder"*, Department of Information Studies, University of Sheffield, 2001
- Miller George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller *"Introduction to WordNet: An On-line Lexical Database"*, 1993
- Mitchell, T.M. *"Machine Learning"*, McGraw-Hill International Editions, 1997
- Pomerleau, D.A., *"ALVINN: An autonomous land vehicle in a neural network"*, Technical Report CMU-CS-89-107. Pittsburg, PA. Carnegie Mellon University, 1989
- 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"*, 2001
- Prager John, Jennifer Chu-Carroll, Krzysztof, *"Use of Wordnet Hypernyms for answering What-Is Questions"*, 2002
- Radev Dragomir R., Prager John, Samn Valerie, *"Ranking suspected answers to natural language questions using predictive annotation"*, 1999
- Reiter E., Dale R., *"Building Natural Language Generation Systems"*, Cambridge University Press, 2000

- Schölkopf Bernhard, Alex Smola, “*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, 2001
- Voorhees Ellen M., “*Overview of the TREC-9 Question Answering Track*”, National Institute of Standards and Technology, 2000
- Voorhees Ellen M., “*The TREC-8 Question Answering Track Report*”, National Institute of Standards and Technology, 1999
- Witten H. Ian, Frank Eibe, “*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*”, Morgan Kaufmann Publishers, 2000